PALGRAVE HANDBOOK OF ECONOMETRICS

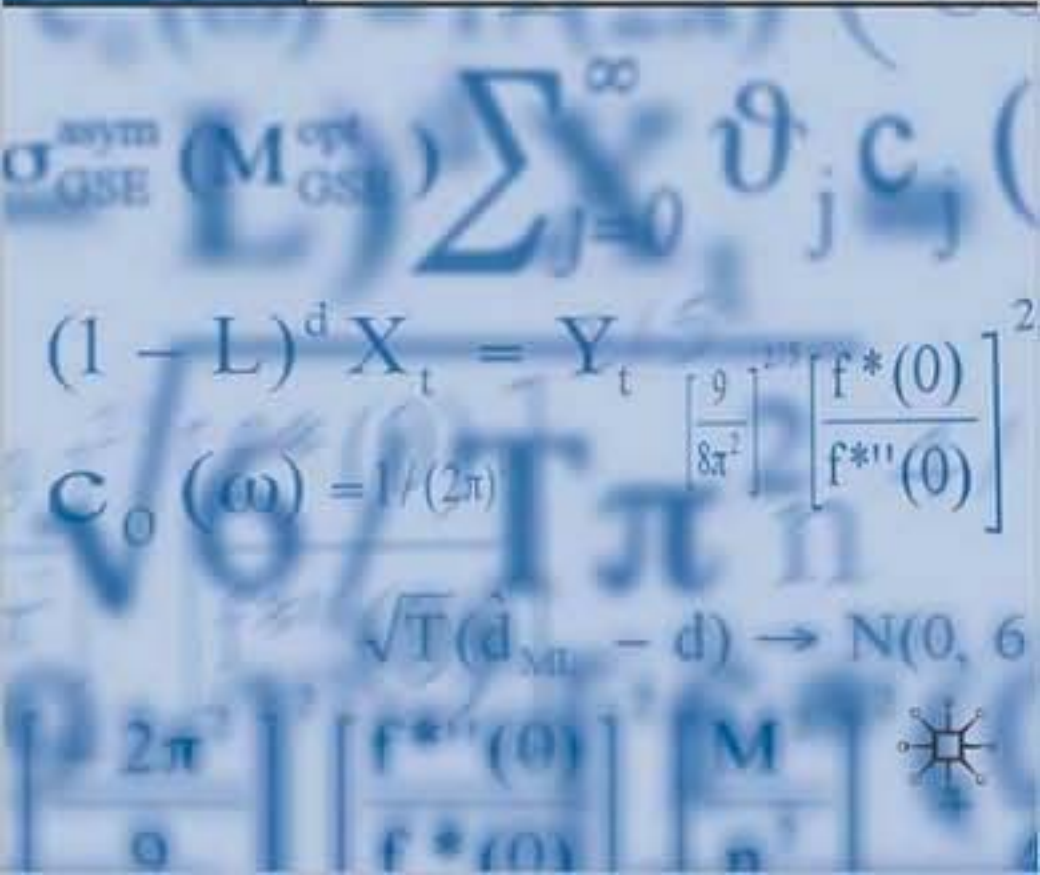
Volume 2

Applied Econometrics

edited by

Terence C. Mills

and Kerry Patterson



Palgrave Handbook of Econometrics

Volume 2: Applied Econometrics

Edited By

Terence C. Mills

and

Kerry Patterson

2009

palgrave
macmillan

Contents

Part IV Cross-section and Panel Data Applications

Part V Microeconometrics

Part VI Applications of Econometrics to Economic Policy

Part VII Applications to Financial Econometrics

Part VIII Growth Development Econometrics

Part IX Spatial Econometrics

Part X Applied Econometrics and Computing

Notes on Contributors

Luc Anselin is Foundation Professor of Geographical Sciences and Director of the School of Geographical Sciences and Urban Planning at Arizona State University, USA.

Aninyda Banerjee is Professor of Econometrics at the University of Birmingham, UK.

Gunnar Bårdsen is Professor of Economics at the Norwegian University of Science and Technology, Norway.

Thorsten Beck is Professor of Economics and Chair at the European Banking Center, Tilburg University and Research Fellow, CEPR.

Colin Cameron is Professor of Economics at the University of California, Davis, USA.

Fabio Canova is ICREA Research Professor in Social Science at Universitat Pompeu Fabra, Barcelona, Spain.

Joe Cardinale is a Manager, Economics at Air Products and Chemicals, Inc., USA.

Michael P. Clements is Professor of Economics at Warwick University, UK.

John DiNardo is Professor of Economics and Public Policy at the University of Michigan, Ann Arbor, USA.

George Dotsis is Lecturer in Finance at the Essex Business School, University of Essex, UK.

Steven N. Durlauf is Professor of Economics at the University of Wisconsin-Madison, USA.

Juan Carlos Escanciano is Assistant Professor of Economics at Indiana University, Bloomington, USA.

Carlo A. Favero is Professor of Economics at IGIER-Bocconi University, Italy.

Julie Le Gallo is Professor of Economics and Econometrics at the Université de Franche-Comté, France.

Luis A. Gil-Alana is Professor of Econometrics at the University of Navarra, Spain.

William Greene is Professor of Economics at the Stern School of Business, New York, USA.

Stephen G. Hall is Professor of Economics at University of Leicester, UK.

David I. Harvey is Reader in Econometrics at the School of Economics, University of Nottingham, UK.

David F. Hendry is Professor of Economics and Fellow, Nuffield College, Oxford University, UK.

Brian Henry is Visiting Fellow at the National Institute of Economic and Social Research, NIESR, UK.

Javier Hualde is Ramon y Cajal Research Fellow in Economics at the Public University of Navarra, Spain.

David Jacho-Chávez is Assistant Professor of Economics at Indiana University, Bloomington, USA.

Paul A. Johnson is Professor of Economics at Vassar College, New York State, USA.

Andrew M. Jones is Professor of Economics at the University of York, UK.

Katarina Juselius is Professor of Empirical Time Series Econometrics at the University of Copenhagen, Denmark.

Ignacio N. Lobato is Professor of Econometrics at the Instituto Tecnológico Autónomo de México, Mexico.

Nancy Lozano-Gracia is Postdoctoral Research Associate in the GeoDa Center for Geospatial Analysis and Computation at Arizona State University, USA.

Raphael N. Markellos is Assistant Professor of Quantitative Finance at the Athens University of Economics and Business (AUEB), Greece.

Bruce D. McCullough is Professor of Decision Sciences and Economics at Drexel University, Philadelphia, USA.

Terence C. Mills is Professor of Applied Statistics and Econometrics at Loughborough University, UK.

James Mitchell is Research Fellow at the National Institute of Economic and Social Research, UK.

Ragnar Nymoen is Professor of Economics at University of Oslo, Norway.

Marius Ooms is Associate Professor of Econometrics at the VU University, Amsterdam, The Netherlands.

Kerry Patterson is Professor of Econometrics at the University of Reading, UK.

Efthymios G. Pavlidis is Lecturer in Economics at the Lancaster University Management School, Lancaster University, UK.

Ivan Paya is Senior Lecturer in Economics at the Lancaster University Management School, Lancaster University, UK.

David A. Peel is Professor in Economics at the Lancaster University Management School, Lancaster University, UK.

D. Stephen G. Pollock is Professor of Economics at the University of Leicester, UK.

Tommaso Proietti is Professor of Economic Statistics at the University of Rome 'Tor Vergata', Italy.

Sergio Rey is Professor of Geographical Sciences at Arizona State University, USA.

Larry W. Taylor is Professor of Economics at the College of Business and Economics, Lehigh University, Pennsylvania, USA.

Jonathan R.W. Temple is Professor of Economics at Bristol University, UK.

Pravin Trivedi is Professor of Economics at Indiana University, Bloomington, USA.

Ruey S. Tsay is Professor of Econometrics and Statistics at the University of Chicago Booth School of Business, USA.

Martin Wagner is Senior Economist at the Institute for Advanced Studies in Vienna, Austria.

Editors' Introduction

Terence C. Mills and Kerry Patterson

The *Palgrave Handbook of Econometrics* was conceived to provide an understanding of major developments in econometrics, both in theory and in application. Over the last twenty-five years or so, econometrics has grown in a way that few could have contemplated, and it became clear to us, as to others, that no single person could have command either of the range of technical knowledge that underpins theoretical econometric developments or the extent of the application of econometrics. In short, econometrics is not, as it used to be considered, a set of techniques that is applied to a previously well-defined problem in economics; it is not a matter of finding the "best" estimator from a field of candidates, applying that estimator and reporting the results. The development of economics is now inextricably entwined with the development of econometrics.

The first Nobel Prize in Economics was awarded to Ragnar Frisch and Jan Tinbergen, both of whom made significant contributions to what we now recognize as applied econometrics. More recently, Nobel Prizes in Economics have been awarded to Clive Granger, Robert Engle, James Heckman and Daniel McFadden, who have all made major contributions to applied econometrics. It is thus clear that the discipline has recognized the influential role of econometrics, both theoretical and applied, in advancing economic knowledge.

The aim of this volume is to make major developments in applied econometrics accessible to those outside their particular field of specialization. The response to Volume 1 was universally encouraging and it has become clear that we were fortunate to be able to provide a source of reference for others for many years to come. We hope that this high standard is continued and achieved here. Typically, applied econometrics, unlike theoretical econometrics, has always been rather poorly served for textbooks, making it difficult for both undergraduate and postgraduate students to get a real "feel" for how econometrics is actually done. To some degree, the econometric textbook market has responded, so that now the leading textbooks include many examples; even so, these examples typically are of an illustrative nature, focusing on simple points, simply exposited, rather than on the complexity that is revealed in practice. Thus our hope is that this volume will provide a genuine entry into the detailed considerations that have to be

made when combining economics and econometrics in order to carry out serious empirical research.

As in the case of Volume 1, the chapters here have been specially commissioned from acknowledged experts in their fields; further, each of the chapters has been reviewed by the editors, one or more of the associate editors and a referee. Thus, the process is akin to submission to a journal; however, whilst ensuring the highest standards in the evaluation process, the chapters have been conceived of as part of a whole rather than as a set of unrelated contributions. It has not, however, been our intention to provide just a series of surveys or overviews of some areas of applied econometrics, although the survey element is directly or indirectly served in part here. By its very nature, this volume is about econometrics as it is applied and, to succeed in its aim, the contributions, conceived as a whole, have to meet this goal.

We have organized the chapters of this volume of the *Handbook* into ten parts. The parts are certainly not watertight, but serve as a useful initial organization of the central themes. Part I contains three chapters under the general heading of "The Methodology and Philosophy of Applied Econometrics." The lead chapter is by David Hendry, who has been making path-breaking contributions in theoretical and applied econometrics for some forty years or so. It is difficult to conceive how econometrics would have developed without David's many contributions. This chapter first places the role of applied econometrics in an historical context and then develops a theory of applied econometrics. As might be expected, the key issues are confronted head-on.

In introducing the first volume we noted that the "growth in econometrics is to be welcomed, for it indicates the vitality and importance of the subject. Indeed, this growth and, arguably, the dominance over the last ten or twenty years of econometric developments in taking economics forward, is a notable change from the situation faced by the subject some twenty-five years or so ago." Yet in Chapter 1, Hendry notes that, next to data measurement, collection and preparation, on the one hand, and teaching, on the other, "Applied Econometrics" does not have a high credibility in the profession. Indeed, whilst courses in theoretical econometrics or econometric techniques are de rigueur for a good undergraduate or Masters degree, courses in applied econometrics have no such general status.

The intricacies, possibly even alchemy (Hendry, 1980), surrounding the mix of techniques and data seem to defy systematization; perhaps they should be kept out of the gaze of querulous students, who may – indeed should – be satisfied with illustrative examples! Often to an undergraduate or Masters student undertaking a project, applied econometrics is the application of econometrics to data, no more, no less, with some relief if the results are at all plausible. Yet, in contrast, leading journals, for example, the *Journal of Econometrics*, the *Journal of Applied Econometrics* and the *Journal of Business and Economic Statistics*, and leading topic journals, such as the *Journal of Monetary Economics*, all publish applied econometric articles having substance and longevity in their impact and which serve to change the direction of the development of econometric theory (for a famous example, see Nelson and Plosser, 1982). To some, applying econometrics seems unsystematic

and so empirical results are open to question; however, as Hendry shows, it is possible to formalize a theory of applied econometrics which provides a coherent basis for empirical work. Chapter 1 is a masterful and accessible synthesis and extension of Hendry's previous ideas and is likely to become compulsory reading for courses in econometrics, both theory and applied; moreover, it is completed by two applications using the Autometrics software (Doornik, 2007). The first application extends the work of Magnus and Morgan (1999) on US food expenditure, which was itself an update of a key study by Tobin (1950) estimating a demand function for food. This application shows the Autometrics algorithm at work in a simple context. The second application extends the context to a multiple equation setting relating industrial output, the number of bankruptcies and patents, and real equity prices. These examples illustrate the previously outlined theory of applied econometrics combined with the power of the Autometrics software.

In Chapter 2, Fabio Canova addresses the question of how much structure there should be in empirical models. This has long been a key issue in econometrics, and some old questions, particularly those of identification and the meaning of structure, resurface here in a modern context. The last twenty years or so have seen two key developments in macroeconometrics. One has been the development of dynamic stochastic general equilibrium (DSGE) models. Initially, such models were calibrated rather than estimated, with the emphasis on "strong" theory in their specification; however, as Canova documents, more recently likelihood-based estimation has become the dominant practice. The other key development has been that of extending the (simple) vector autoregression (VAR) to the structural VAR (SVAR) model. Although both approaches involve some structure, DSGE models, under the presumption that the model is correct, rely more on an underlying theory than do SVARs. So which should be used to analyze a particular set of problems? As Canova notes: "When addressing an empirical problem with a finite amount of data, one has ... to take a stand on how much theory one wants to use to structure the available data prior to estimation." Canova takes the reader through the advantages and shortcomings of these methodologies; he provides guidance on what to do, and what not to do, and outlines a methodology that combines elements of both approaches.

In Chapter 3, John DiNardo addresses some philosophical issues that are at the heart of statistics and econometrics, but which rarely surface in econometric textbooks. As econometricians, we are, for example, used to working within a probabilistic framework, but we rarely consider issues related to what probability actually is. To some degree, we have been prepared to accept the axiomatic or measure theoretic approach to probability, due to Kolgomorov, and this has provided us with a consistent framework that most are happy to work within. Nevertheless, there is one well-known exception to this unanimity: when it comes to the assignment and interpretation of probability measures and, in particular, the interpretation of some key conditional probabilities; this is whether one adopts a Bayesian or non-Bayesian perspective. In part, the debate that DiNardo discusses relates to the role of the Bayesian approach, but it is more than this; it concerns metastatistics and philosophy, because, in a sense, it relates to a discussion of the

theory of theories. This chapter is deliberately thought-provoking and certainly controversial – two characteristics that we wish to encourage in a *Handbook* that aims to be more than just an overview. For balance, the reader can consult Volume 1 of the *Handbook*, which contains two chapters devoted to the Bayesian analysis of econometric models (see Poirier and Tobias, 2006, and Strachan *et al*., 2006). The reader is likely to find familiar concepts here, such as probability and testing, but only as part of a development that takes them into potentially unfamiliar areas. DiNardo's discussion of these issues is wide-ranging, with illustrations taken from gambling and practical examples taken as much from science, especially medicine, as economics. One example from the latter is the much-researched question of the causal effect of union status on wages: put simply, do unions raise wages and, if so, by how much? This example serves as an effective setting in which to raise issues and to show that differences in approach can lead to differences in results.

For some, the proof of the pudding in econometrics is the ability to forecast accurately, and to address some key issues concerning this aspect of econometrics Part II contains two chapters on forecasting. The first, Chapter 4, by Michael Clements and David Harvey, recognizes that quite often several forecasts are available and, rather than considering a selection strategy that removes all but the best on some criterion, it is often more fruitful to consider different ways of combining forecasts, as suggested in the seminal paper by Bates and Granger (1969). In an intuitive sense, one forecast may be better than another, but there could still be some information in the less accurate forecast that is not contained in the more accurate forecast. This is a principle that is finding wider application; for example, in some circumstances, as in unit root testing, there is more than one test available and, indeed, there may be one uniformly powerful test, yet there is still potential merit in combining tests.

In the forecasting context, Clements and Harvey argue that the focus for multiple forecasts should not be on testing the null of equal accuracy, but on testing for encompassing. Thus it is not a question of choosing forecast A over forecast B, but of whether the combination of forecasts A and B is better than either individual forecast. Of course, this may be of little comfort from a structuralist point of view if, for example, the two forecasts come from different underlying models; but it is preferable when the loss function rewards good fit in some sense. Bates and Granger (1969) suggested a simple linear combination of two unbiased forecasts, with weights depending on the relative accuracy of the individual forecasts, and derived the classic result that, even if the forecasts are equally accurate in a mean squared error loss sense, then there will still be a gain in using the linear combination unless the forecasts are perfectly correlated, at least theoretically. Clements and Harvey develop from this base model, covering such issues as biased forecasts, nonlinear combinations, and density or distribution forecasts. The concept of forecast encompassing, which is not unique in practice, is then considered in detail, including complications arising from integrated variables, non-normal errors, serially correlated forecast errors, ARCH errors, the uncertainty implied by model estimation, and the difficulty of achieving tests with the correct actual size. A number of recent developments are examined, including the concept of conditional forecast

evaluation (Giacomini and White, 2006), evaluating quantile forecasts, and relaxing the forecast loss function away from the traditional symmetric squared error. In short, this chapter provides a clear, critical and accessible evaluation of a rapidly developing area of the econometrics literature.

Chapter 5 is by Stephen Hall and James Mitchell, who focus on density forecasting. There has been a great deal of policy interest in forecasting key macroeconomic variables such as output growth and inflation, some of which has been institutionally enshrined by granting central banks independence in inflation targeting. In turn, there has been a movement away from simply reporting point forecasts of inflation and GDP growth in favor of a fan chart representation of the distribution of forecasts. A density forecast gives much more information than a simple point forecast, which is included as just one realization on the outcome axis. As a corollary, forecast evaluation should also include techniques that evaluate the accuracy, in some well-defined sense, of the density forecast. However, given that generally we will only be able to observe one outcome (or event) per period, some thought needs to be given to how the distributional aspect of the forecast is evaluated. Hall and Mitchell discuss a number of possibilities and also consider methods of evaluating competing density forecasts. A further aspect of density forecasting is the ability of the underlying model to generate time variation in the forecast densities. If the underlying model is a VAR, or can be approximated by a VAR, then, subject to some qualifications, the only aspect of the forecast density which is able to exhibit time variation is the mean; consequently, models have been developed that allow more general time variation in the density through, for example, ARCH and GARCH errors and time-varying parameters. This chapter also links in with the previous chapter by considering combinations of density forecasts. There are two central possibilities: the linear opinion pool is a weighted linear combination of the component densities, whereas the logarithmic opinion pool is a multiplicative combination. Hall and Mitchell consider the problem of determining the weights in such combinations and suggest that predictive accuracy improves when the weights reflect shifts in volatility, a characteristic of note for the last decade or so in a number of economies.

Part III contains four chapters under the general heading of "Time Series Applications." A key area in which the concept of a time series is relevant is in characterizing and determining trends and cycles. Chapter 6, by Stephen Pollock, is a tour de force on modeling trends and cycles, and on the possibilities and pitfalls inherent in the different approaches. In the simplest of models, cyclical fluctuations are purely sinusoidal and the trend is exponential; although simple, this is a good base from which to understand the nature of developments that relax these specifications. Such developments include the view that economic time series evolve through the accumulation of stochastic shocks, as in an integrated Weiner process. The special and familiar cases of the Beveridge–Nelson decomposition, the Hodrick–Prescott filter, the Butterworth filter and the unifying place of Weiner–Kolgomorov filtering are all covered with admirable clarity. Other considerations include the complications caused by the limited data that is often available in economic applications, contrary to the convenient assumptions of theory. In an

appealing turn of phrase, Pollock refers to obtaining a decomposition of components based on the periodogram "where components often reside within strictly limited frequency bands which are separated by dead spaces where the spectral ordinates are virtually zeros." The existence of these "spectral dead spaces" is key to a practical decomposition of an economic time series, however achieved. In practice, trend fitting requires judgment and a clear sense of what it is that the trend is capturing. Other critical issues covered in this chapter include the importance of structural breaks, a topic that has been influential elsewhere (for example, in questioning the results of unit root testing: Perron, 1989); and to aid the reader, practical examples are included throughout the exposition.

Chapter 7, by Joe Cardinale and Larry Taylor, continues the time series theme of analyzing economic cycles whilst focusing on asymmetries, persistence and synchronization. This is a particularly timely and somewhat prophetic chapter given that we are currently experiencing what is perhaps the deepest recession in recent economic history. How can we analyze the critical question "When will it end?" This chapter provides the analytical and econometric framework to answer such a question. The central point is that cycles are much more interesting than just marking their peaks and troughs would suggest. Whilst "marking time" is important, it is just the first part of the analysis, and should itself be subjected to methods for distinguishing phases (for example, expansions and contractions of the output cycle). Once phases have been distinguished, their duration and characteristics become of interest; for example, do long expansions have a greater chance of ending than short expansions? Critical to the analysis is the hazard function: "the conditional probability that a phase will terminate in period t, given that it has *lasted* t *or more* periods." Cardinale and Taylor consider different models and methods of estimating the hazard function and testing hypotheses related to particular versions of it. They also consider tests of duration dependence, the amplitudes of cycles, and the synchronization of cycles for different but related variables; for example, output and unemployment. Their theoretical analysis is complemented with a detailed consideration of US unemployment.

No handbook of econometrics could be without a contribution indicating the importance of cointegration analysis for non-stationary data. In Chapter 8, Katerina Juselius considers one of the most enduring puzzles in empirical economics, namely, if purchasing power parity (PPP) is the underlying equilibrium state that determines the relationship between real exchange rates, why is there "pronounced persistence" away from this equilibrium state? This has been a common finding of empirical studies using data from a wide range of countries and different sample periods. Juselius shows how a careful analysis can uncover important structures in the data; however, these structures are only revealed by taking into account the different empirical orders of integration of the component variables, the identification of stationary relationships between non-stationary variables, the dynamic adjustment of the system to disequilibrium states, the appropriate deterministic components, and the statistical properties of the model. As Juselius notes, and in contrast to common approaches, the order of integration is regarded here as an empirical approximation rather than a structural parameter. This opens up a

distinction between, for example, a variable being empirically $I(d)$ rather than structurally $I(d)$; a leading example here is the $I(2)$ case which, unlike the $I(1)$ case, has attracted some "suspicion" when applied in an absolute sense to empirical series. The challenging empirical case considered by Juselius is the relationship between German and US prices and nominal exchange rates within a sample that includes the period of German reunification. The methodology lies firmly within the framework of general-to-specific modeling, in which a general unrestricted model is tested down (see also Hendry, Chapter 1) to gain as much information without empirical distortion. A key distinction in the methodological and empirical analysis is between pushing and pulling forces: in the current context, prices push whereas the exchange rate pulls. PPP implies that there should be just a single stochastic trend in the data, but the empirical analysis suggests that there are two, with the additional source of permanent shocks being related to speculative behaviour in the foreign exchange market.

In an analysis of trends and cycles, economists often characterize the state of the economy in terms of indirect or latent variables, such as the output gap, core inflation and the non-accelerating rate of inflation (NAIRU). These are variables that cannot be measured directly, but are nevertheless critical to policy analysis. For example, the need to take action to curb inflationary pressures is informed by the expansionary potential in the economy; whether or not a public sector budget deficit is a matter for concern is judged by reference to the cyclically adjusted deficit. These concepts are at the heart of Chapter 9 by Tommaso Proietti, entitled "Structural Time Series Models for Business Cycle Analysis," which links with the earlier chapters by Pollock and Cardinale and Taylor. Proietti focuses on the measurement of the output gap, which he illustrates throughout using US GDP. In the simplest case, what is needed is a framework for decomposing a time series into a trend and cycle and Proietti critically reviews a number of methods to achieve such a decomposition, including the random walk plus noise (RWpN) model, the local linear trend model (LLTM), methods based on filtering out frequencies associated with the cycle, multivariate models that bring together related macroeconomic variables, and the production function approach. The estimation and analysis of a number of models enables the reader to see how the theoretical analysis is applied and what kind of questions can be answered. Included here are a bivariate model of output and inflation for the US and a model of mixed data frequency, with quarterly observations for GDP and monthly observations for industrial production, the unemployment rate and CPI inflation. The basic underlying concepts, such as the output gap and core inflation, are latent variables and, hence, not directly observable: to complete the chapter, Proietti also considers how to judge the validity of the corresponding empirical measures of these concepts.

To complete the part of the *Handbook* on Times Series Applications, in Chapter 10 Luis Gil-Alana and Javier Hualde provide an overview of fractional integration and cointegration, with an empirical application in the context of the PPP debate. A time series is said to be integrated of order d, where d is an integer, if d is the minimum number of differences necessary to produce a stationary time series. This is a particular form of non-stationarity and one which dominated the econometrics

literature in the 1980s and early 1990s, especially following the unit root literature. However, the integer restriction on d is not necessary to the definition of an integrated series (see, in particular, Granger and Joyeux, 1980), so that d can be a fraction – hence the term "fractionally integrated" for such series. Once the integer restriction is relaxed for a single series, it is then natural to relax it for the multivariate case, which leads to the idea of fractional cointegration. Gil-Alana and Hualde give an overview of the meaning of fractional integration and fractional cointegration, methods of estimation for these generalized cases, which can be approached in either the time or frequency domains, the underlying rationale for the existence of fractionally integrated series (for example, through the aggregation of micro-relationships), and a summary of the empirical evidence for fractionally integrated univariate series and fractionally cointegrated systems of series. The various issues and possible solutions are illustrated in the context of an analysis of PPP for four bivariate series. It is clear that the extension of integration and cointegration to their corresponding fractional cases is not only an important generalization of the theory, but one which finds a great deal of empirical support.

One of the most significant developments in econometrics over the last twenty years or so has been the increase in the number of econometric applications involving cross-section and panel data (see also Ooms, Chapter 29). Hence Part IV is devoted to this development. One of the key areas of application is to choice situations which have a discrete number of options; examples include the "whether to purchase" decision, which has wide application across consumer goods, and the "whether to participate" decision, as in whether to enter the labor force, to retire, or to join a club. Discrete choice models are the subject of Chapter 11 by Bill Greene, who provides a critical, but accessible, review of a vast literature. The binary choice model is a key building block here and so provides a prototypical model with which to examine such topics as specification, estimation and inference; it also allows the ready extension to more complex models such as bivariate and multivariate binary choice models and multinomial choice models. Models involving count data are also considered as they relate to the discrete choice framework. A starting point for the underlying economic theory is the extension of the classical theory of consumer behavior, involving utility maximization subject to a budget constraint, to the random utility model. The basic model is developed from this point and a host of issues are considered that arise in practical studies, including estimation and inference, specification tests, measuring fit, complications from endogenous right-hand-side variables, random parameters, the use of panel data, and the extension of the familiar fixed and random effects. To provide a motivating context, Greene considers an empirical application involving a bivariate binary choice model. This is where two binary choice decisions are linked; in this case, in the first decision the individual decides whether to visit a physician, which is a binary choice, and the second involves whether to visit the hospital, again a binary choice: together they constitute a bivariate (and ordered) choice. An extension of this model is to consider the number of times that an individual visits the doctor or a hospital. This gives rise to a counts model (the number of visits to the doctor and the number of visits to the hospital) with its own particular specification. Whilst a natural place to

start is the Poisson model, this, as Greene shows, is insufficient as a general framework; the extension is provided and illustrated with panel data from the German health care system. A second application illustrates a mixed logit and error components framework for modeling modes of transport choice (air, train, bus, car). Overall, this chapter provides an indication, through the variety of its applications, as to why discrete choice models have become such a significant part of applied econometrics.

The theme of panel data methods and applications is continued in Chapter 12 by Andrew Jones. The application of econometrics to health economics has been an important area of development over the last decade or so. However, this has not just been a case of applying existing techniques: rather, econometrics has been able to advance the subject itself, asking questions that had not previously been asked – and providing answers. This chapter will be of interest not only to health economics specialists, but also to those seeking to understand how treatment effects in particular are estimated and to those investigating the extent of the development and application of panel data methods (it is complemented by Colin Cameron in Chapter 14). At the center of health economics is the question "What are the impacts of specific health policies?" Given that we do not observe experimental data, what can we learn from non-experimental data? Consider the problem of evaluating a particular treatment; for an individual, the treatment effect is the difference in outcome between the treated and the control, but since an individual is either treated or not at a particular time, the treatment effect cannot be observed. "Treatment" is here a general term that covers not only single medical treatments but also broad policies, and herein lies its generality, since a treatment could equally be a policy to reduce unemployment or to increase the proportion of teenagers receiving higher education. In a masterful understanding of a complex and expanding literature, Jones takes the reader through the theoretical and practical solutions to the problems associated with estimating and evaluating treatment effects, covering, *inter alia*, identification strategies, dynamic models, estimation methods, different kinds of data, and multiple equation models; throughout the chapter the methods and discussion are motivated by practical examples illustrating the breadth of applications.

A key development in econometrics over the last thirty years or so has been the attention given to the properties of the data, as these enlighten the question of whether the underlying probability structure is stationary or not. In a terminological shorthand, we refer to data that is either stationary or non-stationary. Initially, this was a question addressed to individual series (see Nelson and Plosser, 1982); subsequently, the focus expanded, through the work of Engle and Granger (1987) and Johansen (1988), to a multivariate approach to non-stationarity. The next step in the development was to consider a panel of multivariate series. In Chapter 13, Anindya Banerjee and Martin Wagner bring us up to date by considering panel methods to test for unit roots and cointegration. The reader will find in this chapter a theoretical overview and critical assessment of a vast and growing body of methods, combined with practical recommendations based on the insights obtained from a wide base of substantive applications. In part, as is evident in other areas

of econometric techniques and applications, theory has responded to the much richer sources of data that have become available, not only at a micro or individual level, as indicated in Chapter 12, combined with increases in computing power. As Banerjee and Wagner note, we now have long time series on macroeconomic and industry-level data. Compared to just twenty years ago, there is thus a wealth of data on micro, industry and macro-panels. A panel dataset embodies two dimensions: the cross-section dimension and the time-series dimension, so that, in a macro-context, for example, we can consider the question of convergence not just of a single variable (say, of a real exchange rate to a comparator, be that a PPP hypothetical or an alternative actual rate), but of a group of variables, which is representative of the multidimensional nature of growth and cycles. A starting point for such an analysis is to assess the unit root properties of panel data but, as in the univariate case, issues such as dependency, the specification of deterministic terms, and the presence of structural breaks are key practical matters that, if incorrectly handled, can lead to misleading conclusions. Usually, the question of unit roots is a precursor to cointegration analysis, and Banerjee and Wagner guide the reader through the central methods, most of which have been developed in the last decade. Empirical illustrations, based on exchange rate pass-through in the euro-area and the environmental Kuznets curve, complement the theoretical analysis.

Whilst the emphasis in Chapter 13 is on panels of macroeconomic or industry-level data, in Chapter 14, Colin Cameron, in the first of two chapters in Part V, provides a survey of microeconometric methods, with an emphasis on recent developments. The data underlying such developments are at the level of the individual, households and firms. A prototypical question in microeconometrics relates to the identification, estimation and evaluation of marginal effects using individual-level data; for example, the effect on earnings of an additional year of education. This example is often used to motivate some basic estimation methods, such as least squares, maximum likelihood and instrumental variables, in undergraduate and graduate texts in econometrics, so it is instructive to see how recent developments have extended these methods. The development of the basic methods include generalized method of moments (GMM), empirical likelihood, simulation-based methods, quantile regression and nonparametric and semiparametric estimation, whilst developments in inference include robustifying standard tests and bootstrap methods. Apart from estimation and inference, Cameron considers a number of other issues that occur frequently in microeconometric studies: in particular, issues related to causation, as in estimating and evaluating treatment effects; heterogeneity, for example due to regressors or unobservables; and the nature of microeconometric data, such as survey data and the sampling scheme, with problems such as missing data and measurement error.

The development of econometrics in the last decade or so in particular has been symbiotic with the development of advances in computing, particularly that of personal computers. In Chapter 15, David Jacho-Chávez and Pravin Trivedi focus on the relationship between empirical microeconometrics and computational considerations, which they call, rather evocatively, a "matrimony" between computing

and applied econometrics. No longer is it the case that the mainstay of empirical analysis is a set of macroeconomic time series, often quite limited in sample period. Earlier chapters in this part of the volume emphasize that the data sources now available are much richer than this, both in variety and length of sample period. As Jacho-Chávez and Trivedi note, the electronic recording and collection of data has led to substantial growth in the availability of census and survey data. However, the nature of the data leads to problems that require theoretical solutions: for example, problems of sample selection, measurement errors and missing or incomplete data. On the computing side, the scale of the datasets and estimation based upon them implies that there must be reliability in the high-dimensional optimization routines required by the estimation methods and an ability to handle large-scale Monte Carlo simulations. The increase in computing power has meant that techniques that were not previously feasible, such as simulation assisted estimation and resampling, are now practical and in widespread use. Moreover, nonparametric and semiparametric methods that involve the estimation of distributions rather than simple parameters, as in regression models, have been developed through drawing on the improved power of computers. Throughout the chapter, Jacho-Chávez and Trivedi motivate their discussion by the use of examples of practical interest, including modeling hedonic prices of housing attributes, female labor force participation, Medicare expenditure, and number of doctor visits. Interestingly, they conclude that there are important problems, particularly those related to assessing public policy, such as identification and implementation in the context of structural, dynamic and high-dimensional models, which remain to be solved.

In Part VI, the theme of the importance of economic policy is continued, but with the emphasis now on monetary policy and macroeconomic policy, which remain of continued importance. Starting in the 1970s and continuing into the 1990s, the development of macroeconometric models for policy purposes was a highly regarded area; during that period computing power was developing primarily through mainframe computers, allowing not so much the estimation as the simulation of macroeconomic models of a dimension that had not been previously contemplated. Government treasuries, central banks and some non-governmental agencies developed their own empirical macro-models comprising hundreds of equations. Yet, these models failed to live up to their promise, either wholly or in part. For some periods there was an empirical failure, the models simply not being good enough; but, more radically, the theoretical basis of the models was often quite weak, at least relative to the theory of the optimizing and rational agent and ideas of intertemporal general equilibrium.

In Chapter 16, Carlo Favero expands upon this theme, especially as it relates to the econometrics of monetary policy and the force of the critiques by Lucas (1976) and Sims (1980). A key distinction in the dissection of the modeling corpse is between structural identification and statistical identification. The former relates to the relationship between the structural parameters and the statistical parameters in the reduced form, while the latter relates to the properties of the statistical or empirical model which represents the data. Typically, structural identification is achieved

by parametric restrictions seeking to classify some variables as "exogenous," a task that some have regarded as misguided (or indeed even "impossible"). Further, a failure to assess the validity of the reduction process in going from the (unknown) data-generating process to a statistical representation, notwithstanding criticisms related to structural identification, stored up nascent empirical failure awaiting the macreconometric model. Developments in cointegration theory and practice have "tightened" up the specification of empirical macromodels, and DSGE models, preferred theoretically by some, have provided an alternative "modellus operandi." Subsequently, the quasi-independence of some central banks has heightened the practical importance of questions such as "How should a central bank respond to shocks in macroeconomic variables?" (Favero, Chapter 16). In practice, although DSGE models are favored for policy analysis, in their empirical form the VAR reappears, but with their own set of issues. Favero considers such practical developments as calibration and model evaluation, the identification of shocks, impulse responses, structural stability of the parameters, VAR misspecification and factor augmented VARs. A summary and analysis of Sims' (2002) small macroeconomic model (Appendix A) helps the reader to understand the relationship between an optimizing specification and the resultant VAR model.

In Chapter 17, Gunnar Bårdsen and Ragnar Nymoen provide a paradigm for the construction of a dynamic macroeconometric model, which is then illustrated with a small econometric model of the Norwegian economy that is used for policy analysis. Bårdsen and Nymoen note the two central critiques of "failed" macroeconometric models: the Lucas (1976) critique and the Clements and Hendry (1999) analysis of forecast failure involving "location" shifts (rather than behavioral parameter shifts). But these critiques have led to different responses; first, the move to explicit optimizing models (see Chapter 16); and, alternatively, to greater attention to the effects of regime shifts, viewing the Lucas critique as a possibility theorem rather than a truism (Ericsson and Irons, 1995). Whilst it is de rigueur to accept that theory is important, Bårdsen and Nymoen consider whether "theory" provides the (completely) correct specification or whether it simply provides a guideline for the specification of an empirical model. In their approach, the underlying economic model is nonlinear and specified in continuous time; hence, the first practical steps are linearization and discretization, which result in an equilibrium correction model (EqCM). Rather than remove the data trends, for example by applying the HP filter, the common trends are accounted for through a cointegration analysis. The approach is illustrated step by step by building a small-scale econometric model of the Norwegian economy, which incorporates the ability to analyze monetary policy; for example, an increase in the market rate, which shows the channels of the operation of monetary policy. Further empirical analysis of the New Keynesian Phillips curve provides an opportunity to illustrate their approach in another context. In summary, Bårdsen and Nymoen note that cointegration analysis takes into account non-stationarities that arise through unit roots, so that forecast failures are unlikely to be attributable to misspecification for that reason. In contrast to the econometric models of the 1970s, the real challenges arise from non-stationarities in functional relationships due to structural breaks; however,

there are ways to "robustify" the empirical model and forecasts from it so as to mitigate such possibilities, although challenges remain in an area that continues to be of central importance in economic policy.

One of the key developments in monetary policy in the UK and elsewhere in the last decade or so has been the move to give central banks a semi-autonomous status. In part, this was thought to avoid the endogenous "stop–go" cycle driven by political considerations. It also carried with it the implication that it was monetary policy, rather than fiscal policy, which would become the major macroeconomic policy tool, notwithstanding the now apparent practical limitations of such a move. In Chapter 18, Brian Henry provides an overview of the institutional and theoretical developments in the UK in particular, but with implications for other countries that have taken a similar route. The key question that is addressed in this chapter is whether regime changes, such as those associated with labor market reforms, inflation targeting and instrument independence for the Bank of England, have been the key factors in dampening the economic cycle and improving inflation, unemployment and output growth, or whether the explanation is more one of beneficial international events (the "good luck" hypothesis) and monetary policy mistakes. Henry concludes, perhaps controversially, that the reforms to the labor market and to the operation of the central bank are unlikely to have been the fundamental reasons for the improvement in economic performance. He provides an econometric basis for these conclusions, which incorporates a role for international factors such as real oil prices and measures of international competitiveness. Once these factors are taken into account, the "regime change" explanation loses force.

The growth of financial econometrics in the last two decades was noted in the first volume of this *Handbook*. Indeed, this development was recognized in the award of the 2003 Nobel Prize in Economics (jointly with Sir Clive Granger) to Robert Engle for "methods of analyzing economic time series with time-varying volatility (ARCH)." Part VII of this volume reflects this development and is thus devoted to applications in the area of financial econometrics.

In Chapter 19, George Dotsis, Raphael Markellos and Terence Mills consider continuous-time stochastic volatility models. What is stochastic volatility? To answer that question, we start from what it is not. Consider a simple model of an asset price, $Y(t)$, such as geometric Brownian motion, which in continuous time takes the form of the stochastic differential equation $dY(t) = \mu Y(t) + \sigma Y(t)dW(t)$, where $W(t)$ is a standard Brownian motion (BM) input; then σ (or σ^2) is the volatility parameter that scales the stochastic BM contribution to the diffusion of $Y(t)$. In this case the volatility parameter is constant, although the differential equation is stochastic. However, as Dotsis *et al.* note, a more appropriate specification for the accepted characteristics of financial markets is a model in which volatility also evolves stochastically over time. For example, if we introduce the variance function $v(t)$, then the simple model becomes $dY(t) = \mu Y(t) + \sqrt{v(t)}Y(t)dW(t)$, and this embodies stochastic volatility. Quite naturally, one can then couple this equation with one that models the diffusion over time of the variance function. ARCH/GARCH models are one way to model time-varying volatility, but there are

a number of other attractive specifications; for example, jump diffusions, affine diffusions, affine jump diffusions and non-affine diffusions. In motivating alternative specifications, Dotsis *et al*. note some key empirical characteristics in financial markets that underlie the rationale for stochastic volatility models, namely fat tails, volatility clustering, leverage effects, information arrivals, volatility dynamics and implied volatility. The chapter then continues by covering such issues as specification, estimation and inference in stochastic volatility models. A comparative evaluation of five models applied to the S&P 500, for daily data over the period 1990–2007, is provided to enable the reader to see some of the models "in action."

One of the most significant ideas in the area of financial econometrics is that the underlying stochastic process for an asset price is a martingale. Consider a stochastic process $X = (X_t, X_{t-1}, \ldots)$, which is a sequence of random variables; then the martingale property is that the expectation (at time $t-1$) of X_t, conditional on the information set $I_{t-1} = (X_{t-1}, X_{t-2}, \ldots)$, is X_{t-1}; that is, $E(X_t|I_{t-1}) = X_{t-1}$ (almost surely), in which case, X is said to be a martingale (the definition is sometimes phrased in terms of the σ-field generated by I_{t-1}, or indeed some other "filtration"). Next, define the related process $Y = (\Delta X_t, \Delta X_{t-1}, \ldots)$; then Y is said to be a martingale difference sequence (MDS). The martingale property for X translates to the property for Y that $E(Y_t|I_{t-1}) = 0$ (see, for example, Mikosch, 1998, sec. 1.5). This martingale property is attractive from an economic perspective because of its link to efficient markets and rational expectations; for example, in terms of X, the martingale property says that the best predictor, in a minimum mean squared error (MSE) sense, of X_t is X_{t-1}.

In Chapter 20, J. Carlos Escanciano and Ignacio Lobato consider tests of the martingale difference hypothesis (MDH). The MDH generalizes the MDS condition to $E(Y_t|I_{t-1}) = \mu$, where μ is not necessarily zero; it implies that past and current information (as defined in I_t) are of no value, in an MSE sense, in forecasting future values of Y_t. Tests of the MDH can be seen as being translated to the equivalent form given by $E[(Y_t - \mu)w(I_{t-1})]$, where $w(I_{t-1})$ is a weighting function. A useful means of organizing the extant tests of the MDH is in terms of the type of functions $w(.)$ that are used. For example, if $w(I_{t-1}) = Y_{t-j}, j \geq 1$, then the resulting MDH test is of $E[(Y_t - \mu)Y_{t-j}] = 0$, which is just the covariance between Y_t and Y_{t-j}. This is just one of a number of tests, but it serves to highlight some generic issues. In principle, the condition should hold for all $j \geq 1$ but, practically, j has to be truncated to some finite value. Moreover, this is just one choice of $w(I_{t-1})$, whereas the MDH condition is not so restricted. Escanciano and Lobato consider issues such as the nature of the conditioning set (finite or infinite), robustifying standard test statistics (for example, the Ljung–Box and Box–Pierce statistics), and developing tests in both the time and frequency domains; whilst standard tests are usually of linear dependence, for example autocorrelation based tests, it is important to consider tests based on nonlinear dependence. To put the various tests into context, the chapter includes an application to four daily and weekly exchange rates against the US dollar. The background to this is that the jury is out in terms of a judgment on the validity of the MDH for such data; some studies have found against the

MDH, whereas others have found little evidence against it. In this context, applying a range of tests, Escanciano and Lobato find general support for the MDH.

Chapter 19 by Dotsis *et al*. was concerned with models of stochastic volatility, primarily using the variance as a measure of volatility. Another measure of volatility is provided by the range of a price; for example, the trading day range of an asset price. In turn, the range can be related to the interval between consecutive trades, known as the duration. Duration is a concept that is familiar from counting processes, such as the Poisson framework for modeling arrivals (for example, at a supermarket checkout or an airport departure gate).

Chapter 21 by Ruey Tsay provides an introduction to modeling duration that is illustrated with a number of financial examples. That duration can carry information about market behavior is evident not only from stock markets, where a cluster of short durations indicates active trading relating to, for example, information arrival, but from many other markets; for example, durations in the housing market and their relation to banking failure. The interest in durations modeling owes much to Engle and Russell (1998), who introduced the autoregressive conditional duration (ACD) model for irregularly spaced transactions data. Just as the ARCH/GARCH family of models was introduced to capture volatility clusters, the ACD model captures short-duration clusters indicating the persistence of periods of active trading, perhaps uncovering and evaluating information arrivals. To see how an ACD model works, let the ith duration be denoted $x_i = t_i - t_{i-1}$, where t_i is the time of the ith event, and model x_i as $x_i = \psi_i \varepsilon_i$, where $\{\varepsilon_i\}$ is an i.i.d sequence and $\beta(L)\psi_i = \alpha_0 + \alpha(L)x_i$, where $\alpha(L)$ and $\beta(L)$ are lag polynomials; this is the familiar GARCH form, but in this context it is known as the exponential ACD or EACD. To accommodate the criticism that the hazard function of duration is not constant over time, unlike the assumption implicit in the EACD model, alternative innovation distributions have been introduced, specifically the Weibull and the Gamma, leading to the Weibull ACD (WACD) and the Gamma ACD (GACD). The chapter includes some motivating examples. Evidence of duration clusters is shown in Figures 21.1, 21.4 and 21.7a for IBM stock, Apple stock and General Motors stock, respectively. The development and application of duration models can then exploit the development of other forms of time series models, such as (nonlinear) threshold autoregressive (TAR) models. ACD models have also been developed to incorporate explanatory variables; an example is provided, which shows that the change to decimal "tick" sizes in the US stock markets reduced the price volatility of Apple stock.

The determination of exchange rates has long been an interest to econometricians and, as a result, there is an extensive literature that includes two constituencies; on the one hand, there have been contributions from economists who have employed econometric techniques and, on the other, to risk a simple bifurcation, the modeling of exchange rates has become an area to test out advances in nonlinear econometrics. Chapter 22, by Efthymios Pavlidis, Ivan Paya and David Peel, provides an evaluative overview of this very substantial area. As they note, the combination of econometric developments, the availability of high-quality and high-frequency data, and the move to floating exchange rates in 1973, has led

to a considerable number of empirical papers in this area. Thus, the question of "Where are we now?" is not one with a short answer. Perhaps prototypically, the econometrics of exchange rates is an area that has moved in tandem with developments in the economic theory of exchange rates (for the latter, the reader is referred to, for example, Sarno and Taylor, 2002). An enduring question over the last thirty years (at least), and one that is touched upon in two earlier chapters (Juselius, Chapter 8, and Gil-Alana and Hualde, Chapter 10), has been the status of PPP, regarded as a bedrock of economic theory and macroeconomic models. One early finding that has puzzled many is the apparent failure to find PPP supported by a range of different exchange rates and sample periods. Consider a stylized version of the PPP puzzle: there are two countries, with a freely floating exchange rate, flexible prices (for tradable goods and services), no trade constraints, and so on. In such a situation, at least in the long run, the nominal exchange rate should equal the ratio of the (aggregate) price levels, otherwise, as the price ratio moves, the nominal exchange rate does not compensate for such movements and the real exchange rate varies over time, contradicting PPP; indeed, on this basis the exchange rate is not tied to what is happening to prices. Early studies used an essentially linear framework – for example, ARMA models combined with unit root tests – to evaluate PPP, and rarely found that it was supported by the data; moreover, estimated speeds of adjustment to shocks were so slow as to be implausible. Another puzzle, in which tests indicated that the theory (of efficient speculative markets) was not supported, was the "forward bias puzzle." In this case, the prediction was that prices should fully reflect publicly available information, so that it should not be possible to make a systematic (abnormal) return; however, this appeared not to be the case. In this chapter, Pavlidis *et al*. carefully dissect this and other puzzles and show how the move away from simple linear models to a range of essentially nonlinear models, the development and application of multivariate models, and the use of panel data methods, has provided some explanation of the exchange rate "puzzles."

Part VIII of this volume of the *Handbook* is comprised of three chapters related to what has become referred to as "growth econometrics"; broadly speaking, this is the area that is concerned with variations in growth rates and productivity levels across countries or regions. Chapters 23 and 24 are a coordinated pair by Steven Durlauf, Paul Johnson and Jonathan Temple; in addition, looking ahead, Chapter 27 by Serge Rey and Julie Le Gallo takes up aspects of growth econometrics, with an emphasis on spatial connections. In Chapter 23, Durlauf *et al*. focus on the econometrics of convergence. Of course, convergence could and does mean a number of things: first, the convergence of what? Usually this is a measure of income or output but, in principle, the question of whether two (or more) economies are/have converged relates to multiple measures, for example, output, inflation, unemployment rates, and so on, and possibly includes measures of social welfare, such as literacy and mortality rates. The first concept to be considered in Chapter 23 is β-convergence (so-called because the key regression coefficient is referred to as β): consider two countries; there is β-convergence if the one with a lower initial income grows faster than the other and so "catches up" with the higher-income country.

Naturally, underlying the concept of convergence is an economic model, typically a neoclassical growth model (with diminishing returns to capital and labor), which indicates the sources of economic growth and a steady-state which the economy will (eventually) attain. At its simplest, growth econometrics leads to cross-country regressions of output growth rates on variables motivated from the underlying growth model and, usually, some "control" variables that, additionally, are thought to influence the growth rate. It is the wide range of choice for these control variables, and the resultant multiplicity of studies, that has led to the, perhaps pejorative, description of this activity as the "growth regression industry." One response has been the technique of model averaging, so that no single model will necessarily provide the empirical wisdom. A second central convergence concept is σ-convergence. As the notation suggests, this form of convergence relates to the cross-section dispersion of a measure, usually log per capita output, across countries. As Durlauf *et al*. note, whilst many studies use the log variance, other measures, such as the Gini coefficient or those suggested in Atkinson (1970), may be preferred. In this measure of convergence, a reduction in the dispersion measure across countries suggests that they are getting closer together. As in Chapter 22 on exchange rates, an important methodological conclusion of Durlauf *et al*. is that nonlinearity (due in this case to endogenous growth models) is likely to be an important modeling characteristic, which is not well captured in many existing studies, whether based on cross-section or panel data.

Having considered the question of convergence in Chapter 23, in Chapter 24 Durlauf *et al*. turn to the details of the methods of growth econometrics. Whilst concentrating on the methods, they first note some salient facts that inform the structure of the chapter. Broadly, these are that: vast income disparities exist despite the general growth in real income; distinct winners and losers have begun to emerge; for many countries, growth rates have tended to slow, but the dispersion of growth rates has increased. At the heart of the growth literature is the one-sector neoclassical growth model, transformed to yield an empirical form in terms of the growth rate of output per labor unit, such that growth is decomposed into growth due to technical progress and the gap between initial output per worker and the steady-state value. Typically, an error is then added to a deterministic equation derived in this way and this forms the basis of a cross-country regression, usually augmented with "control" variables that are also thought to influence growth rates. However, as Durlauf *et al*. note, there are a number of problems with this approach; for example, the errors are implicitly assumed to be exchangeable, but country dependence of the errors violates this assumption; the plethora of selected control variables leads to a multiplicity of empirical models; and parameter heterogeneity. To assess the question of model uncertainty an extreme bounds analysis (Leamer, 1983) can be carried out, and model averaging as in a Bayesian analysis can be fruitful. Parameter heterogeneity is related to the Harberger (1987) criticism that questions the inclusion of countries with different characteristics in a cross-country regression. The key to criticisms of this nature is the meaning of such regressions: is there a DGP that these regressions can be taken as empirically parameterizing? The chapter continues by providing, *inter alia*, an overview of the

different kinds of data that have been used and an assessment of the econometric problems that have arisen and how they have been solved; the conclusion evaluates the current state of growth econometrics, and suggests directions for future research.

A concern that has long antecedents is the relationship between financial development and growth: is there a causal relationship from the former to the latter? In Chapter 25, Thorsten Beck evaluates how this key question has been approached from an econometric perspective. Do financial institutions facilitate economic growth, for example by reducing information asymmetries and transaction costs? Amongst other functions, as Beck notes, financial institutions provide payment services, pool and allocate savings, evaluate information, exercise corporate governance and diversify risk. It would seem, *a priori*, that the provision of such services must surely move out the aggregate output frontier. However, just finding positive correlations between indicators of financial development, such as monetization measures, the development of banking institutions and stock markets, and economic growth is insufficient evidence from an econometric viewpoint. One of the most fundamental problems in econometrics is the problem of identification: by themselves, the correlations do not provide evidence of a causal direction. Beck takes the reader through the detail of this problem and how it has been approached in the finance-growth econometric literature. A classical method for dealing with endogenous regressors is instrumental variables (IV) and, in this context, some ingenuity has been shown in suggesting such variables, including exogenous country characteristics; for example, settler mortality, latitude and ethnic fractionalization. Early regression-based studies used cross-section data on a number of countries; however, more recent datasets now include dynamic panels and methods include GMM and cointegration. More recent developments have been able to access data at the firm and household level, and this has led to much larger samples being used. For example, Beck, Dermirgüç-Kunt and Makisimovic (2005) use a sample of over 4,000 firms in 54 countries to consider the effect of sales growth as a firm-level financing obstacle as well as other variables, including a country-level financial indicator. As Beck notes, the evidence suggests a strong case for a causal link between financial development and economic growth, but there is still much to be done both in terms of techniques, such as GMM, and exploiting advances at the micro-level.

In Volume 1 of the *Handbook*, we highlighted recent developments in theoretical econometrics as applied to problems with a spatial dimension; this is an area that has grown in application and importance, particularly over the last decade, and it is natural that we should continue to emphasize its developmental importance by including two chapters in Part IX. These chapters show how spatial econometrics can bring into focus the importance of the dimension of space in economic decisions and the particular econometric problems and solutions that result. In Chapter 26, Luc Anselin and Nancy Lozano-Gracia consider spatial hedonic models applied to house prices. Hedonic price models are familiar from microeconomics and, in particular, from the seminal contributions of Lancaster (1966) and Rosen (1974). In the context of house prices, there are key characteristics, such as aspects

of neighborhood, proximity to parks, schools, measures of environmental quality, and so on, that are critical in assigning a value to a house. These characteristics lead to the specification of a hedonic price function to provide an estimate of the marginal willingness to pay (MWTP) for a characteristic; a related aim, but one not so consistently pursued, is to retrieve the implied inverse demand function for house characteristics. Two key problems in the estimation of hedonic house price functions, in particular, are spatial dependence and spatial heterogeneity. As Anselin and Lozano-Gracia note, spatial dependence, or spatial autocorrelation, recognizes the importance of geographical or, more generally, network space in leading to a structure in the covariance matrix between observations. Whilst there is an analogy with temporal autocorrelation, spatial autocorrelation is not simply an extension of that concept, but requires its own conceptualization and methods. Spatial heterogeneity can be viewed as a special case of structural instability; two (of several) examples of heterogeneity are spatial regimes (for example, ethnically based sub-neighorhoods) and spatially varying coefficients (for example, different valuations of housing and neighborhood characteristics). In this chapter, Anselin and Lozano-Gracia provide a critical overview of methods, such as spatial two-stage least squares and spatial feasible GLS, a summary of the literature on spatial dependence and spatial heterogeneity, and discussion of the remaining methodological challenges.

In Chapter 27, Serge Rey and Julie Le Gallo consider an explicitly spatial analysis of economic convergence. Recall that Chapter 23, by Durlauf *et al.*, is concerned with the growing interest in the econometrics of convergence; for example, whether there was an emergence of convergence clubs, perhaps suggesting "winners and losers" in the growth race. There is an explicitly spatial dimension to the evaluation of convergence; witness, for example, the literature on the convergence of European countries or regions, the convergence of US states, and so on. Rey and Le Gallo bring this spatial dimension to the fore. The recognition of the importance of this dimension brings with it a number of problems, such as spatial dependence and spatial heterogeneity; these problems are highlighted in Chapter 26, but in Chapter 27 they are put in the context of the convergence of geographical units. Whilst Rey and Le Gallo consider what might be regarded as purely econometric approaches to these problems, they also show how exploratory data analysis (EDA), extended to the spatial context, has been used to inform the theoretical and empirical analysis of convergence. As an example, a typical focus in a non-spatial context is on σ-convergence, which relates to a cross-sectional dispersion measure, such as the variance of log per capita output, across regions or countries. However, in a broader context, there is interest in the complete distribution of regional incomes and the dynamics of distributional change, leading to, for example, the development of spatial Markov models, with associated concepts such as spatial mobility and spatial transition. EDA can then provide the tools to visualize what is happening over time: see, for example, the space-time paths and the transition of regional income densities shown in Figures 27.5 and 27.6. Rey and Le Gallo suggest that explicit recognition of the spatial dimension of convergence, combined with the use of EDA and its extensions to include the spatial element,

offers a fruitful way of combining different methods to inform the overall view on convergence.

Part X comprises two chapters on applied econometrics and its relationship to computing. In Chapter 28, Bruce McCullough considers the problem of testing econometric software. The importance of this issue is hard to understate. Econometric programs that are inaccurate, for any reason, will produce misleading results not only for the individual researcher but, if published, for the profession more generally, and will lead to applications that are impossible to replicate. The development of sophisticated methods of estimation means that we must also be ever-vigilant in ensuring that software meets established standards of accuracy. A seminal contribution to the development of accuracy benchmarks was Longley (1967). As McCullough notes, Longley worked out by hand the solution to a linear regression problem with a constant and six explanatory variables. When run through the computers of the time, he found that the answers were worryingly different. Of course, the Longley benchmark is now passed by the econometric packages that are familiar to applied econometricians. However, the nature of the problems facing the profession is different (sophisticated estimators, large datasets, simulation-based estimators) and McCullough's results imply that there is no reason for complacency. Many econometric estimators involve problems of a nonlinear nature – for example, the GARCH and multivariate GARCH estimators and the probit estimator – and it is in the case where a nonlinear solver is involved that the user will find problems, especially when relying on the default options. Another area that has seen substantial growth in the last two decades has been the use of Monte Carlo experimentation, an area that makes fundamental use of random numbers, and hence any package must have a reliable random number generator (RNG). But are the numbers so generated actually random? The answer is, not necessarily! (The reader may wish to refer to Volume 1 of this *Handbook*, which includes a chapter by Jurgen Doornik on random number generation.) The importance of maintaining standards of numerical accuracy has been recognised in the National Institute of Standards and Technology's Statistical Reference Datasets, which has resulted in a number of articles using these datasets to evaluate software for econometric problems. To illustrate some of the issues in software evaluation, for example in establishing a benchmark, McCullough includes a study of the accuracy of a number of packages to estimate ARMA models. The central methods for the estimation of such models include unconditional least squares (UCLS), conditional least squares (CLS), and exact maximum likelihood. The questions of interest are not only in the accuracy of the point estimates from these methods in different packages, but also what method of standard error calculation is being used. Overall, McCullough concludes that we, as a profession, have some way to go in ensuring that the software that is being used is accurate, that the underlying methods are well-documented, and that published results are replicable.

In Chapter 29, Marius Ooms takes a historical perspective on the nature of applied econometrics as it has been represented by publications and reviews of econometric and statistical software in the *Journal of Applied Econometrics* (*JAE*).

Over the 14-year review period, 1995–2008, there were 513 research articles published in the *JAE*, of which 253 were categorized as applications in time series, 140 as panel data applications and 105 as cross-section applications. Ooms notes that there has been a gradual shift from macroeconometrics to microeconometrics and applications using panel data. The software review section of the *JAE* has been a regular feature, so enabling an analysis of the programmes that have been in use – and continue to be in use, reflecting the development policy of the providers. This section is likely to be a very useful summary for research and teaching purposes. Ooms also notes the growth of high-level programming languages, such as Gauss, MATLAB, Stata and Ox, and illustrates their use with a simple program. In combination, the profession is now very much better served for econometric software than it was twenty years ago. Of course, these developments have not taken place in isolation but rather as a response to developments in theoretical and applied econometrics. A leading example in this context, noted by Ooms, is the Arellano and Bond (1991) approach to the estimation of applications using panel data (dynamic panel data, or DPD, analysis), which led to the widespread implementation of new code in existing software and many new applications; an example in the area of time series applications is the growth of ARCH and GARCH-based methods and the implantation of estimation routines in econometric software. As noted in Chapter 28 by McCullough, reproducibility of results is a key aspect in the progression and reputation of applied econometrics. Results that are irreproducible by reason of either inaccurate software or unavailability of data will do long-term harm to the profession. In this respect, the *JAE*, through Hashem Pesaran's initiative, has been a leader in the context of requiring authors to provide the data and code which they used. The *JAE* archive is indexed and carefully managed, and provides the standard for other journals.

As a final comment, which we hope is evident from the chapters contained in this volume, one cannot help but be struck by the incredible ingenuity of those involved in pushing forward the frontiers of applied econometrics. Had this volume been compiled even, say, just twenty years ago, how different would it have been! Viewed from above, the landscape of applied econometrics has changed markedly. Time series econometrics and macroeconometrics, whilst still important, are not predominant. The combination of the availability of large datasets of a microeconomic nature, combined with enormous increases in computing power, has meant that econometrics is now applied to a vast range of areas. What will the next twenty years bring?

Finally, thanks are due to many in enabling this volume to appear. First, our thanks go collectively to the authors who have cooperated in contributing chapters; they have, without exception, responded positively to our several and sometimes many requests, especially in meeting deadlines and accommodating editorial suggestions. We hope that the quality of these chapters will be an evident record of the way the vision of the *Handbook* has been embraced. We would also like to record our gratitude to the Advisory Editors for this volume: Bill Greene, Philip Hans Franses, Hashem Pesaran and Aman Ullah, whose support was invaluable, especially at an early stage.

Thanks also go the production team at Palgrave Macmillan, only some of whom can be named individually: Taiba Batool, the commissioning editor, Ray Addicott, the production editor, and Tracey Day, the indefatigable copy-editor. A special mention goes to Lorna Eames, secretary to one of the editors, for her willing and invaluable help at several stages in the project.

References

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 177–97.

Atkinson, A.B. (1970) On the measurement of inequality. *Journal of Economic Theory* **2**, 244–63.

Bates, J.M. and C.W.J. Granger (1969) The combination of forecasts. *Operations Research Quarterly* **20**, 451–68.

Beck, T., A. Demirgüç–Kunt and V. Maksimovic (2005) Financial and legal constraints to firm growth: does firm size matter? *Journal of Finance* **60**, 137–77.

Clements, M.P. and D.F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Doornik, J.A. (2007) Autometrics. Working paper, Economics Department, University of Oxford.

Engle, R.F. and C.W.J. Granger (1987) Co-integration and error-correction: representation, estimation and testing. *Econometrica* **55**, 251–76.

Engle, R.F. and J.R. Russell (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–62.

Ericsson, N.R. and J.S. Irons (1995) The Lucas critique in practice: theory without measurement. In K.D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, Ch. 8. Dordrecht: Kluwer Academic Publishers.

Giacomini, R. and H. White (2006) Tests of conditional predictive ability. *Econometrica* **74**, 1545–78.

Granger, C.W.J. and R. Joyeux (1980) An introduction to long memory time series and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.

Harberger, A. (1987) Comment in S. Fischer (ed.), *Macroeconomics Annual* 1987. Cambridge, Mass.: MIT Press.

Hendry, D.F. (1980) Econometrics: alchemy or science? *Economica* **47**, 387–406.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231–54.

Koop, G., R. Strachan, H. van Dijk and M. Villani (2006) Bayesian approaches to cointegration. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, pp. 871–900.

Lancaster, K.J. (1966) A new approach to consumer theory. *Journal of Political Economy* **74**, 132–56.

Leamer, E. (1983) Let's take the con out of econometrics. *American Economic Review* **73**, 31–43.

Longley, J.W. (1967) An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association* **62**, 819–41.

Lucas, R.E. (1976) Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labour Markets*, Volume 1 of *Carnegie-Rochester Conferences on Public Policy*, pp. 19–46. Amsterdam: North-Holland.

Magnus, J.R. and M.S. Morgan (eds.) (1999) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*. Chichester: John Wiley and Sons.

Mikosch, T. (1998) *Elementary Stochastic Calculus*. London and New Jersey: World Scientific Publishers.

Nelson, C.R. and C.I. Plosser (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* **10**, 139–62.

Perron, P. (1989) The great crash, the oil price shock and the unit root hypothesis. *Econometrica* **57**, 1361–401.

Poirier, D.J. and J.L. Tobias (2006) Bayesian econometrics. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, pp. 841–70. Basingstoke: Palgrave Macmillan.

Rosen, S.M. (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* **82**, 534–57.

Sarno, L. and M. Taylor (2002) *The Economics of Exchange Rates*. Cambridge and New York: Cambridge University Press.

Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* **48**, 1–48.

Sims, C.A. (2002) Solving linear rational expectations models. *Computational Economics* **20**, 1–20.

Tobin, J. (1950) A survey of the theory of rationing. *Econometrica* **26**, 24–36.

1

The Methodology of Empirical Econometric Modeling: Applied Econometrics Through the Looking-Glass

David F. Hendry

Abstract

This chapter considers the methodology of empirical econometric modeling. The historical background is reviewed from before the Cowles Foundation to the rise of economic theory-based econometrics and the decline of data concerns. A theory for "Applied Econometrics" suggests reinterpreting the role of economic theory given that the intrinsic non-stationarity of economic data vitiates analyses of incomplete specifications based on *ceteris paribus*. Instead, the many steps from the data-generation process (DGP) through the local DGP (LDGP) and general unrestricted model to a specific representation allow an evaluation of the main extant approaches. The potential pitfalls confronting empirical research include inadequate theory, data inaccuracy, hidden dependencies, invalid conditioning, inappropriate functional form, non-identification, parameter non-constancy, dependent, heteroskedastic errors, wrong expectations formation, misestimation and incorrect model selection. Recent automatic methods help resolve many of these difficulties. Suggestions on the teaching of "Applied Econometrics" are followed by revisiting and updating the "experiment in applied econometrics" and by automatic modeling of a four-dimensional vector autoregression (VAR) with 25 lags for the numbers of bankruptcies and patents, industrial output per capita and real equity prices over 1757–1989.

1.1 Introduction

> "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!" (Quote from the Red Queen in *Through the Looking-Glass and What Alice Found There*, Lewis Carroll, Macmillan & Co., 1899, [henceforth cited as "Lewis Carroll, 1899"])

Most econometricians feel a bit like Alice did at having to run fast even to stand still. Handbooks are an attempt to alleviate the problem that our discipline moves forward rapidly, and *infoglut* can overwhelm, albeit that one has to run even faster for a short period to also find time to read and digest their contents. That will require some sprinting here, given that the contents of this *Handbook of Econometrics* provide up-to-date coverage of a vast range of material: time series, cross-sections, panels, and spatial; methodology and philosophy; estimation – parametric and nonparametric – testing, modeling, forecasting and policy; macro, micro, finance, growth and development; and computing – although I do not see *teaching*. Such general headings cross-categorize "Applied Econometrics" by types of data and their problems on the one hand – time series, cross-sections, panels, high frequency (see, e.g., Barndorff-Nielsen and Shephard, 2007), limited dependent variables (see, e.g., Heckman, 1976), or count data (excellently surveyed by Cameron and Trivedi, 1998), etc. – and by activities on the other (modeling, theory calibration, theory testing, policy analysis, forecasting, etc.). The editors considered that I had written on sufficiently many of these topics during my career to "overview" the volume, without also noting how markedly all of them had changed over that time. The

main aim of an introductory chapter is often to overview the contents of the volume, but that is manifestly impossible for the *Handbook of Econometrics* given its wide and deep coverage. In any case, since the *Handbook* is itself an attempt to overview Applied Econometrics, such an introduction would be redundant.

Thus, my focus on empirical econometric modeling concerns only one of the activities, but I will also try to present an interpretation of what "Applied Econometrics" is; what those who apply econometrics may be trying to achieve, and how they are doing so; what the key problems confronting such applications are; and how we might hope to resolve at least some of them. Obviously, each aspect is conditional on the previous one: those aiming to calibrate a theory model on a claimed set of "stylized facts" are aiming for very different objectives from those doing data modeling, so how they do so, and what their problems are, naturally differ greatly. This chapter will neither offer a comprehensive coverage, nor will it be an uncontroversial survey. En route, I will consider why "Applied Econometrics" does not have the highest credibility within economics, and why its results are often attacked, as in Summers (1991) among many others (see Juselius, 1993, for a reply). Evidence from the contents of textbooks revealing the marginal role of "Applied Econometrics" and "economic statistics" within the discipline has been provided recently by Qin (2008) and Atkinson (2008) respectively. Since two aspects of our profession with even lower status than "Applied Econometrics" are data (measurement, collection and preparation), and teaching, I will try and address these as well, as they are clearly crucial to sustaining and advancing a viable "Applied Econometrics" community. Economic forecasting and policy are not addressed explicitly, being uses of empirical models, and because the criteria for building and selecting such models differ considerably from those applicable to "modeling for understanding" (see, e.g., Hendry and Mizon, 2000; and for complete volumes on forecasting, see Clements and Hendry, 2002a, 2005; Elliott, Granger and Timmermann, 2006).

Economists have long been concerned with the status of estimated empirical models. How a model is formulated, estimated, selected and evaluated all affect that status, as do data quality and the relation of the empirical model to the initial subject-matter theory. All aspects have been challenged, with many views still extant. And even how to judge that status is itself debated. But current challenges are different from past ones – partly because some of the latter have been successfully rebutted. All empirical approaches face serious problems, yet the story is one of enormous progress across uncharted terrain with many mountains climbed – but many more to surmount. I will recount some of that story, describe roughly where we are presently located, and peer dimly into the future. Why "*Applied Econometrics Through the Looking-Glass*"? Lewis Carroll was the pseudonym for Charles Dodgson, a mathematician who embodied many insights in the book which is cited throughout the present chapter: a Looking-Glass is a mirror, and applied findings in economics can only reflect the underlying reality, so obtaining a robust and reliable reflection should guide its endeavors.

Following the brief section 1.2 on the meaning of the topic, section 1.3 summarizes some of the history of our fallible discipline. Then section 1.4 proposes a

"theory of Applied Econometrics" which highlights some of the problems empirical modeling confronts in a non-stationary environment, where non-stationary is used throughout in the "wide sense" to denote any changes in the distributions of the random variables modeled by economists. Section 1.5 discusses a recent tool for automatic modeling, Autometrics, based on the last decade of research into model selection (see Doornik, 2007a; Hendry and Krolzig, 2005; Hendry, with Doornik and Nielsen, 2007). Section 1.6 comments on teaching Applied Econometrics, and section 1.7 revisits the experiment in applied econometrics conducted by Magnus and Morgan (1999). Section 1.8 then looks at automatic modeling of a four-variable dynamic system related to industrial output since 1700, with 25 lags in its initial formulation and many outliers over more than 250 years. Section 1.9 concludes. Throughout, I draw heavily on a number of my previous papers. Despite there being almost 200 citations to other scholars, I am conscious that documentation is bound to be incomplete, and apologize for omitting many contributions.

1.2 What is "Applied Econometrics"?

> "When I use a word," Humpty Dumpty said in rather a scornful tone, "it means just what I choose it to mean – neither more nor less." (Lewis Carroll, 1899)

At the superficial level, "Applied Econometrics" is "any application of econometrics," as distinct from theoretical econometrics. If it were not for the imperialist tendencies of econometricians, that would suffice, but econometrics has been applied in space science, climatology, political science, sociology, epidemiology, marketing, *inter alia*, not to mention the claim in *How the Laws of Physics Lie* (see Cartwright, 1983) that econometrics is the key methodology for all of science ... Sorry to disappoint the eager reader, but I will not be covering even a wide range of the economic applications, never mind that plethora of outside studies.

Some applied econometricians would include any applications involving analyses of "real economic data" by econometric methods, making "Applied Econometrics" synonymous with empirical econometrics. However, such a view leads to demarcation difficulties from applied economics on the one hand and applied statistics on the other. Defining "econometrics," as in Frisch (1933), to comprise only studies involving the unification of economic theory, economic statistics (data), and mathematics (statistical methods) helps in demarcation, but limits its scope and inadvertently excludes (say) developing econometric theory itself, or just improving data measurement and collection.

Outsiders might have thought that "Applied Econometrics" was just the application of econometrics to data, but that is definitely not so; virtually no journal editor would publish such a piece. Rather, the notion of mutual penetration dominates – but as a one-way street. Economic theory comes first, almost mandatorially. Perhaps this just arises from a false view of science, namely that theory precedes evidence, even though, apart from a few famous occasions, science rarely proceeds by imposing a preconceived theory on evidence, and evidence regularly shapes and

stimulates theory. Yet the latter is rarely the case in applied econometrics – and to see how we arrived at such a state, we need to consider the contingent history of our discipline.

1.3 Historical background

> "The time has come," the Walrus said, "To talk of many things: Of shoes – and ships – and sealing-wax – Of cabbages – and kings – And why the sea is boiling hot – And whether pigs have wings." (Lewis Carroll, 1899)

The histories of statistics and econometrics are now reasonably well documented: on the former, see, e.g., the books by Stigler (1986, 1999) and Hald (1990, 1998); and for the latter, see Epstein (1987), Morgan (1990), Qin (1993), Klein (1997), and Le Gall (2007); also see Christ (1994), Spanos (2006), Farebrother (2006) and Gilbert and Qin (2006); and for reprints of key papers, see Darnell (1994) and Hendry and Morgan (1995), with related material in Caldwell (1993), Hamouda and Rowley (1997), Mills (1999) and Campos, Ericsson and Hendry (2005). These books provide overall bibliographic perspective.

1.3.1 Pre-Cowles

> The shop seemed to be full of all manner of curious things – but the oddest part of it all was, that whenever she looked hard at any shelf, to make out exactly what it had on it, that particular shelf was always quite empty: though the others round it were crowded as full as they could hold. (Lewis Carroll, 1899)

An aspect of that history which is still somewhat under-emphasized, despite being stressed by Hendry and Morgan (1995), is the role that empirical studies have played as a driver of new econometric concepts, theories and methods, standing in some contrast to its direct impact on economics. Certainly, early attempts were replete with what we would now view as blunders – William Stanley Jevons' sunspot, and Henry Moore's Venus, theories of business cycles are regularly trotted out as examples of how silly econometricians can be, yet Jevons (1875) and Moore (1923) respectively need to be contrasted with other careful and insightful empirical analyses in Jevons' research, edited by Foxwell (1884), and in Moore (1911). On the former, see the appraisal in Peart (2001); and on the latter, e.g., Stigler (1962) comments: "Moore's standard of craftsmanship is high: the basic data are fully reported and the work was carefully done." Also, note the cited comment from Alfred Marshall to Moore in 1912 that "the 'ceteris paribus' clause – though formally adequate seems to me impracticable,"a point that will recur below. The "upward sloping demand curve" for pig-iron in Moore (1914) is perhaps the most notorious misinterpretation, but in fact led to many later insights – in particular, reactions to it helped unravel the whole complicated and intertwined issues

of simultaneity, identification, exogeneity, and partial effects (see Wright, 1915, 1929; Working, 1927; Tinbergen, 1930, *inter alia*).

Equally importantly, many "strange" empirical correlations had been found that stimulated the unraveling of both spurious, and later nonsense, regressions in works such as Yule (1897), Hooker (1901), and especially the famous explanation in Yule (1926), leading first to a distinction between short-run and long-term relationships, then unit roots, and eventually cointegration, as in Granger (1981), and dozens of later contributions surveyed in Hendry and Juselius (2000, 2001). Despite obvious progress – Stigler (1962) begins his article about Moore by "If one seeks distinctive traits of modern economics, traits which are not shared to any important degree with the Marshallian or earlier periods, he will find only one: the development of statistical estimation of economic relationships" – trouble lay ahead.

The attack by Robbins (1932) on the empirical studies of Schultz (1928) – portrayed as the feckless Dr. Blank studying the demand for herring (rather than sugar) – was the first of several critiques which sought to deny any substantive role for econometrics in economics.[1] Tinbergen's attempts to build empirical models of investment activity brought down the wrath of John Maynard Keynes (see Tinbergen, 1939, 1940; Keynes, 1939, 1940), who insisted that the economist had to be "in the saddle" with the econometrician as the "patient ass," and sarcastically demanded that Tinbergen satisfy:

> an experiment on his part. It will be remembered that the seventy translators of the Septuagint were shut up in seventy separate rooms with the Hebrew text and brought out with them, when they emerged, seventy identical translations. Would the same miracle be vouchsafed if seventy multiple correlators were shut up with the same statistical material? And anyhow, I suppose, if each had a different economist perched on his *a priori*, that would make a difference to the outcome.

We will return in section 1.4 both to that issue, which may well now be possible, and to Keynes' general claims – one might like to ponder whether 70 economic theorists asked to tackle the same puzzle would derive precisely the same model? As ever, other more constructive outcomes followed from that debate, especially the memorandum by Frisch (1938), and it certainly did not discourage Haavelmo (1944).

1.3.2 War and post-war

> "I'll tell you all my ideas about Looking-glass House. First, there's the room you can see through the glass – that's just the same as our drawing-room, only the things go the other way." (Quote from Alice in Lewis Carroll, 1899)

Despite Koopmans (1937) being a key precursor to the establishment of modern econometrics in Haavelmo (1944), the attack by Koopmans (1947) on Burns and

Mitchell (1946) allowed the assertion of "measurement without theory" to become capable of dismissing empirical work without further serious consideration. The vigorous reply by Vining (1949a, 1949b) still merits reading. With a few honorable exceptions (such as Atkinson, 2005), even the use of the word "measurement" as a title for economics' papers seems to have decreased since (other than in "measurement errors"). Tress (1959) offers a near contemporary analysis of the acrimony between economics and econometrics at that time, and a possible reconciliation.

Keynes (1939) had asserted that a long list of "preconditions" had to be satisfied to validate empirical inferences, implicitly arguing that empirical econometrics must fail unless everything was known in advance (see, e.g., Hendry, 1980). But if it was impossible to empirically uncover things not already known theoretically, then no science could have progressed; rather, scholasticism would still rule. There are several flaws in Keynes' claims, of which three are the most important.

First, "partial knowledge" can be valuable, and can be learnt from evidence, with or without prior theories, albeit being subject to revision later. Our understanding of gravity remains incomplete, but has advanced greatly since Aristotle's early view of objects' natural places (smoke rises, stones fall, as natural to go to heaven or the centre of the earth, respectively), through Ptolemy's epicycle theory of planetary motions, Descartes' vortex theory, and Newtonian inverse-square laws, which Adam Smith (1795) presciently noted was just a model and not the "truth" (as most of his contemporaries assumed). Einstein's relativity theory is still not the "final answer." Retrospectively, Aristotle's theory did not go beyond "explaining" the phenomena themselves, whereas Newtonian theory, while closely based on Kepler's laws of motion of planetary bodies, explained many additional aspects, so was a clear advance in knowledge, even if later it too was found to be incomplete, and at relativistic speeds, incorrect. Moreover, despite neither Aristotle's nor Newton's theories being "true," both were at least consistent with the observed facts of their time. The relevant empirical regularities persisted through many theories, which provided better explanations, often with unanticipated predictions of new phenomena – genuine "mutual penetration." Thus, progress is the key to science, not one-off forging of true laws that hold forever.

Second, if there are invariant features of reality – as in physics and chemistry – then empirical research can discover them without prior knowledge, as happened historically in many branches of science. Conversely, if nothing is invariant (an extreme of Heraclitus of Ephesus supposed view that "reality is change"), neither economic theories nor econometric models would be of any practical value (see Hendry, 1995b). Following Bachelier (1900), equity prices have long been viewed as close to random walks, which may be thought to entail the absence of any invariants, but if correct – as suggested by some modern theories and empirical tests of efficient markets – is actually a powerful invariant, as contrasted with a data generating process whose structure alters every period.

Third, empirical econometrics could still "advance" by rejecting economic theories. This would at least allow economists to focus on theories that were not yet rejected, if any, and improve those that faced discordant evidence. However,

progress might be somewhat inefficient when new theories are easily generated as variants of previous incarnations.

Similar comments apply to Koopmans' claim that "without resort to theory, in the sense indicated, conclusions relevant to the guidance of economic policies cannot be drawn." Such an argument is not sustainable in general. Originally, aspirin was a folk remedy for hangovers, derived from brewing willow-tree bark – of which acetylsalicylic acid, aspirin's active ingredient, is a natural constituent – without any theory as to how or why that policy intervention worked (see Weissmann, 1991). A less well known example is the use in folk medicine of fungal-based products, some of which contain natural antibiotics such as penicillin: over 3,000 years ago, the Chinese had used moldy soybean curd for treating skin infections, again with no theory on which to base that policy. Theories can be invaluable, and can enhance the credibility of proposed policies, but they are not essential, especially when they are incorrect.

1.3.3 The rise of economic theory-based econometrics

> "If you'll tell me what language 'fiddle-de-dee' is, I'll tell you the French for it!" she exclaimed triumphantly. (Quote from Alice to the Red Queen in Lewis Carroll, 1899)

Another critique of empirical modeling follows from the joint dependence of economic events, namely the resulting issues of endogeneity and identification. It seems widely believed that identification restrictions must be given *a priori* by economic theory, especially in simultaneous systems, yet that belief also does not have a substantive basis, as shown in section 1.4.7 on identification.

Together, the cumulative critiques just noted led to an almost monolithic approach to empirical econometric research: first postulate an individualistic, intertemporal optimization theory; next derive a model therefrom; third, find data with the same names as the theory variables; then select a recipe from the econometrics cookbook that appropriately blends the model and the data, or if necessary, develop another estimator; finally report the newly forged empirical law. Thus, we have a partial answer to the issue posed in section 1.2: the contingent history of econometrics suggested that the only viable route for applied research in economics, where all current-dated variables are potentially endogenous, was to provide the quantitative cloth for a completely pre-specified theoretical formulation derived from general economic principles. But that approach too is problematic and not without its critics. Economic theory has progressed dramatically over the past century – imagine being forced to impose 1900's economic theory today as the basis for empirical research. If you recoil in horror at that idea, then you have understood why much past Applied Econometrics research is forgotten: discard the economic theory that it quantified and you discard the empirical evidence. Instead of progress, we find fashions, cycles and "schools" in research. The problem is not that early pioneers ignored economic theory, but that the available theory was seriously incomplete – as it still is today. Indeed, the Cowles Commission research was essentially predicated on the belief that the relevant economic

theory already existed, so complicated issues of model choice could be avoided by imposing valid restrictions derived from correct economic theories: on discovering that such theory was not available, many turned to help develop it (see, e.g., Qin, 2008; Bjerkholt, 2007 (Bjerkholt, 2005, is a useful precursor). Koopmans, Hurwicz and Arrow all made major contributions to economic theory, and to quote Bjerkholt (2007): "Haavelmo stated later on various occasions that economic theory needed further development for the econometric methods to become fully applicable" (also see Moene and Rødseth, 1991). Indeed, to quote Haavelmo (1989) himself:

> The basis of econometrics, the economic theories that we had been led to believe in by our forefathers, were perhaps not good enough. It is quite obvious that if the theories we build to simulate actual economic life are not sufficiently realistic, that is, if the data we get to work on in practice are not produced the way that economic theories suggest, then it is rather meaningless to confront actual observations with relations that describe something else.

He reiterated that view in his presidential address to the Econometric Society (published as Haavelmo, 1958):

> What I believe to be true, however, is this: The training in the technical skills of econometrics can represent a powerful tool for imaginative speculation upon the basic phenomena of economic life; and, furthermore, it would be fruitful to bring the requirements of an econometric "shape" of the models to bear upon the formulation of fundamental economic hypotheses from the very beginning.

Once model choice cannot be avoided, methodology becomes a salient issue, and it would seem every conceivable methodology has at least one advocate. Pagan (1987) considered what he viewed as the three main econometric methodologies, relating mine to Leamer (1978) and Sims (1980), yet the ubiquitous "theory-based" approach was not mentioned, albeit that there are really many variants thereof.

1.3.4 The decline of data concerns

> "You're travelling the wrong way." (Train guard to Alice in Lewis Carroll, 1899)

At about the same time that *a priori* theory-based econometrics became dominant, data measurement and quality issues were also relegated as a central component of empirical publications. Early on, data series were often published in their entirety, with careful caveats about accuracy, but later, at best, were recorded in appendices, or not at all. For example, Clark (1932) allowed the first famous estimate of the size of the Keynesian "multiplier." Although computerized databases have recently started to compensate for the absence of the printed record, electronic data are sometimes revised in situ, making it difficult for later investigators to duplicate previously-published findings. In a detailed study of a number of cases, Atkinson

(2008) emphasizes that the outcomes reported would change substantively if data had been more carefully evaluated prior to the econometric analysis. To quote:

> my concern (is) with the status, within economics, of economic statistics. By "economic statistics," I mean the study of how we create, use and assess economic data – what one might call "data appreciation." ... It is true that economists are using empirical data to an unprecedented extent, and applying tools of great sophistication. Economics is a much more data-driven subject than it was in the past. But, I shall argue, economists have too often come to take data for granted, without critical examination of their strengths and weaknesses.

With that caveat about data firmly in mind, let us turn to methodology: measurement is reconsidered in section 1.4.3, and an illustration of some effects of substantial revisions in section 1.7.1.

The route ahead views all models as arising from reductions of whatever process generated the data, which is a combination of the economic outcome and the measurement system. We discuss these reductions in relation to their impact on the parameters that actually governed the economic decisions of the relevant agents. Most reductions occur implicitly, as investigators usually approach modeling from the opposite perspective, namely what to include in their analysis, although its success or failure will depend on whether the sub-set of variables considered allows a model to capture the salient and constant characteristics of the data-generating process (DGP). What to include and how to include it certainly depends on the economics behind the analysis; but what is found depends on the unknown data-generating process and the losses of information from the reductions that were necessary to derive the postulated model.

1.4 A theory of Applied Econometrics

> "Why, sometimes I've believed as many as six impossible things before breakfast." (Quote from the White Queen in Lewis Carroll, 1899)

If only it were just six! To believe that he or she has ascertained the "truth," an applied econometrician would have to believe at least the following dozen impossible (composite) assumptions:

1. a correct, complete, and immutable underlying economic theory derivation
2. a correct, comprehensive choice of all relevant variables, including all dynamic specifications
3. exact data measurements on every variable
4. the absence of any hidden dependencies, including collinearity and simultaneity
5. the validity and relevance of all conditioning variables (including instruments)
6. the precise functional forms for every variable
7. that all parameters of interest are identified in the resulting model specification

8. that all entities treated as parameters are constant over time, and invariant to all potentially omitted variables and regime changes
9. the errors have "independent," homoskedastic, distributions
10. all expectations formulations are correct, or agents' expectations are accurately measured
11. the choice of estimator is appropriate at relevant sample sizes
12. a valid and non-distortionary method of model selection is used.

If all of these assumptions had to be perfectly correct to produce useful empirical evidence, there would be no hope of ever doing so. In Hendry (1987), I suggested the four "golden prescriptions" of econometrics, abbreviated here as:

(i) think brilliantly: if you think of the right answer before modeling, then the empirical results will be optimal and, of course, confirm your brilliance;
(ii) be infinitely creative: if you do not think of the correct model before commencing, the next best is to think of it as you proceed;
(iii) be outstandingly lucky: if you do not think of the "true model" before starting nor discover it en route, then luckily stumbling over it before completing the study is the final sufficient condition. This may be the most practical of these suggestions. Failing this last prescription:
(iv) stick to doing theory.

Lest the reader thinks the list of a dozen requirements above is overly dramatic, or even new, Hendry and Morgan (1995) record:

> In the thesis as a whole, Koopmans (1937) assembles together and confronts most of the major issues in econometrics, which we have translated into current terminology as:
> 1. the joint occurrence of errors-in-variables and errors-in-equations
> 2. the need for a complete set of determining variables to leave an innovation error
> 3. a reductionist approach of proceeding from general to simple
> 4. the distinctions between the activities of specification, estimation and distribution, as spelt out by R.A. Fisher
> 5. the non-experimental nature of economic data
> 6. the need to condition on systematic components with independently varying error terms
> 7. the choice of functional form, using linearity for convenience
> 8. the formulation of the parameters of interest
> 9. the need to test underlying assumptions
> 10. the importance of incorporating all relevant information
> 11. the avoidance of unnecessary assumptions
> 12. the need for the general model to be estimable
> 13. the need for the model specification to be robust.

We now consider each of the twelve assumptions in turn, devoting the separate section 1.5 to the last, namely model selection.

1.4.1 Economic theory

> "It seems very pretty," she said when she had finished it, "but it's rather hard to understand." (Alice after reading the Jabberwocky poem in Lewis Carroll, 1899)

Economic theory has created many major ideas that have in turn changed the world, from the "invisible hand" in (Smith, 1759, p. 350), understanding the gains from trade and the problems with mercantilism, through the effects of tariffs and taxes, to modern insights into issues such as welfare economics, option pricing, auctions, contracts, principal–agent and game theories, trust and moral hazard, asymmetric information, institutions, and all their attendant impacts on market functioning and industrial, and even political, organization. In doing so, economics has evolved dramatically over time, and will undoubtedly continue doing so, hence at no instant can it be claimed to be correct, complete and immutable. For example, most theories take preferences as a given – sometimes even as "deep parameters" – but there are many endogenous determinants, from learning, adaptation, and advertising among others (see, e.g., von Weizsacker, 2005), with psychological, behavioral, and neuro-economics bidding fair to play key roles in the future (see, *inter alia*, Fehr and Falk, 2002; Fehr, Fischbacher, Kosfeld, 2005; Camerer, 2007).

Theories need to be distinguished in terms of their "levels', where low-level theories are well established and widely accepted (e.g., the optical theory behind the design of telescopes and the interpretation of their evidence), whereas high-level theories usually assume the validity of many lower levels, but are subject to doubt (as in theories of the accelerating expansion of the universe as due to "dark energy"). Facts are items of empirical information which depend only on low-level theories and measurements, and can be reliably replicated. Since all empirical evidence is theory laden to some degree, albeit often just from very low-level theories, "measurement without theory" is trivially impossible, and must relate to the lack of use of high level theories – the appropriate blend of theory and empirical evidence affects research efficiency, not necessarily the validity of any resulting findings (see, e.g., Gilbert and Qin, 2007). Many low-level statements are correct, complete and immutable, such as $1 + 1 = 2$, and although essential to arithmetic, cannot "explain" economic behavior. Conversely, testing theories just by their predictions is problematic, as false assumptions can entail correct conclusions: assume $1 = 2$, then $2 = 1$, so adding both sides, $3 = 3$, which is valid and presumably thereby establishes that $1 = 2$ (also see Ericsson and Hendry, 1999).

General theories that do "explain" the *Gestalt* of empirical evidence are a boon, but are not essential. Similarly, although experimentation can be helpful, it is far from the only source of evidence: observational science is not an oxymoron. The progressivity of science – its cumulation of findings that cohere, are consolidated in theoretical explanations, and suggest where next to investigate – is its most salient

attribute. There is simply no case that we understand no more than (say) Aristotle, or Kepler, etc.: lights work, computers run, planes fly. Moreover, it is possible to "predict" with considerable accuracy what changes to chips will speed up calculations, and what putative aircraft will not fly, which are inferences beyond any local set of experiments and evidence, are not purely inductive, and can be generalized, though doubtless with limits. Science seeks progress, whether by new experiments, new instruments or observations, new theories or refutations of extant ones. We now know that ulcers are caused by helicobacter pylori bacteria – not by stress – so cheap, painless antibiotics can cure ulcers, replacing life-threatening operations or expensive drugs. The path that led to that idea is irrelevant to its validity, and could be serendipity, careful testing, or a theory prediction, whereas stringent evaluation and replicability are crucial. In turn, such advances can lead to radical new thinking, albeit that initially they often face considerable opposition – sometimes even derision: scientists are humans, and rarely change their views until evidence overwhelms. Even then, theories are usually not rejected by evidence, but rather are replaced when "better" theories develop that explain more, especially if they account for previous anomalies.

Statistical analyses become essential in observational sciences, such as astronomy and economics, where "field" experiments are almost impossible to control. Then theory and modeling difficulties both explode and certainty declines, especially when behavioral change is possible: despite rendering previous analyses less than fully relevant to new settings, progress remains the key. It is widely recognized that special factors may intrude on a theory-based model (e.g., changes in credit rationing, nationalization, deregulation, price controls, wars, etc.), but less recognized that such special factors can dominate when accounting for data variability. Morgan (1990), Spanos (1995), Hendry (1995b) and Hendry and Mizon (2000) discuss some of the problems involved in testing theories using observational data.

Economists have not formally justified the principle of deriving empirical models from theory – most seem to assume it is obvious – so a substantial proportion of empirical econometric evidence is "high level" in that its credibility depends on the prior credibility of the theoretical model from which it was derived. Given any conjecture, we can usually test its empirical validity, thereby sustaining a destructive approach (see, e.g., Popper, 1963, on conjectures and refutations), although issues of inference from small and heterogeneous data samples complicate the analysis. If a theory implementation is simply discarded when it is rejected, the process fails to incorporate learning from the evidence. Conversely, if it is not discarded, some or all of the empirical model, the measurements and the theory must be revised, although there is no unique or even structured way of doing so. It is a *non sequitur* to assume that the particular alternative considered is true when the null is rejected. A progressive research approach of successively encompassing congruent (see section 1.4.2.4) models consolidated by empirically-relevant theories offers one possibility.

Alternative approaches to "macroeconomic theory" abound in the literature: Samuelson (1947) initiated a tradition of models based on constrained optimization, implemented by Hall (1978) as Euler equations; Kydland and Prescott (1990,

1991) formulate real-business cycle theories with rational expectations, leading to dynamic stochastic general equilibrium (DSGE) models as in Smets and Wouters (2003), whereas Hildenbrand (1994, 1999) emphasizes heterogeneity of endowments; Stiglitz (2003) stresses that asymmetric information can induce Keynesian effects, and Aghion *et al.* (2002) argue that agents only have imperfect-knowledge expectations. Moreover, many aspects of economic theory models can be chosen freely, such as the units of time and forms of utility functions: indeed, Stigum (1990) views theories as characterizing "toy agents in toy economies." However, when data are non-stationary, few transformations will be able to characterize the evidence in a constant relationship. For example, linear relationships between variables, which often arise in Euler equations, seem unlikely to be good descriptions in growing economies (see Ermini and Hendry, 2008, and Spanos, Hendry and Reade, 2008, for tests of log versus linear dependent variables in I(1) processes).

The absence from many economic theories of some of the main sources of data variability occurs across most research areas in economics, and although it differs in form, is probably part of the reason for the rash of "puzzles" (i.e., anomalous or even contradictory evidence) so beloved of the present generation of journal editors. In microeconomics, low R^2 values reveal that much of the variability is not accounted for by the postulated models. That outcome is usually ascribed to individual heterogeneity and idiosyncrasies, which can indeed generate high levels of unexplained variability, but there has to be some doubt that all the major factors have been included. In panel data studies, much observed data variation is attributed to "individual effects," which are removed by (e.g.) differencing or deviations from individual means. However, if the evidence that most micro-variability is due to individual heterogeneity is correct, then "representative" agent theories cannot be the best basis for macro-behavior, although aggregation could sustain some approaches (see, e.g., Granger, 1987; Blundell and Stoker, 2005), but not others (Granger, 1980). Finally, cross-country studies rarely account for key institutional differences between the constituent economies, and often use averages of data over historical epochs where considerable changes occurred between periods (see, e.g., Sala-i-Martin, 1997, and the criticisms in Hoover and Perez, 2004; Hendry and Krolzig, 2004).

1.4.1.1 Non-stationarity and ceteris paribus

> "You don't know how to manage Looking-glass cakes," the Unicorn remarked. "Hand it round first, and cut it afterwards." (Lewis Carroll, 1899)

A time series process is non-stationary if its moments or distributional form change over time. Two important forms of non-stationarity are unit roots and structural breaks, both of which lead to permanent changes. The former induce stochastic trends, which can be eliminated by differencing, or cointegration can also remove unit roots and retain linear combinations of levels of the variables (however, unit roots and cointegration are only invariant under linear transformations of variables). There is a vast literature, and recent surveys include Hendry and Juselius

(2000, 2001) and Johansen (2006). Structural breaks matter most when they induce location shifts in the processes under analysis, but those can also be removed in part by differencing or co-breaking (see Hendry and Massmann, 2007). Forecast failure – defined as a significant deterioration in forecast performance relative to its anticipated outcome, usually based on historical performance – is all too common, and is almost certainly due to structural breaks (see, e.g., Clements and Hendry, 1998, 1999, 2002b).

Because much observed data variability is due to factors absent from economic theories, a serious gap exists between macroeconomic theory models and applied econometric findings (see Spanos, 1989; Juselius, 1993; Hendry, 1995b; Nymoen, 2002). All economic theories rely on implicit *ceteris paribus* clauses, as "controls in thought experiments," although in a general equilibrium system in which everything depends on everything else, *ceteris paribus* is suspect. In empirical modeling, *ceteris paribus* cannot apply under non-stationarity even if the relevant variables are strongly exogenous, since "other things" will not be "equal." Cartwright (2002) describes *ceteris paribus* as roughly equivalent to "if nothing interferes then . . . some regularity is observed." In non-stationary processes, nothing will interfere only if all other factors are irrelevant, not because they will not change. Many sources of wide-sense non-stationarity impinge on economic data, including technical progress, R&D, new legislation, institutional changes, regime shifts, financial innovation, shifting demography, evolving social and political mores, as well as conflicts and other major catastrophes, inducing both evolution and structural breaks, all of which change the distributional properties of data.

Two resolutions are possible to wide-sense non-stationarity. First, a "minor influence" theorem could show on theoretical or evidential grounds that all omitted factors can be neglected, either because changes in them are of a smaller order of importance than included effects, or because they are orthogonal to all the effects that matter (see Hendry, 2005, and compare Boumans, 2005, who refers to *ceteris neglectis* and *ceteris absentibus*). Neither condition is plausible unless at least all the major influences have been included. Doing so brings us anyway to the second solution, namely including all potentially relevant variables at the outset, embedding theory models in more general systems that also allow for all the empirically-known influences, as well as the many historical contingencies that have occurred. Thus, institutional knowledge and economic history become essential ingredients in Applied Econometrics. Far from diminishing the importance of economic reasoning as a basis for empirical econometrics, including all non-stationarities seems the only way to reveal the underlying economic behavior uncontaminated by excluded changes. Of course, theory models of the likely behavioral reactions of economic agents to major changes would also help. As macro-data are the aggregates of the economic microcosm, these problems must afflict all empirical econometric studies, and are not merely a problem for time series analysts. Since the need to model all non-stationarities if empirical results are to be useful is important for both economics and econometrics, the next section considers its prevalence.

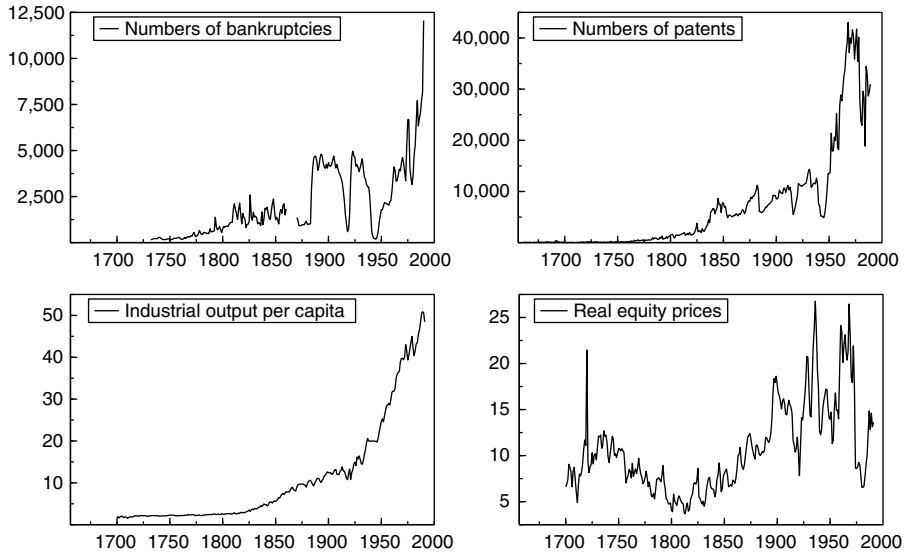


Figure 1.1 Historical time series for the UK

1.4.1.2 Long-run change

> "I see you're admiring my little box," the Knight said in a friendly tone. "It's my own invention – to keep clothes and sandwiches in. You see I carry it upside-down, so that the rain can't get in."
>
> "But the things can get out," Alice gently remarked. "Do you know the lid's open?" (Lewis Carroll, 1899)

Figure 1.1 records some historical time series for the UK over the period from about 1700 to 1991 (the dates differ for the various variables). Many other variables manifesting dramatic non-stationarities are shown in Hendry (2001a, 2001b, 2005) and Clements and Hendry (2001), where the first and last examine UK industrial output in more detail. Here we focus on numbers of bankruptcies and patents, industrial output per capita, and real equity prices (deflated by a cost of living price index) (see Feinstein, 1972; Mitchell, 1988; Crafts and Mills, 1994, *inter alia*).

These four variables were selected from a range of alternatives as being related to advances in technology and medicine, their implementation, incentives for progress through intellectual property, and one source of financing (see Siegel and Wright, 2007, for a recent review and bibliographic perspective). Technological change is sometimes modeled as an "exogenous" random walk. While that is an improvement over a deterministic trend, it is hardly a convincing representation of a process which requires substantial inputs of human and physical capital, as highlighted by endogenous growth models (see, e.g., Crafts, 1997). At the very

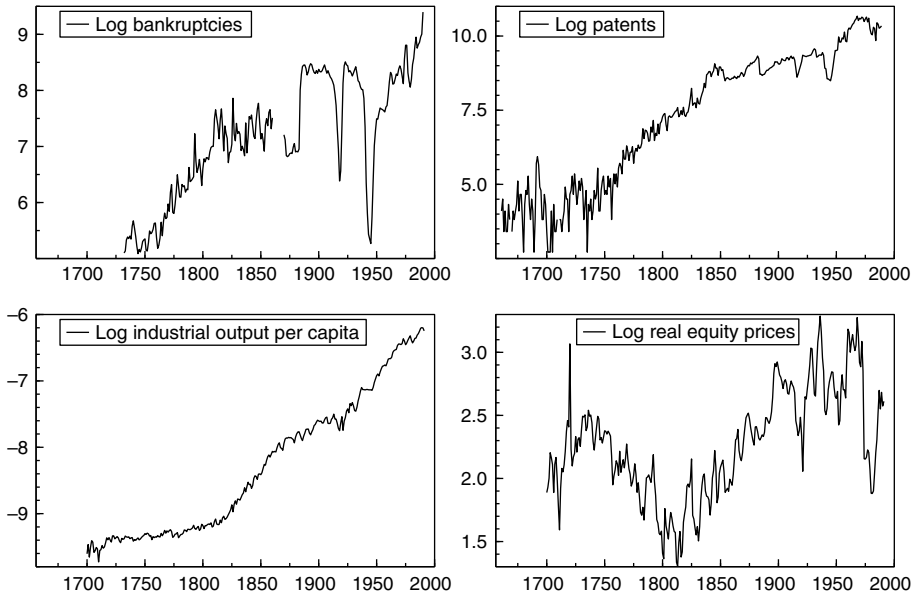


Figure 1.2 Logs of historical UK time series

start of the industrial revolution, Smith (1776, Ch. 1) notes the key development of specialists in R&D: "In the progress of society, philosophy or speculation becomes, like every other employment, the principal or sole trade and occupation of a particular class of citizens."

The non-stationarities in Figure 1.1 are blatant, and reflect more than just unit roots. Major historical events are clear: e.g., real equity prices exploded in the South Sea Bubble, not regaining such levels again for 200 years, collapsed in the Napoleonic and both world wars, as well as the first oil crisis, and today are little above pre-industrial revolution levels. Frankly, it is almost infeasible to build sensible empirical models of these levels series.

Figure 1.2 records the same four series in logs. Non-stationarity remains clear, but one can at least imagine ways of sucessfully modeling such data. Figure 1.3 reports their data densities separately in each of (approximately) the three centuries involved. The shifts in means and variances are marked, even if the presence of heterogeneity and dependence within each century can distort such histograms. Nor is the non-stationarity restricted to the levels of the variables, as Figure 1.4 illustrates for the annual changes in the four variables. Variance changes are clear, which also serves to make the densities of the changes look much more alike, as seen in Figure 1.5. Differencing alone can be insufficient to induce stationarity.

We will analyze the log transformations of these data in section 1.8.

In the absence of complete theoretical guidance on all relevant and irrelevant variables, functional forms, exogeneity, dynamics, and non-stationarities, empirical determination is essential. Consequently, the initially postulated models of

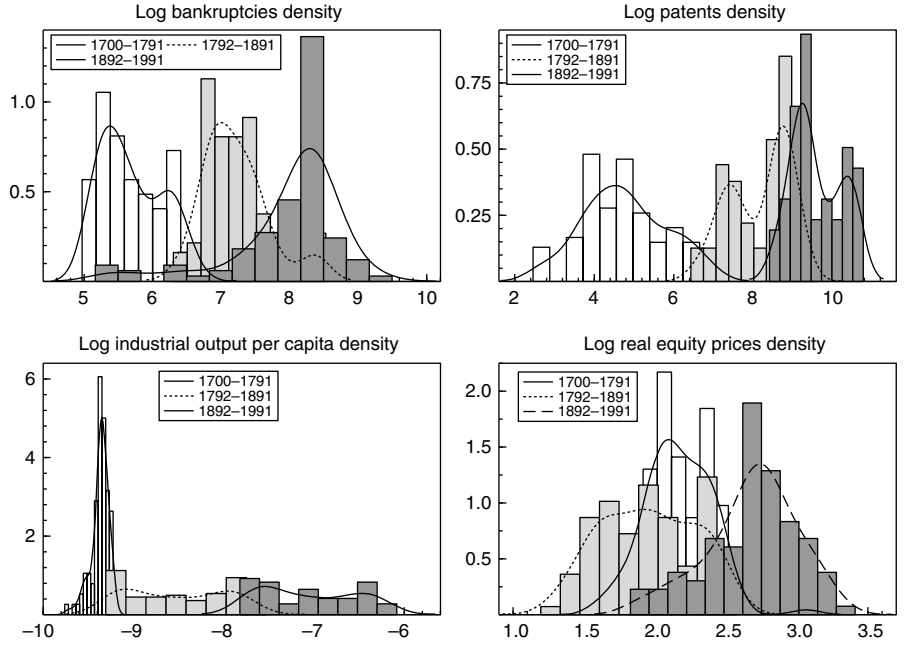


Figure 1.3 Three centuries of data distributions of levels

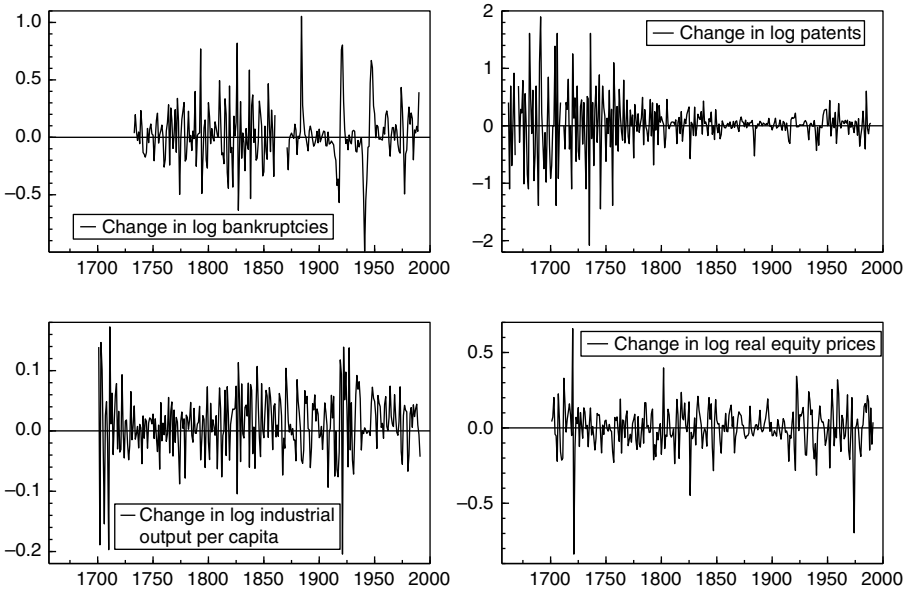


Figure 1.4 Changes in historical time series for the UK

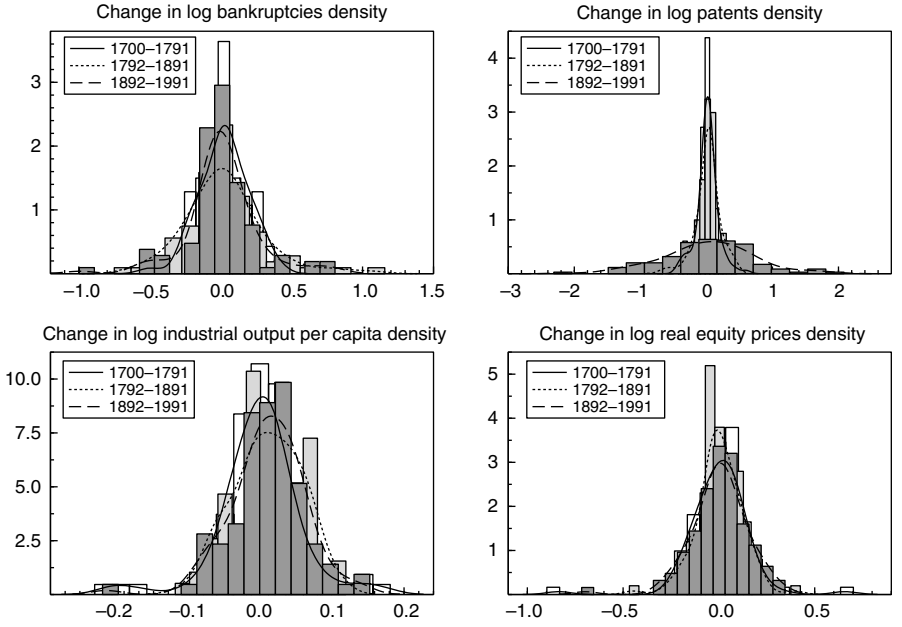


Figure 1.5 Three centuries of data distributions of changes

empirical research should usually involve many variables, although final selections may prove to be parsimonious (an implication is considered in section 1.5): we now consider that route.

1.4.2 Incomplete specifications

> "What am I to do?" exclaimed Alice, looking about in great perplexity as first one round head, and then the other, rolled down from her shoulder, and lay like a heavy lump in her lap. (Lewis Carroll, 1899)

Economies are so high dimensional, interdependent, heterogeneous, and evolving that a comprehensive specification of all events is impossible: the number of economy-wide relevant variables is uncountable in a human lifetime. Reducing that high dimensionality by aggregation over any or all of time, space, commodities, agents, initial endowments, etc., is essential, but precludes any claim to "truth." So if one cannot get at the "truth," what is on offer in economics? Three alternatives are: imposing theory-based models; constructing partial models, which aim to estimate some parameters associated with a theory, usually by generalized method of moments (GMM); or seeking the local DGP guided by economic theory. All three could operate, but depend on different assumptions. We first outline how empirical models must arise, then evaluate the three approaches in general against that basis.

1.4.2.1 From DGP to LDGP

> "The prettiest are always further!" she said at last. (Quote from Alice in Lewis Carroll, 1899)

Granted a stochastic basis for individual agent decision taking, such that any economic transaction can be described as an event in an event space, which could have been different for a myriad of reasons, then outcomes are measurable random variables with (possibly different) distributions at each point in time. Let $\mathbf{U}_T^1 = (\mathbf{u}_1, \ldots, \mathbf{u}_T)$ be the complete set of random variables relevant to the economy under investigation over a time span $t = 1, \ldots T$, defined on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$, where Ω is the sample space, $\mathcal{F}$ the event space and P the probability measure. Denote the vast, complex, and ever-changing joint distribution of $\{\mathbf{u}_t\}$ conditional on the pre-sample outcomes $\mathbf{U}_0$ and all necessary deterministic terms $\mathbf{Q}_T^1 = (\mathbf{q}_1, \ldots, \mathbf{q}_T)$ (like constants, seasonal effects, trends, and shifts) by:

$$\mathsf{D}_{\mathsf{U}}\left(\mathbf{U}_T^1 \mid \mathbf{U}_0, \mathbf{Q}_T^1, \xi_T^1\right), \tag{1.1}$$

where $\xi_T^1 \in \Xi \subseteq \mathcal{R}^k$ are the parameters of the agents' decision rules that led to the outcomes in (1.1). Then $\mathsf{D}_{\mathsf{U}}(\cdot)$ is the unknown, and almost certainly unknowable, data-generation process of the relevant economy. The theory of reduction discussed in, *inter alia*, Hendry (1987), Florens, Mouchart and Rolin (1990) and Hendry (1995a, Ch. 9) shows that a well-defined sequence of operations leads to the "local" DGP (LDGP), which is the actual generating process in the space of the variables under analysis. The resulting LDGP may be complex, non-linear and non-constant from aggregating, marginalizing (following the relevant data partition), and sequential factorization (the order of these reductions below is not a central aspect), so the choice of the set of variables to analyze is crucial if the LDGP is to be viably "captured" by an empirical modeling exercise. In turn, that LDGP can be approximated by a "general unrestricted model" (GUM) based on truncating lag lengths, approximating the functional form (perhaps after data transformations) and specifying which parameters are to be treated as constant in the exercise. Finally, a further series of reductions, involving mapping to non-integrated data, conditioning, and simultaneity, lead to a parsimonious representation of the salient characteristics of the dataset. Tests of losses from all these reductions are feasible, as discussed in section 1.4.2.4.

Aggregation. Almost all econometric data are aggregated in some way, implicitly discarding the disaggregates: although some finance data relate to point individual transactions, their determinants usually depend on aggregates (such as inflation). We represent this mapping as $\mathbf{U}_T^1 \rightarrow \mathbf{V}_T^1$, where the latter matrix is a mix of the data to be analyzed and all other variables. The key issue is the impact of that mapping on $\xi_T^1 \rightarrow \phi_T^1$, where the latter set may include more or fewer constant parameters depending on the benefits or costs of aggregation. Aggregates are linear sums, so their means have variances proportional to population size: if $x_{i,t} \sim \mathsf{IN}\left[\mu_t, \sigma_t^2\right]$

then $\overline{x}_t = N_t^{-1} \sum_{i=1}^{N_t} x_{i,t} \sim \mathsf{IN}\left[\mu_t, N_t^{-1}\sigma_t^2\right]$. Log transforms of totals and means, $\overline{x} > 0$, only differ by that population size as $\ln \sum_{i=1}^{N_t} x_{i,t} = \ln \overline{x}_t + \ln N_t$, so standard deviations of log aggregates are proportional to scaled standard deviations of means: $\mathsf{SD}[\ln \sum_{i=1}^{N_t} x_{i,t}] \simeq N_t^{-1}\sigma_t/\mu_t$ (see, e.g., Hendry, 1995a, Ch. 2). Thus logs of aggregates can be well behaved, independently of the underlying individual economic behavior.

Data transformations. Most econometric models also analyze data after transformations (such as logs, growth rates, etc.), written here as $\mathbf{W}_T^1 = \mathbf{g}(\mathbf{V}_T^1)$. Again, the key impact is on $\phi_T^1 \rightarrow \varphi_T^1$ and the consequences on the constancy of, and cross-links between, the resulting parameters. At this stage we have created:

$$\mathsf{D}_{\mathsf{W}}\left(\mathbf{W}_T^1 \mid \mathbf{U}_0, \mathbf{Q}_T^1, \varphi_T^1\right). \tag{1.2}$$

The functional form of the resulting representation is determined here by the choice of $\mathbf{g}(\cdot)$. Many economic variables are intrinsically positive in levels, a property imposed in models by taking logs, which also ensures that the error standard deviation is proportional to the level.

Data partition. No reduction is involved in specifying that $\mathbf{W}_T^1 = (\overline{\mathbf{W}}_T^1 : \mathbf{R}_T^1)$, where $\mathbf{R}_T^1$ denotes the $n \times T$ data to be analyzed and $\overline{\mathbf{W}}_T^1$ the rest. However, this decision is a fundamental one for the success of the modeling exercise, in that the parameters of whatever process determines $\mathbf{R}_T^1$ must deliver the objectives of the analysis.

Marginalizing. To implement the choice of $\mathbf{R}_T^1$ as the data under analysis necessitates discarding all the other potential variables, which corresponds to the statistical operation of marginalizing (1.2) with respect to $\overline{\mathbf{W}}_T^1$:

$$\mathsf{D}_{\mathsf{W}}\left(\overline{\mathbf{W}}_T^1, \mathbf{R}_T^1 \mid \mathbf{U}_0, \mathbf{Q}_T^1, \varphi_T^1\right) = \mathsf{D}_{\overline{\mathsf{W}}}\left(\overline{\mathbf{W}}_T^1 \mid \mathbf{R}_T^1, \mathbf{U}_0, \mathbf{Q}_T^1, \overline{\varphi}_T^1\right) \mathsf{D}_{\mathsf{R}}\left(\mathbf{R}_T^1 \mid \mathbf{U}_0, \mathbf{Q}_T^1, \omega_T^1\right). \tag{1.3}$$

While such a conditional-marginal factorization is always possible, a viable analysis requires no loss of information from just retaining ω_T^1. That will occur only if $\left(\overline{\varphi}_T^1, \omega_T^1\right)$ satisfy a cut, so their joint parameter space is the cross-product of their individual spaces, precluding links across those parameters. At first sight, such a condition may seem innocuous, but it is very far from being so: implicitly, it entails Granger non-causality of (all lagged values of) $\overline{\mathbf{W}}_T^1$ in $\mathsf{D}_{\mathsf{R}}(\cdot)$, which is obviously a demanding requirement (see Granger, 1969; Hendry and Mizon, 1999). Spanos (1989) calls the marginal distribution $\mathsf{D}_{\mathsf{R}}(\cdot)$ in (1.3) the Haavelmo distribution.

Sequentially factorizing. Next, letting $\mathbf{R}_{t-1}^1 = (\mathbf{r}_1, \ldots, \mathbf{r}_{t-1})$, the retained marginal density from (1.3) can be sequentially factorized as (see, e.g., Doob, 1953):

$$\mathsf{D}_{\mathsf{R}}\left(\mathbf{R}_T^1 \mid \mathbf{U}_0, \mathbf{Q}_T^1, \omega_T^1\right) = \prod_{t=1}^{T} \mathsf{D}_{\mathsf{r}_t}\left(\mathbf{r}_t \mid \mathbf{R}_{t-1}^1, \mathbf{U}_0, \mathbf{q}_t, \lambda_t\right). \tag{1.4}$$

The right-hand side of (1.4) completes the intrinsic reductions from the DGP to the LDGP for the set of variables under analysis (generally, the effects of the initial conditions $\mathbf{U}_0$ are ignored and assumed to be captured by $\mathbf{R}_0$). The sequential densities in (1.4) create a martingale difference (or innovation) process:

$$\epsilon_t = \mathbf{r}_t - \mathsf{E}\left[\mathbf{r}_t \mid \mathbf{R}_{t-1}^1, \mathbf{U}_0, \mathbf{q}_t\right], \tag{1.5}$$

where $\mathsf{E}[\epsilon_t | \mathbf{R}_{t-1}^1, \mathbf{U}_0, \mathbf{q}_t] = \mathbf{0}$ by construction.

Parameters of interest. These are the targets of the modeling exercise, and are hypothesized – on the basis of prior reasoning, past studies, and institutional knowledge – to be the features of interest. We denote them by $\theta \in \Theta$, and any later reduction choices must be consistent with obtaining θ from the final specification. To the extent that the economic theory supporting the empirical analysis is sufficiently comprehensive, the $\{\lambda_t\}$ in (1.4) should still contain the required information about the agents' decision parameters, so $\theta = \mathbf{h}(\omega_T^1)$. The next stage is to formulate a general model of (1.4) that also retains the necessary information.

1.4.2.2 From LDGP to general unrestricted model

The LDGP in (1.4) can be approximated by a model based on a further series of reductions, which we now discuss. Indeed, (1.4) is often the postulated basis of an empirical analysis, as in a vector autoregression, albeit with many additional assumptions to make the study operational. There are no losses when the LDGP also satisfies these reductions, and if not, evidence of departures can be ascertained from appropriate tests discussed in section 1.4.2.4, so that such reductions are then not undertaken.

Lag truncation. The potentially infinite set of lags in (1.4) can usually be reduced to a small number, so $\mathbf{R}_{t-1}^1 \simeq \mathbf{R}_{t-1}^{t-s} = (\mathbf{r}_{t-s} \ldots \mathbf{r}_{t-1})$, where the maximum lag length becomes s periods, with initial conditions $\mathbf{R}_0^{1-s}$. Long-memory and fractional integration processes are considered in, e.g., Granger and Joyeux (1980), Geweke and Porter-Hudak (1983), Robinson (1995) and Baillie (1996). Letting $\mathsf{f}_{\mathsf{r}_t}(\cdot)$ denote the resulting statistical model of the $\{\mathbf{r}_t\}$, which could coincide with the LDGP when the reduction is without loss, then the mapping is:

$$\prod_{t=1}^{T} \mathsf{D}_{\mathsf{r}_t}\left(\mathbf{r}_t \mid \mathbf{R}_{t-1}^1, \mathbf{U}_0, \mathbf{q}_t, \lambda_t\right) \Rightarrow \prod_{t=1}^{T} \mathsf{f}_{\mathsf{r}_t}\left(\mathbf{r}_t \mid \mathbf{R}_{t-1}^{t-s}, \mathbf{R}_0^{1-s}, \mathbf{q}_t, \psi_t\right). \tag{1.6}$$

The obvious check on the validity of such a reduction is whether longer lags matter; and as before, the key criterion is the impact on $\{\psi_t\}$.

Parameter constancy. The parameters in question are those that characterize the distribution $\mathsf{f}_{\mathsf{r}}(\cdot)$ in (1.6). Then their constancy entails that the $\{\psi_t\}$ depend on a smaller set of parameters that are constant, at least within regimes. Complete constancy requires $\psi_t = \psi_0 \ \forall t$, and while unlikely in economics, is often the assumption made, at least until there is contrary evidence. When there is no loss, $\theta = \mathbf{f}(\psi_0)$, so all parameters of interest can be recovered from the model.

Linearity. The distribution in (1.6) may correspond to the linear Normal when the functional form is chosen appropriately to ensure that a homoskedastic process also results:

$$\mathsf{f}_{\mathsf{r}_t}\left(\mathbf{r}_t \mid \mathbf{R}_{t-1}^{t-s}, \mathbf{R}_0^{1-s}, \mathbf{q}_t, \psi_0\right) \underset{app}{\sim} \mathsf{IN}_k\left[\sum_{i=1}^{s} \Pi_i \mathbf{r}_{t-i} + \Pi_{s+1}\mathbf{q}_t, \Omega\right]. \tag{1.7}$$

The LDGP distribution need not be Normal, but that is partly dependent on the specification of $\mathbf{q}_t$, especially whether breaks in deterministic terms are modeled therein. The constancy of the coefficients of any model also depends on the functional forms chosen for all the data transformations, and an operational GUM presumes that $\{\mathbf{r}_t\}$ has been transformed appropriately, based on theoretical and empirical evidence. Checks for various nonlinear alternatives and homoskedasticity are merited.

1.4.2.3 From the general to the specific

Providing that a viable set of basic parameters is postulated (and below we will allow for the possibility of many shifts), then a variant of (1.7) can act as the GUM for a statistical analysis. When the LDGP is nested in the GUM, so none of the reductions above led to important losses, a well-specified model which embeds the economic theory and can deliver the parameters of interest should result. When the LDGP is not nested in the GUM, so the reductions in the previous sub-section entail losses, it is difficult to establish what properties the final specific model will have, although a well-specified approximation at least will have been found. Because wide-sense non-stationarity of economic variables is such an important problem, and within that class, location shifts are the most pernicious feature, section 1.5 considers the recent approach of impulse saturation (see Hendry, Johansen and Santos, 2008; Johansen and Nielsen, 2008).

Mapping to a non-integrated representation. Many economic variables appear to be integrated of at least first order (denoted I(1)), so there is a mapping $\mathbf{r}_t \rightarrow (\Delta \mathbf{r}_{p,t} : \beta' \mathbf{r}_t) = \mathbf{x}_t$, where there are $n - p$ cointegrating relations and p unit roots, so $\mathbf{x}_t$ is now I(0). Processes that are I(2) can be handled by mapping to second differences as well (see, e.g., Johansen, 1995). This reduction to I(0) transforms ψ_0 to ρ_0 (say)

and leads from (1.6) to:

$$\prod_{t=1}^{T} \mathsf{f}_{\mathbf{x}_t}\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \rho_0\right). \tag{1.8}$$

VARs like (1.7) are often formulated for $\mathbf{r}_t$, rather than $\mathbf{x}_t$, as occurs in the first stage of some cointegration analyses.

Contemporaneous conditioning. Conditioning concerns both contemporaneous variables in models and current-dated instrumental variables (IVs), so let $\mathbf{x}_t' = (\mathbf{y}_t' : \mathbf{z}_t')$, where the former are the k variables to be modeled and the latter $n-k$ are taken as given. Then for $\rho_0 = (\kappa_1 : \kappa_2)$:

$$\begin{aligned}\mathsf{f}_{\mathbf{x}_t}\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \rho_0\right) = \mathsf{f}_{\mathbf{y}_t|\mathbf{z}_t}\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_1\right) \\ \mathsf{f}_{\mathbf{z}_t}\left(\mathbf{z}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_2\right).\end{aligned} \tag{1.9}$$

A viable analysis from the conditional distribution alone in (1.9) requires that $\theta = \mathbf{h}_1(\kappa_1)$; and there will be no loss of information only if (κ_1, κ_2) satisfy a cut so $(\kappa_1, \kappa_2) \in \mathcal{K}_1 \times \mathcal{K}_2$, in which case $\mathbf{z}_t$ is weakly exogenous for θ. When (1.7) holds, both conditional and marginal distributions in (1.9) will be Normal, and the relationships linear. The former leads to VAR-type modeling as noted, whereas the conditional representation in (1.9) underpins more "structural" approaches when the $\mathbf{z}_t$ are instruments: we return to conditioning in section 1.4.5 below.

Simultaneity. Finally, at least for the order of reductions considered here, simultaneity can allow a more parsimonious representation of the conditional distribution by modeling in terms of $\Gamma\mathbf{y}_t$, where Γ is a non-singular matrix that captures the current-dated interdependencies. If $\mathbf{z}_t$ does not enter the conditional distribution, $\Gamma\mathbf{x}_t$ could be modeled directly relative to lagged information (see, e.g., Demiralp and Hoover, 2003).

1.4.2.4 Implications

Five important issues are clarified by these reductions from the DGP down to a specific model of a sub-set of the variables.

Econometric concepts. First, there exists an LDGP as in (1.4) for whatever choices are made of $\mathbf{x}_t$. When all reductions are without loss, the statistical model $\mathsf{f}_{\mathbf{y}_t|\mathbf{z}_t}(\cdot)$ in (1.9) could also be the LDGP. Although most empirical analyses seem to commence by specifying what is included, rather than what is eliminated, almost all the central concepts in econometrics (in italics below) correspond to when reductions (in bold face) can be achieved without loss of relevant information:

- **Aggregation** entails no loss of information on marginalizing with respect to disaggregates when the formulation retains *sufficient statistics* for θ.
- **Data transformations** have no associated reduction, but relate to *parameters of interest*, θ, and hence the need for these to be *invariant and identifiable*.

- **Data partition** determines which variables to include and which to omit in the model *specification*, a decision that is dependent on the purpose of the modeling exercise, but is fundamental to the success of the empirical model.
- **Marginalizing** with respect to $\mathbf{v}_t$ is without loss if $\mathbf{X}_T^1$ is *sufficient* for θ; and marginalizing with respect to $\mathbf{V}_{t-1}^1$ is without loss if it is *Granger non-causal* for $\mathbf{x}_t$ and the conditional-marginal parameters satisfy a *cut*.
- **Sequential factorization** induces no loss as ϵ_t from (1.5) is an *innovation* relative to $\mathbf{R}_{t-1}^1$.
- **Parameter constancy** over time is fundamental to most uses of a model, and *invariance* (constancy across interventions to the marginal process) is essential for policy.
- **Lag truncation** leads to no loss if ϵ_t remains an *innovation* against $\mathbf{X}_{t-1}^1$.
- **Integrated data** can be reduced to I(0) by *cointegration* and *differencing*, sustaining a more *parsimonious* representation, and supporting *conventional inference*.
- **Functional form** specification may or may not entail a reduction, and does not when the two densities are equivalent (e.g., logs of log-normal variables are normal).
- **Conditional factorizations** entail no loss of information when $\mathbf{z}_t$ is *weakly exogenous* for θ, addressed in section 1.4.5.
- **Simultaneity** can allow one to *parsimoniously* capture *joint dependence*.

Testing reductions. Second, reductions are testable against any preceding, less reduced, distributions. Indeed, there is an accompanying taxonomy of evaluation information that seeks to ascertain the statistical significance of the losses imposed by the various reductions. This leads to six major null hypotheses about the final model's specification: homoskedastic innovations $\{\epsilon_t\}$; $\mathbf{z}_t$ weakly exogenous for θ; constant, invariant θ; data-admissible formulations on accurate observations; theory consistent, identifiable structures; encompassing rival models. While this exhausts the nulls to test, there are many alternatives to each. Models which satisfy the first and third are well specified on the available information, and if satisfying the first three are said to be (empirically) congruent. One model (parsimoniously) variance dominates another if it has a smaller unexplained variance (and no more parameters): the notion of one model explaining the results of other models extends variance dominance to account for all other parameters. The principle of encompassing was formalized in Hendry and Richard (1982), and the theory of testing developed by Mizon (1984) and Mizon and Richard (1986) (see Hendry and Richard, 1989, and Hendry, Marcellino, and Mizon, 2008, for surveys). An admissible, theory-consistent, encompassing, congruent model satisfies all six criteria.

Choosing the Haavelmo distribution. Third, knowledge of the LDGP is the "optimum" one can achieve for the given set of variables. Different choices of $\{\mathbf{r}_t\}$, and hence the Haavelmo distribution, will lead to different LDGPs with more or less constancy and congruence with the available evidence. If (1.7) were indeed the LDGP, then model selection could target its variables. The congruence of

an empirical model corresponds to its encompassing the LDGP (so not deviating significantly from it in any of the first five directions) (see Bontemps and Mizon, 2003). Testing the selected model against all extant models of the same variables allows a rigorous evaluation of its "closeness" to the LDGP (see, *inter alia*, White, 1990; Mayo and Spanos, 2006).

Parameter dependence. Fourth, the resulting coefficients in (1.7) or (1.9) remain dependent on the initial DGP parameters. If those DGP parameters change, induced shifts can occur in the parameters of the LDGP. The extent to which these shifts occur, and when they do so, whether they can be anticipated, modeled or even understood, will depend on how usefully the reduced representation captures the structure of the relevant sub-set of the economy under analysis. Here, "structure" denotes invariance under extensions of the information set over (i) time (i.e., constancy), (ii) regimes (i.e., changes to marginal distributions or policy variables) and (iii) variables (so the reductions did not eliminate any important explanatory factors). When the initial economic analysis that led to the specification of $\{\mathbf{x}_t\}$ (i.e., the transformed sub-set of data under analysis) actually captured the main features of the behavior of the agents involved, then ρ_0, or κ_1, should be an invariant that also throws light on the agents' decision parameters underlying φ_T^1 in (1.2). Thus, properly embedded in a general congruent model, the economics should carry through.

Minimizing reductions. Finally, given the inertial dynamics of a high dimensional, interdependent and non-stationary system like an economy, reductions seem likely to be costly in practice and involve real information losses. These will manifest themselves through non-constant models, empirical "puzzles" and poor forecasts, so general systems seem generically preferable. "Errors" on empirical models are created by reductions, so will be highly composite, reflecting many components. It is unclear whether that also favors disaggregation, given problems posed by measurement errors and heterogeneity difficulties as disaggregation increases, or whether a "law of large numbers" may induce substantial offsets (as discussed above).

1.4.2.5 Evaluating the three main approaches

We now consider how the three basic approaches fare against the above analysis. Given their assumptions, each would of course work well; and with sufficiently rigorous testing, the choice of approach becomes a matter of research efficiency (see White, 1990). But efficiency is important, as the assumptions may not adequately characterize reality, and rigorous attempts to reject are not always undertaken.

Imposing economic theory. First, if one simply imposes an *a priori* theory on the data, then the outcome will be excellent when the theory is complete (relative to the issue under analysis) and "correct" (in that all omissions are relatively negligible). Otherwise, it is difficult to ascertain in general how poor the outcome will be (see, e.g., Juselius and Franchi, 2007). If no testing occurs, that strategy is both highly

risky and theory dependent. The risk is that a major discrepancy is not detected, leading to a poor description of the underlying agents' behavior: we addressed the issue of "*ceteris paribus*" in section 1.4.1.1, but if economies are inherently wide-sense non-stationary, then other things will not stay constant. When theories lack precise formulations of lag lengths, functional dependencies, other potential determinants, breaks, and non-economic factors of importance, such difficulties seem all too likely. The problem with theory dependence is that since no economic analysis has yet proved immutable, the empirical results will be discarded when the theory is altered, so there is no progressive knowledge accumulation. This is the real reason that Summers (1991) finds little contribution from empirical econometrics – it was not really allowed to make one, being restricted to providing empirical cloth for a pre-designed framework.

Partial use of economic theory. Second, a partial use of economic theory often leads to pre-specified moment conditions linking variables, $\mathbf{x}_t$, and parameters, φ, usually being zero for the "true" value of the parameter, φ_0, in the form (sometimes without conditioning):

$$\mathsf{E}\left[\mathbf{h}\left(\mathbf{x}_t, \varphi_0 \mid \mathbf{X}_{t-1}^1\right)\right] = \mathbf{0} \; \forall t, \tag{1.10}$$

enabling GMM estimation of φ (see, e.g., Smith, 2007) (also, Smith, 1992, develops non-nested tests applicable after GMM). Equally often, inference has to be based on heteroskedastic and autocorrelation-consistent covariance (HAC) matrices (see White, 1980a; Andrews, 1991), which assume that the residuals reflect precisely those problems in the errors. Unfortunately, residuals can be heteroskedastic and autocorrelated for many other reasons, including unmodeled breaks, measurement errors, incorrect dynamics, omitted variables, or an inappropriate functional form *inter alia*, most of which would invalidate the estimates derived from (1.10), and possibly refute the underlying theory. Thus, rigorous testing against that range of hypotheses would seem necessary, leading to three important difficulties. First, unless a joint test is used, non-rejection of each may occur when there are several failures. Second, if any test rejects at the chosen significance level (controlling for the number of tests undertaken), the validity of all the other tests is cast into doubt. Third, if rejection does occur, it remains a *non sequitur* to assume that the hypothesis which was rejected was the source of the failure, so model revision may require a theory revamp. Again, there seem to be distinct advantages to beginning with general formulations that can be simplified when evidence permits, subject to maintaining identifiability – which can also be a problem with GMM (section 1.4.7 discusses identification).

Economic theory guidelines. Finally, seeking a congruent model of the LDGP based on economic theory guidelines by embedding the theory-based model in a more general GUM for the set of candidate variables, with a range of possible specifications of lags, functional forms, breaks, etc., offers many advantages, not least avoiding restrictive assumptions dependent on hope rather than evidence. Such a general-to-specific (Gets) approach can be demanding, and while it can

help mitigate problems of under-specification, it is no free lunch, as it leads to a different set of possible problems relating to data-based model selection, discussed in section 1.5. However, the criticism that the LDGP is too complicated for Gets to work well must also apply to all other approaches, as they will not fare any better in such a state of nature unless some remarkable requirements chance to hold (e.g., the complexity of the LDGP happens to lie in directions completely unrelated to the aspects under study). In general, even if one simply wants to test an economic hypothesis as to whether some effect is present, partial inference cannot be conducted alone, unless one is sure about the complete absence of all contaminating influences.

1.4.3 Data exactitude

> "She can't do Addition," the Red Queen interrupted. "Can you do Subtraction? Take nine from eight." "Nine from eight I can't, you know," Alice replied very readily. (Lewis Carroll, 1899)

No agency produces perfect data measures on every variable, and although some observations may be both accurate and precise (e.g., specific stock market, or foreign exchange, transactions), most are subject to measurement errors. These can be difficult to handle, especially when there are both revisions and changes in exactitude over time, which thereby introduce an additional source of non-stationarity. Moreover, in any given sample of time series, more recent data will be subject to potentially larger later revisions: section 1.7.1 considers the impact of one example of considerable data revisions.

Mapping theoretical constructs to data counterparts and measuring (or modeling) latent variables both raise further issues. Many commonly used macro variables do not have established measurements, e.g., output gaps, business cycles, capacity utilization, trade union power, etc. Even those that do, such as constructs for consumption, user costs, etc., are open to doubt. These types of measurement errors are not directly caused by inaccurate data collection, but both impinge on empirical studies, and can change over time.

Incentives to improve data quality, coverage and accuracy were noted in section 1.3.4 (see Boumans, 2007, for recent discussions of various measurement issues). In the absence of exact data, there must remain trade-offs between using theory to impose restrictions on badly-measured data, using such data to reject theory specifications, or building data-based models. Again, a balance utilizing both theory and evidence in a progressive process seems advisable.

1.4.4 Hidden dependencies

> "Why, it's a Looking-glass book of course! And if I hold it up to a glass, the words will all go the right way again." (Quote from Alice in Lewis Carroll, 1899)

Hidden dependencies abound in all data forms, including cross-sections, time series and panels. An important aspect of sequential conditioning in time series

is to explicitly remove temporal dependence, as (1.4) showed, where a martingale difference process is created by the sequential conditioning. In principle, the same concepts apply to cross sections. It must be stressed that "random sampling" by itself does not justify factorizing a joint density or likelihood function. As an extreme form of cross-section dependence, put 1,000 copies of the number "1" in a hat, then draw a random sample of 100 therefrom: one learns nothing after the first draw, although all are "randomly drawn." Sequential factorization correctly reveals that difficulty. Denote the randomly-drawn data sample by $(r_1 \dots r_N)$; then for any ordering when τ is the mean value of all the numbers in the hat:

$$\mathsf{D}_{\mathsf{r}}\left(r_1 \dots r_N \mid \tau\right) = \prod_{i=1}^{N} \mathsf{D}_{\mathsf{r}_i}\left(r_i \mid r_{i-1} \dots r_1; \tau\right) = \mathsf{D}_{\mathsf{r}_1}\left(r_1; \tau\right), \tag{1.11}$$

since all the other probabilities are precisely unity. As $r_1 = 1$, we correctly deduce $\tau = 1$. Certainly, the other $N - 1$ draws add the information that all the numbers are unity, but would do so even if not randomly drawn.

More generally, the order of an independent sample does not matter, so unlike (1.11), for any ordering the joint density should factorize as:

$$\mathsf{D}_{\mathsf{r}}\left(r_1 \dots r_N \mid \tau\right) = \prod_{i=1}^{N} \mathsf{D}_{\mathsf{r}_i}\left(r_i \mid r_{i-1} \dots r_1; \tau\right) = \prod_{i=1}^{N} \mathsf{D}_{\mathsf{r}_i}\left(r_i \mid \tau\right). \tag{1.12}$$

Consequently, potential dependence is testable by conditioning on s "neighbors" after a suitable exogenous ordering to check if their influence is non-zero; i.e., to see whether:

$$\prod_{i=1}^{N} \mathsf{D}_{\mathsf{r}_i}\left(r_i \mid r_{i-1} \dots r_{i-s}; \tau\right) \neq \prod_{i=1}^{N} \mathsf{D}_{\mathsf{r}_i}\left(r_i \mid \tau\right). \tag{1.13}$$

Suitable tests for the absence of dependence would seem essential before too great a weight is placed on results that base (1.12) on the claim of random sampling, especially when the units are large entities like countries. More generally, when all units are affected in part by macro-forces and their attendant non-stationarities, dependence like (1.13) is likely. If an ordering based on an outside variable is available, then models of $\mathsf{D}_{\mathsf{r}_i}\left(r_i \mid r_{i-1} \dots r_{i-s}; \tau\right)$ could be investigated directly, similar to some cases of spatial dependence (see Anselin, 2006).

There is a large literature on panel data analysis recently discussed in Choi (2006) and Baltagi (2006).

1.4.5 Conditioning variables

> "I'm afraid he'll catch cold with lying on the damp grass," said Alice, who was a very thoughtful little girl. (Lewis Carroll, 1899)

Instrumental variables are a key part of any conditioning set, so require weak exogeneity as well as correlation with the relevant endogenous variables (or the auxiliary assumptions of orthogonality to any unknown vector of excluded influences and independence from the "true" model's errors).

1.4.5.1 Weak exogeneity

The notion of exogeneity, or synonyms thereof, in relation to econometric modeling dates back to the origins of the discipline (see, e.g., Morgan, 1990; Hendry and Morgan, 1995), with key contributions by Koopmans (1950) and Phillips (1957). Weak exogeneity was formalized by Engle, Hendry and Richard (1983), building on Richard (1980) (see Ericsson, 1992, for an exposition), and is a fundamental requirement for efficient conditional inference, which transpires to be at least as important in integrated systems as in stationary processes (see Phillips and Loretan, 1991). Weak exogeneity is equally relevant to instrumental variables estimation, since the marginal density of $\mathbf{z}_t$ then relates to the distribution of the claimed instruments: asserting orthogonality to the error term is often inadequate, as shown by the counter-examples in Hendry (1995a).

Further, $\mathbf{z}_t$ is strongly exogenous for θ if $\mathbf{z}_t$ is weakly exogenous for θ, and:

$$\mathsf{D}_{z_t}\left(\mathbf{z}_t \mid \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_2\right) = \mathsf{D}_{z_t}\left(\mathbf{z}_t \mid \mathbf{Z}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_2\right). \tag{1.14}$$

When (1.14) is satisfied, $\mathbf{z}_t$ does not depend upon $\mathbf{Y}_{t-1}$ so $\mathbf{y}$ does not Granger-cause $\mathbf{z}$, following Granger (1969). This requirement sustains marginalizing $\mathsf{D}_{z_t}(\mathbf{z}_t|\mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_2)$ with respect to $\mathbf{Y}_{t-1}^1$, but does not concern conditioning. Consequently, Granger causality alone is neither necessary nor sufficient for weak exogeneity, and cannot validate inference procedures (see Hendry and Mizon, 1999).

The consequences of failures of weak exogeneity can vary from just a loss of estimation efficiency through to a loss of parameter constancy, depending on the source of the problem (see Hendry, 1995a, Ch. 5). We now illustrate both extreme cases and one intermediate example.

Outperforming Gauss–Markov. First, consider a standard regression setting where Gauss–Markov conditions seem satisfied:

$$\mathbf{y} = \mathbf{Z}\beta + \epsilon \;\text{ with }\; \epsilon \sim \mathsf{N}_T\left[\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right], \tag{1.15}$$

when $\mathbf{Z} = (\mathbf{z}_1 \dots \mathbf{z}_T)'$ is a $T \times k$ matrix, $\text{rank}(\mathbf{Z}) = k$, and $\epsilon' = (\epsilon_1 \dots \epsilon_T)$, with:

$$\mathsf{E}[\mathbf{y} \mid \mathbf{Z}] = \mathbf{Z}\beta,$$

and hence $\mathsf{E}[\mathbf{Z}'\epsilon] = \mathbf{0}$. OLS estimates of β, the parameter of interest here, are:

$$\widehat{\beta} = \beta + \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\epsilon \sim \mathsf{N}_k\left[\beta, \sigma_\epsilon^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right].$$

However, ordinary least squares (OLS) need not be the most efficient unbiased estimator of β, and an explicit weak exogeneity condition is required to preclude that possibility when $\mathbf{Z}$ is stochastic. For example, let:

$$\mathbf{z}_t = \beta + \nu_t \;\text{ where }\; \nu_t \sim \mathsf{IN}_k\left[\mathbf{0}, \Sigma\right],$$

estimated by the mean vector:

$$\overline{\beta} = \beta + \overline{\nu} \sim \mathsf{N}_k\left[\beta, T^{-1}\Sigma\right],$$

then it is easy to construct scenarios where $\overline{\beta}$ is much more efficient than $\widehat{\beta}$. Consequently, even in simple regression, for the Gauss–Markov theorem to be of operational use one needs the condition that β cannot be learned from the marginal distribution.

Weak exogeneity in cointegrated systems. Second, cointegrated systems provide a major forum for testing one aspect of exogeneity. Formulations of weak exogeneity conditions and tests for various parameters of interest in cointegrated systems are discussed in, *inter alia*, Johansen and Juselius (1990), Phillips and Loretan (1991), Hunter (1992), Urbain (1992), Johansen (1992), Dolado (1992), Boswijk (1992) and Paruolo and Rahbek (1999). Equilibrium-correction mechanisms which cross-link equations violate long-run weak exogeneity, confirming that weak exogeneity cannot necessarily be obtained merely by choosing the "parameters of interest." Conversely, the presence of a given disequilibrium term in more than one equation is testable. Consider an apparently well-defined setting with the following bivariate DGP for the I(1) vector $\mathbf{x}_t = (y_t : z_t)'$ from Hendry (1995c):

$$y_t = \beta z_t + u_{1,t} \tag{1.16}$$

$$z_t = \lambda y_{t-1} + u_{2,t}, \tag{1.17}$$

where:

$$\begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \rho & 1 \end{pmatrix} \begin{pmatrix} u_{1,t-1} \\ u_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}, \tag{1.18}$$

and:

$$\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \sim \mathsf{IN}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \gamma\sigma_1\sigma_2 \\ \gamma\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right] = \mathsf{IN}_2 \left[\mathbf{0}, \Sigma \right]. \tag{1.19}$$

The DGP in (1.16)–(1.19) defines a cointegrated vector process in triangular form (see Phillips and Loretan, 1991) which can be written in many ways, of which the following equilibrium-correction form is perhaps the most useful:

$$\begin{aligned} y_t &= \beta z_t + \epsilon_{1,t} \\ \Delta z_t &= \lambda \Delta y_{t-1} + \rho \left(y_{t-1} - \beta z_{t-1} \right) + \epsilon_{2,t}, \end{aligned} \tag{1.20}$$

where $\epsilon_t = \left(\epsilon_{1,t} : \epsilon_{2,t} \right)'$ is distributed as in (1.19).

The parameters of the DGP are $(\beta, \lambda, \rho, \gamma, \sigma_1, \sigma_2)$. When cointegration holds, β and σ_1 can be normalized at unity without loss of generality, and we also set $\sigma_2 = 1$. The parameter of interest is β, which characterizes the long-run relationship between y_t and z_t. Let $\mathcal{I}_{t-1}$ denote available lagged information (the σ-field generated by $\mathbf{X}_{t-1}$). Then, from (1.19) and (1.20), the conditional expectation of y_t given $(z_t, \mathcal{I}_{t-1})$ is:

$$\mathsf{E}\left[y_t \mid z_t, \mathcal{I}_{t-1} \right] = \beta z_t + \gamma \Delta z_t - \gamma\rho \left(y_{t-1} - \beta z_{t-1} \right) - \gamma\lambda \Delta y_{t-1}. \tag{1.21}$$

For some parameter values in the DGP, the conditional expectation will coincide with (1.16), whereas for other parameter configurations, (1.16) and (1.21) will differ, in which case it is unsurprising that (1.16) is not fully informative. However, an exact match between the equation to be estimated and the conditional expectation of the dependent variable given $\mathcal{I}_{t-1}$ is not sufficient to justify least squares estimation, even when the error is an innovation against $\mathcal{I}_{t-1}$. Indeed, when $\lambda = \gamma = 0$, but $\rho \neq 0$, there is a failure of weak exogeneity of z_t for β, even though the conditional expectation is:

$$\mathsf{E}\left[y_t \mid z_t, \mathcal{I}_{t-1}\right] = \beta z_t. \tag{1.22}$$

Nevertheless, z_t is not weakly exogenous for β when $\rho \neq 0$ since:

$$\Delta z_t = \rho \left(y_{t-1} - \beta z_{t-1}\right) + \epsilon_{2t}, \tag{1.23}$$

so a more efficient analysis is feasible by jointly estimating (1.16) (or (1.22)) and (1.23). Here the model coincides with both the conditional expectation and the DGP equation, but as shown in Phillips and Loretan (1991) and Hendry (1995c), the violation of weak exogeneity can lead to important distortions to inference when estimating the parameters of (1.16), highlighting the important role of weak exogeneity in conditional inference.

1.4.5.2 Super exogeneity and structural breaks

Next, processes subject to structural breaks sustain tests for super exogeneity and the Lucas (1976) critique (following Frisch, 1938): (see, e.g., Hendry, 1988; Fischer, 1989; Favero and Hendry, 1992; Engle and Hendry, 1993; Hendry and Santos, (2009). Formally, super exogeneity augments weak exogeneity with the requirement that the parameters of the marginal process can change (usually over some set) without altering the parameters of the conditional. Reconsider (1.9), written with potentially non-constant parameters as:

$$\begin{aligned}\mathsf{D}_{\mathbf{x}_t}\left(\mathbf{x}_t \mid \mathbf{X}_{t-1}^{t-s}; \mathbf{X}_0^{1-s}, \mathbf{q}_t, \rho_t\right) = \mathsf{D}_{\mathbf{y}_t|\mathbf{z}_t}\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{X}_{t-1}^{t-s}, \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_{1,t}\right) \\ \mathsf{D}_{\mathbf{z}_t}\left(\mathbf{z}_t \mid \mathbf{X}_{t-1}^{t-s}; \mathbf{X}_0^{1-s}, \mathbf{q}_t, \kappa_{2,t}\right).\end{aligned} \tag{1.24}$$

When θ enters both $\kappa_{1,t}$ and $\kappa_{2,t}$ in (1.24), inference can again be distorted if weak exogeneity is falsely asserted. When conditional models are constant despite data moments changing considerably, there is *prima facie* evidence of super exogeneity for that model's parameters; whereas, if the model as formulated does not have constant parameters, resolving that failure ought to take precedence over issues of exogeneity. However, while super exogeneity tests are powerful in detecting location shifts, changes to "reaction parameters" of mean-zero stochastic variables are difficult to detect (see, e.g., Hendry, 2000b). Hendry and Santos (2009) propose a test for super exogeneity based on impulse saturation (see Hendry, Johansen and Santos, 2008) to automatically select breaks in the marginal processes, then test their relevance in the conditional. When none of the breaks enters the conditional model, that provides evidence in favor of $\mathbf{z}_t$ causing y_t, since the same response

occurs to changes across different "regimes" (see, e.g., Heckman, 2000; Hendry, 2004; and the references therein).

1.4.5.3 Weak exogeneity and economic theory

Much economic theory concerns relationships between means such as:

$$\mu_y = \beta' \mu_z. \tag{1.25}$$

A famous example is the permanent income hypothesis (PIH), where μ_y is permanent consumption and μ_z is permanent income, so the income elasticity of consumption is unity $\forall \beta$. Most demand and supply functions relate to expected plans of agents; expectations and Euler equation models involve conditional first moments, as do GMM approaches; policy relates planned instruments to expected targets, etc. Since constructs like μ_y and μ_z are inherently unobservable, additional assumptions are needed to complete the model. For example, Friedman (1957) uses:

$$y_t = \mu_y + \epsilon_{y,t} \text{ and } z_t = \mu_z + \epsilon_{z,t} \text{ where } \mathsf{E}\left[\epsilon_{y,t}\epsilon_{z,t}\right] = 0, \tag{1.26}$$

which precludes weak exogeneity of z_t for β given the dependence between the means in (1.25). Allowing μ_y to also depend on second moments would not alter the thrust of the following analysis.

Econometrics, however, depends on second moments of observables. Consider the regression:

$$y_t = \gamma' \mathbf{z}_t + v_t \text{ where } v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right]. \tag{1.27}$$

For $\mathbf{z}_t \sim \mathsf{IN}_n\left[\mu_z, \Sigma_{zz}\right]$ with $\mathsf{E}\left[\mathbf{z}_t v_t\right] = \mathbf{0}\ \forall t$:

$$\mathsf{E}\left[y_t \mid \mathbf{z}_t\right] = \gamma' \mathbf{z}_t, \tag{1.28}$$

then, for $\mathbf{y}' = (y_1 \dots y_T)$ and $\mathbf{Z}' = (\mathbf{z}_1 \dots \mathbf{z}_T)$:

$$\widehat{\gamma} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \mathbf{Z}'\mathbf{y}, \tag{1.29}$$

so that second moments are used to estimate γ. Here, (1.27) entails $\mathsf{E}[y_t] = \gamma' \mathsf{E}[\mathbf{z}_t]$, and from (1.28):

$$\mathsf{E}\left[\mathbf{z}_t y_t\right] = \mathsf{E}\left[\mathbf{z}_t \mathbf{z}_t'\right] \gamma \text{ or } \sigma_{yz} = \Sigma_{zz}\gamma, \tag{1.30}$$

both of which involve γ. Thus, there seems to be no difference between how means and variances are related, which is why second moments can be used to infer about links between first moments. However, when any relation like (1.25) holds, then σ_{yz} and Σ_{zz} in (1.30) must be connected by β, not γ, if valid inferences are to result about the parameters of interest β. Weak exogeneity is needed, either directly in (1.27), or indirectly for "instrumental variables." This is more easily seen from the joint distribution:

$$\begin{pmatrix} y_t \\ \mathbf{z}_t \end{pmatrix} \sim \mathsf{IN}_{n+1}\left[\begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix}, \begin{pmatrix} \sigma_{yy} & \sigma_{yz}' \\ \sigma_{zy} & \Sigma_{zz} \end{pmatrix}\right], \tag{1.31}$$

so given (1.25):

$$\mathsf{E}\left[y_t \mid \mathbf{z}_t\right] = \mu_y - \gamma' \mu_z + \gamma' \mathbf{z}_t = (\beta - \gamma)' \mu_z + \gamma' \mathbf{z}_t, \tag{1.32}$$

which coincides with (1.28) only if $\gamma = \beta$, so means and variances are then related by identical parameters.

Note from (1.31):

$$\mathbf{z}_t = \mu_z + \mathbf{u}_t, \tag{1.33}$$

so impulse responses cannot be identified uniquely as originating from perturbing $\mathbf{u}_t$ or μ_z. But from (1.32), the response of y_t to these perturbations in (1.33) will differ unless $\gamma = \beta$, so weak exogeneity is essential for unique impulse responses, which cannot be based on an arbitrary choice of Cholesky decompositions (only one variant could coincide with valid conditioning).

1.4.6 Functional form

> Alice began to remember that she was a Pawn, and that it would soon be time to move. (Lewis Carroll, 1899)

In practice, one cannot expect every functional form specification to coincide with that which generated the data, however well-based its logic or theory credentials. There are theories of what various linear and other approximations deliver (see, e.g., White, 1980b, 2008), but such approximations cannot ensure non-systematic residuals. Automatic model selection has been extensively applied to select functional forms from quite large classes using data evidence (see, *inter alia*, Perez-Amaral, Gallo and White, 2003, 2005; Castle, 2005; Castle and Hendry, 2005, and section 1.5). In low-dimensional models, semiparametric and nonparametric methods are often used to avoid specifying the functional form, but can be susceptible to unmodeled outliers and breaks.

1.4.7 Identification

> . . . watching one of them that was bustling about among the flowers, poking its proboscis into them, "just as if it was a regular bee," thought Alice. However, this was anything but a regular bee: in fact, it was an elephant. (Lewis Carroll, 1899)

Identification has three attributes of uniqueness, interpretation, and correspondence to reality (see, e.g., Hendry, 1995a), which we discuss in turn. Since unidentified parameters entail a non-unique model specification – so what is estimated need not match the parameters of the generating process – identification is a fundamental attribute of a parametric specification.

First, a general understanding of identification as uniqueness has been developed (see, e.g., Fisher, 1966; Rothenberg, 1971; Sargan, 1983: Hsiao, 1983, provides an overview), building on the rank and order conditions so well known to be necessary

and sufficient in simultaneous systems when the restrictions are given by subject-matter theory: we could call this a technical issue. Cowles Commission researchers showed that the "reduced form" (or statistical system) was always identified in their formulation, and that all just-identified models were isomorphic to that statistical system, hence tests of overidentified "structural forms" could be derived by comparing their two likelihoods. Their analysis, therefore, entailed that the "structural form" is actually a *reduction* of the statistical system, so logically can be obtained from it without any prior knowledge of the relevant restrictions. Thus, when a model is identified relative to an identified system, the identification restrictions in question do not have to be known *a priori*, but can be found by a suitable algorithm (see, e.g., Hendry and Krolzig, 2005) – indeed, several overidentified, but distinct, representations can coexist (see, e.g., Hendry, Lu and Mizon, 2008). Such a conclusion is predicated on the statistical system itself being identified, which requires sufficient explanatory variables – the vexed topic of "exogeneity" discussed in section 1.4.5 above. Moreover, for an overidentification test to be valid, the statistical system must be well specified, so needs to be modeled and evaluated first (see, e.g., Spanos, 1990), after which it can be reduced to a "structural form." Nevertheless, that prior identification restrictions must be known in advance remains the dominant belief, which if true, would preclude empirical modeling not preceded by a rigorous theory derivation that entailed sufficient restrictions.

Second, interpretation is a regular seminar question along the lines: "How do you know you have identified the demand curve" (as opposed to some other entity)? This is essentially an economic theory issue, and only substantive theory can resolve such a debate. It is separate from uniqueness: a regression of price on quantity is always unique, but hardly qualifies as a demand curve just because the regression coefficient is negative.

Third, even if both uniqueness and interpretation are confirmed, the result still need not correspond to reality, which is an empirical issue (and related to the usage of the word "identification" in, say, Box and Jenkins', 1976, analysis, as well as the quote above). An estimated equation may be unique and interpretable but not the relevant relation. Thus, all aspects of model building are involved in establishing satisfactory identification.

Recently, problems of weak instruments, and the resulting issue of identification, have become salient (see, among others, Staiger and Stock, 1997; Stock and Wright, 2000; Stock, Wright and Yogo, 2002; Kleibergen, 2002; Mavroeidis, 2004).

1.4.8 Parameter constancy

> "Yes, all his horses and all his men," Humpty Dumpty went on. "They'd pick me up again in a minute, they would!" (Lewis Carroll, 1899)

Parameters are the entities which must be constant if the specified model is to be a useful characterization of reality. However, that does not preclude the coefficients in any model formulation from changing, as in "random coefficients" models or "structural time series" (see, e.g., Hildreth and Houck, 1968; Harvey,

1993). The main problem for economic forecasting using econometric models is that coefficients of deterministic terms do not seem to stay constant, but suffer location shifts, which in turn induce forecast failure (see, e.g., Clements and Hendry, 2005, 2006). While changes in zero-mean variables seem less damaging to forecasts (see, e.g., Hendry and Doornik, 1997), such breaks nevertheless remain pernicious for policy analyses.

1.4.9 "Independent" homoskedastic errors

> "Contrariwise," continued Tweedledee, "if it was so, it might be; and if it were so, it would be: but as it isn't, it ain't. That's logic." (Lewis Carroll, 1899)

Joint densities can always be factorized into sequential forms, as with martingale difference sequences. Moreover, equations can often be standardized to be homoskedastic by dividing by contemporaneous error variances (when these exist), so this category may be one of the least stringent requirements.

1.4.10 Expectations formation

> "What sort of things do you remember best?" Alice ventured to ask. "Oh, things that happened the week after next" the Queen replied in a careless tone. (Lewis Carroll, 1899)

Surprisingly little is known about how economic agents actually form their expectations for variables relevant to their decisions. Almost no accurate expectations data exist outside financial market traders, so resort is usually needed to proxies for the unobserved expectations, or to untested assumptions, such as "rational" expectations (RE), namely the correct conditional expectation $\mathsf{E}[\cdot]$ of the variable in question (y_{t+1}) given the available information ($\mathcal{I}_t$). There is a large gap between economic theory models of expectations – which often postulate that agents hold RE – and the realities of economic forecasting, where forecast failure is not a rare occurrence. The "rational" expectation is often written as (see Muth, 1961):

$$y^{re}_{t+1} = \mathsf{E}\left[y_{t+1} \mid \mathcal{I}_t\right], \tag{1.34}$$

which implicitly assumes free information and free computing power as available information is vast. The usual argument, perhaps loosely worded to avoid contradictions, is that otherwise there would be arbitrage opportunities, or agents would suffer unnecessary losses. But expectations are instrumental to agents' decisions, and the accuracy thereof is not an end in itself, so agents should just equate the marginal benefits of improved forecast accuracy against the extra costs of achieving that, leading to "economically rational expectations" (ERE) (see Aghion *et al.*, (2002)). "Model consistent expectations" instead impose the expectations formation process as the solved estimated model specification, so – unless the model is perfect – suffer the additional drawback of imposing invalid restrictions.

While ERE may be more realistic than RE, it still assumes knowledge of the form of dependence of y_{t+1} on the information used: as expressed in $\mathsf{E}\left[y_{t+1}|\mathcal{I}_t\right]$ in (1.34),

that assumption corresponds to agents knowing precisely what the conditioning operator is. In a stationary world, one could imagine learning mechanisms that eventually led to its discovery (see, e.g., Evans and Honkapohja, 2001). However, in a wide-sense non-stationary environment, an explicit statement of the form of (1.34) is:

$$y^{re}_{t+1} = \mathsf{E}_{t+1}\left[y_{t+1} \mid \mathcal{I}_t\right] = \int y_{t+1} f_{t+1}\left(y_{t+1} \mid \mathcal{I}_t\right) \mathrm{d}y_{t+1}. \tag{1.35}$$

Thus, when $f_t(\cdot) \neq f_{t+1}(\cdot)$, agents need to *know the future* conditional density function $f_{t+1}(y_{t+1}|\mathcal{I}_t)$, given present information, to obtain the appropriate conditioning relation, since only then will y^{re}_{t+1} be an unbiased predictor of y_{t+1}. That $f_t(\cdot) \neq f_{t+1}(\cdot)$ is precisely why forecasting is so prone to problems. Unfortunately, knowing $f_{t+1}(\cdot)$ virtually requires agents to have crystal balls that genuinely "see into the future." When distributions are changing over time, agents can at best form "sensible expectations," y^{se}_{t+1}, based on forecasting $f_{t+1}(\cdot)$ by $\widehat{f}_{t+1}(\cdot)$ from some rule, such that:

$$y^{se}_{t+1} = \int y_{t+1} \widehat{f}_{t+1}\left(y_{t+1} \mid \mathcal{I}_t\right) \mathrm{d}y_{t+1}. \tag{1.36}$$

There are no guaranteed good rules for estimating $f_{t+1}(y_{t+1}|\mathcal{I}_t)$ when $\{y_t\}$ is wide-sense non-stationary. In particular, when the conditional moments of $f_{t+1}(y_{t+1}|\mathcal{I}_t)$ are changing in unanticipated ways, setting $\widehat{f}_{t+1}(\cdot) = f_t(\cdot)$ could be a poor choice, yet that underlies most of the formal derivations of RE, which rarely distinguish between $f_t(\cdot)$ and $f_{t+1}(\cdot)$. Outside a stationary environment, agents cannot solve (1.34), or often even (1.35). The drawbacks of (1.34) and (1.35), and the relative success of robust forecasting rules (see, e.g., Clements and Hendry, 1999; Hendry, 2006), suggest agents should use them, an example of imperfect-knowledge expectations (IKE) (see Aghion *et al.*, 2002; Frydman and Goldberg, 2007). IKE acknowledges that agents cannot know how $\mathcal{I}_t$ enters $f_t(\cdot)$ when processes are evolving in a non-stationary manner, let alone $f_{t+1}(\cdot)$, which still lies in the future. Collecting systematic evidence on agents' expectations to replace the unobservables by estimates, rather than postulates, deserves greater investment (see, e.g., Nerlove, 1983).

Finally, take expectations conditional on the available information set $\mathcal{I}_{t-1}$ in a regression model with valid weak exogeneity:

$$y_t = \beta' \mathbf{z}_t + \epsilon_t, \tag{1.37}$$

so that:

$$\mathsf{E}\left[y_t \mid \mathcal{I}_{t-1}\right] = \beta' \mathsf{E}\left[\mathbf{z}_t \mid \mathcal{I}_{t-1}\right], \tag{1.38}$$

as $\mathsf{E}\left[\epsilon_t|\mathcal{I}_{t-1}\right] = 0$. Writing (1.38) as $y^e_t = \beta' \mathbf{z}^e_t$, the conditional model (1.37) always has an expectations representation, although the converse is false. Importantly, therefore, contemporaneous conditioning variables can also be expectations variables, and some robust forecasting rules like $\Delta\widehat{p}_{t+1} = \Delta p_t$ have that property.

The New Keynesian Phillips curve is perhaps the best-known model which includes expected inflation to explain current inflation. Models of this type are usually estimated by replacing the expected value by the actual future outcome, then using IV or GMM to estimate the resulting parameters, as in, say, Galí, Gertler and Lopez-Salido (2001). As shown in Castle *et al.* (2008), since breaks and regime shifts are relatively common, full-sample estimates of equations with future values can deliver spuriously significant outcomes when breaks are not modeled, a situation detectable by impulse saturation (see section 1.5).

1.4.11 Estimation

> "I was wondering what the mouse-trap was for," said Alice. "It isn't very likely there would be any mice on the horse's back."
>
> "Not very likely, perhaps," said the Knight; "but if they do come, I don't choose to have them running all about." (Lewis Carroll, 1899)

Developing appropriate estimators comprises a major component of extant econometric theory, and given any model specification, may seem an uncontentious task. However, only in recent decades has it been clear how to avoid (say) nonsense correlations in non-stationarity data, or tackle panel dependencies, so unknown pitfalls may still lurk.

1.5 Model selection

> In another moment Alice was through the glass, and had jumped lightly down into the Looking-glass room. (Lewis Carroll, 1899)

Model selection is the empirical route whereby many of the simplifications in sections 1.4.2.2 and 1.4.2.3 are implemented in practice. In that sense, it is not a distinct step *per se*, but a way of carrying out some of the earlier steps, hence our treating the topic in a separate section.

Selection remains a highly controversial topic. It must be granted that the best approaches cannot be expected to select the LDGP on every occasion, even when the GUM nests the LDGP, and clearly cannot do so ever when the LDGP is not a nested special case. However, that statement remains true when the GUM is exactly the LDGP, but conventional inference is nevertheless undertaken to check that claim. If the LDGP were known at the outset of a study, apart from the unknown values of its parameters, then if any specification or misspecification testing was undertaken, one could only end by doubting the claim that the initial formulation was indeed the LDGP. The least worst outcome would be weak confirmation of the prior specification, and otherwise either some included variables will be found insignificant, or some assumptions will get rejected, casting doubt on the claim. That is the risk of undertaking statistical inference. The alternative of not testing claimed models is even less appealing, namely never learning which ones are useless. To quote Sir Francis Bacon: "If a man will begin with certainties, he

shall end in doubts; but if he will be content to begin with doubts he shall end in certainties."

Conversely, the list at the beginning of section 1.4 makes it clear that "model uncertainty" comprises much more than whether one selected the "correct model" from some set of candidate variables that nested the LDGP. If, say, 1,000 possibly lagged, nonlinear functions of a set of candidate exogenous variables in a model with many breaks are checked for relevance at a significance level of 0.1%, and all are indeed irrelevant, then on average *one* will be retained adventitiously, so uncertainty is greatly reduced by eliminating about 999 potential influences. The entire point of model selection is to reduce some of the uncertainties about the many aspects involved in model specification, and the cost for doing so is a "local increase" in uncertainty as to precisely which influences should be included and which excluded around the margin of significance. Thus, embedding the claimed theory in a more general specification that is congruent with all the available evidence offers a chance to both utilize the best available theory insights and learn from the empirical evidence. Since such embedding can increase the initial model size to a scale where a human has intellectual difficulty handling the required reductions, we next consider computerized, or automatic, methods for model selection.

1.5.1 Automatic model selection

> "Does – the one – that wins – get the crown?" she asked, as well as she could, for the long run was putting her quite out of breath.
>
> "Dear me, no!" said the King. "What an idea!" (Alice to the White King in Lewis Carroll, 1899)

The many alternatives now available include, but are not restricted to, Phillips (1994, 1995, 1996), Tibshirani (1996), Hoover and Perez (1999, 2004), Hendry and Krolzig (1999, 2001), White (2000), Krolzig (2003), Kurcewicz and Mycielski (2003), Demiralp and Hoover (2003), and Perez-Amaral *et al*. (2003); also see the special issue on model selection edited by Haldrup, van Dijk and Hendry (2003) (the references cited therein provide bibliographic perspective on this huge literature). Complaints about model selection have a long pedigree, from Keynes (1939) about "data-based modeling" and Koopmans (1947) on "measurement without theory," through "pre-test biases" from test-based selection in Judge and Bock (1978); "repeated testing" inducing adventitious significance in Leamer (1978, 1983) and Lovell (1983) criticizing selection rules seeking "significance," to Pagan (1987) on the potential "path dependence of any selection"; Hendry, Leamer and Poirier (1990) debating "arbitrary significance levels"; Chatfield (1995) criticizing "ignoring selection effects" as misrepresenting uncertainty, and Faust and Whiteman (1997) on "lack of identification," but most have now been rebutted (see, e.g., Hendry, 2000a). Concerning Keynes' comment quoted above, not only should everyone get the same answer from an automatic algorithm applied to the same GUM using the same selection criteria, investigators with different GUMs, which differed only by irrelevant variables, could also end with the same model.

Here we consider Autometrics, an Ox package (see Doornik, 2006, 2007a) implementing automatic Gets modeling based on the theory of reduction discussed above. The present implementation of Autometrics is primarily for linear regression models, but extensions have been derived theoretically to automatically model dynamic, cointegrated, simultaneous systems; nonlinear equations; structural breaks; more variables (N) than observations (T); and testing exogeneity (see, e.g., Hendry and Krolzig, 2005; Castle and Hendry, 2005; Hendry, *et al.*, 2008; Johansen and Nielsen, 2008; Doornik, 2007b; and Hendry and Santos, 2009, respectively). Given any available theoretical, historical, institutional, and measurement information, as well as previous empirical evidence, a GUM must be carefully formulated, preferably with a relatively orthogonal parameterization, a subject-matter basis, and must encompass existing models. When $T \gg N$, the GUM can be estimated from all the available evidence, and rigorously tested for congruence. If congruence fails, a new formulation is required: but at least one has learned the general inadequacy of a class of models. If congruence is accepted, it is then maintained throughout the selection process by not following simplification paths which are rejected on diagnostic checking (using the same statistics), ensuring a congruent final model. When $N > T$, as must happen when impulse saturation is used and can occur more generally (discussed below), misspecification testing can only be undertaken once a feasible model size $n < T$ has been reached.

Statistically insignificant variables are eliminated by selection tests, using a tree-path search in Autometrics, which improves on the multi-path procedures in Hoover and Perez (1999) and Hendry and Krolzig (2001). Checking many paths prevents the algorithm from becoming stuck in a sequence that inadvertently eliminates a variable which actually matters, and thereby retains other variables as proxies (as in stepwise regression). Path searches terminate when no variable meets the elimination criteria. Non-rejected (terminal) models are collected, then tested against each other by encompassing: if several remain acceptable, so are congruent, undominated, mutually encompassing representations, the search is terminated using, e.g., the Schwarz (1978) information criterion, although all are reported and can be used in, say, forecast combinations.

To understand why an automatic search procedure might work, consider a case where the complete set of N candidate regressors is mutually orthogonal, but which ones are relevant is unknown *a priori*, and $T \gg N$. The postulated GUM nests the LDGP. Estimate the GUM, then, squaring to eliminate signs, rank the resulting t_i^2 statistics from the largest to the smallest. When c_α is the criterion for retention, let n be such that $t_n^2 \geq c_\alpha$ when $t_{n+1}^2 < c_\alpha$. Then select the model with those n regressors. That required precisely *one* decision – what to include, and hence what to exclude. No issues of search, repeated testing, path dependence, etc., arise. Goodness-of-fit is not directly used to select models; and no attempt is made to "prove" that a given number of variables matters. In practice, the role of the tree search is to ascertain "true" relevance when orthogonality does not hold; and the choice of c_α affects R^2 and n through retention of t_n^2. Generalizations to other maximum likelihood estimators, or approximations thereto such as IV, are feasible (see Hendry and Krolzig, 2005; Doornik, 2007a).

However, it does matter that selection occurs: the selected model's estimates do not have the same properties as if the LDGP equation had been estimated without any testing. Sampling vagaries entail that some variables which enter the LDGP will by chance have a sample $t^2 < c_\alpha$ (low power). Since they are only retained when $t^2 \geq c_\alpha$, their estimated magnitudes will be biased away from the origin, and hence selected coefficients need to be bias corrected, which is relatively straightforward (see Hendry and Krolzig, 2005). Some variables which are irrelevant will have $t^2 \geq c_\alpha$ (adventitiously significant), where the probability of that event is $\alpha\left(N - n^*\right)$ when n^* variables actually matter. Fortunately, bias correction will also drive such estimates sharply towards the origin. Thus, despite selecting from a large set of potential variables, nearly unbiased estimates of coefficients and equation standard errors can be obtained with little loss of efficiency from testing many irrelevant variables, and some loss for relevant, from the increased value of c_α. The normal distribution has "thin tails," so the power loss from tighter significance levels is usually not substantial, whereas financial variables may have fat tails, so power loss could be more costly at tighter α.

Impulse saturation is described in Hendry *et al.* (2008) and Johansen and Nielsen (2008) as including an indicator for every observation, entered (in the simplest case) in blocks of $T/2$, with the significant outcomes retained. This approach both helps remove outliers, and is a good example of why testing large numbers of candidate regressors does not cost much efficiency loss under the null that they are irrelevant. Setting $c_\alpha \leq 1/T$ maintains the average false null retention at one "outlier," and that is equivalent to omitting one observation, so is a tiny efficiency loss despite testing for the relevance of T variables. Since all regressors are exact linear functions of T impulses, that effect carries over directly in the independent and identically distributed (i.i.d.) setting, and in similar ways more generally. Thus, $N > T$ is not problematic for automatic model selection, opening the door to large numbers of new applications.

Since an automatic selection procedure is algorithmic, simulation studies of its operational properties are straightforward. In the Monte Carlo experiments reported in Hendry and Krolzig (2005), commencing from highly over-parameterized GUMs (between 8 and 40 irrelevant variables; zero and 8 relevant), PcGets recovered the LDGP with an accuracy close to what one would expect if the LDGP specification were known initially, but nevertheless coefficient tests were conducted. To summarize its simulation-based properties, false rejection frequencies of null hypotheses (measured as retention rates for irrelevant variables) can be controlled at approximately the desired level; correct rejections of alternatives are close to the theoretical upper bound of power (measured as retention rates for relevant variables); model selection is consistent for a finite model size as the sample size grows without bound; nearly unbiased parameter estimates can be obtained for all variables by bias-correction formulae, which also reduce the mean square errors of adventitiously retained irrelevant variables; and reported equation standard errors are nearly unbiased estimates of those of the correct specification (see, e.g., Hendry and Krolzig, 2005). Empirically, automatic Gets selects

(in seconds) models at least as good as those developed over several years by their authors (see Ericsson, 2007, for several examples). Although automatic model selection is in its infancy, exceptional progress has already been achieved (see Hoover and Perez, 1999; Hoover and Perez, 2004, provide additional evidence).

1.5.2 Costs of inference and costs of search

> "Don't keep him waiting, child! Why, his time is worth a thousand pounds a minute!" (Train passengers to Alice in Lewis Carroll, 1899)

Costs of inference are inevitable when tests have non-zero size and non-unit power, even if investigators commence from the LDGP – but do not know that is the correct specification, so have to test for congruence and significance. Costs of search are due to commencing from any GUM that is over-parameterized relative to the LDGP. Under-specification ensures that an invalid model of the LDGP will result. Given the many criticisms of model selection, it may surprise readers that costs of search are small in comparison to costs of inference: the main difficulty is not selection *per se*, but the vagaries of sampling. In selecting a model from a GUM, there are two possible mistakes. The first is including irrelevant variables (ones not in the LDGP), the second is omitting relevant variables. Since the first group are absent when the DGP is the GUM, that is purely a cost of search. The second is primarily a cost of inference, with possible additional search costs if there are lower probabilities of retaining relevant variables when commencing from the GUM.

When the nominal rejection frequency of individual selection tests is set at $\alpha \leq 1/N \rightarrow 0$ as $T \rightarrow \infty$, on average at most one irrelevant variable will be retained as adventitiously significant out of N candidates. Thus, there is little difficulty in eliminating almost all of the irrelevant variables when starting from the GUM (a small cost of search). The so-called overall "size" of the selection procedure, namely $1-(1-\alpha)^N$, can be large, but is uninformative about the success of a simplification process that on average correctly eliminates $(1-\alpha)N$ irrelevant variables.

Conversely, even for a loose significance level like $\alpha = 0.05$, and commencing from the LDGP, there is only a 50% chance of keeping a relevant variable where the t-test on its coefficient has a non-centrality of 2 (a high cost of inference). A more stringent critical value (say $\alpha = 0.01$, so $c_\alpha \simeq 2.63$) worsens the costs of inference as the retention probability falls to 27% despite the correct specification being postulated. Costs of inference usually exceed costs of search, the exception being when all relevant variables have large non-central t-statistics (in excess of about ± 5), so there are no costs of inference. The probabilities of locating the LDGP commencing from the GUM are reasonably close to the corresponding outcomes when the search commences from the LDGP. Since the LDGP is sometimes never retained even when it is the initial specification, the apparent problem of a search algorithm may be a cost of inference.

The limits of automatic model selection must also be clarified. If the LDGP equation would not be reliably selected by the given inference rules applied to itself as the initial specification, then selection methods cannot rectify that. Many

apparent criticisms of selection have failed to note that key limitation. In the simulations described above, the same algorithm and selection criteria were always applied to commencing from both the GUM and the LDGP, and only the additional costs attributable to starting from the former comprise search costs. Also, when there are relevant variables with small t-statistics because the parameters are $\mathsf{O}(1/\sqrt{T})$, especially if they are highly correlated with other regressors (see Pötscher, 1991; Leeb and Pötscher, 2003, 2005), then selection is not going to work well: one cannot expect success in selection if a parameter cannot be consistently estimated. Thus, although uniform convergence seems infeasible, selection works for parameters larger than $\mathsf{O}(1/\sqrt{T})$ (as they are consistently estimable) or smaller than $\mathsf{O}(1/T)$ (as they vanish), yet $1/\sqrt{T}$ and $1/T$ both converge to zero as $T \to \infty$, so "most" parameter values are unproblematic. If the LDGP would always be retained by the algorithm when commencing from it, then a close approximation will generally be selected when starting from a GUM which nests that LDGP.

Additional problems for any empirical modeling exercise arise when the LDGP is not nested in the GUM, due to the regressor set being incomplete, the functional form misspecified or structural breaks and other non-stationarities not being fully accommodated, as well as serious measurement errors contaminating the data or endogenous variables being incorrectly treated as regressors. For very high levels of collinearity between relevant and irrelevant variables, the selected approximation may use the incorrect choice if that is undominated, but in a progressive research strategy when there are intermittent structural breaks in both relevant and irrelevant variables, such a selection will soon be dominated. Phillips (2003) provides an insightful analysis of the limits of econometrics.

1.6 Teaching "Applied Econometrics"

> "Manners are not taught in lessons," said Alice. "Lessons teach you to do sums, and things of that sort."
>
> "And you do Addition?" the White Queen asked. "What's one and one and one and one and one and one and one and one and one and one?"
>
> "I don't know," said Alice. "I lost count." (Lewis Carroll, 1899)

Both economic theory and theoretical econometrics are relatively structured subjects to teach, whereas applied econometrics is not, so many approaches are extant. The obvious way might be to include substantive empirical findings in the relevant subject-matter part of other economics courses, and so effectively abolish the need to teach what applied econometrics has established. This certainly happens in part, usually with a lag after the relevant study was published, but seems less common than courses specifically oriented to applied econometrics. I was taught in such a course, the bulk of which concerned studying how the "masters" had conducted their investigations, and what they found – essentially an apprenticeship. Other courses focus more on the economic and econometric theory behind key studies, with less attention to their empirical outcomes: systems of demand equations seem to be addressed that way. Presumably the aim is to explicate the relation between

economics and its applications in particular cases. Finally, some courses require students to undertake empirical work themselves, often replicating or evaluating existing studies rather than novel research. Combinations of some or all of these also happen.

If the objective is one where completing students are to be able to reliably tackle a new application, then teaching applied econometrics becomes very demanding. A wide range of skills and insights need to be conveyed, many of which concern "auxiliary" issues such as data availability, its quality and its correspondence to the target of the analysis, including frequency, seasonality, transformations, etc.; institutions and policy agencies that impinge on the economic process; important historical and non-economic contingencies that occurred; the specification of the candidate list, their dynamics, exogeneity, functional forms and constancy of possible models; and the use of software. When the first attempt fails on the desired criteria, a revision process is needed, so difficulties of expanding searches and sequentially correcting problems with a model must be confronted, all too often leaving the student floundering.

A key job of an applied econometrician is to formulate the general model that underpins the analysis, which includes the specification of all candidate variables that might influence the "target" variables of interest in the modeling problem, their general functional forms (e.g., logs), and putative exogeneity assumptions. General economic reasoning plays a substantive part at this stage. Further, one must collect and carefully check all the data series to be modeled, and investigate their historical context. Finally, the easier part is using appropriate software to congruently model the relevant series. Yet, many studies by "experts" remain clever "detective exercises" in which a feel for the evidence helped point towards a viable conclusion. The approach in Hendry and Nielsen (2007a), summarized in Hendry and Nielsen (2007b), is to first prepare students to understand the elements of likelihood theory, using likelihood ratio tests for inference and evaluation – testing the assumptions for the validity of those inferences – leading to model selection in the econometric theory part of the course. A sequence of increasingly realistic theoretical models is developed from i.i.d. binary data through to cointegrated equations with structural breaks. On the applied part of the course, we thoroughly examine an agreed data set, and after teaching the relevant software, students can rapidly move from simple models to general ones using automatic methods. Our focus is on developing well-specified empirical models of interesting economic issues. Given a new problem, students then have a structured approach to follow in their investigations. We consider there has been a marked improvement in their resulting empirical studies.

An example of my own approach is recorded in Hendry (1999): there may be some "tacit knowledge" therein (and I hope there is value added), but most of the above steps can be formalized without the need for an extensive apprenticeship. The next section focuses on comparing how well automatic model selection does without any "prior" historical knowledge or empirical modeling experience. The results reported in section 1.7 took about 20 minutes of real time, including the write-up: even granted that the data were pre-prepared and the log

transformations were entailed by the pre-analysis, the efficiency gains over my previous "handcrafted" study are huge. In addition, hypotheses that were previously imposed or sequentially investigated can be evaluated jointly.

1.7 Revisiting the "experiment in applied econometrics"

> "If I wasn't real," Alice said – half-laughing though her tears, it all seemed so ridiculous – "I shouldn't be able to cry."
>
> "I hope you don't suppose those are real tears?" Tweedledum interrupted in a tone of great contempt. (Lewis Carroll, 1899)

The recent huge increases in the prices of many foods makes the exercise of re-examining US food expenditure over 1931–89 (based on the update of Tobin, 1950, by Magnus and Morgan, 1999) of more than just historical interest. If the various price and income elasticities estimated below are approximately correct, then substantial responses can be anticipated (indeed, could this be the long-sought solution to society's burgeoning obesity problem?). Hendry (1999) sought to explain why other contributors to Magnus and Morgan (1999) had found their models were non-constant over the combined inter-war and post-war samples, so had eschewed modeling the 1930s data. Impulse dummies for a food program and post-war de-rationing allowed a constant equation to be developed over the sample 1931–89. While an indicator variable is a crude level of measurement, the converse strategy of not modeling major institutional interventions seems even less attractive. Theory and common sense suggest that food programs and switches in rationing matter; but few theory models allow for such factors in a way suitable for empirical implementation (although the original analyst of this data also published on rationing in Tobin, 1952).

The per capita variables are as follows (lower case denotes logs):

e_f: constant-price expenditure on food
$p_f - p$: real price of food
e: total constant-price expenditure
$s = \log Y - \log E$: (an approximation to the savings ratio)
a: average family size.

Figure 1.6 shows the time series, and reveals considerable changes over the period. After falling sharply at the commencement of the Great Depression, both e_f and e rise substantially till World War II, fall after, then resume a gentle rise (panels (a) and (c)), so $\Delta e_{f,t}$ is vastly more volatile pre-war (panel (e)) (Δe_t has a similar but less pronounced pattern). Next, $p_f - p$ is quite volatile till after the war, then is relatively stable (panel (b)), whereas the dramatic rise in s from "forced saving" during the war is manifest (panel (d)).

The earlier study of a cointegrated VAR for the system established that e, s, and $p_f - p$ were weakly exogenous in the food demand equation. Here, the general conditional model allowed for two lags on each of e_f, e, $p_f - p$, s and one lag

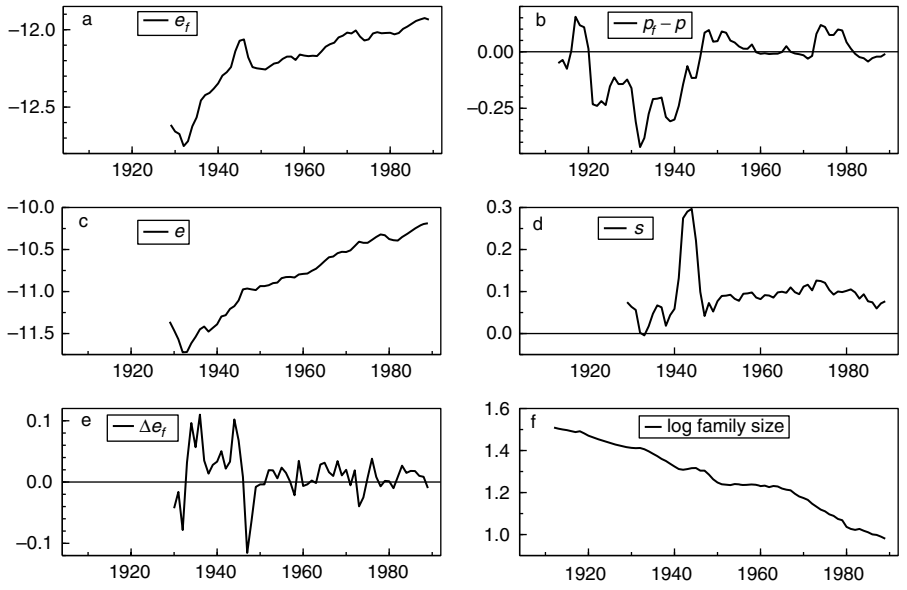


Figure 1.6 Food expenditure and related time series

on a, and was selected by Autometrics at 1% for all candidate variables, including impulse saturation. All diagnostic tests were insignificant, and the PcGive unit root test strongly rejected the null of no cointegration ($t_{ur} = -11.37^{**}$: (see Banerjee and Hendry, 1992; Ericsson and MacKinnon, 2002) with the long-run solution:

$$c_0 = e_f + 7.99 - 0.4e + 0.36(p_f - p). \tag{1.39}$$

Transforming to differences and the equilibrium-correction term from (1.39), Autometrics selected over 1931–89 (at 2.5%, again including impulse saturation):

$$\begin{aligned}
\Delta e_{f,t} = {} & \underset{(0.02)}{0.34}\, s_{t-1} - \underset{(0.02)}{0.32}\, c_{0,t-1} + \underset{(0.04)}{0.67}\, \Delta e_t + \underset{(0.03)}{0.13}\, \Delta e_{t-1} \\
& - \underset{(0.03)}{0.64}\, \Delta(p_f - p)_t - \underset{(0.01)}{0.09}\, \mathrm{I}_{31} - \underset{(0.01)}{0.10}\, \mathrm{I}_{32} + \underset{(0.01)}{0.04}\, \mathrm{I}_{34} \\
& + \underset{(0.01)}{0.03}\, \mathrm{I}_{41} + \underset{(0.01)}{0.05}\, \mathrm{I}_{42} + \underset{(0.01)}{0.03}\, \mathrm{I}_{51} + \underset{(0.01)}{0.02}\, \mathrm{I}_{52} + \underset{(0.01)}{0.03}\, \mathrm{I}_{70} \\
& \left(\mathsf{R}^*\right)^2 = 0.96 \;\; \mathsf{F}_{\mathsf{M}}(13, 45) = 94.9^{**} \;\; \widehat{\sigma} = 0.0078 \;\; \mathsf{F}_{\mathsf{ar}}(2, 44) = 1.34 \\
& \chi^2(2) = 1.04 \;\; \mathsf{F}_{\mathsf{arch}}(1, 44) = 2.25 \;\; \mathsf{F}_{\mathsf{reset}}(1, 45) = 0.35 \\
& \mathsf{F}_{\mathsf{het}}(18, 27) = 0.48 \;\; \mathsf{F}_{\mathsf{Chow}}(9, 37) = 0.99.
\end{aligned} \tag{1.40}$$

In (1.40), $\left(\mathsf{R}^*\right)^2$ is the squared multiple correlation when a constant is added, $\mathsf{F}_{\mathsf{M}}(13, 45)$ is the associated test of the null, and $\widehat{\sigma}$ is the residual standard deviation, with coefficient standard errors shown in parentheses. The diagnostic tests are of the form $\mathsf{F}_{\mathsf{j}}(k, T - l)$, which denotes an approximate F-test against the alternative hypothesis j for: kth-order serial correlation (F_{ar}: see Godfrey, 1978), kth-order autoregressive conditional heteroskedasticity ($\mathsf{F}_{\mathsf{arch}}$: see Engle, 1982), heteroskedasticity ($\mathsf{F}_{\mathsf{het}}$: see White, 1980a); the RESET test ($\mathsf{F}_{\mathsf{reset}}$: see Ramsey, 1969); parameter constancy ($\mathsf{F}_{\mathsf{Chow}}$ (see Chow, 1960) over k periods; and a chi-square test for normality ($\chi^2_{\mathsf{nd}}(2)$ (see Doornik and Hansen, 2008). No misspecification test rejects.

The result in (1.40) is to be contrasted with the equation reported earlier, which had a similar equilibrium correction term based on Johansen (1988):

$$c_2 = e_f + 7.88 - 0.4e + 0.4\left(p_f - p\right), \tag{1.41}$$

leading to (D3133 and D4446 are dummies with the value unity over the periods 1931–33 and 1944–46 respectively):

$$\begin{aligned}\Delta e_{f,t} = {} & \underset{(0.04)}{0.27}\, s_{t-1} - \underset{(0.02)}{0.34}\, c_{2,t-1} - \underset{(0.004)}{0.019} + \underset{(0.05)}{0.24}\, \Delta s_t \\ & + \underset{(0.05)}{0.53}\, \Delta e_t - \underset{(0.04)}{0.46}\, \Delta(p_f - p)_t - \underset{(0.01)}{0.12}\, \mathrm{D3133} + \underset{(0.010)}{0.038}\, \mathrm{D4446} \\ & \mathsf{R}^2 = 0.936 \;\; \mathsf{F}_{\mathsf{M}}(7, 51) = 107.2^{**} \;\; \widehat{\sigma} = 0.0098 \;\; \mathsf{F}_{\mathsf{ar}}(2, 49) = 0.18 \\ & \chi^2(2) = 0.21 \;\; \mathsf{F}_{\mathsf{arch}}(1, 49) = 0.59 \;\; \mathsf{F}_{\mathsf{reset}}(1, 50) = 0.26 \;\; \mathsf{F}_{\mathsf{het}}(13, 37) = 0.47. \end{aligned} \tag{1.42}$$

Thus, six additional outliers have been detected in (1.40), whereas none of the components of D4446 was found, nor was I_{33}: neither dummy is remotely significant if added to (1.40). Consistent with that result, when (1.42) and (1.40) are denoted models 1 and 2 on encompassing tests, $\mathsf{F}_{\mathsf{Enc}_{1,2}}(10, 41) = 4.83^{**}$ and $\mathsf{F}_{\mathsf{Enc}_{2,1}}(5, 41) = 2.18$, so (1.42) is encompassed by (1.40) but not vice versa. Nevertheless, both models are rejected against the other on Cox (1961) and Ericsson (1983) non-nested tests with $\widehat{\sigma}_J = 0.0074$.

Another recent development that can be implemented based on impulse saturation is to test for the super exogeneity of the parameters of the conditional model in response to changes in the LDGPs of the two main conditioning variables, Δe_t and $\Delta(p_f - p)_t$ (see section 1.4.5 and Hendry and Santos, 2009). The latter's equation revealed no significant breaks, but commencing from one lag of Δe, $\Delta(p_f - p)$, Δs and Δa, the former produced:

$$\begin{aligned}\Delta e_t = {} & \underset{(0.003)}{0.016} + \underset{(0.081)}{0.256}\, \Delta e_{t-1} - \underset{(0.083)}{0.302}\, \Delta(p_f - p)_{t-1} - \underset{(0.02)}{0.11}\, \mathrm{I}_{31} \\ & - \underset{(0.02)}{0.19}\, \mathrm{I}_{32} + \underset{(0.02)}{0.10}\, \mathrm{I}_{34}\end{aligned}$$

$$+ \underset{(0.02)}{0.06}\,\mathrm{I}_{35} + \underset{(0.02)}{0.07}\,\mathrm{I}_{36} - \underset{(0.02)}{0.08}\,\mathrm{I}_{38} + \underset{(0.02)}{0.07}\,\mathrm{I}_{41} + \underset{(0.02)}{0.08}\,\mathrm{I}_{43}$$

$$+ \underset{(0.02)}{0.10}\,\mathrm{I}_{46} - \underset{(0.02)}{0.06}\,\mathrm{I}_{80}$$

$$\mathsf{R}^2 = 0.86 \;\; \mathsf{F}_{\mathsf{M}}(12, 46) = 23.9^{**} \;\; \widehat{\sigma} = 0.019 \;\; \mathsf{F}_{\mathsf{ar}}(2, 44) = 0.14$$

$$\chi^2(2) = 1.27 \;\; \mathsf{F}_{\mathsf{arch}}(1, 44) = 4.42^{*} \;\; \mathsf{F}_{\mathsf{reset}}(1, 45) = 0.01 \;\; \mathsf{F}_{\mathsf{het}}(14, 31) = 0.69. \quad (1.43)$$

Almost all the inter-war and war years are revealed as discrepant, with four impulses in common with (1.40). Adding the 6 additional impulses found in (1.43) to (1.40) and testing their significance yields $\mathsf{F}_{\mathsf{SExog}}(6, 40) = 1.14$, not rejecting. Nevertheless, that there are four impulses in common is strongly against the exogeneity of Δe_t in (1.40), especially as their signs all match, and even the magnitudes are not too far from 0.67 times those in (1.43). It is not surprising that major shifts in total expenditure are associated with shifts in expenditure on a subcomponent, but since Δe_t is included in (1.40), the conclusion must be that agents altered their decision rules more than that effect. Since a food program was in place for several of the common impulses and rationing for the other, additional shifts do not necessarily invalidate the economics behind the equation, so the overall outcome is inconclusive.

An alternative check on the commonality of the inter-war and post-war periods is to use the former to predict the latter, given the actual outcomes for the regressors. We have implicitly done so via impulse saturation, which revealed only one post-war outlier in 1970. The F-test of constancy, $\mathsf{F}_{\mathsf{Chow}}(37, 10) = 2.02$, does not reject. Figure 1.7 shows the outcomes: panel (a) reports the fitted and actual values till 1952 and the predicted thereafter, with the full-sample fit shown immediately below in panel (c), the residuals and forecast errors in panel (b), and one-step 95% forecast intervals in panel (d). The outlier in 1970 is obvious, and otherwise there is little difference between the sub-sample and full-sample fit. Such constancy in the face of changing data behavior supports both the specification in (1.40) and the use of the whole sample to estimate and evaluate these models.

1.7.1 An update

> Everything was happening so oddly that she didn't feel a bit surprised. (Lewis Carroll, 1899)

The obvious extension is to update the data, and test the model on the extended information. Unfortunately, Applied Econometrics is never that easy: the data have been extensively revised. It came as a surprise even to an experienced empirical modeler that data back to 1929 could differ so much between a 1989-based set (denoted by a subscript $_0$ in the graphs) and a 2008 update when extending the data to 2000 (denoted $_1$), but Figures 1.8 (data) and 1.9 (deviations) show the extent of the revisions. Both food and total real expenditure have changed, the latter by up to 15%, and savings have shifted by up to 5%, whereas the relative price of

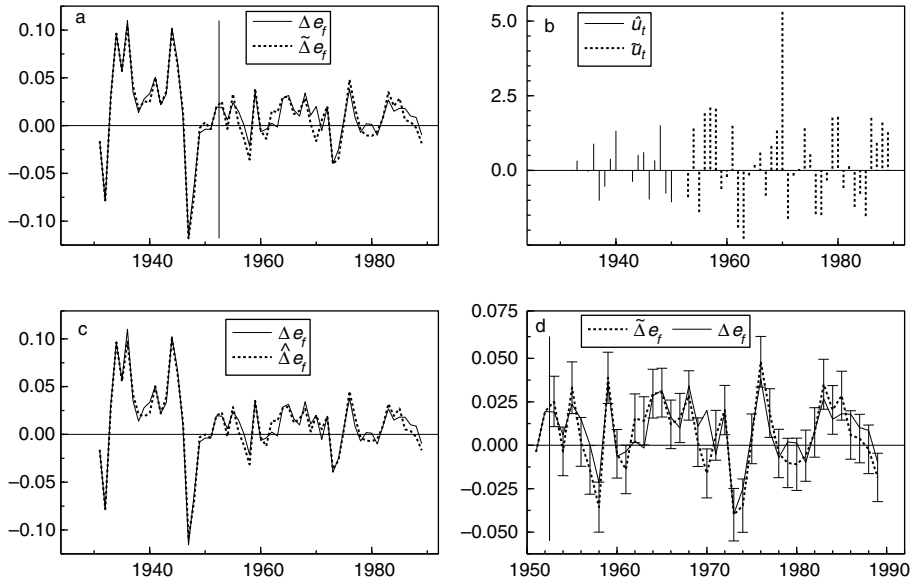


Figure 1.7 Fitted and actual values, residuals and forecasts for $\Delta e_{f,t}$

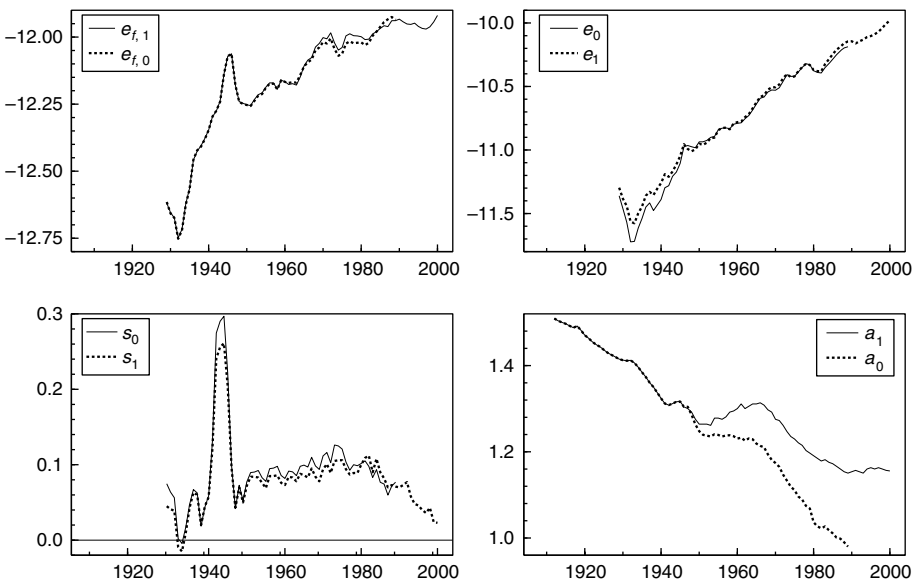


Figure 1.8 Revised data on food expenditure time series

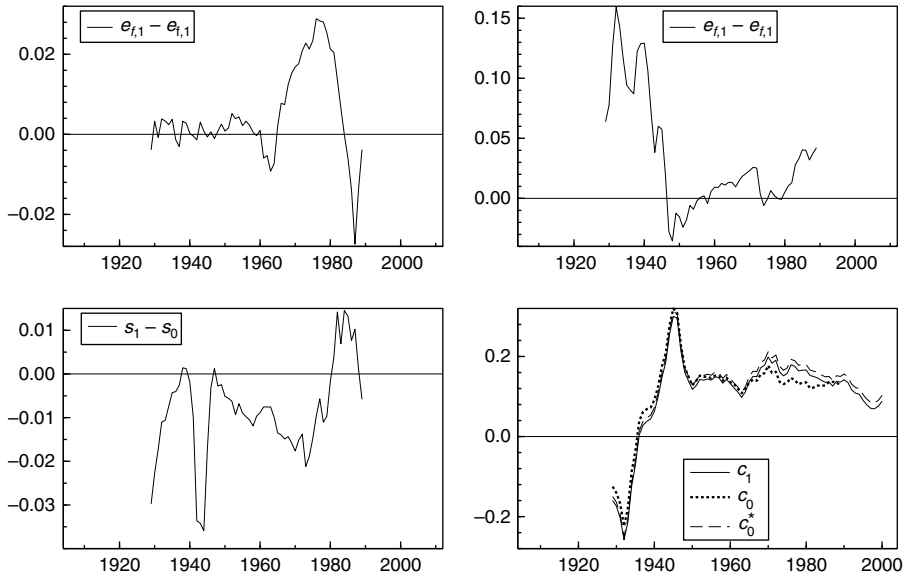


Figure 1.9 Deviations between old and revised data on food expenditure time series

food is unaltered – yet family size is unrecognizably different. The impacts on the equilibrium-correction terms, c_0 in (1.39), that calculated for the revised data c_0^*, and c_1 in (1.45) below, are also shown (see Hendry, 1994; Cook, 2008, on possible approaches for cross-data-vintage encompassing).

First, enforcing the identical specification to (1.40) but on the revised data over 1930–89, testing on the 11 new years led to:

$$\begin{aligned}\Delta e_{f,t} = {} & \underset{(0.04)}{0.34}\, s_{t-1} - \underset{(0.02)}{0.27}\, c_{0,t-1} + \underset{(0.06)}{0.57}\, \Delta e_t + \underset{(0.05)}{0.09}\, \Delta e_{t-1} \\ & - \underset{(0.04)}{0.40}\, \Delta(p_f - p)_t - \underset{(0.01)}{0.10}\, \mathrm{I}_{31} - \underset{(0.01)}{0.12}\, \mathrm{I}_{32} + \underset{(0.01)}{0.04}\, \mathrm{I}_{34} \\ & + \underset{(0.01)}{0.02}\, \mathrm{I}_{41} + \underset{(0.01)}{0.05}\, \mathrm{I}_{42} + \underset{(0.01)}{0.03}\, \mathrm{I}_{51} + \underset{(0.01)}{0.02}\, \mathrm{I}_{52} + \underset{(0.01)}{0.04}\, \mathrm{I}_{70} \\ & \left(\mathsf{R}^*\right)^2 = 0.93 \;\; \mathsf{F}_{\mathsf{M}}(13,45) = 94.9^{**} \;\; \widehat{\sigma} = 0.011 \;\; \mathsf{F}_{\mathsf{ar}}(2,44) = 5.74^{**} \\ & \chi^2(2) = 2.40 \;\; \mathsf{F}_{\mathsf{arch}}(1,44) = 2.79 \;\; \mathsf{F}_{\mathsf{reset}}(1,45) = 0.01 \\ & \mathsf{F}_{\mathsf{het}}(18,27) = 0.82 \;\; \mathsf{F}_{\mathsf{Chow}}(11,46) = 1.42. \end{aligned} \tag{1.44}$$

The revisions have altered the coefficients to some extent, the fit is poorer and there is significant residual autocorrelation, but (without "correcting" the standard errors for that problem), the Chow test does not reject, although as Figure 1.10 reveals, the forecast errors are clearly autocorrelated.

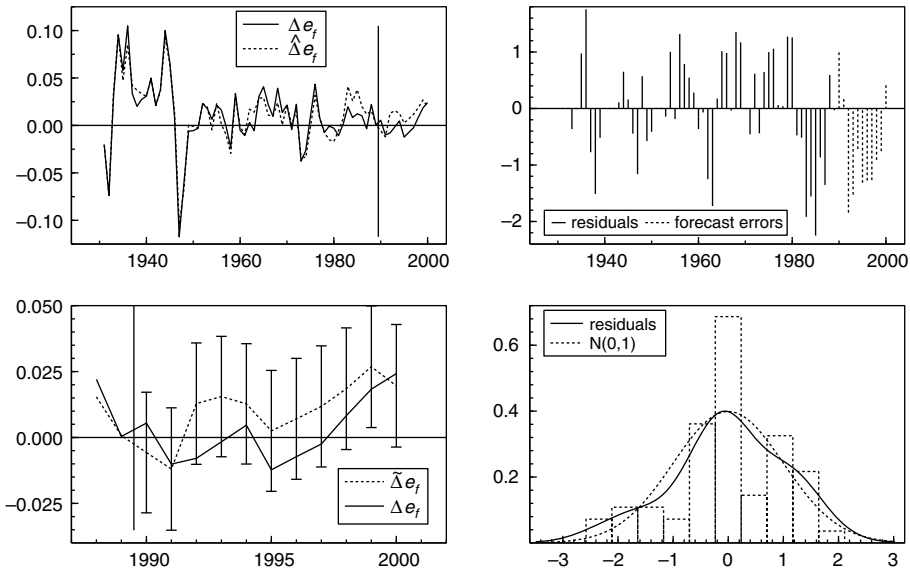


Figure 1.10 Old model revised data fitted and actual values, residuals and forecasts for $\Delta e_{f,t}$

Next, automatic remodeling at 1% on the revised data up to 1989 (with impulse saturation to remove the outliers) led to:

$$c_1 = e_f + 8.49 - 0.35e + 0.21(p_f - p), \tag{1.45}$$

with a much simpler final equation being selected:

$$\begin{aligned}\Delta e_{f,t} = {} & \underset{(0.032)}{0.35}\, s_t - \underset{(0.02)}{0.25}\, c_{1,t-1} + \underset{(0.05)}{0.65}\, \Delta e_t - \underset{(0.04)}{0.29}\, \Delta(p_f - p)_t \\ & - \underset{(0.01)}{0.05}\, \mathrm{I}_{30} - \underset{(0.01)}{0.08}\, \mathrm{I}_{31} - \underset{(0.01)}{0.08}\, \mathrm{I}_{32} + \underset{(0.01)}{0.03}\, \mathrm{I}_{70}\end{aligned}$$

$$\begin{aligned}\left(\mathsf{R}^*\right)^2 = 0.90 \;\; \mathsf{F}_{\mathsf{M}}(8,51) = 60.4^{**} \;\; \widehat{\sigma} = 0.012 \;\; \mathsf{F}_{\mathsf{ar}}(2,50) = 0.65 \\ \chi^2(2) = 2.09 \;\; \mathsf{F}_{\mathsf{arch}}(1,50) = 0.13 \;\; \mathsf{F}_{\mathsf{reset}}(1,51) = 0.11 \\ \mathsf{F}_{\mathsf{het}}(18,33) = 0.89 \;\; \mathsf{F}_{\mathsf{Chow}}(11,52) = 0.68 \; (1990 - 2000).\end{aligned} \tag{1.46}$$

Nevertheless, despite the revisions, the model in (1.46) has many features in common with both its predecessors, and is constant over the next 11 years as Figure 1.11 reports, and $\mathsf{F}_{\mathsf{Chow}}(11,52)$ confirms. The short-run elasticities still exceed their long-run counterparts, but by less than previously.

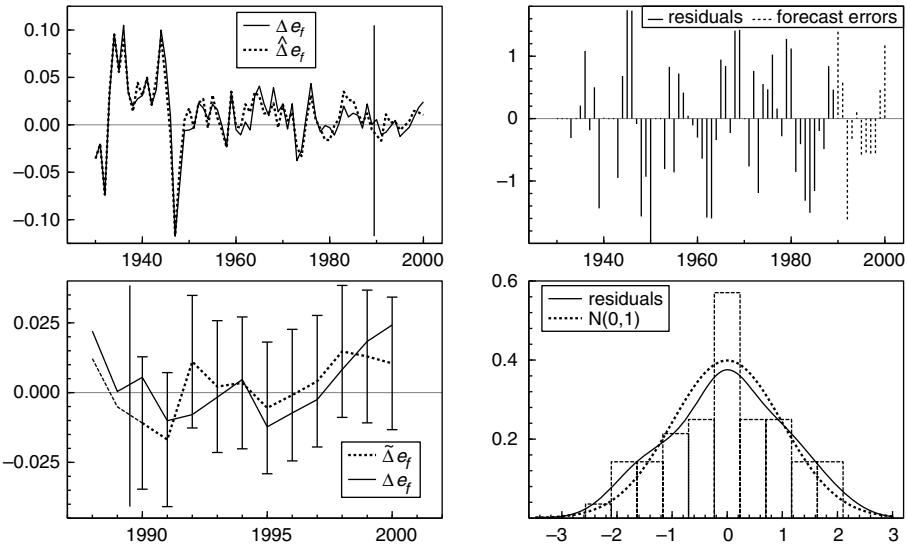


Figure 1.11 New model on revised data fitted and actual values, residuals and forecasts for $\Delta e_{f,t}$

1.8 Automatic modeling of a VAR_4 (25)

> "The Eighth Square at last!" she cried as she bounded across ... "Oh, how glad I am to get here! And what is this on my head?" she exclaimed ... It was a golden crown. (Quote from Alice in Lewis Carroll, 1899)

To illustrate that automatic modeling is not restricted to single equations (see, e.g., Krolzig, 2003), we now model the four variables in section 1.4.1.1, namely industrial output per capita, $y_{c,t}$, numbers of bankruptcies, b_t, and patents, p_t, and real equity prices (deflated by a cost of living index), e_t, using a VAR with 25 lags, augmented by impulse saturation over the common sample T =1757–1989 at $\alpha = 0.0025$ (so on average about one variable will be retained by chance as there are 337 candidates in the initial general model). The marginal critical t-ratio is about 3.1, and only about 3 regressors (other than impulses) were near or below that in the four finally-selected models. Most diagnostic tests were insignificant in those final models (but not computable at the start). The entire exercise took under two hours, including this write-up: technical progress in undertaking empirical econometrics is huge, as such an analysis would have been simply impossible (conceptually and practically) when I first started empirical modeling in 1967.

$$\Delta y_{c,t} = \underset{}{-} \underset{(0.031)}{0.128}\, y_{c,t-1} + \underset{(0.037)}{0.143}\, y_{c,t-5} - \underset{(0.037)}{0.152}\, y_{c,t-19} + \underset{(0.034)}{0.135}\, y_{c,t-23}$$

$$+ \underset{(0.004)}{0.016}\, p_t - \underset{(0.005)}{0.016}\, b_{t-3} + \underset{(0.013)}{0.081}\, e_t - \underset{(0.013)}{0.084}\, e_{t-2}$$

$$+ \underset{(0.038)}{0.129}\ \mathrm{I}_{1827} - \underset{(0.039)}{0.206}\ \mathrm{I}_{1921} + \underset{(0.037)}{0.119}\ \mathrm{I}_{1927}$$

$$\widehat{\sigma} = 0.0368 \quad \mathsf{F}_{\mathsf{AR1-2}}(2, 220) = 1.08 \quad \mathsf{F}_{\mathsf{ARCH1}}(1, 220) = 0.14$$

$$\chi^2_{\mathsf{nd}}(2) = 1.09 \quad \mathsf{F}_{\mathsf{Het}}(19, 202) = 2.24^{**} \quad \mathsf{F}_{\mathsf{RESET}}(1, 221) = 5.97^{*}$$

$$\begin{aligned} p_t = {} & \underset{(0.65)}{3.32} + \underset{(0.03)}{0.87}\ p_{t-1} - \underset{(0.22)}{0.46}\ y_{c,t-4} + \underset{(0.22)}{0.68}\ y_{c,t-5} \\ & - \underset{(0.04)}{0.19}\ b_{t-11} + \underset{(0.04)}{0.18}\ b_{t-12} \\ & - \underset{(0.05)}{0.16}\ e_{t-17} + \underset{(0.16)}{0.67}\ \mathrm{I}_{1757} - \underset{(0.15)}{0.51}\ \mathrm{I}_{1759} - \underset{(0.15)}{0.57}\ \mathrm{I}_{1761} + \underset{(0.15)}{0.58}\ \mathrm{I}_{1766} \\ & - \underset{(0.15)}{0.43}\ \mathrm{I}_{1771} - \underset{(0.15)}{0.63}\ \mathrm{I}_{1775} + \underset{(0.15)}{0.43}\ \mathrm{I}_{1783} - \underset{(0.15)}{0.65}\ \mathrm{I}_{1793} - \underset{(0.15)}{0.46}\ \mathrm{I}_{1826} \\ & - \underset{(0.15)}{0.49}\ \mathrm{I}_{1884} - \underset{(0.15)}{0.45}\ \mathrm{I}_{1940} - \underset{(0.15)}{0.46}\ \mathrm{I}_{1942} - \underset{(0.15)}{0.52}\ \mathrm{I}_{1984} + \underset{(0.15)}{0.60}\ \mathrm{I}_{1985} \end{aligned}$$

$$\widehat{\sigma} = 0.148 \quad \mathsf{F}_{\mathsf{AR1-2}}(2, 210) = 1.46 \quad \mathsf{F}_{\mathsf{ARCH1}}(1, 210) = 0.57$$

$$\chi^2_{\mathsf{nd}}(2) = 4.88 \quad \mathsf{F}_{\mathsf{Het}}(26, 185) = 1.41 \quad \mathsf{F}_{\mathsf{RESET}}(1, 211) = 0.19$$

$$\begin{aligned} b_t = {} & \underset{(0.05)}{1.10}\ b_{t-1} - \underset{(0.05)}{0.29}\ b_{t-2} - \underset{(0.10)}{0.41}\ y_{c,t-2} - \underset{(0.18)}{0.41}\ y_{c,t-17} + \underset{(0.16)}{0.80}\ y_{c,t-23} \\ & + \underset{(0.06)}{0.42}\ p_{t-1} - \underset{(0.06)}{0.23}\ p_{t-2} - \underset{(0.09)}{0.46}\ e_t + \underset{(0.10)}{0.55}\ e_{t-1} - \underset{(0.05)}{0.23}\ e_{t-5} \\ & + \underset{(0.21)}{0.64}\ \mathrm{I}_{1757} + \underset{(0.19)}{0.50}\ \mathrm{I}_{1766} - \underset{(0.18)}{0.56}\ \mathrm{I}_{1822} - \underset{(0.19)}{0.65}\ \mathrm{I}_{1838} + \underset{(0.18)}{0.83}\ \mathrm{I}_{1884} \end{aligned}$$

$$\widehat{\sigma} = 0.18 \quad \mathsf{F}_{\mathsf{AR1-2}}(2, 216) = 1.60 \quad \mathsf{F}_{\mathsf{ARCH1}}(1, 216) = 2.48$$

$$\chi^2_{\mathsf{nd}}(2) = 5.25 \quad \mathsf{F}_{\mathsf{Het}}(25, 192) = 1.64^{*} \quad \mathsf{F}_{\mathsf{RESET}}(1, 217) = 1.59$$

$$\begin{aligned} e_t = {} & \underset{(0.06)}{1.12}\ e_{t-1} - \underset{(0.06)}{0.17}\ e_{t-2} + \underset{(0.16)}{0.69}\ y_{c,t} - \underset{(0.16)}{0.69}\ y_{c,t-1} \\ & - \underset{(0.03)}{0.10}\ b_t + \underset{(0.03)}{0.12}\ b_{t-1} + \underset{(0.10)}{0.35}\ \mathrm{I}_{1802} + \underset{(0.11)}{0.31}\ \mathrm{I}_{1922} \\ & + \underset{(0.10)}{0.31}\ \mathrm{I}_{1959} - \underset{(0.10)}{0.31}\ \mathrm{I}_{1973} - \underset{(0.11)}{0.58}\ \mathrm{I}_{1974} \end{aligned}$$

$$\widehat{\sigma} = 0.10 \;\; \mathsf{F}_{\mathsf{AR1-2}}(2,220) = 2.59 \;\; \mathsf{F}_{\mathsf{ARCH1}}(1,220) = 0.01$$
$$\chi^2_{\mathsf{nd}}(2) = 3.72 \;\; \mathsf{F}_{\mathsf{Het}}(17,204) = 1.21 \;\; \mathsf{F}_{\mathsf{RESET}}(1,221) = 3.72.$$

Most of the effects found make economic sense in the context of the limited information set used here as an illustration. In reverse order, real equity prices are near a random walk, but respond positively to changes in output, and negatively to changes in bankruptcies. In turn, bankruptcies fall with increased output or equity prices, but rise with patent grants. Neither equation has many outliers, whereas the patents equation does, especially in the eighteenth century. Patents fall initially as output, equity prices, bankruptcies rise, but adjust back later. Finally, changes in output respond positively to patents and changes in equity prices, but negatively to bankruptcies.

A substantive exercise would involve additional variables like interest rates and human and physical capital; would check whether bankruptcies and patents should also be per capita; and investigate cointegration reductions. Are the long lags 'spurious'? The general historical record suggests that major innovations are both creative and destructive of output, the former by the enlargement of the production frontier, and the latter through the negative impact on those already engaged in the occupations concerned (spinners, weavers, etc., initially; clerks and secretaries in more modern times), so a "generation" is required for the new state to dominate – that motivated the original choice of 25 lags. Innovations take time to develop and be adopted; and the seeds for bankruptcy are often sown well before the reaping, even if the span is not quite "clogs to clogs in three generations." Notably, the equation for equity prices still has short lags despite the "opportunity" to find other correlations.

1.9 Conclusion

> Ever drifting down the stream –
> Lingering in the golden gleam –
> Life, what is it but a dream? (Lewis Carroll, 1899)

"Applied Econometrics" has a vast range of empirical issues to investigate: the very non-stationarity of economies keeps creating new topics for analysis. However, so long as "Applied Econometrics" is just a calibration of extant economic theory, it will never make much of an independent contribution: in that sense, one must agree with Summers (1991) but for completely opposite reasons. Much of the observed data variability in economics is due to features that are absent from most economic theories, but which empirical models have to tackle. *Ceteris paribus* conditions can sometimes be justified for theoretical reasoning, but do not provide a viable basis for empirical modeling: only a "minor influence" theorem, which must be established empirically, will suffice.

This implication is not a tract for mindless modeling of data in the absence of economic analysis, but instead suggests formulating more general initial models

that embed the available economic theory as a special case, consistent with our knowledge of the institutional framework, historical record, and the data properties. Once a congruent encompassing general model is established, an automatic model selection approach based on general-to-simple principles could help bring objectivity and credibility to empirical econometric modeling.

Economic observations are far from perfect, being subject to revision, and even to conceptual changes, with important variables unobserved, and available proxies of unknown quality. Theory constructs (such as "consumption," or "user cost of capital") and their measured counterparts (consumers' expenditure or after-tax real interest rates adjusted for depreciation) can differ markedly, especially after aggregation. Thus, a "pure" data-based approach can lack substance.

Economics has delivered a range of invaluable insights into individual decision taking, market functioning, and system-wide economies, with a vast body of theory, which has made rapid technical and intellectual progress – and will continue to do so. Applied econometrics cannot be conducted without an economic theoretic framework to guide its endeavors and and help interpret its findings. Nevertheless, since economic theory is not complete, correct, and immutable, and never will be, one also cannot justify an insistence on deriving empirical models from theory alone. That paradigm encourages covert data mining, so the credibility of existing evidence is unclear.

Data "mining" does not have pernicious properties when using a structured approach, using appropriate significance levels that decline with both the number of candidate variables and sample size: at 1% significance, one irrelevant variable in 100 will be significant by chance, at the cost of raising the selection t-ratio from around ± 2.0 to ± 2.7. Parsimony is not a justification for arbitrarily excluding many potentially relevant contenders, not even when doing so to avoid more initial variables than observations. While it is essential that the final model is much smaller than the sample size, that does not preclude starting general and making the maximum use of our best available theory and econometrics to guide our empirical endeavors and then interpret their outcomes. Thus, Frisch (1933) remains our best advice: "mutual penetration', which entails using economic analysis to guide an applied study, but letting the empirical evidence play a real role.

Acknowledgments

Financial support from the ESRC under Research Grant RES-062-23-0061 is gratefully acknowledged. I am indebted to Gunnar Baardsen, Julia Campos, Jennifer Castle, Guillaume Chevillon, Jurgen Doornik, Neil Ericsson, Katarina Juselius, Søren Johansen, Bobby Mariano, Jaime Marquez, Terry Mills, Mary Morgan, Bent Nielsen, Ragnar Nymoen, Kerry Patterson, Duo Qin, James Reade, Aris Spanos and Pravin Trivedi for helpful comments.

Note

1. Atkinson (2008) notes Robbins' apparent dismissal of Richard Stone (1951) as "not economics."

References

Aghion, P., R. Frydman, J. Stiglitz and M. Woodford (eds.) (2002) *Knowledge, Information and Expectations in Modern Macroeconomics*. Princeton: Princeton University Press.

Andrews, D.W.K. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–58.

Anselin, L. (2006) Spatial econometrics. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory*, pp. 901–69. Basingstoke: Palgrave Macmillan.

Atkinson, A.B. (2005) *Measurement of Government Output and Productivity for the National Accounts: The Atkinson Review Final Report*. London: Palgrave Macmillan.

Atkinson, A.B. (2008) Economic data and the distribution of income. Stone lectures, 2008, Department of Economics, University of Oxford.

Bachelier, L. (1900) Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure* **3**, 21–86.

Baillie, R.T. (1996) Long memory processes and fractional integration in econometrics. *Journal of Econometrics* **73**, 5–59.

Baltagi, B.H. (2006) Panel data models. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory*, pp. 633–61. Basingstoke: Palgrave Macmillan.

Banerjee, A. and D.F. Hendry (1992) Testing integration and cointegration: an overview. *Oxford Bulletin of Economics and Statistics* **54**, 225–55.

Barndorff-Nielsen, O.E. and N. Shephard (2007) *Financial Volatility in Continuous Time*. Cambridge: Cambridge University Press. Forthcoming.

Bjerkholt, O. (2005) Frisch's econometric laboratory and the rise of Trygve Haavelmo's probability approach. *Econometric Theory* **21**, 491–533.

Bjerkholt, O. (2007) Writing "the probability approach" with nowhere to go: Haavelmo in the United States, 1939–1944. *Econometric Theory* **23**, 775–837.

Blundell, R. and T.M. Stoker (2005) Heterogeneity and aggregation. *Journal of Economic Literature* **43**, 347–91.

Bontemps, C. and G.E. Mizon (2003) Congruence and encompassing. In B.P. Stigum (ed.), *Econometrics and the Philosophy of Economics*, pp. 354–78. Princeton: Princeton University Press.

Boswijk, H.P. (1992) *Cointegration, Identification and Exogeneity*, Volume 37 of *Tinbergen Institute Research Series*. Amsterdam: Thesis Publishers.

Boumans, M.A. (2005) Measurement in economic systems. *Measurement* **38**, 275–84.

Boumans, M.A. (ed.) (2007) *Measurement in Economics: A Handbook*. Amsterdam: Elsevier.

Box, G.E.P. and G.M. Jenkins, (1976) *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day (first published 1970).

Burns, A.F. and W.C. Mitchell (1946) *Measuring Business Cycles*. New York: NBER.

Caldwell, B.J. (ed.) (1993) *The Philosophy and Methodology of Economics*, *Volume II*. Aldershot: Edward Elgar.

Camerer, C.F. (2007) Neuroeconomics: using neuroscience to make economic predictions. *Economic Journal* **117**, C26–42.

Cameron, A.C. and P.K. Trivedi (1998) *The Analysis of Count Data*. Cambridge: Cambridge University Press.

Campos, J., N.R. Ericsson and D.F. Hendry (eds.) (2005) *Readings on General-to-Specific Modeling*. Cheltenham: Edward Elgar.

Carroll, L. (1899) *Through the Looking-Glass and What Alice Found There*. London: Macmillan & Co.

Cartwright, N. (1983) *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, N. (2002) In favor of laws that are not ceteris paribus after all. *Erkenntnis* **57**, 425–39.

Castle, J.L. (2005) Evaluating PcGets and RETINA as automatic model selection algorithms. *Oxford Bulletin of Economics and Statistics* **67**, 837–80.

Castle, J.L., J.A. Doornik, D.F. Hendry and R. Nymoen (2008) Testing the invariance of expectations models of inflation. Working Paper, Economics Department, Oxford University.

Castle, J.L. and D.F. Hendry (2005) Extending the boundaries of automatic Gets to non-linear models. Mimeo, Economics Department, Oxford University.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, A* **158**, 419–66. With discussion.

Choi, I. (2006) Non-stationary panels. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory,* pp. 511–39. Basingstoke: Palgrave Macmillan.

Chow, G.C. (1960) Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28**, 591–605.

Christ, C.F. (1994) The Cowles Commission's contributions to econometrics at Chicago, 1939–1955. *Journal of Economic Literature* **32**, 30–59.

Clark, C. (1932) *The National Income 1924–31*. London: Macmillan.

Clements, M.P. and D.F. Hendry (1998) *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M.P. and D.F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Clements, M.P. and D.F. Hendry (2001) An historical perspective on forecast errors. *National Institute Economic Review* **177**, 100–12.

Clements, M.P. and D.F. Hendry (eds.) (2002a) *A Companion to Economic Forecasting*. Oxford: Blackwell.

Clements, M.P. and D.F. Hendry (2002b) Explaining forecast failure in macroeconomics. In M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, pp. 539–71. Oxford: Blackwell.

Clements, M.P. and D.F. Hendry (2005) Guest Editors' introduction: Information in economic forecasting. *Oxford Bulletin of Economics and Statistics* **67**, 713–53.

Clements, M.P. and D.F. Hendry (2006) Forecasting with breaks in data processes. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Ecometrics on Forecasting*, pp. 605–57. Amsterdam: Elsevier.

Cook, S. (2008) Cross-data vintage encompassing. *Oxford Bulletin of Economics and Statistics*. Forthcoming.

Cox, D.R. (1961) Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *Volume 1*, pp. 105–23. Berkeley: University of California Press.

Crafts, N.F.R. (1997) Endogenous growth: lessons for and from economic history. In D.M. Kreps and K.F. Wallis (eds.), *Advances in Economics and Econometrics: Theory and Applications. Seventh World Congress*, *Volume 2*. Cambridge: Cambridge University Press.

Crafts, N.F.R. and T.C. Mills (1994) Trends in real wages in Britain, 1750–1913. *Explorations in Economic History* **31**, 176–94.

Darnell, A. (ed.) (1994) *The History of Econometrics*. Aldershot: Edward Elgar.

Demiralp, S. and K.D. Hoover (2003) Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics* **65**, 745–67.

Dolado, J.J. (1992) A note on weak exogeneity in VAR cointegrated systems. *Economics Letters* **38**, 139–43.

Doob, J.L. (1953) *Stochastic Processes* (1990 edition). New York: John Wiley Classics Library.

Doornik, J.A. (2006) *Object-Oriented Matrix Programming using Ox* (fourth edition). London: Timberlake Consultants Press.

Doornik, J.A. (2007a) Autometrics. Working Paper, Economics Department, University of Oxford.

Doornik, J.A. (2007b) Econometric modelling when there are more variables than observations. Working paper, Economics Department, University of Oxford.

Doornik, J.A. and H. Hansen (2008) A practical test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*. Forthcoming.

Elliott, G., C.W.J. Granger and A. Timmermann (eds.) (2006) *Handbook of Econometrics on Forecasting*. Amsterdam: Elsevier.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

Engle, R.F. and D.F. Hendry (1993) Testing super exogeneity and invariance in regression models. *Journal of Econometrics* **56**, 119–39.

Engle, R.F., D.F. Hendry and J.-F. Richard (1983) Exogeneity. *Econometrica* **51**, 277–304.

Epstein, R.J. (1987) *A History of Econometrics*. Amsterdam: North-Holland.

Ericsson, N.R. (1983) Asymptotic properties of instrumental variables statistics for testing non-nested hypotheses. *Review of Economic Studies* **50**, 287–303.

Ericsson, N.R. (1992) Cointegration, exogeneity and policy analysis: an overview. *Journal of Policy Modeling* **14**, 251–80.

Ericsson, N.R. (2007) *Econometric Modeling*. Oxford: Oxford University Press. Forthcoming.

Ericsson, N.R. and D.F. Hendry (1999) Encompassing and rational expectations: how sequential corroboration can imply refutation. *Empirical Economics* **24**, 1–21.

Ericsson, N.R. and J.G. MacKinnon (2002) Distributions of error correction tests for cointegration. *Econometrics Journal* **5**, 285–318.

Ermini, L. and D.F. Hendry (2008) Log income versus linear income: an application of the encompassing principle. *Oxford Bulletin of Economics and Statistics*. Forthcoming.

Evans, G.W. and S. Honkapohja (2001) *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Farebrother, R.W. (2006) Early explorations in econometrics. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Ecometrics, Volume I: Econometric Theory,* pp. 88–116. Basingstoke: Palgrave Macmillan.

Faust, J. and C.H. Whiteman (1997) General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model of the DGP: a translation and critique. *Carnegie–Rochester Conference Series on Public Policy* **47**, 121–61.

Favero, C. and D.F. Hendry (1992) Testing the Lucas critique: a review. *Econometric Reviews* **11**, 265–306.

Fehr, E. and A. Falk (2002) Psychological foundations of incentives. *European Economic Review* **46**, 687–724.

Fehr, E., U. Fischbacher and M. Kosfeld (2005) Neuroeconomic foundation of trust and social preferences. Dp. 5127, CEPR, London.

Feige, E.L. and D.K. Pearce, (1976) Economically rational expectations. *Journal of Political Economy* **84**, 499–522.

Feinstein, C.H. (1972) *National Income, Expenditure and Output of the United Kingdom, 1855–1965*. Cambridge: Cambridge University Press.

Fischer, A.M. (1989) Policy regime changes and monetary expectations: testing for super exogeneity. *Journal of Monetary Economics* **24**, 423–36.

Fisher, F.M. (1966) *The Identification Problem in Econometrics*. New York: McGraw-Hill.

Florens, J.-P., M. Mouchart and J.-M. Rolin (1990) *Elements of Bayesian Statistics*. New York: Marcel Dekker.

Foxwell, H.S. (ed.) (1884) *Investigations in Currency and Finance*. London: Macmillan.

Friedman, M. (1957) *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Frisch, R. (1933) Editorial. *Econometrica* **1**, 1–4.

Frisch, R. (1938) Statistical versus theoretical relations in economic macrodynamics. Mimeograph dated July 17, 1938, League of Nations Memorandum. Reprinted in D.F. Hendry

and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.
Frydman, R. and M.D. Goldberg (eds.) (2007) *Imperfect Knowledge Economics: Exchange Rates and Risk*. Princeton: Princeton University Press.
Galí, J., M. Gertler and J.D. Lopez-Salido (2001) European inflation dynamics. *European Economic Review* **45**, 1237–70.
Geweke, J.F. and S. Porter-Hudak (1983) The estimation and application of long memory time series models. *Journal of Time Series Analysis* **4**, 221–38.
Gilbert, C.L. and D. Qin (2006) The first fifty years of modern econometrics. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory,* pp. 117–55. Basingstoke: Palgrave Macmillan.
Gilbert, C.L. and D. Qin (2007) Representation in econometrics: a historical perspective. In M.A. Boumans (ed.), *Measurement in Economics: A Handbook,* Ch. 10. Amsterdam: Elsevier.
Godfrey, L.G. (1978) Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* **46**, 1303–13.
Granger, C.W.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–38.
Granger, C.W.J. (1980) Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* **14**, 227–38.
Granger, C.W.J. (1981) Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* **16**, 121–30.
Granger, C.W.J. (1987) Implications of aggregation with common factors. *Econometric Theory* **3**, 208–22.
Granger, C.W.J. and R. Joyeux (1980) An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–30.
Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica* **12**, 1–118. Supplement.
Haavelmo, T. (1958) The role of the econometrician in the advancement of economic theory. *Econometrica* **26**, 351–7.
Haavelmo, T. (1989) *Prize Lecture*. Sveriges Riksbank: Prize in Economic Sciences in Memory of Alfred Nobel.
Hald, A. (1990) *A History of Probability and Statistics and Their Applications before 1750*. New York: Wiley.
Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.
Haldrup, N., H. H. van Dijk and D.F. Hendry (eds.) (2003) Model selection and evaluation. *Oxford Bulletin of Economics and Statistics* **65**.
Hall, R.E. (1978) Stochastic implications of the life cycle-permanent income hypothesis: evidence. *Journal of Political Economy* **86**, 971–87.
Hamouda, O.F. and J.C.R. Rowley (1997) *The Reappraisal of Econometrics*. Aldershot: Edward Elgar.
Harvey, A.C. (1993) *Time Series Models* (second edition; first edition 1981). Hemel Hempstead: Harvester Wheatsheaf.
Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475–92.
Heckman, J.J. (2000) Causal parameters and policy analysis in economics: a twentieth century retrospective. *Quarterly Journal of Economics* **115**, 45–97.
Hendry, D.F. (1980) Econometrics: alchemy or science? *Economica* **47**, 387–406.
Hendry, D.F. (1987) Econometric methodology: a personal perspective. In T.F. Bewley (ed.), *Advances in Econometrics*, pp. 29–48. Cambridge: Cambridge University Press.
Hendry, D.F. (1988) The encompassing implications of feedback versus feedforward mechanisms in econometrics. *Oxford Economic Papers* **40**, 132–49.

Hendry, D.F. (1994) HUS revisited. *Oxford Review of Economic Policy* **10**, 86–106.
Hendry, D.F. (1995a) *Dynamic Econometrics*. Oxford: Oxford University Press.
Hendry, D.F. (1995b) Econometrics and business cycle empirics. *Economic Journal* **105**, 1622–36.
Hendry, D.F. (1995c) On the interactions of unit roots and exogeneity. *Econometric Reviews* **14**, 383–419.
Hendry, D.F. (1999) An econometric analysis of US food expenditure, 1931–1989. In J.R. Magnus and M.S. Morgan (eds.), *Methodology and Tacit Knowledge: Two Experiments in Econometrics,* pp. 341–61. Chichester: John Wiley and Sons.
Hendry, D.F. (2000a) Epilogue: The success of general-to-specific model selection. In *Econometrics: Alchemy or Science?* (new edition), pp. 467–90. Oxford: Oxford University Press.
Hendry, D.F. (2000b) On detectable and non-detectable structural change. *Structural Change and Economic Dynamics* **11**, 45–65.
Hendry, D.F. (2001a) How economists forecast. In D.F. Hendry and N.R. Ericsson (eds.), *Understanding Economic Forecasts*, pp. 15–41. Cambridge, Mass.: MIT Press.
Hendry, D.F. (2001b) Modelling UK inflation, 1875–1991. *Journal of Applied Econometrics* **16**, 255–75.
Hendry, D.F. (2004) Causality and exogeneity in non-stationary economic time series. In A. Welfe (ed.), *New Directions in Macromodelling*, pp. 21–48. Amsterdam: North-Holland.
Hendry, D.F. (2005) Bridging the gap: linking economics and econometrics. In C. Diebolt and C. Kyrtsou (eds.), *New Trends in Macroeconomics*, pp. 53–77. Berlin: Springer Verlag.
Hendry, D.F. (2006) Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics* **135**, 399–426. Special Issue in Honor of Clive Granger.
Hendry, D.F. and J.A. Doornik (1997) The implications for econometric modelling of forecast failure. *Scottish Journal of Political Economy* **44**, 437–61. Special Issue.
Hendry, D.F. (with J.A. Doornik and B. Nielsen) (2007) *Econometric Model Selection: Arne Ryde Lectures*. Lund University, Sweden.
Hendry, D.F., S. Johansen and C. Santos (2008) Automatic selection of indicators in a fully saturated regression. *Computational Statistics* **33**, 317–35. Erratum, 337–9.
Hendry, D.F. and K. Juselius (2000) Explaining cointegration analysis: Part I. *Energy Journal* **21**, 1–42.
Hendry, D.F. and K. Juselius (2001) Explaining cointegration analysis: Part II. *Energy Journal* **22**, 75–120.
Hendry, D.F. and H.-M. Krolzig (1999) Improving on "Data mining reconsidered" by K.D. Hoover and S.J. Perez. *Econometrics Journal* **2**, 202–19.
Hendry, D.F. and H.-M. Krolzig (2001) *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.
Hendry, D.F. and H.M. Krolzig (2004) We ran one regression. *Oxford Bulletin of Economics and Statistics* **66**, 799–810.
Hendry, D.F. and H.-M. Krolzig (2005) The properties of automatic Gets modelling. *Economic Journal* **115**, C32–61.
Hendry, D.F., E.E. Leamer and D.J. Poirier (1990) A conversation on econometric methodology. *Econometric Theory* **6**, 171–261.
Hendry, D.F., M. Lu and G.E. Mizon (2008) Model identification and non-unique structure. In J.L. Castle and N. Shephard (eds.), *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press. Forthcoming.
Hendry, D.F., M. Marcellino and G.E. Mizon (eds.) (2008) *Encompassing*. *Oxford Bulletin of Economics and Statistics*. Special Issue. Forthcoming.
Hendry, D.F. and M. Massmann (2007) Co-breaking: recent advances and a synopsis of the literature. *Journal of Business and Economic Statistics* **25**, 33–51.

Hendry, D.F. and G.E. Mizon (1999) The pervasiveness of Granger causality in econometrics. In R.F. Engle and H. White (eds.), *Cointegration, Causality and Forecasting*. Oxford: Oxford University Press.

Hendry, D.F. and G.E. Mizon (2000) Reformulating empirical macro-econometric modelling. *Oxford Review of Economic Policy* **16**, 138–59.

Hendry, D.F. and M.S. Morgan (eds.) (1995) *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.

Hendry, D.F. and B. Nielsen (2007a) *Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University Press.

Hendry, D.F. and B. Nielsen (2007b) Teaching undergraduate econometrics using *OxMetrics*. Mimeo, Economics Department, University of Oxford.

Hendry, D.F. and J.-F. Richard (1982) On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* **20**, 3–33.

Hendry, D.F. and J.-F. Richard (1989) Recent developments in the theory of encompassing. In B. Cornet and H. Tulkens (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, Mass.: MIT Press.

Hendry, D.F. and C. Santos (2009) An automatic test of super exogeneity. In M.W. Watson, T. Bollerslev and J. Russell (eds.), *Volatility and Time Series Econometrics*. Oxford: Oxford University Press. Forthcoming.

Hildenbrand, W. (1994) *Market Demand: Theory and Empirical Evidence*. Princeton: Princeton University Press.

Hildenbrand, W. (1998) How relevant are the specifications of behavioural relations on the micro-level for modelling the time path of population aggregates? *European Economic Review* **42**, 437–58.

Hildreth, C. and J.P. Houck (1968) Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association* **63**, 584–95.

Hooker, R.H. (1901) Correlation of the marriage rate with trade. *Journal of the Royal Statistical Society* **64**, 485–92. Reprinted in D.F. Hendry and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.

Hoover, K.D. and S.J. Perez (1999) Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal* **2**, 167–91.

Hoover, K.D. and S.J. Perez (2004) Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics* **66**, 765–98.

Hsiao, C. (1983) Identification. In Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, *Volume* 1, Ch. 4. Amsterdam: North-Holland.

Hunter, J. (1992) Cointegrating exogeneity. *Economics Letters* **34**, 33–5.

Jevons, W.S. (1875) The solar period and the price of corn. In H.S. Foxwell (ed.), *Investigations in Currency and Finance*, pp. 194–205. London: Macmillan.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231–54.

Johansen, S. (1992) Testing weak exogeneity and the order of cointegration in UK money demand. *Journal of Policy Modeling* **14**, 313–34.

Johansen, S. (1995) *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Johansen, S. (2006) Cointegration: an overview. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory*, pp. 540–77. Basingstoke: Palgrave Macmillan.

Johansen, S. and K. Juselius (1990) Maximum likelihood estimation and inference on co-integration – with application to the demand for money. *Oxford Bulletin of Economics and Statistics* **52**, 169–210.

Johansen, S. and B. Nielsen (2008) An analysis of the indicator saturation estimator as a robust regression estimator. In J.L. Castle and N. Shephard (eds.), *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press. Forthcoming.

Judge, G.G. and M.E. Bock (1978) *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Juselius, K. (1993) VAR modelling and Haavelmo's probability approach to econometrics. *Empirical Economics* **18**, 595–622.

Juselius, K. and M. Franchi (2007) Taking a DSGE model to the data meaningfully. *Economics: The Open-Access, Open-Assessment E-Journal* **1**, 4.

Keynes, J.M. (1939) Professor Tinbergen's method. *Economic Journal* **44**, 558–68. Reprinted in D.F. Hendry and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.

Keynes, J.M. (1940) Statistical business-cycle research: Comment. *Economic Journal* **50**, 154–6.

Kleibergen, F. (2002) Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* **70**, 1781–803.

Klein, J.L. (1997) *Statistical Visions in Time*. Cambridge: Cambridge University Press.

Koopmans, T.C. (1937) *Linear Regression Analysis of Economic Time Series*. Haarlem: Netherlands Economic Institute.

Koopmans, T.C. (1947) Measurement without theory. *Review of Economics and Statistics* **29**, 161–79.

Koopmans, T.C. (1950) When is an equation system complete for statistical purposes? In T.C. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*, No. 10 in Cowles Commission Monograph, Ch. 17. New York: John Wiley & Sons.

Krolzig, H.-M. (2003) General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics* **65**, 769–802.

Kurcewicz, M. and J. Mycielski (2003) A specification search algorithm for cointegrated systems. Discussion paper, Statistics Department, Warsaw University.

Kydland, F.E. and E.C. Prescott (1990) Business cycles: real facts and a monetary myth. *Federal Reserve Bank of Minneapolis, Quarterly Review* **14**, 3–18.

Kydland, F.E. and E.C. Prescott (1991) The econometrics of the general equilibrium approach to business cycles. *Scandinavian Journal of Economics* **93**, 161–78.

Le Gall, P. (2007) *A History of Econometrics in France: From Nature to Models*. London: Routledge.

Leamer, E.E. (1978) *Specification Searches. Ad-Hoc Inference with Non-Experimental Data*. New York: John Wiley.

Leamer, E.E. (1983) Let's take the con out of econometrics. *American Economic Review* **73**, 31–43.

Leeb, H. and B.M. Pötscher (2003) The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory* **19**, 100–42.

Leeb, H. and B.M. Pötscher (2005) Model selection and inference: facts and fiction. *Econometric Theory* **21**, 21–59.

Lovell, M.C. (1983) Data mining. *Review of Economics and Statistics* **65**, 1–12.

Lucas, R.E. (1976) Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, Volume 1 of *Carnegie-Rochester Conferences on Public Policy*, pp. 19–46. Amsterdam: North-Holland.

Magnus, J.R. and M.S. Morgan (eds.) (1999) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*. Chichester: John Wiley and Sons.

Mavroeidis, S. (2004) Weak identification of forward-looking models in monetary economics. *Oxford Bulletin of Economics and Statistics* **66**, 609–35.

Mayo, D.G. and A. Spanos (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science* **57**, 323–57.

Mills, T.C. (ed.) (1999) *Economic Forecasting* (two volumes). Cheltenham: Edward Elgar.

Mills, T.C. and K.D. Patterson (eds.) (2006) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Basingstoke: Palgrave Macmillan.

Mitchell, B.R. (1988) *British Historical Statistics*. Cambridge: Cambridge University Press.

Mizon, G.E. (1984) The encompassing approach in econometrics. In D.F. Hendry and K.F. Wallis (eds.), *Econometrics and Quantitative Economics*, pp. 135–72. Oxford: Basil Blackwell.
Mizon, G.E. and J.-F. Richard (1986) The encompassing principle and its application to non-nested hypothesis tests. *Econometrica* **54**, 657–78.
Moene, K.A. and A. Rødseth (1991) Nobel Laureate: Trygve Haavelmo. *Journal of Economic Perspectives* **5**, 175–92.
Moore, H.L. (1911) *Laws of Wages: An Essay in Statistical Economics*. New York: Macmillan.
Moore, H.L. (1914) *Economic Cycles – Their Law and Cause*. New York: Macmillan.
Moore, H.L. (1923) *Generating Economic Cycles*. New York: Macmillan.
Morgan, M.S. (1990) *The History of Econometric Ideas*. Cambridge: Cambridge University Press.
Muth, J.F. (1961) Rational expectations and the theory of price movements. *Econometrica* **29**, 315–35.
Nerlove, M. (1983) Expectations, plans, and realizations in theory and practice. *Econometrica* **51**, 1251–79.
Nymoen, R. (2002) Faulty watch towers – "structural" models in Norwegian monetary policy analysis. Unpublished paper, University of Oslo.
Pagan, A.R. (1987) Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys* **1**, 3–24.
Paruolo, P. and A. Rahbek (1999) Weak exogeneity in I(2) systems. *Journal of Econometrics* **93**, 281–308.
Peart, S.J. (2001) "Facts carefully marshalled" in the empirical studies of William Stanley Jevons. *History of Political Economy* **33**, 252–76.
Perez-Amaral, T., G.M. Gallo and H. White (2003) A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics* **65**, 821–38.
Perez-Amaral, T., G.M. Gallo and H. White (2005) A comparison of complementary automatic modelling methods: RETINA and PcGets. *Econometric Theory* **21**, 262–77.
Phillips, A.W.H. (1957) Stabilization policy and the time form of lagged response. *Economic Journal* **67**, 265–77.
Phillips, P.C.B. (1994) Bayes models and forecasts of Australian macroeconomic time series. In C. Hargreaves (ed.), *Non-stationary Time-Series Analyses and Cointegration*. Oxford: Oxford University Press.
Phillips, P.C.B. (1995) Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review* **1**, 92–102.
Phillips, P.C.B. (1996) Econometric model determination. *Econometrica* **64**, 763–812.
Phillips, P.C.B. (2003) Laws and limits of econometrics. *Economic Journal* **113**, C26–52.
Phillips, P.C.B. and M. Loretan (1991) Estimating long-run economic equilibria. *Review of Economic Studies* **58**, 407–36.
Popper, K.R. (1963) *Conjectures and Refutations*. New York: Basic Books.
Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory* **7**, 163–85.
Qin, D. (1993) *The Formation of Econometrics: A Historical Perspective*. Oxford: Clarendon Press.
Qin, D. (2008) Consolidation of the Cowles Commission paradigm. Unpublished paper, Queen Mary College, London.
Ramsey, J.B. (1969) Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B* **31**, 350–71.
Richard, J.-F. (1980) Models with several regimes and changes in exogeneity. *Review of Economic Studies* **47**, 1–20.
Robbins, L. (1932) *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.
Robinson, P.M. (1995) Log-periodogram regression of time series with long range dependence. *Annals of Statistics* **23**, 1048–72.
Rothenberg, T.J. (1971) Identification in parametric models. *Econometrica* **39**, 577–92.

Sala-i-Martin, X.X. (1997) I have just run two million regressions. *American Economic Review* **87**, 178–83.
Samuelson, P.A. (1947) *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.
Sargan, J.D. (1983) Identification and lack of identification. *Econometrica* **51**, 1605–33.
Schultz, H. (1928) *The Theory and Measurement of Demand*: Chicago: University of Chicago Press.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461–4.
Siegel, D.S. and M. Wright (2007) Intellectual property: the assessment. *Oxford Review of Economic Policy* **23**, 529–40.
Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* **48**, 1–48.
Smets, F. and R. Wouters (2003) An estimated stochastic dynamic general equilibrium model of the Euro Area. *Journal of the European Economic Association* **1**, 1123–75.
Smith, A. (1759) *Theory of Moral Sentiments*. Edinburgh: A. Kincaid & J. Bell.
Smith, A. (1776) *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan & T. Cadell.
Smith, A. (1795) The history of astronomy. In D. Stewart (ed.), *Essays on Philosophical Subjects by Adam Smith*, pp. 33–105. Edinburgh: W. Creech. Liberty Classics edition, by I.S. Ross, 1982.
Smith, R.J. (1992) Non-nested tests for competing models estimated by generalised method of moments. *Econometrica* **60**, 973–80.
Smith, R.J. (2007) Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics* **138**, 430–60.
Spanos, A. (1989) On re-reading Haavelmo: a retrospective view of econometric modeling. *Econometric Theory* **5**, 405–29.
Spanos, A. (1990) The simultaneous equations model revisited: statistical adequacy and identification. *Journal of Econometrics* **44**, 87–105.
Spanos, A. (1995) On theory testing in econometric modelling with non-experimental data. *Journal of Econometrics* **67**, 189–226.
Spanos, A. (2006) Econometrics in retrospect and prospect. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory,* pp. 3–58. Basingstoke: Palgrave Macmillan.
Spanos, A., D.F. Hendry and J.J. Reade (2008) Linear vs. log-linear unit root specification: an application of mis-specification encompassing. *Oxford Bulletin of Economics and Statistics*. Forthcoming.
Staiger, D. and J.H. Stock (1997) Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–86.
Stigler, G.J. (1962) Henry L. Moore and statistical economics. *Econometrica* **30**, 1–21.
Stigler, S.M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, Mass.: Harvard University Press.
Stigler, S.M. (1999) *Statistics on the Table*. Cambridge, Mass.: Harvard University Press.
Stiglitz, J. (2003) Is Keynes dead? Reviving a sensible macroeconomics. Clarendon lectures, Department of Economics, University of Oxford.
Stigum, B.P. (1990) *Towards a Formal Science of Economics*. Cambridge, Mass.: MIT Press.
Stock, J.H. and J.H. Wright (2000) GMM with weak identification. *Econometrica* **68**, 1055–96.
Stock, J.H., J.H. Wright and M. Yogo (2002) A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* **20**, 518–29.
Stone, J.R.N. (1951) *The Role of Measurement in Economics*. Cambridge: Cambridge University Press.
Summers, L.H. (1991) The scientific illusion in empirical macroeconomics. *Scandinavian Journal of Economics* **93**, 129–48.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B, **58**, 267–88.

Tinbergen, J. (1930) Determination and interpretation of supply curves: an example [Bestimmung und Deutung von Angebotskurven: ein Beispiel]. *Zeitschrift fur Nationalökonomie* **1**, 669–79. Reprinted in D.F. Hendry and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.

Tinbergen, J. (1939) *Statistical Testing of Business-Cycle Theories. Volume I: A Method and its Application to Investment Activity*. Geneva: League of Nations.

Tinbergen, J. (1940) *Statistical Testing of Business-Cycle Theories. Volume II: Business Cycles in the United States of America, 1919–1932*. Geneva: League of Nations.

Tobin, J. (1950) A statistical demand function for food in the U.S.A. *Journal of the Royal Statistical Society A* **113**(2), 113–41.

Tobin, J. (1952) A survey of the theory of rationing. *Econometrica* **26**, 24–36.

Tress, R.C. (1959) The contribution of economic theory to economic prognostication. *Economica* **26**, 194–211.

Urbain, J.-P. (1992) On weak exogeneity in error correction models. *Oxford Bulletin of Economics and Statistics* **54**, 187–207.

Vining, R. (1949a) Koopmans on the choice of variables to be studied and of methods of measurement. *Review of Economics and Statistics* **31**, 77–86.

Vining, R. (1949b) A rejoinder. *Review of Economics and Statistics* **31**, 91–4.

von Weizsacker, C.C. (2005) The welfare economics of adaptive preferences. Preprint 2005/11, Max Planck Institute, Bonn.

Weissmann, G. (1991) Asprin. *Scientific American*, January, 58–64.

White, H. (1980a) A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–38.

White, H. (1980b) Using least squares to approximate unknown regression functions. *International Economic Review* **21**, 149–70.

White, H. (1990) A consistent model selection. In C.W.J. Granger (ed.), *Modelling Economic Series*, pp. 369–83. Oxford: Clarendon Press.

White, H. (2000) A reality check for data snooping. *Econometrica* **68**, 1097–126.

White, H. (2008) Approximate nonlinear forecasting methods. Mimeo, Economics Department, University of California at San Diego.

Working, E.J. (1927) What do statistical demand curves show? *Quarterly Journal of Economics* **41**, 212–35.

Wright, P.G. (1915) Review of Moore, "*Economic Cycles*" (1915). *Quarterly Journal of Economics* **29**, 631–41. Reprinted in D.F. Hendry and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.

Wright, P.G. (1929) Review of H. Schultz: "*Statistical Laws of Demand and Supply*" (1915). *Journal of the American Statistical Association* **24**, 207–15.

Yule, G.U. (1897) On the theory of correlation. *Journal of the Royal Statistical Society* **60**, 812–38.

Yule, G.U. (1926) Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time series (with discussion). *Journal of the Royal Statistical Society* **89**, 1–64. Reprinted in D.F. Hendry and M.S. Morgan (eds.), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press, 1995.

2
How much Structure in Empirical Models?

Fabio Canova

Abstract

This chapter highlights the problems that structural methods and SVAR approaches have when estimating DSGE models and examining their ability to capture important features of the data. We show that structural methods are subject to severe identification problems due, in large part, to the nature of DSGE models. The problems can be patched up in a number of ways, but solved only if DSGEs are completely reparameterized or respecified. The potential misspecification of the structural relationships gives Bayesian methods an edge over classical ones in structural estimation. SVAR approaches may face invertibility problems but simple diagnostics can help to detect and remedy these problems. A pragmatic empirical approach ought to use the flexibility of SVARs against potential misspecification of the structural relationships but must firmly tie SVARs to the class of DSGE models which could have generated the data.

2.1 Introduction

The 1990s witnessed a remarkable development in the specification of stochastic general equilibrium models. The literature has added considerable realism to the popular workhorses of the 1980s; a number of shocks and frictions have been introduced into first-generation Real Business Cycle (RBC) models driven by a single

technological disturbance; and our understanding of the propagation mechanism of structural shocks has been considerably enhanced. Steps forward have also been made in comparing the quality of the models' approximation to the data. While a few years ago it was standard to calibrate the parameters of a model and informally evaluate the quality of its fit to the data, now full information likelihood-based estimation of the structural parameters has become common practice (see, for example, Smets and Wouters, 2003; Ireland, 2004; Canova, 2005; Rabanal and Rubio-Ramirez, 2005; Gali and Rabanal, 2005) and new techniques have been introduced for model evaluation purposes (see Del Negro *et al*., 2006). Given the complexities involved in estimating stochastic general equilibrium models and the difficulties in designing criteria which are informative about their discrepancy with the data, a portion of the literature has also considered less demanding limited information methods and focused on whether a model matches the data only along certain dimensions. For example, following Rotemberg and Woodford (1997) and Christiano, Eichenbaum and Evans (2005), it is now common to estimate structural parameters by quantitatively matching the conditional dynamics in response to certain structural shocks. Regardless of the approach a researcher selects, the stochastic general equilibrium model one uses to restrict the data is taken very seriously: in both estimation and testing, it is in fact implicitly assumed that the model is the data-generating process (DGP) of the actual data, up to a set of serially uncorrelated measurement errors. Despite the above-mentioned progress, such an assumption is still too heroic to be credibly entertained. As a consequence, estimates of the parameters may reflect this primitive misspecification and, as the sample size grows, parameter estimates need not converge to those of the true DGP.

The 1990s also witnessed an extraordinary development of vector autoregressive (VAR) techniques: from simple reduced form models, VARs have evolved into tools to analyze questions of interest to academics and policy makers. Structural VARs have enjoyed an increasing success in the profession for two reasons: they are easy to estimate and the computational complexities are of infinitesimal order relative to those of structural techniques; structural inference can be performed without conditioning on a single, and possibly misspecified, model. Clearly, there is no free lunch and robustness against misspecification comes at the cost of limiting the type of policy exercises one can entertain. One additional advantage of structural VARs needs to be mentioned. While techniques to deal with parameter variations are sufficiently well developed in this literature (see Cogley and Sargent, 2005; Primiceri, 2005; Canova and Gambetti, 2007), they are still at an infant stage when it comes to structurally estimating time variations in the parameters of a stochastic general equilibrium model (see Justiniano and Primiceri, 2008; Fernandez-Villaverde and Rubio-Ramirez, 2007).

When addressing an empirical problem with a finite amount of data, one has therefore to take a stand on how much theory one wants to use to structure the available data prior to estimation. If the former approach is taken (which we will call "structural" for simplicity), model-based estimation can be performed, but

inference is valid only to the extent that the model correctly represents the DGP of the data. If the latter approach is taken (which we call "SVAR" for simplicity), one can work with a class of structural models and use implications that are common to the members of this class to identify shocks and trace out their effects on the endogenous variables of the system, but cannot say much about preference or production function parameters, nor conduct certain policy exercises that involve changes in expectation formation. The choice between the two alternatives is easy in two extreme and unlikely situations: the stochastic models one writes down are in fact the DGP of the actual data; there is a unique mapping between the structural models to reduced form ones. Under these two conditions, direct (structural) or indirect (SVAR) estimation will give similar answers to a set of core questions investigators like to study (transmission of certain disturbances, effects of shocks to certain policy rules, and so on) and for these questions, accuracy and computational time become the most important factors that determine the choice of technique.

Unfortunately, the reality is far from the ideal and both approaches have important shortcomings. Current dynamic stochastic general equilibrium (DSGE) models, even in the large-scale versions that are now used in central banks and international institutions, are still too simple to capture the complexities of the macro-data. In addition, because they are highly nonlinear in the structural parameters and the mapping between structural parameters and the coefficients of the aggregate decision rules is analytically unknown – the exact mapping is known only in a few but uninteresting cases – the identification of the structural parameters from the data is far from clear. Structural VAR estimation also faces identification problems. The identification restrictions researchers use are often conventional, have little economic content, and are not derived from any class of models that macroeconomists use to interpret the results. Furthermore, there are DSGE models which do not admit a finite order VAR representation and others which cannot be recovered when the Wold decomposition is used to set up a VAR. Omitted variables may play an important role in SVAR results and the use of small-scale systems may distort the conclusions one draws from the exercise. In both cases, small samples, or samples which contain different regimes, may further complicate the inferential problem. All in all, the issues of misspecification, identification, low signal-to-noise ratio, invertibility, omitted variables and reduced number of shocks and, last but not least, small samples, should always be in the back of the mind of an investigator who is interested in studying an applied problem and/or suggesting policy recommendations from his/her analysis.

The scope of this chapter is to highlight the problems one faces when using either of the two methodologies to conduct policy analyses, and to address questions concerning the validity of models and their ability to capture features of the data and, in general, empirical issues of interest to academics and to policy makers. In particular, we discuss identification problems and problems connected with the potential non-representability of the aggregate decision rules with VARs. The problems we describe do not have a solution yet and standard approaches to deal with them may make the problems worse. We provide a list of "dos and

don'ts" which applied investigators may want to keep in mind in their work and outline a methodology, combining ideas from both types of approaches, which can potentially avoid some of the problems we discuss and allow useful inference on interesting economic questions. Nevertheless, it should be clear that asking too much from a model is equivalent to asking for trouble. One should use theory as a flexible mechanism to organize the data and to avoid questions that the data, the nature of the model, or the estimation approach employed cannot answer.

2.2 DSGE models

DSGE models are consistent theoretical laboratories where the preferences and the objective functions of the agents are fully specified, the general equilibrium interactions are taken into account, the stochastic structure of the driving forces is exactly defined, the expectations of the agents are consistently treated and the equilibrium of the economy is clearly spelled out. The economic decisions of the agents are derived under the assumption that they maximize their objectives in a rational, forward-looking manner. Individual optimality conditions are highly nonlinear functions of the parameters of agents' objective functions and constraints and of the variables that are predetermined and exogenous to their actions. Given the complicated nature of these conditions, explicit decision rules, expressing the choice variables as a function of the predetermined and exogenous variables and the parameters, are not generally available in a closed form. Hence, it is typical to use numerical procedures to approximate these functions, either locally or globally. The solutions to the individual problem are then aggregated into total demand and supply curves, the equilibrium for the economy is computed, and perturbations produced by selected disturbances are analyzed to understand both the mechanics and the timing of the adjustments back to the original equilibrium.

Under regularity conditions, we know that a solution to agents' optimization problems exists and is unique. Hence, one typically guesses the form of the solution, uses a particular functional form to approximate the guess and calculates the coefficients of the approximating function which, given the stationarity of the problem, must be the same for every t. For most situations of interest, (log-)linear or second-order approximations, computed around a carefully selected pivotal point, suffice. The optimality conditions of agents' problems in (log)-linearized deviations from the steady-state are:

$$0 = E_t[A(\theta)x_{t+1} + B(\theta)x_t + C(\theta)x_{t-1} + D(\theta)z_{t+1} + F(\theta)z_t] \tag{2.1}$$

$$0 = z_{t+1} - G(\theta)z_t - e_t, \tag{2.2}$$

where θ is a vector which includes the parameters of preferences, technologies, and policies; $A(\theta), B(\theta), C(\theta), D(\theta), F(\theta), G(\theta)$ are continuous and differentiable functions of θ; x_t are the endogenous variables of the model; and z_t the uncontrollable driving forces, which are typically assumed to follow an AR(1) process with possibly contemporaneously correlated errors e_t. These approximate individual optimality conditions are numerically solved to produce individual decisions rules which can

be equivalently written in a restricted state space format:

$$x_{1t} = J(\theta)x_{1t-1} + K(\theta)e_t$$
$$x_{2t} = G(\theta)x_{1t}, \tag{2.3}$$

where x_{1t} are the predetermined and exogenous variables and x_{2t} are the choice variables of the agents, or in a restricted VAR format:

$$A_0(\theta)x_t = H_1(\theta)x_{t-1} + H_2(\theta)E_t, \tag{2.4}$$

where:

$$A_0(\theta) = \begin{bmatrix} I & 0 \\ I & -G(\theta) \end{bmatrix},\ H_1(\theta) = \begin{bmatrix} J(\theta) & 0 \\ 0 & 0 \end{bmatrix},\ H_2(\theta) = \begin{bmatrix} K(\theta) & 0 \\ 0 & 0 \end{bmatrix},\ E_t = \begin{bmatrix} e_t \\ 0 \end{bmatrix}. \tag{2.5}$$

The solution of a log-linearized DSGE model therefore has the same format as well-known time series models and this makes it particularly attractive to applied macroeconomists with some time series background. However, several unique features of the individual decision rules produced by DSGE models need to be noted. First, (2.3)–(2.4) are nonlinear in the structural parameters θ, and it is θ and not J, K or G that a researcher is typically interested in. Second, the decision rules feature cross-equation restrictions, in the sense that the $\theta_i, i = 1, 2, \ldots$, may appear in several of the elements of the matrices J, K and G. Third, the number of structural shocks is typically smaller than the number of endogenous variables that the model generates. This implies singularities in the covariance of the x_ts, which are unlikely to hold in the data. Finally, H_1 and H_2 are of reduced rank. Note that if A_0 is invertible, (2.4) can be transformed into:

$$x_t = M_1(\theta)x_{t-1} + v_t, \tag{2.6}$$

where $M_1(\theta) = A_0(\theta)^{-1}H_1(\theta), v_t = A_0(\theta)^{-1}H_2(\theta)E_t$ and a (reduced form) VAR representation for the theoretical model could be derived. As we will see, the nonlinearity in the mapping between θ and J, K, G makes identification and estimation difficult, even when cross-equation restrictions are present. System singularity, on the other hand, is typically avoided by adding measurement errors to the decision rules or by considering only the implications of the model for a restricted number of variables – in this case the number of variables is equal to the number of exogenous variables. Finally, rank failures are generally avoided by integrating variables out of (2.4) and obtaining a new representation featuring invertible matrices. As we will see, such an integration exercise is not harmless. In fact, this reduction process will in general produce a VAR moving average (VARMA) representation for the individual decision rules of the DSGE model. Hence, aggregate decision rules may not be always representable with a finite order VAR.

Given the linearity of (2.3) or (2.4) in the predetermined and exogenous variables, aggregate decision rules will also be linear in predetermined and exogenous variables. Therefore, given values for the θ vector, time series can be easily simulated, responses to exogenous impulses calculated and sources of business cycle fluctuations examined.

How does one select the θ vector used in simulation exercises? Until a few years ago, it was common to calibrate θ so that selected statistics of the actual and simulated data were close to each other. This informal selection process was justified by the fact that DSGE models were too simple and stylized to be faced with rigorous statistical estimation. In recent years the complexity of models has increased; a number of frictions have been introduced on the real, the monetary and, at times, the financial side of the economy; a larger number of disturbances has been considered and a number of more realistic features added. Therefore, it has become more common to attempt structural estimation of the θ using limited information approaches, such as impulse response matching exercises, or full information ones, such as likelihood-based methods.

A clear precondition for any structural estimation approach to be successful is that the parameters of interest are identifiable from the chosen objective function. In the next sub-section we discuss why parameter identifiability may be hard to obtain in the context of DSGE models and why, perhaps, calibration was originally preferred by DSGE modelers.

2.2.1 Identification

Identification problems can emerge in three distinct situations. First, a model may face identification problems in population, that is, the mapping between the structural parameters and the parameters of the aggregate equilibrium decision rule is ill-conditioned. We call this phenomenon the "solution identification" problem. Since the objective functions are typically a deterministic transformation of either (2.3) or (2.4), failure to identify θ from the entries of the aggregated versions of the $J(\theta), K(\theta), G(\theta)$ matrices (or from the aggregate versions of the $A_0(\theta), H_1(\theta), H_2(\theta)$ matrices) is sufficient for having population identification problems for all possible choices of objective functions.

Second, it could be that identification pathologies emerge because the selected objective function neglects important model information – for example, the steady-states or the variance-covariance matrix of the shocks. In other words, one can conceive situations where all the structural parameters are identifiable if the whole model is considered, but some of them cannot be recovered from, say, a sub-set of the equations of the model or a sub-set of the responses to shocks. We call this phenomenon the "limited information identification" problem. As an example of why this may happen, suppose you have two variables, say output and inflation, and two shocks, say technology and monetary shocks. Obviously, the responses to technology shocks carry little information for the autoregressive parameter of the monetary shock. Hence, this parameter is unlikely to be identified from the dynamics induced by technology shocks. It should also be clear that limited information and solution identification problems are independent of each other and therefore may appear in isolation or jointly.

Finally, difficulties in identifying parameters may be the result of small samples. That is to say, even if the mapping between the structural parameters and the parameters of the aggregate decision rules is well behaved and even if the objective function considers all the implications of the model, it may be difficult to recover

structural parameters because the sample does not contain enough information to invert the mapping from $J(\theta)$, $K(\theta)$, and $G(\theta)$ or from the objective function to θ. To understand why this problem may emerge, consider the likelihood function of one parameter for a given dataset. It is well known that, as the sample size increases, the shape of the likelihood function changes, becoming more sharply peaked around the mode. Therefore, when the sample is small, the likelihood function may feature large flat areas in a relevant portion of the parameter space and this may make it difficult to infer the parameter vector which may have been generating the data.

Econometricians have long been concerned with identification problems (see, for example, Liu, 1960; Sims, 1980, among others). When models are linear in the parameters, and no expectations are involved, it is relatively straightforward to check whether the first two types of problems are present: it is sufficient to use rank and order conditions and look at the mapping between structural parameters and the aggregate decision rules. It is also easy to measure the extent of small sample issues – the size of the estimated standard errors or an ill-conditioned matrix of second-order derivatives of the objective function evaluated at parameter estimates give us an indication of the importance of this problem. For DSGE models none of these diagnostics can really be used. Since the mapping between θ and the parameters of (2.3) or (2.4) is nonlinear, traditional rank and order conditions do not apply. Furthermore, the size of estimated standard errors is insufficient to inform us about identification problems.

If identification problems are detected, what can one do? While for the first type of problems there is very little to be done, except going back to the drawing board and respecifying or reparametrizing the model, the latter two problems could in principle be alleviated by specifying a full-information objective function and by adding external information. If one insists on using a limited information criteria, one then needs to experiment with the sub-set of the model's implications to be used in estimation. Such experimentation is far from straightforward because economic theory offers little guidance in the search, and because certain variables produced by the model are non-observable (for example, effort) or non-measurable (for example, capital) by the applied researcher. Information from external sources may not always be available; it may be plagued by measurement errors or not very informative about the parameters of interest (see Boivin and Giannoni, 2005).

DSGE models face a large number of population identification problems. Canova and Sala (2005) provide an exhaustive list of potentially interesting pathologies. To summarize their taxonomy: a number of DSGE models, with potentially different economic implications, may be observationally equivalent in the sense that the aggregate decision rules they produce will be indistinguishable; they may be subject to under- or partial identification of their parameters, that is, a set of parameters may disappear from the aggregate decision rules or enter only in a particular functional form; and they may face weak identification problems – the mapping between structural parameters and the coefficients of the aggregate decision rules may display little curvature or be asymmetric in some direction. All these problems could occur locally or globally in the parameter space. However, given the common

practice of obtaining estimates using optimization routines which constrain the search of the maximum to an interval, we will consider only local problems in what follows. Also, while the econometric literature has often considered the latter as a small sample problem, weak identification problems easily occur in the population. In other words, while it is generally true that when the sample size is small the curvature of the mapping may not be sufficient to recover the underlying vector of structural parameters from the coefficients of the aggregate decision rules, there is nothing that ensures that such a mapping in DSGE models will be better behaved with an infinitely large sample.

Next, we present two examples which show the pervasiveness of population identification problems in standard DSGE models. While the models are of small scale, it should be remembered that most of the larger-scale DSGE models used in the literature feature the equations of these models as building blocks. Therefore, the problems we highlight are likely to emerge also in more complex set-ups.

2.2.1.1 Example 1: observational equivalence

Consider the following three equations:

$$y_t = \frac{1}{\lambda_2 + \lambda_1} E_t y_{t+1} + \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} y_{t-1} + v_t \tag{2.7}$$

$$y_t = \lambda_1 y_{t-1} + w_t \tag{2.8}$$

$$y_t = \frac{1}{\lambda_1} E_t y_{t+1} \text{ where } y_{t+1} = E_t y_{t+1} + e_t, \tag{2.9}$$

where $\lambda_2 \geq 1 \geq \lambda_1 \geq 0$ and v_t, w_t and e_t are independent and identically distributed (i.i.d.) processes with zero mean and variance σ_v^2, σ_w^2, σ_e^2 respectively. It is well known that the unique stable rational expectations solution of (2.6) is $y_t = \lambda_1 y_{t-1} + \frac{\lambda_2+\lambda_1}{\lambda_2} v_t$ and that the stable solution of (2.8) is $y_t = \lambda_1 y_{t-1} + e_t$. Therefore, if $\sigma_w = \sigma_e = \frac{\lambda_2+\lambda_1}{\lambda_2}\sigma_v$, a unitary impulse in the three innovations will produce the same responses for $y_{t+j}, j = 0, 1, \ldots$, in the three equations, and these are given by $[\frac{\lambda_2+\lambda_1}{\lambda_2}, \lambda_1 \frac{\lambda_2+\lambda_1}{\lambda_2}, \lambda_1^2 \frac{\lambda_2+\lambda_1}{\lambda_2}, \ldots]$.

What makes the three processes equivalent in terms of impulse responses? Clearly, the unstable root λ_2 in (2.6) enters the solution only contemporaneously. Since the variance of the shocks is not estimable from normalized impulse responses (any value simply implies a proportional increase in all the elements of the impulse response function), it becomes a free parameter which we can arbitrarily select to capture the effects of the unstable root. Turning things around, the dynamics produced by normalized impulses to these three processes will be observationally equivalent because λ_2 is left underidentified in the exercise.

While equations (2.6)–(2.8) are stylized, it should be kept in mind that many refinements of currently used DSGE models add some backward-looking component to a standard forward-looking one, and that the current Great Moderation debate in the US hinges on the existence of determinate versus sunspot solutions (see, for example, Lubik and Schorfheide, 2004). What this example suggests is

that these features may be indistinguishable when one looks just at normalized impulse responses.

How can one avoid observational equivalence? Clearly, part of the problem emerges because normalized impulse responses carry no information for the unstable root λ_2. However, the variance of the shocks does have this information and, for example, the likelihood function of the first process will be different from those of the other two. Hence, adding information could help limit the extent of observational equivalence problems. In the case where one is not willing to or cannot use this information and only employs the dynamics in response to normalized shocks to recover structural parameters, information external to the models needs to be brought in to disentangle various structural representations (as it is done, for example, in Boivin and Giannoni, 2006).

2.2.1.2 Example 2: identification problems in a New Keynesian model

Throughout this sub-section we assume that the investigator knows the correct model and the restrictions needed to identify the shocks. Initially, we assume that he/she chooses as an objective function the distance between the responses in the model and in the data. Later on, we examine how identification is affected when additional information is brought into the estimation process.

We consider a well-known small-scale New Keynesian (NK) model, which has become the workhorse in academic and policy discussions and constitutes the building block of larger-scale models currently estimated in the literature. Several authors, including Ma (2002), Beyer and Farmer (2004), Nason and Smith (2005) and Canova and Sala (2005), have pointed out that such a structure is liable to identification problems. Here we discuss where and how these problems emerge.

The log-linearized version of the model consists of the following three equations for the output gap y_t, inflation π_t and the nominal rate r_t:

$$y_t = \frac{h}{1+h}y_{t-1} + \frac{1}{1+h}E_t y_{t+1} + \frac{1}{\phi}(i_t - E_t\pi_{t+1}) + \nu_{1t} \tag{2.10}$$

$$\pi_t = \frac{\omega}{1+\omega\beta}\pi_{t-1} + \frac{\beta}{1+\omega\beta}E_t\pi_{t+1} + \frac{(\phi+\nu)(1-\zeta\beta)(1-\zeta)}{(1+\omega\beta)\zeta}y_t + \nu_{2t} \tag{2.11}$$

$$i_t = \lambda_r i_{t-1} + (1-\lambda_r)(\lambda_\pi\pi_{t-1} + \lambda_y y_{t-1}) + \nu_{3t}, \tag{2.12}$$

where h is the degree of habit persistence, ϕ is the relative risk aversion coefficient, β is the discount factor, ω is the degree of price indexation, ζ is the degree of price stickiness, and ν is the inverse elasticity of labor supply, while $\lambda_r, \lambda_\pi, \lambda_y$ are monetary policy parameters. The first two shocks follow an AR(1) process with parameters ρ_1, ρ_2, while ν_{3t} is i.i.d. The variances of the shocks are denoted by $\sigma_i^2, i = 1, 2, 3$. For the sake of presentation, we assume that the shocks are contemporaneously uncorrelated even though, in theory, some correlation must be allowed for.

Since the model features three shocks and three endogenous variables, we can construct several limited information objective functions, obtained by considering

the distances of all the responses to only one type of shock, the distance of the responses of a sub-set of the endogenous variables to all shocks, and the distance of the responses of all variables to all shocks.

The model has 14 parameters: $\theta_1 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ is underidentified from scaled impulse responses, just as in the previous example, the parameters of $\theta_2 = (\nu, \zeta)$ cannot be identified separately as they enter only in the slope of the Phillips curve (2.10) and in a multiplicative fashion, while $\theta_3 = (\beta, \phi, h, \omega, \lambda_r, \lambda_\pi, \lambda_y, \rho_1, \rho_2)$ contains the parameters of interest.

To construct aggregate decisions rules numerically, we set $\beta = 0.985, \phi = 2.0, \nu = 1.0, \zeta = 0.68, \omega = 0.75, h = 0.85, \lambda_r = 0.2, \lambda_\pi = 1.55, \lambda_y = 1.1, \rho_1 = 0.65, \rho_2 = 0.65$. With the aggregate decision rules we compute population responses and use 20 equally weighted responses to construct the distance function. We explore the shape of the distance function in the neighborhood of this parameter vector by tracing out how it changes when we change either one or two parameters belonging to θ_3 at a time, keeping the others fixed at their chosen values. As we have mentioned, identification problems could be due to solution or objective function pathologies. Here we convolute the two mappings, and directly examine how the shape of the objective function varies with θ, because the graphical presentation of these separate mappings is cumbersome.

Figure 2.1 plots the shape of the distance function when we vary β, ϕ, ω, h. Column 1 presents the distance function obtained using the responses of all three variables to monetary shocks; column 2 the distance function obtained using the responses of inflation to all shocks; and column 3 the distance function obtained using the responses of all variables to all the shocks. The range for the parameters considered is on the x-axis, while the height of the distance function for each parameter value is on the y-axis.

It is easy to see that monetary shocks have a hard time to identify the four structural parameters over the chosen intervals (the distance function is extremely flat in each of the parameters), that considering the responses of inflation to all shocks still leaves the coefficient of relative risk aversion pretty much underidentified, and that considering all the responses to all the shocks makes the distance function much better behaved. Still, asymmetries in the mapping between the risk aversion coefficient and the distance function remain even in this latter specification. Hence, taking a limited information approach, either in the sense of considering the responses of all variables to one shock or of one variable to all shocks, is problematic from an identification point of view.

One may wonder if this behavior is due to the choice of the parameters around which we map the distance function. The answer is negative. Canova and Sala (2005) construct the concentration statistic, defined as $C_{\theta_0}(i) = \int_{j \neq i} \frac{g(\theta) - g(\theta_0) d\theta}{\int (\theta - \theta_0) d\theta}, i = 1, \ldots, 9$, where g represents the distance function and θ_0 the pivot point, and let θ_0 vary over a reasonable range. Such a statistic synthetically measures how the multidimensional slope of the distance function changes around the selected parameter vector (see Stock, Wright and Yogo, 2002). Canova and Sala show that the minimum and maximum of this statistic in the range of θ_0 they consider varies

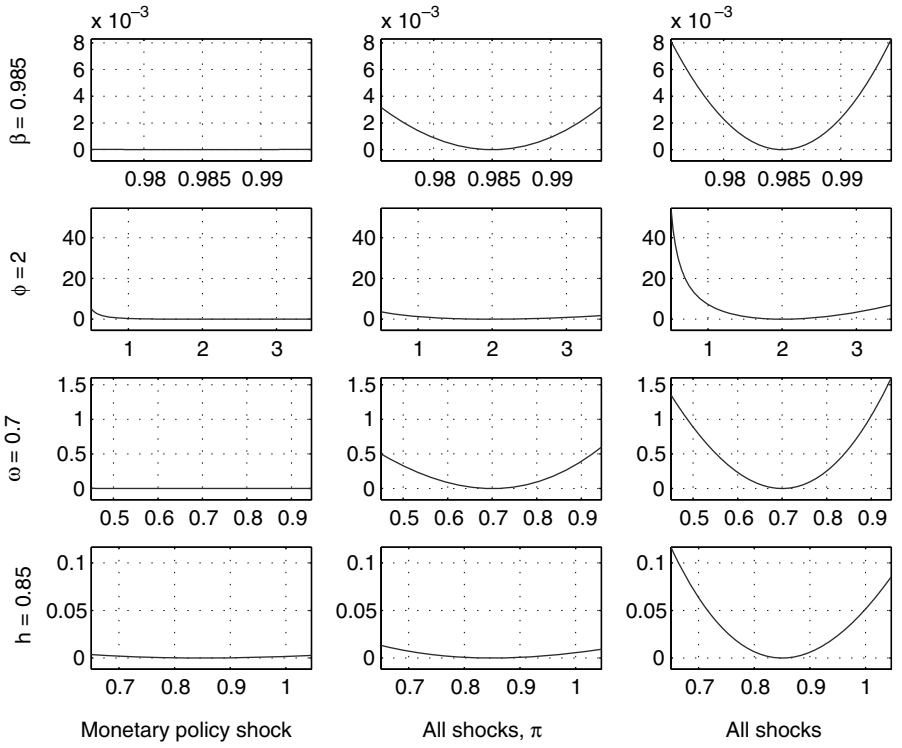


Figure 2.1 Shape of the distance function

very little, suggesting that the problems present in Figure 2.1 are not specific to the selected parameter vectors.

Since Figure 2.1 considers one dimension at a time, partial identification problems, where only combinations of parameters are identifiable, cannot be detected. Figure 2.2 shows that ridges indeed exist: for example, responses to monetary shocks carry little information about the correct combination of λ_y and λ_π; IS shocks cannot separately identify the risk aversion coefficient ϕ and the habit persistence parameter h, while Phillips curve shocks have little information about the discount factor β. What is interesting is that when the responses to all shocks are considered, some problems are reduced. For example, there appear to be fewer difficulties in identifying the parameters of the policy rule when all the responses to all shocks are considered – the distance function is more bell-shaped even though there is a significantly large flat area. However, even in this case, the true values of β, ϕ and h are difficult to pin down.

This model, in addition to partial, weak and underidentification problems, faces generic observational equivalence problems. For example, it would be hard to detect whether the data are generated by an indeterminate version of the model (which would be the case if $\lambda_\pi < 1$) or a determinate one ($\lambda_\pi > 1$), so long

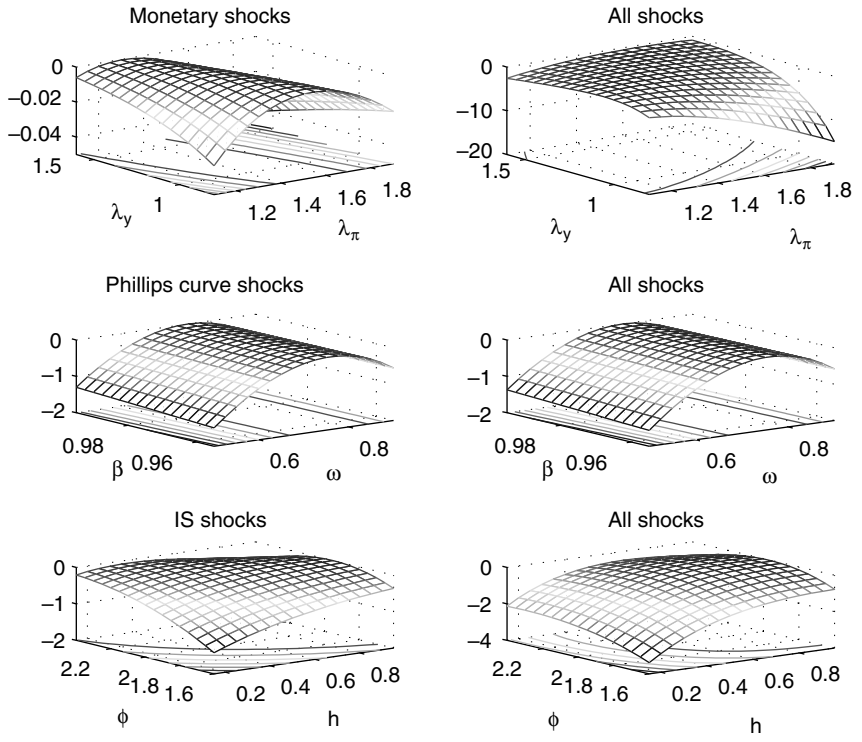


Figure 2.2 Distance function and contour plots

as the other parameters are allowed to be adjusted. Figure 2.3, which is reproduced from Canova and Gambetti (2007), shows that the shape and, in many cases, the size of the responses at almost all horizons to the three shocks are similar in the two regimes. Hence, if this were the only information available to the investigator, it would be difficult to detect which regime has generated the data.

This latter problem is a special case of a general pathology that applied investigators often face when dealing with DSGE models: the objective functions that one constructs from the aggregate decision rules may display multiple peaks, which may be clearly separated (as is the case in the above example; see also Lubik and Schorfheide, 2004) or not (see the example discussed in section 5 of Canova and Sala, 2005). Observational equivalence, probably more than any other identification problem, prevents attaching any meaningful economic interpretation to the outcomes of the estimation process and, obviously, conducting any meaningful policy analysis with the estimated model.

What generates the identification problems we have detected? All the nonlinear transformations, which are necessary to go from the structural parameters to the distance function, contribute. For example, consider the case of the price indexation parameter ω, which enters nonlinearly in the model and in several of

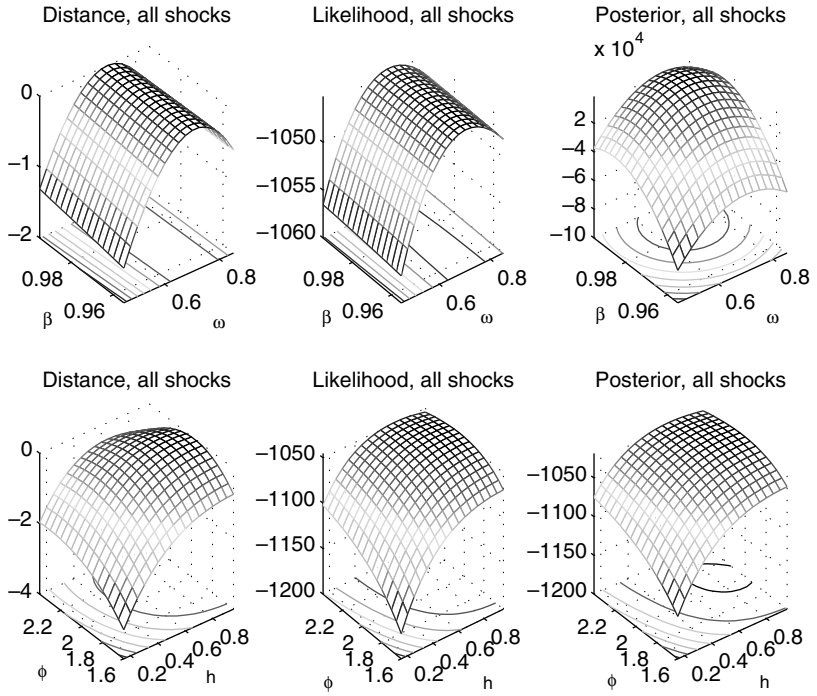


Figure 2.3 Impulse responses: determinate versus indeterminate equilibrium

the coefficients of the aggregate decision rules, but always in combination with other parameters. The coefficients of the restricted VAR solution are inverted to compute impulse responses and their distance from the "truth" is then squared and summed. One would guess that it is just by chance that such a complex set of operations will allow the mapping from ω to the objective function to be well behaved.

The standard answer to the problems shown in Figures 2.1 and 2.2 is to fix parameters with difficult identification features (after all, it does not matter what value we select) and estimate the remaining ones. While this approach is common, there is no guarantee that it will give meaningful answers to the questions of interest. In fact, while such a mixed calibration-estimation approach will be successful, at least in population, if the parameters that are treated as fixed are set at their true value, setting them at values which are only slightly different from the true ones may lead estimation astray. Intuitively this happens because, for example, setting β to the wrong value implies adjustments in parameters which enter jointly with β in the coefficients of the aggregate decision rules and this may move the minimum of the function in a somewhat unpredictable way. Canova and Sala (2005) show, in the context of a simple RBC example, that these shifts may be significant and may drive inference the wrong way.

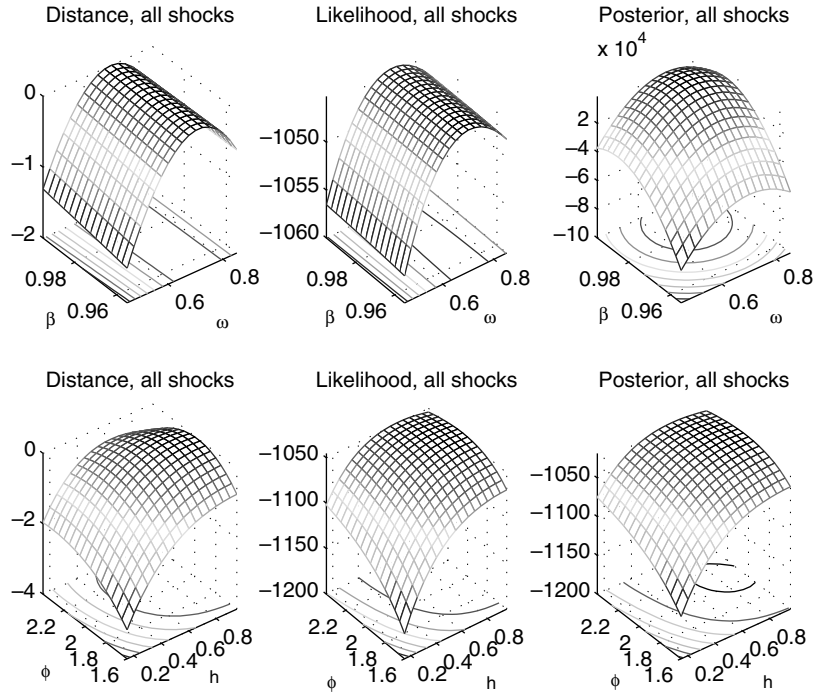


Figure 2.4 Distance function, likelihood and posterior plots

What can one then do to conduct structural estimation? The distance function we have employed can be obtained by approximating the likelihood function of the model. Therefore, the resulting estimators can be thought of as quasi-maximum likelihood (ML) estimators of the structural parameters. However, there is no reason to use such an approximation. Once the decision rules are written in a state space format, the likelihood function can be easily and efficiently computed with the Kalman filter. Therefore, identification problems could be reduced if information about the covariance matrix of the shocks or the steady-states of the model – which are not used when normalized impulse response matching is performed – are brought into the estimation. Figure 2.4, which plots the distance function when all the shocks are considered and the likelihood function in β and ω, and ϕ and h, indeed suggests that these parameters could be better identified from the likelihood than from the distance function – the curvature of the latter is much larger than the curvature of the former. Nevertheless, the problem with ridges remains. Since the likelihood has all the information that the model delivers, one can conclude that it is the solution mapping, rather than the objective function mapping, that induces under- and partial identification problems in this example.

It has become quite common to estimate the parameters of a DSGE model by Bayesian methods. Bayesian methods attempt to trace out the shape of the posterior

distribution of the structural parameters, which is proportional to the likelihood times the prior. The use of prior information could add curvature to the likelihood function, therefore making identification problems apparently disappear. We show how this can happen in the last column of Figure 2.4. A sufficiently tight prior has given the posterior a nice bell-shaped appearance with round contours in (β, ω). Clearly, the use of Bayesian methods are not the solution to the identification problems we have highlighted in this sub-section – it could, however, help when identification problems are caused by small samples. Achieving identification via prior restrictions does not change the fact that the likelihood function constructed through the lenses of the aggregate decision rules of the model has little information about the structural parameters. In this case the shape of the posterior distribution will, to a large extent, mimic the shape of the prior, so that structural estimation is nothing more than sophisticated calibration – rather than calibrating to a point, we calibrate to an interval, and within the interval we assume that some parameter values are more likely than others. When population identification problems exist and a researcher is interested in estimating the structural parameters, it is necessary to reparametrize the model. If this is infeasible or undesirable, then informal calibration is one simple and internally consistent device to make the model operative for inference and forecasting. The deep issue here is that DSGE models are not typically designed with an eye to the estimation of their parameters and this is clearly reflected in the identification problems they display.

Prior information on the parameters of macroeconomic models may come from different sources. It may be accumulated knowledge about a phenomenon repeatedly studied in the literature (for example, the properties of the transmission of monetary policy shocks), evidence obtained from micro-studies, or from the experience of other countries. All this information may be valuable to the applied investigator and should be formally introduced in the structural estimation of the model, if available. However, if the likelihood has little information about the structural parameters, and this additional information was all that was available to identify the parameters, structural estimation would not be particularly useful – it would resemble confirmatory analysis where prior expectations are verified *a posteriori*. In this situation, policy exercises are difficult to interpret, and the alternative of measuring the range of outcomes produced by the model using a selected range of parameters, as suggested in Canova (1995), is a feasible and more plausible approach.

What are the consequences of the identification problems we have described? For the sake of presentation, we will focus on estimates obtained by matching responses to monetary policy shocks, which appear to produce the distance function with the worst identification properties, and are those on which the literature has paid most attention. In this exercise we still assume that shocks are correctly identified – in our model, reduced-form interest rate innovations are the true monetary policy shocks. If this were not true, additional problems, such as those discussed in, for example, Canova and Pina (2005), would be compounded by those discussed here. We consider different sample sizes, on the one hand, to highlight some of the properties of the estimates of parameters

with problematic identification features and, on the other, to examine whether additional identification problems may emerge just because of small samples.

We simulate 200 time series for interest rates, the output gap and inflation for $T = 120, 200, 1000$, fixing $\nu = 1$ and $\sigma_i^2 = 1.0$ in all cases; we estimate an unrestricted VAR(2), which is the correct empirical reduced form model to use in this case, and compute impulse responses and bootstrap confidence bands which are then used to build a diagonal matrix of weights: the weights are inversely proportional to the uncertainty in the estimates. Table 2.1 presents a summary of the estimation results. It reports the true parameters, the mean estimate, the numerical standard errors computed across replications (in parentheses) and the percentage bias (in square brackets).

A few features of the table are worth commenting upon. First, biases are evident in the estimates of the partially identified parameters (λ_π, λ_y), the weakly identified parameters (ζ, ω, h), and the underidentified parameters $(\rho_1,\ \rho_2)$. Note that even with 250 years of quarterly data, major biases remain. Second, numerical standard errors are large for the partially identified parameters and invariant to sample size for the underidentified ones. Third, parameter estimates do not converge to population values as T increases. Finally, and concentrating on $T = 200$, estimates suggest economic behavior which is somewhat different from that characterizing the DGP. For example, it appears that price stickiness is stronger and central bank reaction to the output gap and inflation is equally strong.

In sum, identification problems lead to biased estimates of certain structural parameters (see also Choi and Phillips, 1992), to inappropriate inference when conventional asymptotic theory is used to judge the significance of estimated parameters, and, possibly, to wrong economic interpretations. For unconditional forecasting, identification problems are unimportant: as long as the fit and the forecasting performance is the same, the true nature of the DGP does not matter. However, policy analyses and conditional forecasting exercises conducted with estimated parameters may lead to conclusions which are very different from

Table 2.1 NK model: matching monetary policy shocks

	True	$T = 120$			$T = 200$			$T = 1000$		
β	0.985	0.984	(0.007)	[0.6]	0.985	(0.007)	[0.7]	0.986	(0.008)	[0.7]
ϕ	2.00	2.39	(2.81)	[95.2]	2.26	(2.17)	[70.6]	1.41	(1.19)	[48.6]
ζ	0.68	0.76	(0.14)	[19.3]	0.76	(0.12)	[17.5]	0.83	(0.10)	[23.5]
λ_r	0.20	0.47	(0.29)	[172.0]	0.43	(0.27)	[152.6]	0.41	(0.24)	[132.7]
λ_π	1.55	2.60	(1.71)	[98.7]	2.22	(1.51)	[78.4]	2.18	(1.38)	[74.5]
λ_y	1.1	2.82	(2.03)	[201.6]	2.56	(2.01)	[176.5]	2.16	(1.68)	[126.5]
ρ_1	0.65	0.52	(0.20)	[30.4]	0.49	(0.21)	[34.3]	0.50	(0.19)	[31.0]
ρ_2	0.65	0.49	(0.20)	[32.9]	0.48	(0.21)	[34.8]	0.48	(0.21)	[34.7]
ω	0.25	0.76	(0.39)	[238.9]	0.73	(0.40)	[232.3]	0.65	(0.38)	[198.1]
h	0.85	0.79	(0.35)	[30.9]	0.76	(0.37)	[32.4]	0.90	(0.21)	[21.3]

those obtained with the true one. Hence, it is generally unwise to attach any economic interpretation to the estimates or draw conclusions about how the economy works from structural exercises which are plagued by identification problems.

What is left for the applied investigator to do? Apart from attempting to reparametrize the model, not much. One interesting issue still unexplored in the literature is to take population identification problems as being the norm rather than the exception, and try to find estimation techniques or objective functions which, given a sample size, are able to minimize the distortions produced by identification pathologies. While some progress has been made in the context of moment estimation (see Stock and Wright, 2000; Rosen, 2006), these procedures are applicable only in restrictive situations (the weighting matrix must be chosen in a particular way) and are awkward to use in DSGE models, which are highly parameterized and nonlinear.

How does one detect identification problems? The univariate and bivariate exploratory analysis we have presented, for example, in Figures 2.1 and 2.2 can definitely help in spotting potential problems and this analysis could easily be complemented with local derivatives of the objective function in the dimensions of interest. Alternatively, numerically computing the Hessian of the objective function around particular parameter values and calculating the size of its eigenvalues can give more formal indications on how flat or how information-deficient the objective function is locally. For example, if the rank of the Hessian is less than the number of structural parameters, one of its eigenvalues is zero and at least one parameter is underidentified. If the rank of the Hessian is close to deficient, one or more of its eigenvalues is close to zero and either weak or partial identification problems, or both, are likely to be present. Experimentation with the number of shocks used to construct the objective function and the number of variables can also give useful information about what statistic may identify a particular structural parameter, as is experimentation with different objective functions and with different features of the data (for example, steady-state versus dynamics).

Clearly, diagnostics of this type have to be run prior to estimation, but such an exercise is not much more complicated or time consuming than the type of exercises one performs to measure the sensitivity of the results to the selection of calibrated parameters. In general, the following rules of thumb are useful to limit the extent of identification problems: given a model, always choose a likelihood-based objective function, which has the highest informational content; given a model and the likelihood function, and if it is the sample which is problematic, add information in the form of additional data or prior restrictions, which synthetically reproduces it.

It is important to stress that looking at the minimized value of the objective function, at standard errors of the estimates or at the resulting impulse responses, is not generally useful as an *ex post* device to detect identification problems. The distance function is within the tolerance level (10^{-7}) for all the parameter combinations generating Table 2.1, and the practice of blowing up the objective function by appropriately choosing the matrix of weights will not change the fact that the

gradient or the Hessian display problematic features. Furthermore, it can be shown that population responses fall within a 68% band centered around the estimates of the responses to monetary shocks computed with the parameter estimates, even when the sample size is $T = 120$. Therefore, the practice of showing that the model's responses computed using the estimated parameters lie within the confidence bands of the responses estimated from the data is not particular informative as far as identification problems are concerned. Large standard errors do emerge when identification failures exist, but also when other problems are present (for example, very noisy data or regime switches). Hence, associating large standard errors with identification issues is, in general, incorrect.

It is also important to stress that the addition of measurement errors for estimation purposes can distort the identification properties of structural parameters. It is not particularly difficult to conceive situations where a parameter that was identified by certain features of the model becomes free to move and fit other properties of the data it was not designed for, once measurement error is added. Therefore, while there is some logic in adding measurement errors to link the model variables to the observables, one should be careful and investigate the consequences that such a process has on the identification properties of the parameters.

2.3 Structural VARs

Structural VAR inference is typically perceived to be at the extreme opposite to structural model-based inference. SVAR models take a minimalist approach to the estimation problem and consider only a very limited sub-set of the large number of restrictions that DSGE models impose on the data. For example, the fact that the matrices H_1 and H_2 in (2.4) depend on θ is typically neglected and only a part of the information present in $A_0(\theta)$ is used. Furthermore, the singularity that the model imposes on the data is completely disregarded.

This minimalist approach has one obvious disadvantage: if less structure is imposed on the data, fewer interesting economic questions can be asked. However, such a limited information approach is advantageous when some of the model's restrictions are dubious, which would be the case if the model is misspecified in some dimensions, or fragile, which would be the case if the restrictions depend on the functional forms or the parameter values one specifies. In this case, neglecting these restrictions can robustify estimation and inference.

As we have mentioned in the introduction, and despite recent attempts to make them more realistic, the current generation of DSGE models is still far from reproducing the DGP of the actual data in many respects: models fail to capture, for example, the heterogeneities present in the actual world; important relationships are modeled with black-box frictions; timing restrictions are used to make them compatible with the dynamics observed in the data; and *ad hoc* shocks are often employed to dynamically span the probabilistic space of the data. Since misspecification is likely to be pervasive, system-wide and even limited-information classical structural methods are problematic, even when identification problems are absent.

Bayesian methods have an edge in structural estimation when model misspecification is present. Inference in this context, in fact, does not require the asymptotic correctedness of the model under the null. Furthermore, these methods can potentially deal with model misspecification, for example, by imposing prior distributions over models and weighting the posterior information contained in each of them by their posterior probability. However, this potential advantage of Bayesian methods is often unexpressed: except for Schorfheide (2000), it is very unusual for researchers to consider an array of models, all of which can potentially be useful to answer the question of interest. In this situation, one is often left wondering what posterior estimates obtained from a misspecified model mean in practice and whether policy makers could and should trust these estimates when taking important policy decisions.

The difficulties of the current generation of DSGE models in representing the DGP of the data have been highlighted by Del Negro *et al.* (2006), who take a workhorse model, popular among academics and central bankers, and show that it is possible to improve its fit by considerably relaxing the cross-equation restrictions that it imposes on the matrices $H_1(\theta)$ and $H_2(\theta)$. Their approach, which uses a DSGE model as a prior for a VAR, is useful for designing a metric to assess the distance between the model and the VAR of the data, and represents a promising way to evaluate model fit, to suggest ways to bring models in closer contact with the data and, in general, to conduct structural inference in misspecified models.

If one takes the inherent misspecification that the current generation of DSGE models display seriously and heavily weights inferential mistakes, one may then want to proceed in a more agnostic way. Rather than conditioning on one model and attempting to estimate its structural parameters, one could be much less demanding in the estimation process, and employ a sub-set of the model restrictions, which are either uncontroversial or likely to be shared by a class of economies with potentially different features, to identify structural shocks. One way of doing this is to neglect the restrictions present in the matrices H_1 and H_2, which are often not robust, and use some of those present in $A_0(\theta)$, for which a strong *a priori* consensus can be found in theory, and then trace out the dynamics of the variables of interest in response to disturbances or measure the relative importance of each shock for business cycle fluctuations. Therefore, with such an approach, most of the detailed cross-equation restrictions imposed by a model will be eschewed from the estimation process and only constraints which are likely to hold in many models are used to identify structural shocks. Unfortunately, it has become common in the literature to employ constraints which are unrelated to any specific class of models or are so generic that they lack economic content. While 20 years ago researchers spent considerable time and effort justifying their identification restrictions from a theoretical point of view (see, for example, Sims, 1986; Bernanke, 1986), now it is often the case that these restrictions are not even spelled out in detail, and the only justification for them a reader can find is that they are used because someone else in the literature has used them before. In general, delay-type restrictions, which use the flow of information accrual to agents in the economy, and placing zeros in the impact matrix of shocks, are the preferred identification devices.

Canova and Pina (2005) have shown that delay-type restrictions do not naturally arise in general equilibrium models, are often inconsistent with their logic, and one has to work hard to cook up general equilibrium environments with such features (see, for example, Rotemberg and Woodford, 1997). Long-run restrictions have been hailed in the past as the answer to these problems, since restrictions of this type are common to a variety of theories (for example, money neutrality or the idea that technological progress explains the long-run path of variables are features which are shared by macro-models with different micro-fundations) and allow inference without tying one's hand to a particular specification for the short-run dynamics around these long-run paths. However, this alternative identification approach is non-operative: long-run restrictions are scarce relative to the number of shocks researchers are interested in recovering. Therefore, when four or five shocks need to be identified, one is forced to use a mixture of long-run and delay restrictions. Furthermore, as pointed out by Faust and Leeper (1997), long-run restrictions are weak and prone to observational equivalence problems.

The sign and shape approach, suggested in Canova and De Nicoló (2002) and Uhlig (2005), is advocated in the next section and can bridge SVAR and DSGE models in a more solid way and provide a constructive answer to the quest for identification restrictions. Unfortunately, such an approach does not yet have widespread use in the profession (exceptions are, among others, Dedola and Neri, 2007; Pappa, 2005) and the science of identification in SVARs is still very much the craft of finding restrictions that would not bother anyone in the profession.

Apart from identification issues, which have received attention in the VAR literature since, at least, Cooley and LeRoy (1985), a number of authors have recently questioned the ability of structural VARs to recover the true DGP of the data, even when an appropriate identification approach is used. To see why this could be the case, consider the following alternative restricted state space representation for the log-linearized aggregate decision rules of a DSGE model:

$$\begin{aligned} x_{1t} &= J(\theta)x_{1t-1} + K(\theta)e_t \\ x_{2t} &= N(\theta)x_{1t-1} + M(\theta)e_t, \end{aligned} \tag{2.13}$$

where $e_t \sim iid(0, \Sigma_e)$. The questions we ask are the following: (i) Does a VAR representation for a subset of the variables of the model, say x_{2t}, exist? (ii) Would the resulting VAR be of finite order? (iii) What would happen to inference if only a sample of limited size is available? We have already mentioned that, if both x_1 and x_2 were observable, (2.12) is simply a restricted, though reduced rank, VAR(1). However, this is not a very realistic set-up: usually x_{1t} includes non-observable variables; furthermore, only a sub-set of the variables appearing in x_{2t} may be of interest, could be reasonably measured, or have relevant information for the exercises one may want to conduct. Therefore, it is legitimate to ask what the process of integrating out non-observable, uninteresting or badly measured variables would imply for the restricted time series representation of the aggregate decision rules of the model.

2.3.1 Invertibility

If $M(\theta)$ is a square matrix, and if $J(\theta) - K(\theta)M(\theta)^{-1}N(\theta)$ has a convergent inverse (for example, if all its eigenvalues are less than 1 in absolute value), it is easy to show that:

$$x_{2t} = N(\theta)\{[1 - (J(\theta) - K(\theta)M(\theta)^{-1}N(\theta)]^{-1}K(\theta)M(\theta)^{-1}\}x_{2t-1} + u_t, \qquad (2.14)$$

where $u_t \sim (0, M(\theta)'\Sigma_e M(\theta))$. Therefore, if only x_{2t} is observable, the aggregate decision rules have a restricted VAR(∞) representation. If instead $N(\theta)$ is a square matrix, then:

$$x_{2t} = N(\theta)J(\theta)N(\theta)^{-1}x_{2t-1} + (I + (N(\theta)K(\theta)M(\theta)^{-1} - N(\theta)J(\theta)N(\theta)^{-1})\ell)u_t, \quad (2.15)$$

where ℓ is the lag operator. Under this alternative assumption, the aggregate decision rule for x_{2t} therefore has a VARMA(1,1) representation.

Hence, if a reduced number of variables is considered, the aggregate decision rules of the model have a much more complicated structure than a restricted VAR(1). The question of interest is whether we can still use a VAR with a finite number of lags to approximate the aggregate decision rules for x_{2t}. Straightforward algebra can be used to show that if the exogenous driving forces are AR(1) and if both the predetermined states and x_{2t} are observed, then the correct representation for the vector of predetermined states and choice variables is a restricted VAR(2) with singular covariance matrix. On the other hand, if only x_{2t} is observable and the dimension of x_{2t} is the same as the dimension of e_t, Ravenna (2006) has shown that the aggregate decision rules for x_{2t} can be approximated with a finite order VAR if and only if the determinant of $\{I - [J(\theta)K(\theta)M(\theta)^{-1}N(\theta)]\ell\}$ is of degree zero in ℓ.

What does this all mean? It means that the aggregate decision rules for a sub-set of the variables of the model can be represented with a finite order VAR only under a set of restrictive conditions. These conditions include invertibility of the mapping between structural shocks and the variables included in the VAR, a fundamentalness condition, which implies that the information contained in the observables is the same as the information contained in disturbances of the model, and the condition that random perturbations produce fluctuations around the steady-state that are not too persistent.

Note that the condition we have used to derive (2.13), is never satisfied in practice. One can think, at best, of four or five truly structural sources of disturbances and this typically is much less than the size of the vector x_{2t}. Therefore, it is only after *ad hoc* disturbances and/or measurement errors are *ex post* included that $M(\theta)$ is a square matrix. Similarly, the restriction that $N(\theta)$ is a square matrix is difficult to satisfy in practice – the number of states is typically smaller than the number of endogenous variables. The other conditions clearly depend on the structure of the model but, for example, specifications in which agents react to news that may materialize in the future fail to satisfy the first condition – the resulting MA representation of the model is nonfundamental. Finally, the convergence of the

economy to its steady-state when perturbed by shocks depends on the details of the specification. Therefore, it is difficult to assess how important in practice this assumption is. Given that many DSGE models have fairly weak internal propagation mechanisms, and as long as the structural shocks are stationary, such a condition is likely to be satisfied in practice.

In sum, one should not be surprised to find DSGE models featuring aggregate decision rules for a sub-set of the variables that are not representable with a finite-order VAR (see Fernandez-Villaverde *et al.*, 2007, for examples). Nevertheless, a large class of models does have aggregate decision rules with these properties. To be sure that SVAR inference is valid, one must first select a class of models which could have generated the data and check whether the required conditions are satisfied for alternative parameterizations. While this requires a SVAR investigator to take a certain class of models much more seriously before drawing any inference from his/her analysis, it also makes SVAR estimation less straightforward and more time consuming since the number of parameters, functional form and friction permutations that need to be checked before the analysis is conducted is large. Furthermore, since bizarre counter-examples can always be found, it may become difficult for an applied macroeconomist to assess in practice whether a finite order VAR is a good approximation to the class of DSGE models one is interested in or not.

For the final question, Chari, Kehoe and McGrattan (2006) have recently shown that one may be led astray when evaluating the relevance of economic theories using SVARs simply because, with small samples, the population properties of the aggregate decision rules may be very poorly approximated with a VAR. That is to say, even when there exists a VAR representation for the variables in x_{2t}, when this representation is of finite order, and when identification of shocks is properly performed, small sample biases in the estimates of the reduced form parameters and the covariance matrix of the shocks may make inference whimsical. For example, they show that a short sample of data simulated from an RBC model driven by a neutral technology shock may lead a researcher to believe that it could have been generated by a model with different microfundations – in the population, hours worked increases in response to a technology shock, but in small samples hours may fall in response to the correctly identified technology shocks.

An applied investigator has to live with small sample biases. Long samples, even when they are available, are rarely used because causal relationships are often subject to important regime shifts, and when regime shifts are absent, changes in the definition or in the way the data is sampled or computed make empirical analysis difficult. Econometrics can help here: it is well known that in a variety of experimental designs and with samples of about 100 observations, estimates of the AR(1) coefficient are downward biased by up to 30%. While this type of analysis could be easily extended to more realistic and interesting economic models – for example, to measuring the size of the bias in the largest autoregressive root of the aggregate decision rule (which roughly determines the dynamics of the system) and in the eigenvalues of the covariance matrix of reduced form shocks (which determines the size of the impact effects) – one needs to consider models where the impact effect is fairly weak to have important sign reversals in small samples. Therefore,

while such an issue should be kept in mind, its practical relevance appears to be limited.

There is another way of seeing these representation problems from a different and probably more informative viewpoint – that of omitted variables and shock misaggregation, which have a long tradition in the VAR literature (see, for example, Braun and Mittnik, 1993; Faust and Leeper, 1997). Suppose the aggregate decision rules for the endogenous variables of a DSGE model can be written as a VAR(1):

$$\begin{bmatrix} I - A_{11}\ell & A_{12}\ell \\ A_{21}\ell & I - A_{22}\ell \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} e_t,$$

where y_{1t} are the variables included and y_{2t} the variables excluded from the empirical model and where these vectors do not necessarily coincide with those of the state variables x_{1t} and the choice variables x_{2t}. Then the representation for y_{2t} is:

$$(I - A_{22}\ell - A_{21}A_{12}(1 - A_{11}\ell)^{-1}\ell^2)y_{2t} = [B_2 - (A_{21}(1 - A_{11}\ell)^{-1}B_1\ell]e_t \equiv \upsilon_t. \quad (2.16)$$

When y_{1t} and y_{2t} are of the same dimensions, this simplifies to:

$$[I - (A_{11} + A_{22})\ell + (A_{11}A_{22} - A_{21}A_{12})\ell^2]y_{2t} = [B_2 + (A_{21}B_1 - A_{11}B_2)\ell]e_t \equiv \upsilon_t. \quad (2.17)$$

What does this reduced system representation imply? First, it is easy to see that the model for y_{2t} is an ARMA(∞, ∞) and the lagged effect of the disturbances mixes up the contemporaneous effects of different structural shocks ($B_1 e_{t-1}$ has smaller dimension than e_{t-1}). Furthermore, it is clear that even if the e_ts are contemporaneously and serially uncorrelated, the υ_ts are contemporaneously and serially correlated and that two small-scale VARs featuring different y_{2t}s will have different υ_ts. Finally, since υ_t is a linear combination of current and past e_t, the timing of the innovations in y_{2t} is not preserved unless A_{11} and A_{21} are both identically equal to zero, which is true, for example, if y_{2t} includes the states and y_{1t} the controls of the problem.

In other words, (2.16) implies that shocks extracted from a SVAR featuring a reduced number of the model's variables are likely not only to confound structural shocks of different types, but also to display time series properties which are different from those of the true shocks to these variables. Hence, even if the correct identifying restrictions are used, VAR models which are small relative to the universe of variables potentially produced by a DSGE model are unlikely to be able to capture either its primitive structural disturbances or the dynamics they induce unless some strong and not very practically relevant conditions hold.

Contrary to the previous representation of the invertibility problem, which provides little guidance on how to check for failures, this latter representation does give researchers a way to measure the importance of potentially omitted variables. In fact, if omitted variables are important, reduced form VAR residuals will be correlated with them. Therefore, for any given set of variables included in the VAR, it is sufficient to check whether variables potentially belonging to y_{1t} display significant correlation with the residuals. If so, they should be included in the VAR and estimation repeated; if not, they can be omitted without further ado.

To conclude, we present two examples below illustrating the issues we have discussed in this section. In the first example, noninvertibility emerges because the model has a nonfundamental representation. In the second, the MA of the model is invertible, but the dynamics of the reduced system are different from those of the full one.

2.3.1.1 Example 3: a Blanchard and Quah economy

The example we present belongs to the class of partial equilibrium models popular in the late 1980s. While it is not difficult to build general equilibrium models with the required features, the stark nature of this model clearly highlights how invertibility problems could occur in practice. The model that Blanchard and Quah (1989) consider has implications for four variables (gross domestic product (GDP), inflation, hours and real wages) but the solution is typically collapsed into two equations, one for GDP growth (ΔGDP), the other for the unemployment rate (UN_t), of the form:

$$\Delta GDP_t = \epsilon_{3t} - \epsilon_{3t-1} + a(\epsilon_{1t} - \epsilon_{1t-1}) + \epsilon_{1t} \tag{2.18}$$

$$UN_t = -\epsilon_{3t} - a\epsilon_{1t}, \tag{2.19}$$

where ϵ_{1t} is a supply shocks, ϵ_{3t} a money supply shock and a is a parameter measuring the impact of supply shocks on aggregate demand. Hence, the aggregate decision rule for these two variables is a VMA(1). It is easy to check that a finite-order VAR may approximate the theoretical dynamics of this model only if $a > 1$.

To see this, we set $a = 0.1$ and plot in Figure 2.5 the theoretical responses of output and unemployment to the two shocks and the responses obtained using a VAR(1) and a VAR(4), where the econometrician uses the correct (but truncated) vector autoregressive representation of the model. Note that, while the signs of the responses are correct, the dynamics are very different. Also, while there is some improvement in moving from a VAR(1) to a VAR(4), some of the theoretical responses are very poorly approximated even with a VAR(4). Since a VAR(q), $q > 4$, has responses which are indistinguishable from those of a VAR(4) – as the matrices on longer VAR lags are all very close to zero – no finite-order VAR can capture (2.17) and (2.18).

What generates this result? When $a < 1$ the aggregate decision rules of the model are nonfundamental, that is, innovations to output growth and unemployment do not have the same information as the variables themselves. Therefore, there is no convergent VAR representation for these two variables where the roots of the VAR are all less than one in absolute value, and this is true even when an infinite lag length is allowed for.

2.3.1.2 Example 4: an RBC model

We work with the simplest version of the model since more complicated structures do not bring additional insights into the problem. The social planner maximizes:

$$E_0 \sum_{t=0}^{\infty} \beta^t \frac{c_t^{1-\phi}}{1-\phi} - AN_t, \tag{2.20}$$

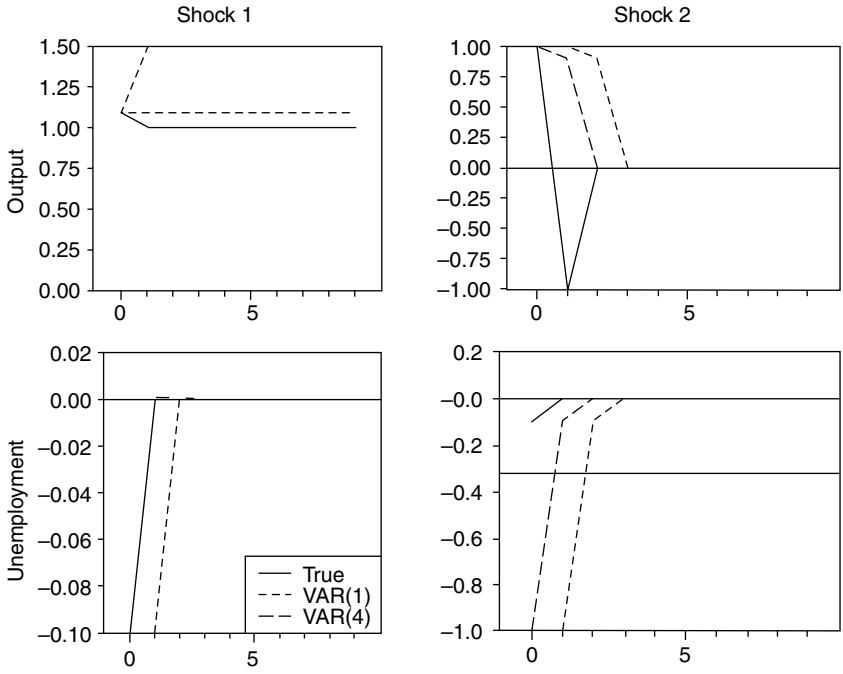


Figure 2.5 Responses in the Blanchard and Quah model

and the resource constraint is:

$$c_t + k_t + g_t = k_{t-1}^{\eta} N_t^{1-\eta} z_t + (1-\delta)k_{t-1}, \tag{2.21}$$

where c_t is consumption and ϕ is the risk aversion coefficient, A is a constant and N_t are hours worked; z_t is a first-order autoregressive process with persistence ρ_z, steady-state value z^{ss} and variance σ_z^2; g_t is a first-order autoregressive process with persistence ρ_g, steady-state value g^{ss} and variance σ_g^2; k_{t-1} is the current capital stock; η is the share of capital in production, and δ the depreciation rate of capital. Using the method of undetermined coefficients, and letting output be $y_t \equiv k_{t-1}^{\eta} N_t^{1-\eta} z_t$, and investment be $i_t = k_t - (1-\delta)k_{t-1}$, the aggregate decision rules for $(k_t, c_t, N_t, y_t, r_t, i_t)$, where r_t is the real rate, imply standard dynamics in response to the two shocks. In particular, as z_t increases, hours, consumption, output, the real rate and investment increase contemporaneously while the dynamics of the capital stock have a hump-shaped pattern. On the other hand, as g_t increases, consumption falls, hours, output, the real rate and investment increase contemporaneously and the capital stock has a hump-shaped pattern.

What would the dynamics induced by the two shocks in a system which includes only the interest rate and investment look like? That is, what would happen if we integrate out the effect of the shocks on the other four variables? Figure 2.6 plots the responses of the two variables of interest to the two shocks in the full and the

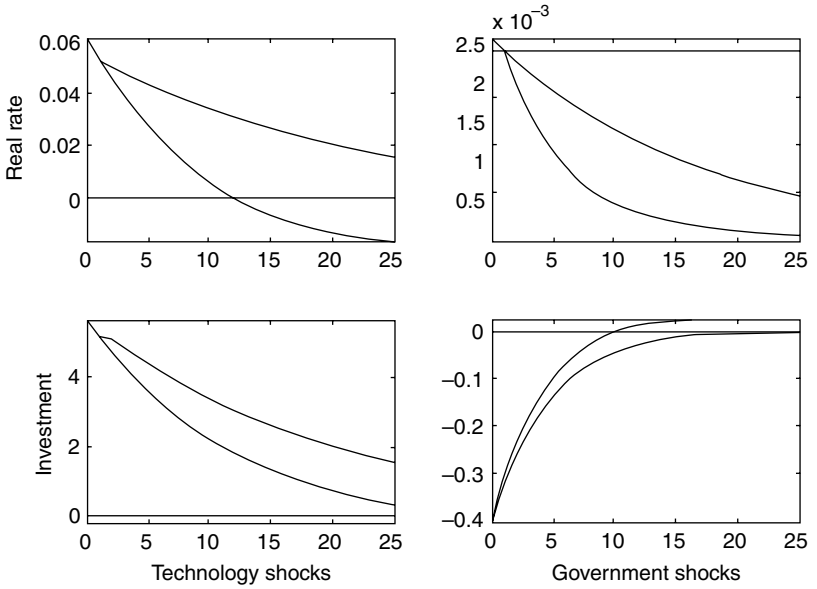


Figure 2.6 Dynamics in an RBC model

reduced systems. Clearly, while the impact effect is identical, lagged dynamics are very different.

What is the reason for this result? Mechanically, since A_{11} and A_{21} are not small, shocks last more than one period and persist for a number of periods. Notice that the persistence in the reduced system is strong (see, for example, the effect of technology shocks on the real rate), suggesting that the process of marginalizing part of the system has serious consequences on the responses of the variables to shocks, at least in this example.

It goes without saying that it makes a lot of difference which of the two systems one uses as a benchmark to represent the DSGE model and in trying to see whether actual and simulated data are similar or not.

2.4 Some final thoughts

The previous two sections may have given the reader a rather pessimistic view about the possibility of conducting meaningful inference with DSGE models and the impression that not many alternatives are left to the applied investigator. If structural estimation is pursued, misspecification of the structural relationships may make the interpretation of estimates difficult; identification problems are likely to be widespread and even in the unlikely case when they are not present, a number of additional statistical and specification assumptions need to be made, making it very difficult to judge what is causing what. The alternative of using SVARs seems to be equally problematic. While VARs are less prone to misspecification of the

structural relationships, identification problems are still present and noninvertibility of the DSGE models aggregate decision rules may also make SVAR analyses uninterpretable.

Chari, Kehoe and McGrattan (2007) have suggested using the so-called business accounting method to evaluate DSGE models, but the logic of the approach represents a step backward relative to what we discuss here – only reduced form relationships are used to judge what is missing from the model – and it is hard to avoid important observational equivalence problems when judging different structural models of the business cycle.

What, then, should one do? No matter which approach one takes, one should be very careful and learn how to interpret the information contained in the diagnostics obtained from experimenting with the structure of the model and investigating the properties of the data. If structural estimation is performed, methods which allow for misspecification should be preferred and extra information, in the form of micro-data or data from other countries, may help to break the deadlock of parameter identification when problems are due to small samples. We have suggested that to solve population identification problems it is necessary to reparameterize or respecify DSGE models, but obviously this is a more long-term goal, since such an approach brings us back to the very basic foundation of DSGE-based exercises. Nevertheless, if theorists would build models bearing in mind that they will be estimated, certain issues could be completely avoided.

If SVAR analysis is preferred, one should link the empirical model to DSGE theories much better than has been done so far, explicitly write down the class of models one will employ to interpret its results (as done, for example, in Canova and De Nicoló, 2002), and perform the preliminary analysis necessary to check whether the aggregate decision rules of such a class of models do have a finite-order VAR format for the sub-set of relevant variables used in the VAR. Identification should also be clearly linked to the class of structural models of interest and artificial delay restrictions avoided. One way of doing this is described in Canova (2002), where robust restrictions on the sign of responses to shocks derived from a class of models are used to identify shocks, and the results of the analysis are discussed through the lenses of such models. Canova and Paustian (2007) show that such an approach has good size and power properties against local alternatives and gives reasonable results in inappropriately marginalized systems.

Integrating structural and VAR analyses, as suggested by Del Negro and Schorfheide (2004, 2005, 2006), also provides an interesting avenue for future research, where structural models and empirical analyses can cross-fertilize each other.

From the point of view of policy makers, DSGE models are useful if they can forecast well, since it is much easier to tell stories with estimates of their parameters than with SVAR estimates or estimates of pure time series models. However, to forecast at least as well as more unrestricted models, the DSGE models popular in the academic literature must produce restrictions which are not rejected in the data, and this is pretty hard to do when one considers, for example, prices rather than

quantities and financial or monetary variables rather than real ones. In addition, to test the quality of these restrictions one needs substantial "cosmetic surgery" in the form of additional shocks, frictions and other black-box jingles, which are difficult to justify from a theoretical point of view and make any hypothesis a joint test of the restrictions and the chosen add-ons. Realizing these facts should probably lead academics and policy makers to be less demanding of the models they write down and use. Typically, small models forecast better than larger ones and different models can be used for different purposes. Having an array of models at one's disposal, which are built to answer different economic questions, and averaging their forecasting results may not only robustify the outcomes of the investigation but also give an entirely different perspective on the reasons driving certain economic phenomena.

While one can envision the disappearance of the "model" of the economy as conceived in the 1970s, constructed by patching up pieces of theoretical structures and a lot of empirical wisdom, and used to answer all possible questions policy makers may have, it is very likely that smaller scale, more or less structurally oriented models will coexist in the portfolio of research departments of central banks and international institutions for a while, serving different purposes and different objectives.

To go back to the main question of this chapter, how much structure should there be in an empirical model? The solomonic and, probably, obvious answer, is that it depends on the scope of the analysis and the information available in the data. Different models can have different structural content if they serve different purposes. Nevertheless, it should be clear that certain policy exercises can be conducted only in models where expectations and general equilibrium features are fully taken into account and where the predictive content of pure time series models is close to nonexistent as the horizon of the forecast surpasses one year. Small-scale structural models that allow a large number of policy exercises and at the same time offer some indication of the potential developments one to two years ahead are probably the ones that will survive the dust of time in the longer run.

Acknowledgments

Conversations with L. Sala and C. Michelacci are gratefully acknowledged. Financial support from the CREI and the Spanish Ministry of Education through the grant SEJ-2004-21682-E is gratefully acknowledged.

References

An, S. and F. Schorfheide (2007) Bayesian analysis of DSGE models. *Econometric Reviews* **26**, 113–72.

Beyer, A. and R. Farmer (2004) On the indeterminacy of New Keynesian economics. ECB Working Paper 323.

Blanchard, O. and D. Quah (1989) The dynamic effect of aggregate demand and supply disturbances. *American Economic Review* **79**, 655–73.

Bernanke, B. (1986) Alternative explanations of the money income correlation. *Carnegie-Rochester Conference Series on Public Policy* **25**, 49–100.
Boivin, J. and M. Giannoni (2005) DSGE models in data rich environments. NBER Working Paper 12272.
Boivin, J. and M. Giannoni (2006) Has monetary policy become more effective? *Review of Economics and Statistics* **88**(3), 445–62.
Braun, P. and S. Mittnik (1993) Misspecification in VAR and their effects on impulse responses and variance decompositions. *Journal of Econometrics* **59**, 319–41.
Canova, F. (1995) Sensitivity analysis and model evaluation in simulated dynamic general equilibrium economies. *International Economic Review* **36**, 477–501.
Canova, F. and G. De Nicoló (2002) Money matters for business cycle fluctuations in the G7. *Journal of Monetary Economics* **49**, 1131–59.
Canova, F. (2002) Validating monetary DSGE models through VARs. CEPR Working Paper 3442.
Canova, F. (2005) Structural changes in the US economy. *Journal of the European Economic Association*. Forthcoming.
Canova, F. and L. Gambetti (2007) Do inflation expectations matter? The Great Moderation revisited, manuscript.
Canova, F. and M. Paustian (2007) Measurement with theory: using sign restriction to evaluate business cycle models, manuscript.
Canova, F. and J. Pina (2005) What VARs tell us about DSGE models. In C. Diebolt and C. Krystou (eds.), *New Trends in Macroeconomics*. New York: Springer Verlag.
Canova, F. and L. Sala (2005) Back to square one: identification issues in DSGE models. ECB Working Paper 583.
Chari, V.V., P. Kehoe and E. McGrattan (2006) A critique of SVAR using real business cycle theory. Federal Reserve Bank of Minneapolis, Working Paper 631.
Chari, V.V., P. Kehoe and E. McGrattan (2007) Business cycle accounting. *Econometrica*, **75**, 781–836.
Choi, I. and P.C. Phillips (1992) Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations. *Journal of Econometrics* **51**, 113–50.
Christiano, L., M. Eichenbaum and C. Evans (2005) Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* **113**, 1–45.
Cogley, T. and T.J. Sargent (2005) Drifts and volatilities: monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics* **8**, 262–302.
Cooley, T. and S. LeRoy (1985) Atheoretical macroeconomics: a critique. *Journal of Monetary Economics* **16**, 283–308.
Dedola, L. and S. Neri (2007) What does a technology shock do? A VAR analysis with model-based sign restrictions. *Journal of Monetary Economics* 54, 512–49.
Del Negro, M. and F. Schorfheide (2004) Priors from general equilibrium models for VARs. *International Economic Review* **95**, 643–73.
Del Negro, M. and F. Schorfheide (2005) Policy predictions if the model does not fit. *Journal of the European Economic Association* **3**, 434–43.
Del Negro, M. and F. Schorfheide (2006) Monetary policy analysis with potentially misspecified models. NBER Working Paper 13099.
Del Negro, M., F. Schorfheide, F. Smets and R. Wouters (2006) On the fit of New Keynesian models, *Journal of Business and Economic Statistics* **25**, 143–62.
Faust, J. and E. Leeper (1997) Do long run restrictions really identify anything? *Journal of Business and Economic Statistics* **15**, 345–53.
Fernandez-Villaverde, J. and J.F. Rubio-Ramirez (2007) How structural are structural parameters? *NBER Macroeconomic Annual* **22**, 83–137.
Fernandez-Villaverde, J., J. Rubio-Ramirez, T. Sargent and M. Watson (2007) The ABC (and D) of structural VARs. *American Economic Review* **97**, 1021–6.

Gali, J. and P. Rabanal (2005) Technology shocks and aggregate fluctuations: how well does the RBC model fit postwar US data? *NBER Macroeconomic Annual*, **20**.
Ireland, P. (2004) Technology shocks in the New Keynesian model. *Review of Economics and Statistics* **86**, 923–36.
Justiniano, A. and G. Primiceri (2008) Time varying volatilities of macroeconomics fluctuations. *American Economic Review* **98**(3), 604–41.
Liu, T. (1960) Underidentification, structural estimation, and forecasting. *Econometrica* **28**, 855–65.
Lubik, T. and F. Schorfheide (2004) Testing for indeterminacy: an application to US monetary policy. *American Economic Review* **94**, 190–217.
Ma, A. (2002) GMM estimation of the new Phillips curve. *Economic Letters* **76**, 411–17.
Nason, J. and G. Smith (2005) Identifying the New Keynesian Phillips curve. Federal Reserve Bank of Atlanta, Working Paper 2005–1.
Pappa, P. (2005) RBC or New-Keynesian transmission: the labor market effects of government expenditure shocks. *International Economic Review.* Forthcoming.
Primiceri, G. (2005), Time varying structural VAR and monetary policy. *Review of Economic Studies* **72**, 453–72.
Rabanal, P. and J. Rubio-Ramirez (2005) Comparing New-Keynesian models of the business cycle: a Bayesian approach. *Journal of Monetary Economics* **52**, 1150–62.
Ravenna, F. (2006) VAR and reduced form representation of DSGE models. *Journal of Monetary Economics* **54**, 2048–64.
Rosen, A. (2006) Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. UCL Working Paper.
Rotemberg, J. and M. Woodford (1997) An optimization based econometric framework for the evaluation of monetary policy. *NBER Macroeconomic Annual* **12**, 297–346.
Schorfheide, F. (2000) Loss function based evaluation of DSGE models. *Journal of Applied Econometrics* **15**, 645–70.
Sims, C. (1980) Macroeconomic and reality. *Econometrica* **48**, 1–48.
Sims, C. (1986) Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review*, Winter, 1–16.
Smets, F. and R. Wouters (2003) An estimated dynamic stochastic general equilibrium models of the Euro area. *Journal of the European Economic Association* **I**, 1123–75.
Stock, J. and J. Wright (2000) GMM with weak identification. *Econometrica* **68**, 1055–96.
Stock, J., J. Wright and M. Yogo (2002) A survey of weak instruments and weak identification in generalized methods of moments. *Journal of Business and Economics Statistics* **20**, 518–29.
Uhlig, H. (2005) What are the effects of monetary policy? Results from an agnostic identification procedure. *Journal of Monetary Economics* **52**, 381–419.

3

Introductory Remarks on Metastatistics for the Practically Minded Non-Bayesian Regression Runner

John DiNardo

Abstract

It would appear that much debate among practically minded researchers in economics, social science, and in other fields, is rooted in (frequently) unstated assumptions about the underlying philosophical justification for the statistical procedures being debated. In this chapter, I try to provide a simple non-technical introduction to some long-standing debates about "metastatistical" questions, especially those that divide (some) "Bayesians" from (some) non-Bayesians while attempting to draw out some implications for the "practically minded non-Bayesian regression runner." Some of the issues which have prompted the most raucous debate in philosophical circles include: the meaning of "probability," the importance or unimportance of pre-designation (pre-specified research design), the role of "models," and the practical value of hypothesis testing and other common statistical practices. I discuss some of the links between these philosophical views and actual practice and consider two different case studies – one from medicine and another from labor economics.

3.1 Introduction

> "Everything has already been said, but perhaps not *by* everyone and *to* everyone."[1]

The purpose of the somewhat silly title of this chapter is to warn the reader what not to expect. This is not intended as a "proper" introduction to metastatistics, which I could not write, of which there are several very good ones.[2] Given the enormous amount of writing on the subject, it is not surprising then that none of the ideas or arguments will be original.[3]

An even sillier title that some might use to describe the following is: "A jaundiced appraisal of some extreme Bayesian views by someone who just doesn't get it."

That is, of course, not my intent. Rather, I think that it is sometimes useful for the practically minded non-Bayesian regression runner (like myself) to consider some of the basic "philosophical" issues at the heart of statistics and econometrics. My purpose is also to bring some of the issues debated in metastatistics or the "philosophy of induction" literature "back to earth" from the somewhat airy realms in which they often dwell and toward the more messy realms of the low sciences, addressing them to an audience, like myself, who aren't philosophers but don't think that philosophy is necessarily a synonym for "useless." It is true that the discussion is often very mathematical, sometimes filled with obscure polysyllabic words of Greek or Latin origin, and pages and pages of definitions where the reader is expected to suspend disbelief before something of practical import seems to enter the discussion. Partly as a consequence, I will talk about ideas that a philosopher would define with much more precision: if you have philosophical inclinations, consider yourself forewarned.

That discussion about metastatistics often dwells in the "airy realms" is unfortunate. First, many of the issues discussed in this literature are of practical import. In economics, Bayesian approaches are becoming increasingly popular: in the United States, for example, the Food and Drug Administration (FDA) issued a call for comments on a proposal about increased use of Bayesian methods (Food and Drug Administration, US Department of Health and Human Services, 2006). Second, it seems to me that much of the debate among practically minded researchers is rooted in (frequently) unstated assumptions about the underlying philosophical justification for statistical procedures being debated. Consider the following statement of the advantages of adopting a Bayesian approach to FDA testing:

1. If we turn to Bayesian methods, difficult issues will be discussed in the right way by the right people.
2. Some of the dilemmas that FDA decision makers face are artifacts of the (non-Bayesian) statistical methods they use, and not due to demands of the scientific method.
3. The Bayesian perspective provides the best way to think about evidence (Goodman, 2004).
4. (In contrast to the usual approach) the Bayesian approach is ideally suited to adapting to information that accrues during a trial, potentially allowing for smaller, more informative, trials and for patients to receive better treatment. Accumulating results can be assessed at any time, including continually, with the possibility of modifying the design of the trial: for example, by slowing (or stopping) or expanding accrual, imbalancing randomization to favour better-performing therapies, dropping or adding treatment arms, and changing the trial population to focus on patient subsets that are responding better to the experimental therapies (Berry, 2006).

Such arguments are becoming increasingly common in domains outside of medicine and are most easily understood by using some of the metastatistical background.

3.1.1 Life, death, and statistical philosophy: an example

The issue of whether to use Bayesian or non-Bayesian methods has sometimes quite literally involved life or death issues. The case of ECMO (extracorporeal membrane oxygenation) is a useful example for those skeptical of the potential importance of the debate.

ECMO was a therapy developed for use with infants with persistent pulmonary hypertension: an ECMO machine circulates blood through an artificial lung back into the bloodstream. The idea is described as providing adequate oxygen to the baby while allowing time for the lungs and heart to rest or heal. The mortality rate using conventional therapy was believed to be 40% (Ware, 1989), although there is debate about whether that number was reasonable.[4] A possibly important consideration is that the notion of providing additional oxygen for infants was not obviously "safe." See the *British Journal of Ophthalmology* (1974) and Silverman

(1980), for example, for a discussion of the case of "oxygen therapy" for infants which, far from being harmless, caused blindness.[5]

Concern about the ethics of a conventional randomized trial (RCT), where half the patients are randomized into treatment and half to control, led the surgeons who had developed the therapy to use a "randomized play-the-winner" statistical method to evaluate the treatment. The purpose of this convoluted randomization scheme was to evade

> the ethical problem aris[ing] from the fact that during a "successful" randomized clinical trial (i.e., one that demonstrates a significant advantage to one treatment) about half of the trial subjects will receive a treatment which, at the end of the trial, will be known to be inferior. The recipients of the inferior treatment are individuals whose own outcomes are, in some sense, sacrificed to the greater good of knowing, with far more certainty than before the trial, the value, lack of value, or actual harm of the treatments under investigation. (Paneth and Wallenstein, 1985)

The randomization procedure is too elaborate to be described fully, but this gloss should be sufficient.[6] The essence of their "modified randomized play-the-winner" method is that "the chance of randomly assigning an infant to one treatment or the other is influenced by the outcome of treatment of each patient in the study. If one treatment is more successful, more patients are randomly assigned to that treatment" (Bartlett *et al.,* 1985).

Call ECMO "Treatment A" and conventional treatment "Treatment B." Initially, a group of biostatisticians prepared a sequence of blinded random treatment assignments. When the outcome of a treatment was known, this information would be sent to the biostatisticians, who would then create another sequence of blinded random treatment assignments; the probability of being assigned to A or B, however, was now a function of the success or failure of the treatment. In their study,

1. The first infant – with even odds – was randomly assigned to ECMO and survived.
2. The second infant – again with even odds – was randomly assigned to conventional treatment and died.
3. The third infant – with better-than-even odds in favor of being placed in ECMO as a result of the first two experiences – was randomized to ECMO and survived.
4. With now even higher odds the next infant was randomized to ECMO and survived.

This continued until there was a (pre-specified) total of 12 events. The result of this unusual randomization was that only one child was randomized to the conventional treatment and the 11 others received the ECMO treatment.

The outcome of this experiment was that the 11 infants randomized to ECMO treatment survived; the one infant randomized to conventional treatment died.

The debate revolved around whether the evidence from that trial and the previous history of non-randomized studies was "sufficient" or whether any other studies involving randomization were necessary. The researchers were reluctant to conclude that the single trial and the previous studies using "historical controls" were enough. Ware (1989), among others, observed that the randomization wasn't satisfactory and that one couldn't rule out other explanations for the observed outcomes. For instance, the sole infant not randomized to treatment was, coincidentally, the most severely ill patient in the study. The implication was that, had this one patient been randomized to ECMO, it is quite likely the child still wouldn't have survived.

Berry (1989), an advocate of Bayesian methods, harshly condemned the decision to continue further study as unethical.[7]

Most of the debate focused on the structure of the randomization, and revolved around a very narrow "binary" question: "Did ECMO work?" or, possibly, "What was the probability that ECMO works?" Both sides focused on whether the answer was "yes" or "no." The debate did not include, for example, a heated discussion about the necessary prerequisites to be considered "eligible" for treatment. Even if the researchers had used a more conventional randomization scheme, the study would not have been able to provide a good answer to *that* much more difficult question.[8]

3.1.2 The metastatistics literature

It is likely that much of the philosophical discussion on "induction" or "metastatistics" is somewhat unfamiliar to regression runners – it was to me. Moreover, often the metastatistics debate seems to involve few participants of the practical sort. As a consequence, many of the case study examples debated by philosophers of induction or statistics are drawn from physics; I am sure this is true in large part because physics has had some success – it is easier to debate "how to get the answer right" in a science when a consensus exists that, at some point, someone got it right. Such cases are rare (non-existent?) for low sciences like medicine and economics. As part and parcel of this general tendency, the types of problems considered in the metastatistics literature often seem far removed from the types of problems confronted by economists of my stripe – the "practically minded regression runner."[9]

When (by accident) I began reading about the philosophy of statistics I was surprised to discover:

1. the vehemence of the debate, and
2. the almost near-unanimous consensus that almost everything someone like me – a "practically minded non-Bayesian regression runner" – understands about statistics is wrong or profoundly misguided at best.

Concerning (2), consider the "stupid" inferences people like myself are "supposed to draw" on account of not adopting a Bayesian point of view. One example is inspired by an example from Berger and Wolpert (1988). Consider computing the

standard error of a measurement that occurs in the following way:

Flip a fair coin.

- If heads, use measuring device A for which the measurement is distributed normally with variance one and expected value equal to the truth.
- If tails, use measuring device B which has zero measurement error.

What is the *right* standard error if B is chosen? Although I had not given the matter a lot of thought before, it seemed obvious to me, a non-Bayesian, that the answer would be zero. Thus it came as a surprise to learn that, on some accounts, a non-Bayesian is "supposed" to give an answer of $\frac{1+0}{2}$.[10] By way of contrast, the Bayesian is described as someone who "naturally" avoids this inference, being "allowed" to "condition" on whether the measurement was made with machine A or B.[11]

As to the vehemence of the debate, LeCam (1977), a thoughtful non-Bayesian, prefaced his (rare) published remarks on "metastatistics"[12] by observing:

> Discussions about foundations are typically accompanied by much unnecessary proselytism, name calling and personal animosities. Since they rarely contribute to the advancement of the debated discipline one may be strongly tempted to brush them aside in the direction of the appropriate philosophers. However, there is always a ghost of a chance that some new development might be spurred by the arguments. Also the possibly desirable side effects of the squabbles on the teaching and on the standing of the debated disciplines cannot be entirely ignored. This partly explains why the present author reluctantly agreed to add to the extensive literature on the subject.

It is also a literature which (until recently) seemed almost entirely dominated by "Bayesians" of various stripes – the iconoclastic Bayesian I.J. Good (Good, 1971) once enumerated 45,656 different varieties of Bayesianism. There are also "objective" Bayesians and radical subjectivists. We might also choose to distinguish between "full-dress Bayesians" (for whom estimation and testing is fully embedded in a decision-theoretic framework) as well as a "Bayesian approach in mufti" (Good and Gaskins, 1971). As I discuss below, this variety and depth results in part from the view that probability and statistics are tools that *can* and *should* be used in a much broader variety of situations than dreamed of by the usual non-Bayesian regression runner: "Probability is the very guide to life." The non-Bayesian rarely thinks of statistics as being an *all-purpose* way to *think* (see Surprising Idea 3 in section 3.2).

This is not to suggest that there is *no* non-Bayesian philosophy involving statistics. Most notably, Mayo (1996) has recently stepped in to present a broader view of the philosophical underpinnings of non-Bayesian statistics that I find helpful, especially her notion of "severe testing." And there is an older tradition as well: Peirce (1878a, 1878b) and Venn (1888) are notable examples. The latter still remains an exceptionally clear exposition of non-Bayesian ideas; the articles by Peirce in *Popular Science Monthly* are insightful as well, but probably a slightly more difficult read. Nonetheless, such examples are few and far between.

In what follows, when I describe something as "Bayesian" I do not mean to suggest any writer in particular holds all the views so attributed here. There is considerable heterogeneity: some view concepts like "the weight of evidence" as important, others do not. Some view expected utility as important, others do not. This is not intended to be a "primer" on Bayesian statistics. Neither is it intended to be a "critique" of Bayesian views. There are several very good ones, some dating as far back as Venn (1888) (although some of these arguments will appear in what follows). Indeed, I will admit that, given the types of questions *I* typically find interesting, I don't find Bayesian ideas particularly helpful (and sometimes harmful). On the other hand, I can imagine situations where others might find formal Bayesian reasoning helpful. Indeed, given the prominent role that "models" play in economics, I am frankly a bit surprised that Bayesian techniques are not more popular than they are.

My purpose is not to do Bayesian ideas justice (or injustice!) but, rather, to try to selectively choose some implications of various strands of Bayesianism and non-Bayesianism for actual statistical practice that highlight their differences so as to be clear to a non-Bayesian perspective.

After having surveyed the metastatistics literature, one feels it is almost impossible to use the English language to label or describe the practically minded non-Bayesian regression runner.[13] When not being dismissed as belaboring under fallacious reasoning (Howson, 1997), she has been variously described as a "frequentist" – someone who is congenial to the notion of probability being about "relative frequency," or an NP (Neyman–Pearson) statistician – even though, as Mayo and Spanos (2006) observe, there is a great deal of confusion about what this means. Indeed, in my experience, most regression runners are not entirely sure what it means to be a user of "NP theory" (which is not surprising given that it is not clear that either Neyman or Pearson practiced or believed NP statistical theory!) Most congenial is Mayo's (1996) term "error statistician" – someone engaged in "severe testing." On the other hand, as a firm adherent of LeCam's Basic Principle Zero – "Do not trust any principle" – I will settle on the term "non-Bayesian."[14]

My hope is that consideration of some of the underlying metastatistics will make it easier to detect some sources of methodological disagreement. Put differently, one focus of what follows is to consider a claim, from Mayo and Kruse (2002), that "principles of inference have consequences" for actual practice.

More on this subsequently, but to ground the discussion, let me list the types of research questions I would like to consider as the aims of the practically minded regression runner:

1. What is the "causal effect" of some new medical treatment?
2. What are the the iatrogenic effects of morphine use? Does the use of pain medicine cause more pain?
3. Does (US) unionization lead to business failures?
4. Do "unions raise wages?",

as well as the types of questions I am *not* going to consider:

1. What is a good estimate of next quarter's GDP?
2. Does this structural model of the US labor market provide representation adequate enough for the purposes of evaluating potential policies?
3. What are the causes and consequences of black culture?

In my experience, what type of questions one is interested in asking often suggests what type of statistics one finds useful. While both types of questions are routinely asked by economists, the types of problems entailed seem very different to me (even if they do not appear this way to some Bayesians). This is not to imply that the second set of questions are necessarily illegitimate: I wouldn't want to suggest that people stop trying to estimate next quarter's GDP!

Indeed, when and where probability and statistics are most "useful" is one subject which divides many Bayesian and non-Bayesians and one that we explore in section 3.3.2.

3.2 Six surprising ideas and one puzzle

It may seem hard to believe that one's views on the metaphysics of statistics have consequences. In this section I enumerate six "surprising ideas" that I think go to the heart of many differences between non-Bayesians and Bayesians. For my purposes, I will focus on suggestions for practice that are most frequently invoked by Bayesians or radical subjectivists that are at furthest remove from *my own* non-Bayesian views. Despite this, my goal isn't to criticize them. Indeed, if they strike *you* as sensible, perhaps you are a (closet) Bayesian!

3.2.1 Six surprising ideas

1. The absence or presence of data-mining strategies, specification mining, non-random sampling, or non-random assignment are (should be) irrelevant to the inference of a set of data. Put differently, what could have happened, but didn't, in an experiment should make no difference to the evidential import of the experiment:

   > considerations about samples that have *not* been observed, are simply not relevant to the problem of how we should reason from the one that has been observed. (Jaynes, 1976, p. 200)
   >
   > Unbiased estimates, minimum variance properties, sampling distributions, significance levels, power, all depend on something ... that is irrelevant in Bayesian inference – sample space. (Lindley, 1971, p. 426)

2. Pre-specified research design is a waste of time:

   > In general, suppose that you collect data of any kind whatsoever – not necessarily Bernoullian, nor identically distributed, nor independent of each other – stopping only when the data thus far collected satisfy some criterion of a

> sort that is sure to be satisfied sooner or later [such as the requirement that a "t-statistic" exceed some critical value], then the import of the sequence of n data actually observed will be exactly the same as it would be had you planned to take exactly n observations in the first place. (Edwards *et al.*, 1963, pp. 238–9).

3. The problem of "how to reason" has been solved:

> Determining which underlying truth is most likely on the basis of the data is a problem in inverse probability, or inductive inference, that was solved quantitatively more than 200 years ago by the Reverend Thomas Bayes. (Goodman, 1999)
>
> [They are mistaken,] those who have insinuated that the Doctrine of Chances . . . cannot have a place in any serious inquiry . . . [it can] shew what reason we have for believing that there in the constitution of things fixt laws according to which things happen, and that, therefore the frame of the world must be to the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity. It will be easy to show that the problem solved in this essay [by the Reverend Bayes] is more directly applicable to this purpose. (Bayes, 1958)

4. Usual (non-Bayesian) practice is very badly wrong:

> . . . almost every frequentist [non-Bayesian] technique has been shown to be flawed, the flaws arising because of the lack of a coherent underpinning that can only come through probability, not as frequency, but as belief. (Lindley, 2000)
>
> Why it is taking the statistics community so long to recognize the essentially fallacious nature of NP [Neyman–Pearson, or non-Bayesian] logic is difficult to say, but I am reasonably confident in predicting that it will not last much longer. Indeed, the tide already seems strongly on the turn. (Howson, 1997)
>
> I explore the historical and logical foundations of the dominant school of medical statistics, sometimes referred to as frequentist statistics, which might be described as error-based. I explicate the logical fallacy at the heart of this system. (Goodman, 1999)

5. Randomization rarely makes sense in those contexts where it is most often employed:

> Physicists do not conduct experiments as Fisher would have them do. For instance, a simple experiment to determine the acceleration due to gravity might, say, require a heavy object to be dropped close to the earth. The conditions would be controlled by ensuring that the air is still, that the space between the object and the ground is free of impediments, and so on for other factors that are thought to interfere with the rate at which the object

> descends. What no scientist would do is to divide the earth's surface into small plots and select some of these at random for the places to perform the experiments. Randomizers might take one of two attitudes to this behavior of scientists. They could either say it is irrational and ought to be changed or else claim that experiments in physics and chemistry are, in some crucial respect, unlike those in biology and psychology, neither of which would appear to be very promising lines of defence. (Urbach, 1985, p. 273)

6. Probability does not exist:

> The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, ... or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception .... (deFinetti, 1974, p. 3)

3.2.2 An introductory puzzle

One of the most unusual aspects of metastatistics is that people on different sides of the debate cite *the same example* to make the case that the other side is wrong.

Consider the following example. Mayo (1979) and Mayo and Kruse (2002) have cited it as an example of a flaw in the usefulness of Bayesian reasoning while Bayesians routinely cite such examples (see Poirier, 1995) to argue that this is evidence of a flaw in non-Bayesian reasoning! It consists of a comparison of what inferences are justified in two different "experiments."

In both cases, suppose you are interested in the fraction of black balls μ in a huge urn (we ignore the complications arising from issues of sampling with or without replacement) that is "well-mixed" and has only red and black balls. Denote the null hypothesis as $\mathcal{H}_0 : \mu = 0.5$ and the alternative as $\mathcal{H}_1 : \mu > 0.5$. Denote the random variable "number of black balls" by X and the sample size as n.

Experiment A	**Experiment B**
Method: Declare in advance that you are going to pick 12 balls randomly from the urn.	*Method:* Instead of predesignating or deciding *in advance of the experiment* that you are going to draw 12 observations, you decide that you are going to keep drawing balls from the urn until you get at least 3 red balls.
Result: 9 of the 12 balls are black. The usual estimate is $\mu = \frac{3}{4}$.	*Result:* You draw the third red ball on the 12th attempt. 9 of the 12 are black and the usual estimate is $\mu = \frac{3}{4}$.

In both experiments, 12 balls were drawn. In both experiments, 9 of the 12 were black. There are several different "loaded" questions one can ask when comparing

the two experiments:

1. Are the two "experiments" different?
2. Does the "evidential import" of the two experiments for your beliefs about the true value μ differ when presented with either experiment A or B?
3. Does your evaluation of the experiment depend on the "mental state" of the investigator?

If your instinct is that "the evidential import" of both "experiments" is the same, you may be Bayesian. To many Bayesians such an example is a demonstration of a logical flaw in non-Bayesian statistics: in both cases someone has drawn 9 black balls and 3 red balls. Why should I bother to consider which experiment was being performed? If the "mental state" of the experimenter is "locked up" in his/her head and, say, inaccessible by someone else analyzing the data, doesn't such a case represent a fundamental problem for the non-Bayesian? I will return to this problem below, but before I do it will be helpful to sketch out some generalizations about the differences between Bayesians and non-Bayesians regarding the *role* of statistics.

3.3 What is statistics good for?

First, the Bayesian is typically more ambitious about the goals of statistics: "According to the Bayesian view, scientific and indeed much of everyday *reasoning* is conducted in probabilistic terms" (Howson and Urbach, 1993, p. 17).

John Maynard Keynes, for example, an exponent of "logical probability," deployed statistics to a very diverse range of subjects, including teleological questions – whether perceived order could be used to provide evidence of the existence of God. He concluded that, although such questions were well suited to study by Bayes' law, the problem was that such evidence could only make the existence of God more credible if it were supported by *other* evidence for God's existence (Keynes, 1921, p. 267).

To understand this point of view it is helpful to think of probability and statistics, for the Bayesian, as tools to bridge the gap between deductive and inductive logic.[15]

Deductive logic is about the validity of "risk-free" arguments.

All men are mortal.
John is a man.
Conclusion: John is mortal. (*)

Such an argument is deductively valid since *if* the premises are true, then so is the conclusion. A sound argument is a valid argument that has true premises. There are many types of risky arguments. Consider the following example. Imagine you are given the option of randomly selecting an orange from a box known to contain mostly good oranges and a few bad oranges.

Most of the oranges in the box are good.
Conclusion: The orange I randomly select will be good. (**)

This argument (**) is risky. Even if the premise is true, the conclusion may be wrong; you may be unlucky and draw one of the few bad oranges.

While probability seems of little value for non-risky arguments such as (*), even a non-Bayesian can easily see how probability might be *helpful* for arguments such as (**). For example, if we know 90% of the oranges in the box are good, the conclusion "There is a 90% chance that the orange I select will be good" seems less risky than the conclusion "There is a 90% chance that the orange will be bad." Probability and statistics for the Bayesian can be viewed as a way to tame risky arguments and make them amenable to the types of reasoning more commonly found in situations requiring merely deductive logic.

As I discuss in section 3.4.2, a Bayesian is typically more comfortable thinking about the probability of most *propositions* – which can be true, false, or uncertain – than a non-Bayesian. The non-Bayesian is most comfortable thinking about probability as the relative frequency of *events*. In the above example, neither the Bayesian nor the non-Bayesian is that uncomfortable talking about the event of a randomly chosen orange being good or bad. On the other hand, a non-Bayesian is more likely to feel unclear about a statement like "There is a 90% chance that an asteroid shower is the source of the Chicxulub impactor that produced the Cretaceous/Tertiary (K/T) mass extinction of the dinosaurs 65 million years ago."[16] The *proposition* that "The mass extinction of the dinosaurs was caused by a piece of an asteroid" is either true or false.[17] It is not a statement about relative frequency, or the fraction of times that the proposition is true in different "worlds."

The divergence between the two points of view becomes clearest when we begin discussing propositions much more generally. If probability is understood as being useful in induction – one version of the argument goes – it is a small step from this example to considering probability as useful *whenever* one is faced with making a risky decision. By these sorts of notions, most decisions in *life* become subject to the probability calculus because most *propositions* that are risky can and should be reasoned about using probability.

Indeed, once you've moved from reasoning about *beliefs* to reasoning about *decisions*, notions of "utility" can often become important. Many (including some Bayesians) have difficulty with this step: the relationship between "beliefs" and "actions" is not always obvious. I, for example, tend to think of them as rather distinct.[18] I think of Voltaire's quip – "I am very fond of truth, but not at all of martyrdom" – as a (perhaps extreme) example of the possible divergence between beliefs and actions. Hacking (1965, p. 16) observes that:

> beliefs do not have consequences in the same way in which actions do ... [For example] we say that a man did something in consequence of his having certain beliefs, or because he believed he was alone. But I think there is pretty plainly a crucial difference between the way in which his opening the safe is a consequence of his believing he was unobserved, and the way in which the safe's

> opening is a consequence of his dialing the right numbers on the combination. It might be expressed thus: simply having the belief, and doing nothing further, has in general no consequences, while simply performing the action, and doing nothing further does have consequences.

While the connections between Bayesian probability and Bayesian decision theory are a matter of debate as well, the connections seem tighter.[19] More importantly, an example from "decision theory" will, I think, highlight an important difference between Bayesians and non-Bayesians.

A useful case study comes from L.J. Savage, an important figure in the development of Bayesian ideas, who argued that the role of a mathematical theory of probability "is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it" (Savage, 1972, p. 57). Indeed, the first seven chapters of Savage (1972) are an introduction to the "personalistic" tradition in probability and utility.

3.3.1 What's utility got to do with it?

To me, the idea of probability as primarily a tool for detecting inconsistencies sounds strange; nonetheless, it appears to be a view held by many. Savage himself provides an interesting example of "detecting an inconsistency" and then removing it. This case study was the result of a "French" complaint about crazy "American" ideas in economics. The Frenchman issuing the complaint, Allais (1953), wrote a hotly contested article arguing against the "American" School's view of a "rational man."[20]

Savage, like some Bayesians, argued that maximizing expected utility is good *normative* advice. Although the ideas will probably be familiar as the "Allais paradox," it may be a good idea to sketch the main idea. If we consider $x_1, x_2, \ldots x_k$ mutually exclusive acts that occur with probability $p_1, p_2, \ldots p_k$, respectively, where $\sum_{i=1}^{k} p_i = 1$, and we can define utility over these acts with a single utility function $U(x)$ with the "usual" properties (increasing in x, and so on), we can define expected utility as:

$$E[U] = \sum_{i=1}^{k} U(x_i)p_i.$$

If utility is, say, increasing in money, then a "rational" person "should" prefer the gamble that yields the highest expected utility. (Note we postpone a discussion of what probability is until the next section.)

One of the gambles Allais devised to demonstrate that maximization of Expected Utility (what Allais referred to as the "Principle of Bernoulli") wasn't necessarily a good idea went as follows:

Imagine 100 well-shuffled cards, numbered from 1 to 100, and consider the two following pairs of bets and determine which you prefer.

First gambling situation

[A.] You win \$500,000 if you draw a card numbered 1–11 (11% chance). If you draw a number from 12–100, you get the status quo (89% chance).
[B.] You win \$2,500,000 if you draw a card numbered 2–11 (10% chance.) Draw a number from 12–100 or 1 and you get the status quo (90% chance).

The second pair of gambles

[C.] You win \$500,000 for certain.
[D.] You win \$2,500,000 if you draw a card numbered 1–10 (10% chance), \$500,000 if you draw a card from 11–99 (89% chance), and the status quo if you draw the card numbered 100.

As Allais found (and has been found repeatedly in surveys posing such gambles), for most people $B \succcurlyeq A$ (B is preferred to A) and $C \succcurlyeq D$ and, as Savage reports, the same was true for him (Savage, 1972, pp. 101–4)!

As most economists will recognize, this is a "paradox" since, from $C \succcurlyeq D$:

$$U(500,000) \quad > \quad 0.1U(2,500,000) + 0.89U(500,000) + 0.01U(0),$$

and from $B \succcurlyeq A$:

$$0.1U(2,500,000) + 0.9U(0) \quad > \quad 0.11U(500,000) + 0.89U(0),$$

and it is obvious that both inequalities can't be true.[21] There are two ways to handle this "paradox."

1. One possibility (the one that appeals to me) is that – even after continued reflection – my original preferences are just fine. For me, the fact that at the stated sums of money, etc., the comparison is inconsistent with Expected Utility Theory is merely too bad for the theory, however plausible it sounds. Indeed, as is well-known, it is possible to axiomatize preferences so that Allais paradox behavior is consistent with "rational" behavior (Chew, 1983).
2. A second possibility is to conclude that something is "wrong" with your "preferences." That was Savage's conclusion; his solution was to "correct himself."

Indeed, as befits a Bayesian, Savage analyzed the situation by rewriting the problem in an equivalent, but different way:

		Ticket number		
		1	2–11	12–100
First pair	Gamble A	5	5	0
	Gamble B	0	25	0
Second pair	Gamble C	5	5	5
	Gamble D	0	25	5

After writing down the problem this way, he then observed that if he were to draw a number from 12 to 100 he would be indifferent between the outcomes, so he decided to "focus" on what would happen if he should draw between 1 and 11. By doing so, he decided that, in the case of a subsidiary problem – ignoring outcomes higher than 11 – the correct answer depended on whether he would "sell an outright gift of $500,000 for a 10 to 1 chance to win $2,500,000 – a conclusion that I think has a claim to universality, or objectivity." He then concluded that, while it was still true that $C \succcurlyeq D$, upon reflection $A \succcurlyeq B$, not the other way around.

As Savage himself noted: "There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be."

We put aside the frequently knotty subject of "prior beliefs" for the moment, and contrast this Bayesian view with a typical non-Bayesian view about "what statistics is good for."

3.3.2 What is statistics good for? A non-Bayesian view

> In its most restricted form [statistical] theory seems to be well adapted to the following type of problem. If two persons disagree about the validity, correctness or adequacy of certain statements about nature they may still be able to agree about conducting an experiment "to find out". For this purpose they will have to debate which experiment should be carried out and which rule should be applied to settle the debate. If one of them modifies his requirements after the experiment, if the experiment cannot be carried out, or if another experiment is used instead, or if something occurs that nobody had anticipated, the original contract becomes void. Since the classical theory is essentially mathematical and clearly not normative it is rather unconcerned about how one interprets the probability measures ... The easiest interpretation is probably that certain experiments such as tossing a coin, drawing a ball out of a bag, spinning a roulette wheel, etc., have in common a number of features which are fairly reasonably described by probability measures. To elaborate a theory or a model of a physical phenomenon in the form of probability measures is then simply to argue by analogy with the properties of the standard "random" experiments.
>
> The classical statistician will argue about whether a certain mechanism of tossing coins or dice is in fact adequately representable by an "experiment" in the technical stochastic sense and he will do that in much the same manner and with the same misgivings as a physicist asking whether a particular mechanical system is in fact isolated or not. (LeCam, 1977, p. 142)

A non-Bayesian doesn't view probability as a singular mechanism for deciding the probability that a proposition is true. Rather, it is a system that is helpful for studying "experimental" situations where it might be reasonable to assume that the experiment is well described by some chance set-up. Even when attempting to use "non-experimental" data, a non-Bayesian feels more comfortable when he/she has reason to believe that the non-experimental situation "resembles" a chance set-up. Indeed, from a strict Bayesian viewpoint it is hard to understand why, in the

low sciences, there is a great deal of interest in "natural experiments." Put another way, one wants to try to draw a contrast between "experience" and "experiment." In the case of the former, statistical tools may or may not be particularly helpful, and other methods for gaining insight might easily dominate. In the latter case, one generally feels more hopeful that statistical reasoning might help.

3.4 A few points of agreement, then ...

Statistics and probability, as we understand them today, got a surprisingly late start in the (European) history of ideas.[22] Before the seventeenth century a major use of the word "probability" in English was to describe a characteristic of an opinion and dealt with the authority of the person who issued the opinion. "Thus [it could be said] Livy had more of probability but Polybius had more of truth." Or, "Such a fact is probable but undoubtedly false," relying on the implicit reference of what is "probable" to authority or consensus (Barnouw, 1979).

A theme that will recur frequently is the notion that *everything* in metastatistics is a topic of debate. As I discuss in section 3.4.2, even the definition of probability is the subject of considerable debate. However, it will be helpful to have at least some terminology to work with before enjoining the metaphysics.

3.4.1 Kolmogorov's axioms

One place to begin is a review of a few of Kolmogorov's axioms which Bayesians and non-Bayesians (generally) accept, although they interpret the meaning of "probability" very differently. Though they can be defined with much more care and generality, we will define them crudely for the discrete case:

1. Given a sample space Ω of possible events $A_1, A_2, \ldots A_k$ such that:

$$\Omega \equiv \sum_{i=1}^{k} \bigcup A_i \quad \text{for} \quad i = 1, 2, \ \ldots \ k.$$

2. The probability of an event A_i is a number which lies between 0 and 1.

$$0 < P(A_i) < 1.$$

   An event which cannot happen has a probability of zero, and a certain outcome has a probability of 1.[23] Two events, A_1 and A_2, are mutually exclusive if $P(A_1 \bigcap A_2) = 0$.
3. For any two mutually exclusive events probability is additive:

$$P(A_1 \bigcup A_2) = P(A_1) + P(A_2).$$

The same is true for pairwise mutually exclusive events so, for example, we can write:

$$\begin{aligned} P(A_1 \bigcup A_2 \bigcup \ldots \bigcup A_k) &= \sum_{j=1}^{k} P(A_j) \\ &= 1. \end{aligned} \tag{3.1}$$

If we were intending a proper introduction to probability, even here complications arise. Is k finite, for example? To address that issue properly one would introduce a set of measure theoretic considerations, but there is no need to cavil about such issues at present.[24]

The most important observation to make is that these are, so far, simply *axioms*. At this level, they are mere statements of mathematics. Indeed, we don't even have to consider them to be "probabilities." They may or may not be readily associated with anything "real" in the world. As Feller (1950, p. 1) explains:

> Axiomatically, mathematics is concerned solely with relations among undefined things. *This property* is well illustrated by the game of chess. It is impossible to "define" chess otherwise than by stating a set of rules ... The essential thing is to know how the pieces move and act. It is meaningless to talk about the "definition" or the "true nature" of a pawn or a king. Similarly, geometry does not care what a point and a straight line "really are." They remain undefined notions, and the axioms of geometry specify the relations among them: two points determine a line, etc. These are rules, and there is nothing sacred about them. We change the axioms to study different forms of geometry, and the logical structure of the several non-Euclidean geometries is independent of their relation to reality. Physicists have studied the motion of bodies under laws of attraction different from Newton's, and such studies are meaningful if Newton's law of attraction is accepted as true in nature.[25]

In sum, these axioms don't commit you to believing anything in particular. One reason you might adopt such axioms (and the reason I do) is because they seem convenient and useful if you are interested in the properties of chance set-ups or things that resemble chance set-ups.

I belabor this obvious point because I think it useful to consider that we *could* begin with different axioms. A nice example comes from Hacking (2001). Consider representing the probability of a certain event, A, as $P(A) = \infty$, and if A were impossible, $P(A) = -\infty$. In such a system:

- If the event A and $\tilde{}A$ (the event "not A") have the same probability, then $P(A) = P(\tilde{}A) = 0$
- If the event A is more probable than $\tilde{}A$, then $P(A) > 0$
- If the event $\tilde{}A$ is more probable than A, then $P(A) < 0$.

This too could form the basis of a theory of probability, but it is one we choose not to adopt because it seems "inconvenient" to work with and makes it more difficult to study the behavior of chance set-ups.

3.4.2 Definitions of probability

DeFinetti's declaration in Surprising Idea 6 that "PROBABILITY DOES NOT EXIST" may, at minimum, appear to be a bit intemperate. Indeed, it presupposes that many practically minded non-Bayesian regression runners are in the grips of some bizarre hallucination. It will help to consider two broad classes of definitions of probability that are sometimes referred to as:

1. "aleatory" or frequency-type probabilities
2. "epistemic" or belief-type probabilities.

Aleatory probabilities are perhaps what is most familiar to the non-Bayesian. For many, the notion of any other type of probability may not have been seriously entertained. It is interesting to observe that criticism of aleatory probability began at the inception of modern statistics and, as Hacking (1975, p. 15) observes, "philosophers seem singularly unable to put asunder the aleatory and the epistemological side of probability. This suggests that we are in the grip of darker powers than are admitted into the positivist ontology."

I began my presentation with Kolomogorov's axioms since everyone seems to agree on something like these; disputants disagree on what they are useful for, or what, precisely, they are "about." I won't do a complete survey, but a few moments of reflection may be all that is required to consider how slippery a notion probability could be.[26]

3.4.3 Aleatory or frequency-type probabilities

When we say "the probability that a fair coin will land as heads is $\frac{1}{2}$" we could take it as a statement of fact, which is either true or not. When we do so, we are generally thinking about probability as describing something that results from a mechanism that tosses coins and the geometry of the coin, perhaps. This mechanism can be described as a "chance set-up." We might go on to describe the physics of the place containing our coin-toss mechanism. A mechanism that would be perfectly useful in Ann Arbor, Michigan, might not work somewhere in the deep reaches of outer-space.

Nonetheless, most non-Bayesians, it would seem, are content to harbor little doubt that, at some fundamental level – whether we know the truth or not – it is meaningful to talk about the probability of a tossed coin falling heads. When pressed to explain what they mean when they say that the probability is $\frac{1}{2}$ that a fair coin will turn up heads, such a person might say "In the long run, if I were to repeatedly toss the coin in the same way, the relative frequency of heads would be $\frac{1}{2}$." We've yet to worry about "the long run" but, even at this level, for example, we would like to exclude the following deterministic but infinite series as being a

prototype for what we have in mind:

$$H \quad T \quad H \quad T \quad H \quad T \quad \ldots H \quad T \ldots$$

In such an example, if we know the last coin toss was "H" we are certain that the next coin toss will be "T." An "intuitive" definition that excludes such a possibility was given by Venn (1888), who talked about a probability as a characteristic of a *series* as "one which exhibits individual irregularity along with aggregate regularity." If we denote the number of "Trials" by N and the number of times the event "Heads" occurs as $m(N)$, we might go on to define the probability of "Heads" as:

$$P(\text{Head}) = \lim_{N \to \infty} \frac{m(N)}{N}.$$

An apparent weakness of this definition is, of course, that infinity is rarely observed. The derisive term about such thought exercises is sometimes referred to as "asymptopia" – which suggests something both unrealistic and unattainable.[27]

3.4.4 Objective, subjective, or "it depends"

Whether such a concept corresponds to something "real" or "objective," or whether it is "in the mind," is a subject on which much has been written. Sometimes such probabilities have been described as "objective" in order to contrast them with Bayesian "probabilities." However, one Bayesian objection is that there is no such thing as an "objective probability" – any such probability depends on purely subjective beliefs:

> To calculate a frequency, it is necessary to consider a repetitive phenomenon of a standardized variety. The appropriate meaning of "repetitive" and "standardized" is not obvious. To calculate a relative frequency it is necessary to subjectively define (possibly only conceptually) a class of events (known as a *collective*) over which to count the frequency. The relative frequency of the event ["Heads"] should also tend to the same limit for all subsequences that can be picked out in advance.[28]
>
> To the extent that individuals agree on a class of events, they share an objective frequency. The objectivity, however, is in themselves, not in nature. (Poirier, 1995)

Poirier, a Bayesian, stresses the (implicit) "subjectivity" of the frequentist notion of probability, specifically the notion of a "collective." The non-Bayesian von Mises (1957, p. 12), for example, defines a collective as a "sequence of uniform events or processes which differs by certain observable attributes, say colours, numbers, or anything else. Only when such a collective is defined, then a probability can be defined. If it is impossible to conceive of such a collective, then it is impossible to talk about probability." For von Mises, the notion of collectives with infinite numbers of entities was an *abstraction* to make the mathematical representation of reality "tractable" (Gillies, 2000, p. 90). While an extensive discussion of a

"collective" is beyond our scope, it is important to acknowledge that there can be "legitimate" disagreements about whether certain probabilities can be said to "exist." Von Mises argues that the reason it is possible to talk about the probability of a tossed coin turning up "Heads" is because it is easy to think of the "collective"; it is not possible, he says, to consider "the probability of winning a battle . . . [which] has no place in our theory of probability because we cannot think of a collective to which it belongs."

I personally share von Mises discomfort with defining the "probability of winning a battle," but I imagine others do not. Whether or not it would be "meaningful" to do so, or whether it "has no place in our theory of probability," the ultimate criterion in the non-Bayesian context is: "Would doing so help in understanding?" The salient issue is not that different, in principle, from the qualms a physicist might feel about "whether a particular mechanical system is in fact isolated or not." Whether that is a "defect" of the theory of probability or whether it introduces an undisciplined element of "subjectivity" is a subject upon which there has been much philosophical debate.[29]

3.4.5 Epistemic probability

The difficulties that a non-Bayesian might feel about conceiving of the appropriate collective are largely avoided/evaded when we consider a different notion of probability – epistemic. A nice place to start is a description from Savage, often called a "radical subjectivist:"

> You may be asking, "If a probability is not a relative frequency or a hypothetical limiting relative frequency, what is it? If, when I evaluate the probability of getting heads when flipping a certain coin as .5, I do not mean that if the coin were flipped very often the relative frequency of heads to total flips would be arbitrarily close to .5, then what do I mean?" We think you mean something about yourself as well as about the coin. Would you not say, "Heads on the next flip has probability 0.5" if and only if you would as soon guess heads as not, even if there were some important reward for being right? If so, your sense of "probability" is ours; even if you would not, you begin to see from this example what we mean by "probability." (Savage, 1972)

What is also interesting is that instead of Kolomogorov's axioms reflecting a (possibly) arbitrary set of axioms about unknown concepts which (one hopes) resemble some real world situation, they can also be derived from "betting rules." Again quoting Savage:

> For you, now, the probability $P(A)$ of an event A is the price you would just be willing to pay in exchange for a dollar to be paid to you in case A is true. Thus, rain tomorrow has probability 1/3 for you if you would pay just \$0.33 now in exchange for \$1.00 payable to you in the event of rain tomorrow. (*ibid.*)

As when we encountered Expected Utility, viewing probability as a device that allows one to make sensible "bets" is not *necessary*. The important distinction between aleatory and epistemic probability is that epistemic probabilities are numbers which obey something like Kolmogorov's axioms but do not refer to anything "real" in the world, but to a (possibly) subjective "degree of belief." Here's one definition from Poirier (1995, p. 19):

> Let κ denote the body of knowledge, experience or information that *an individual* has accumulated about the situation of concern, and let A denote an uncertain event (not necessarily repetitive). Then the *probability* afforded by κ is the "degree of belief" in A held *by the individual* in the face of κ.

Given this definition of probability, stating that the probability that a fair coin lands heads is *not* stating some property of a chance set-up – rather, it is an expression of belief about what the coin will do.[30] It is important to point out that opinions about this subject vary amongst Bayesians. I.G. Good, for instance, maintains that "true" probabilities exist but that we can only learn about them by using subjective probabilities. DeFinetti, as we saw, believes that it is unhelpful to postulate the existence of "true" probabilities.

How does this differ from the aleatory- or frequency-type probability we discussed above? Again quoting from Poirier (1995, p. 19):

> According to the subjective ... interpretation, probability is a property of an individual's perception of reality, whereas according to the ... frequency interpretations, probability is a property of reality itself.

Among other things, in this view the probability that a fair coin-toss is heads differs across individuals.[31]

3.4.6 Conditional probability, Bayes' rule, theorem, law?

Is it Bayes' rule, law, or theorem? Is it one of the most powerful ideas of all time, or the source of much mischief? Dennis Lindley (as cited in Simon, 1997) observes that "[Bayes'] theorem must stand with Einstein's $E = mc^2$ as one of the great, simple truths." Putting aside the intractable issue of what the Reverend Bayes *meant*, this has been the subject of considerable controversy and study.[32]

For the typical non-Bayesian, R.A. Fisher and William Feller, for example, Bayes' rule is nothing but a manipulation of the law of conditional probability.

Everyone starts with a *definition* of conditional probability:

$$P(A_i|B) = \frac{P(A_i \bigcap B)}{P(B)} \text{ if } P(B) > 0. \tag{3.2}$$

Provided the necessary probabilities exist, we can do the same thing in reverse:

$$P(B|A_i) = \frac{P(A_i \bigcap B)}{P(A_i)} \text{ if } P(A_i) > 0. \tag{3.3}$$

Then there is the "conditional" version of the law of total probability: as before, let the A_j be mutually exclusive events, for $j = 1 \ldots k$ and $\sum_{j=1}^{k} P(A_j) = 1$ and, if $0 < P(B) < 1$:

$$P(B) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

What this says is that if $P(B)$ is the probability of some event, and it can be accompanied by some of the k mutually exclusive events A_j in some way, then the probability that $P(B)$ occurs is merely the sum of the different ways B can occur with A_j times the probability of $P(A_j)$.

Using equations (3.2), (3.3) and (3.1), rearranging, and applying this last operation to the denominator yields:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{3.4}$$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}. \tag{3.5}$$

So far, there seems nothing particularly remarkable. However, here the agreement ends. Consider a "Note on Bayes' rule" by the non-Bayesian Feller (1950, p. 125):

> In [the above formulas] we have calculated certain conditional probabilities directly from the definition. The beginner is advised always to do so and not to memorize the formula [Bayes' rule, equation (3.5)] ... Mathematically, [Bayes' rule] is a special way of writing [the definition of conditional probability] and nothing more. The formula is useful in many statistical applications of the type described in [the above] examples and we have used it there. Unfortunately, Bayes's rule has been somewhat discredited by metaphysical applications ... In routine practice this kind of argument can be dangerous. A quality control engineer is concerned with one particular machine and not with an infinite population of machines from which one was chosen at random. He has been advised to use Bayes's rule on the grounds that it is logically acceptable and corresponds to our way of thinking. Plato used this type of argument to prove the existence of Atlantis, and philosophers used it to prove the absurdity of Newton's mechanics. But for our engineer the argument overlooks the circumstance that he desires success and that he will do better by estimating and minimizing the sources of various types of errors in prediction and guessing.

Feller's suggestion that the engineer will do better by minimizing the various types of *errors* is one issue where, at least rhetorically, non-Bayesians differ from Bayesians. For Feller, the focus is on using statistics (or other methods) to put ideas to the test, rejecting those that fail and advancing provisionally with those that survive. Bayes rule is a formula about *revising* one's epistemic probabilities incrementally. This distinction will become apparent when we apply Bayes' rule to estimation.

3.4.7 Reasoning or estimating with Bayes' rule?

Not surprisingly, Bayes' rule is viewed differently by Bayesians: it is a multi- (or all-) purpose tool of reasoning. Consider first the version given by equation (3.4). To fix ideas, let us consider one example of "Bayesian inference." In the above notation, let A_i be a specific hypothesis about the world and let B refer to some "data" that has somehow come in to our possession. For example, A_i might be the hypothesis that a coin is fair and B is the fact that you observed a single toss of the coin and it landed "heads." *Your* job is to ascertain how you should revise your beliefs in light of the data.

1. The "model" or likelihood for the behavior of N tosses of a coin is given by the following likelihood:

$$\mathcal{L}(\theta|N,h) = \left(\begin{array}{c} N \\ h \end{array} \right) \theta^h (1-\theta)^{N-h}. \tag{3.6}$$

   As described by Poirier (1995), $\mathcal{L}$ is a "window" by which to view the world – perhaps an "approximation" to the truth. We might debate what window is appropriate but, in the usual context, it isn't something to be "tested" or "evaluated." Moreover, the likelihood is a function which tells us "how likely we were to have observed the data we did (N, h)" given the truth of the model and a specific value of θ. (NB: here the likelihood is a device that tells you, given the parameter θ, what is the probability of observing the occurrence of h heads in N tosses of a coin.)

   Instead of using the coin toss mechanism to help you randomize, you are going to study the coin (and the mechanism) and learn about it.
2. The next step is to specify a prior distribution – one particularly *convenient* choice is the beta distribution. Priors are subtle things, but let us consider our beliefs about the value of θ to be describable by the following two parameter distribution:

$$\begin{aligned} f(\theta;\alpha,\delta) &= \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)} \theta^{\alpha-1}(1-\theta)^{\delta-1} \\ &= \frac{1}{B(\alpha,\delta)} \theta^{\alpha-1}(1-\theta)^{\delta-1}, \end{aligned} \tag{3.7}$$

   where $\Gamma(\cdot)$ is the gamma function and $B(\cdot)$ is the beta function. This is a very flexible distribution which can put weight on all values between 0 and 1. Figure 3.1 displays some of the wide variety of shapes the prior distribution can take for different values of α and δ.

   Different values of α and δ correspond to different beliefs. One way to get some intuition about what type of beliefs the parameters correspond to is to observe, for example, that the mode of the prior distribution (when it exists) occurs at:

$$\frac{\alpha-1}{\alpha+\delta-2}.$$

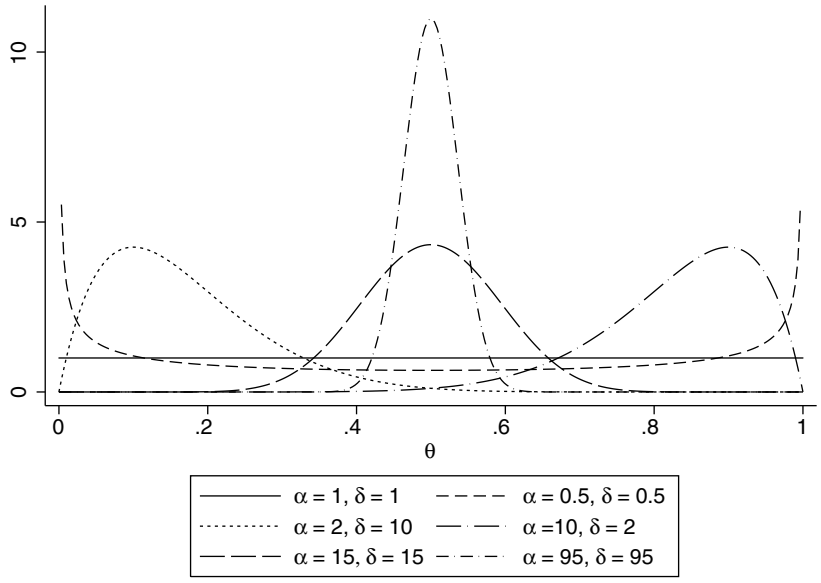


Figure 3.1 Different priors using the Beta distribution

It is sometimes helpful to think of $\alpha - 1$ as the number of heads "previously" observed, $\delta - 1$ the number of tails, and $\alpha + \delta - 2$ as the total number of coin flips previously observed from the experiment. On the other hand, it is not clear how someone could verify that a particular choice of prior was a good or bad description of one's beliefs.

3. In the third step, we merely plug our prior and our likelihood into Bayes' rule and what we come up with is[33]

$$\frac{1}{B((a+h),(\delta+(N-h)))}\theta^{\alpha+h-1}(1-\theta)^{\delta-1+N}. \tag{3.8}$$

Given the usual caveats, equation (3.8) is a statement of your personal beliefs about the value of θ, modified in light of the observed coin-toss. The beta distribution is a nice example because it is easier than usual to characterize the resulting "beliefs."

The left panel of Figure 3.2 shows two different prior distributions – one labeled "less informative" and the other "very informative." The first prior distribution corresponds to Beta(199,1) and the second to Beta(2,2). A convenient fiction to appreciate these prior beliefs is to imagine that, in the first case, you have previously observed 200 observations, 199 of which were heads. In the second case, you have previously observed four observations, two of which were heads. The first case corresponds to having "more prior information" than the second.

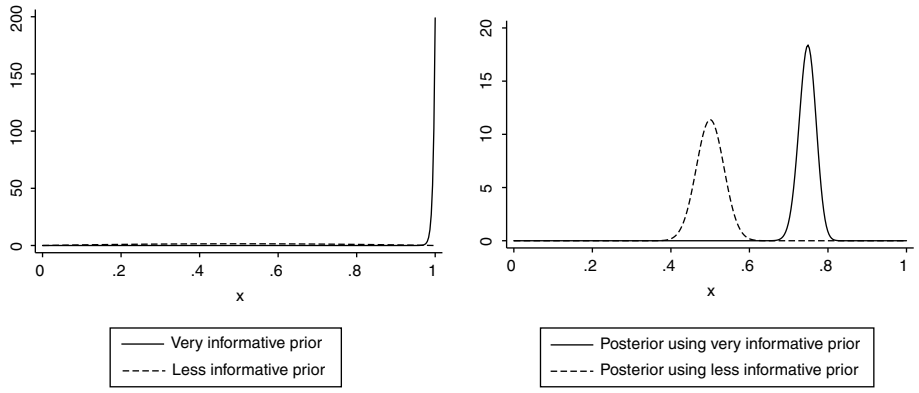


Figure 3.2 Different prior and different posterior distributions

The mode of the posterior distribution occurs at:

$$\frac{\alpha + h - 1}{\alpha + \delta + N - 2}.$$

This can be fruitfully compared to the usual non-Bayesian maximum likelihood (or method of moments) estimator, which is merely the sample mean:

$$\frac{h}{N}.$$

The difference between the posterior mode and the usual non-Bayesian estimator is that the former "adds" $\alpha - 1$ heads to the numerator and "adds" $\alpha + \delta - 2$ observations to the denominator.[34]

To see what effect this has, the right-hand panel of Figure 3.3 shows the resulting posterior distributions updated with 200 coin tosses, 100 of which are heads.

For a slightly different type of comparison one can consider two situations:

Prior	*Data*
Beta(99,99)	2 heads, 2 tails
Beta(2,2)	99 heads, 99 tails

In this case, although the experiments are very different, our conclusions are exactly the same (see Figure 3.3).

The role of the prior distribution and the sufficiency of the posterior distribution or likelihood are among the longest-standing debates in metastatistics. While a complete review is impossible, some of the most frequently enumerated difficulties are:

1. There is no way to verify whether the prior one has chosen adequately characterizes one's beliefs. Also, there is no unique way to translate ignorance or

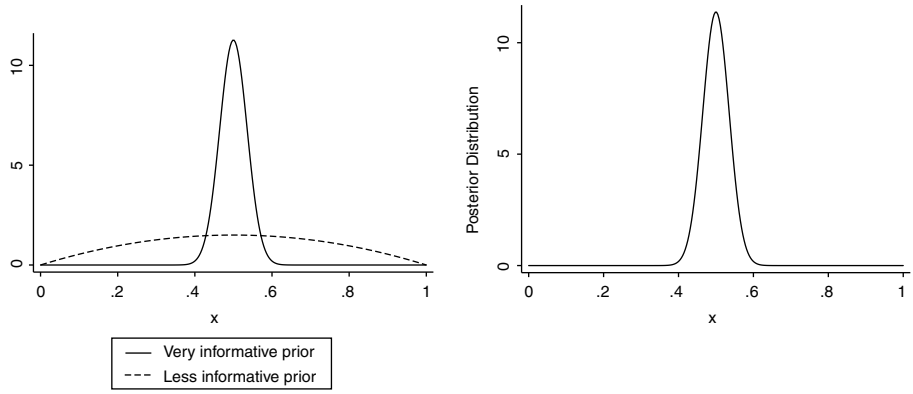


Figure 3.3 Different prior distributions, same posterior

"no information" into a prior distribution.[35] Consider the problem of estimating the length of a square garden which has sides of length between 1 and 5 feet. Based on this information, it seems "natural" to say that there is a 0.5 probability that the garden has sides of *length* between 1 and 3 feet. Equivalently, the information could be cast as saying that the area of the garden is between 1 and 25 square feet. In that case, it would appear just as natural to say that the probability is 0.5 that the *area* of the garden is between 1 and 13 square feet. This natural assignment of probability, however, implies that the probability is 0.5 that the length of the sides is between 1 and $\approx$ 3.61 feet ($\sqrt{13}$). However, it would be personally inconsistent to believe both claims and there is no principled method to reconcile the two different priors.

2. Even if a prior distribution is useful to the person holding it, it is not clear that it is useful to anyone else. LeCam (1977) observes that, for the binomial experiment, for arbitrary positive constant C, "if we follow the theory and communicate to another person a density $C\theta^{100}(1-\theta)^{100}$ this person has no way of knowing whether (1) an experiment with 200 trials has taken place or (2) no experiment took place and this is simply an *a priori* expression of opinion. Since some of us would argue that the case with 200 trials is more 'reliable' than the other, something is missing in the transmission of information."

3.5 The importance of the data-generation process

3.5.1 An idealized hypothesis test

Ultimately we would like to return to the "introductory puzzle," but before we do, let us introduce some context. The value of hypothesis testing has been frequently debated among non-Bayesians, but it may help to consider an idealized notion of how it is *supposed* to be done – this version is from Kmenta (2000) – when wishing

to make a statement about a *population* from a "random sample":

Preamble	State the maintained hypothesis [for example, the random variable X is normally distributed with σ^2 equal to ...].
Step 1	State the null hypothesis and the alternative hypothesis [for example, $\mathcal{H}_0 : \mu = \mu_0$ and $\mathcal{H}_A : \mu \neq \mu_0$].
Step 2	Select the test statistics [for example, $\overline{X}$ based on sample size $n = \ldots$].
Step 3	Determine the distribution of the test statistic under the null hypothesis [for example, $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ is distributed $N(0, 1)$ – normal, with mean zero and variance 1].
Step 4	Choose the level of significance and determine the acceptance and the rejection region [for example, *"do not reject* $\mathcal{H}_0$ *if* $-1.96 \leq \sqrt{n}\frac{(\overline{X}-\mu_0)}{\sigma} \leq 1.96$*; otherwise reject it"*].
Step 5	Draw a sample and evaluate the results [for example, *"the value of* $\overline{X}$ *is ... which lies inside (outside) the acceptance region"*].
Step 6	Reach a conclusion [for example, *"the sample does (does not) provide evidence against the null hypothesis"*]. To distinguish between 5% and 1% levels of significance we may add the word "strong" before "evidence" when using the 1% level.

It will be worth noting Kmenta's observations about the procedure: "According to the above scheme, the planning of the test and the decision strategy are set *before* the actual drawing of the sample observations, which does not occur until step 5. This prevents rejudging the verdict to suit the investigator's wishes."

This observation comes up frequently in non-Bayesian discourse, but less frequently among Bayesians: does the investigator want to ensure him/herself against "rejudging the verdict?" Perhaps they should "rejudge the verdict?" As we will see, this points to a notion of **severity** as being primary, as opposed to merely a concern about the correctness of the various statistical tests (although the two are not unrelated).

3.5.2 The introductory puzzle revisited

With this in mind, we can now reintroduce the puzzle. Specifically, the puzzle arises because, by using some variant of the above procedure, under one experiment observing 9 black of 12 balls allows one to *reject* the null hypothesis; in the other, observing 9 black of 12 balls would not permit the researcher to reject the null. Some Bayesians point to this example as evidence of a flaw in non-Bayesian reasoning: why should what is "locked up in the head" of the researcher – his/her intentions about what he/she was going to do – matter? In both cases, he/she has the same "data." This problem appears in many guises: in clinical trials there is a debate about what should be done if, for example, "early" evidence from a trial suggests that a drug is effective. The non-Bayesian response is that the Bayesian view misconstrues the purpose of error probabilities.

First, let's illustrate the problem. In experiment A, the question is: "How often would we expect to see 9 black balls out of 12 balls under the null hypothesis?":

$$\begin{aligned} P(\widehat{\mu} \geq \frac{3}{4}|\mathcal{H}_0) &\equiv P(X \geq 9|\mathcal{H}_0) \\ &= \sum_{x=9}^{12} \left(\begin{array}{c} 12 \\ x \end{array} \right) \mu^x (1-\mu)^{12-x} \\ &= \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \frac{1}{2}^9 (1-\frac{1}{2})^3 + \left(\begin{array}{c} 12 \\ 10 \end{array} \right) \frac{1}{2}^{10} (1-\frac{1}{2})^2 + \ldots \\ &= \frac{220+66+12+1}{2^{12}} \\ &= \frac{299}{2^{12}} \\ &= 0.073. \end{aligned}$$

In experiment B, the question is: "Under the null hypothesis, what is the probability of drawing 9 or more black balls before drawing a third red ball?" Let $r = 3$ be the pre-specified number of red balls to be drawn before the experiment is to be stopped. Let x index the number of black balls drawn, and let $n = x + r$.

This is a straightforward application of the negative binomial distribution where:

$$\begin{aligned} P(X \geq 9|\mathcal{H}_0) &= \sum_{x=9}^{\infty} \left(\begin{array}{c} r+x-1 \\ r-1 \end{array} \right) \mu^x (1-\mu)^r \\ &= \sum_{x=9}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) \mu^x (1-\mu)^r. \end{aligned}$$

It is very helpful to observe in doing the calculation that:

$$\sum_{x=j}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) (\frac{1}{2})^x = \frac{8+5j+j^2}{2^j}.$$

We can then write:

$$\begin{aligned} &= \sum_{x=9}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) \mu^x (1-\mu)^3 \\ &= \left(\frac{1}{2}\right)^3 \frac{8+5(9)+9^2}{2^9} \\ &= \frac{1}{8}\left(\frac{134}{512}\right) \\ &= 0.0327. \end{aligned}$$

There are several points to make about these "experiments" from a non-Bayesian perspective.

1. One point to emphasize is that in experiment A, the sample size is fixed. In experiment B, it was *possible* that the same experimenter would have continued to draw balls from the urn if a third red ball had not been drawn.
2. In neither case is it correct to make a statement such as "Given the experimental results (of 9 black and 3 red) there is a 7.3% probability in experiment A (3.3% probability in experiment B) that the null hypothesis is true." The hypothesis is presumably either true or false. The probability statements are statements about one particular "property" of a procedure. Whether it is a "good" procedure depends on a great deal more.
3. For many purposes, neither experiment is particularly "good." It depends on the alternative hypothesis that is the salient rival, but it is easy to come up with cases where Type I and II errors are going to be rather large. Figure 3.4 displays the sampling distribution of the two estimators. Neither experiment is going to be good, for example, at detecting the difference between a true mean of 0.5 and 0.51.

   Indeed, this was the non-Bayesian reaction to our earlier examination of ECMO: these experiments aren't likely to settle a well-meaning debate. Sometimes one is faced with a situation where one is trying to squeeze some inferential

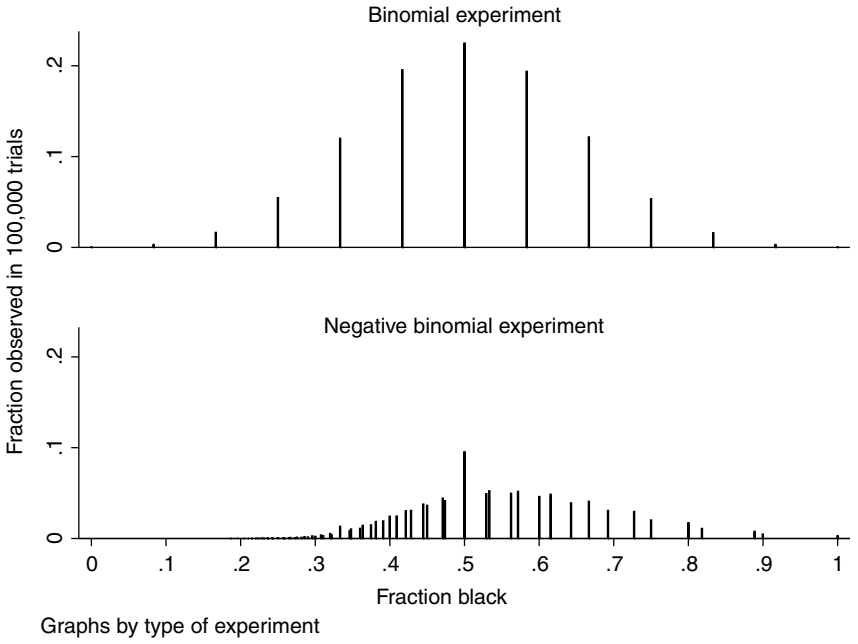


Figure 3.4 The introductory puzzle – which data-generation process?

blood from an experimental rock. In many such cases, we will not be able to put any proposition to a "severe test."

For the Bayesian the resolution of the problem is quite different – the data-generation process (DGP) doesn't (and shouldn't) matter. This if often referred to as "the likelihood principle." To see how this works, recall the statement of Bayes' rule in equation (3.5):

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

Consider two different likelihoods such that:

$$zP(B|A_i) = P^*(B|A_i) \quad \forall A_i, z > 0.$$

Now use Bayes' rule to show that one's inference is unaffected by use of $P^*(B|A_i)$ instead of $P(B|A_i)$:

$$\begin{aligned} P(A_i|B) &= \frac{P^*(B|A_i)P(A_i)}{\sum_{j=1}^{k} P^*(B|A_j)P(A_j)} \\ &= \frac{zP(B|A_i)P(A_i)}{\sum_{j=1}^{k} zP(B|A_j)P(A_j)} \\ &= \frac{zP(B|A_i)P(A_i)}{z\sum_{j=1}^{k} P(B|A_j)P(A_j)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}. \end{aligned}$$

Indeed, because of this property, Bayes' rule is often written as:

$$\underbrace{P(A_i|B)}_{\text{Posterior}} \propto \underbrace{P(B|A_i)}_{\text{Likelihood}} \underbrace{P(A_i)}_{\text{Prior}}. \tag{3.9}$$

Consequently, how the data was generated does not matter for the typical Bayesian analysis. It is also why a Bayesian would view the information in the binomial versus negative binomial experiment as being the "same."

This property of Bayesian inference has been frequently cited as being one of the most significant differences between Bayesians and non-Bayesians.[36]

3.5.3 If the DGP is irrelevant is the likelihood really everything?

A great deal more follows from the Bayesian approach. Unlike the previous example, which might discomfit some non-Bayesians, another implication seems a bit more problematic. One significant difficulty with the simplest versions of Bayesian analysis concerns the distinction between "theorizing after the fact" and "predesignation."

There exist many discussions of this problem. Our discussion follows Sober (2002), who poses a problem involving a deck of cards with 52 different types of cards. Suppose 5 cards are randomly drawn from a typical 52-card deck. Call the configuration of cards that results X. We are now going to use data on X to revise our beliefs about various theories of the world.

Two theories that can "explain" X include:

1. *Theory A*. The particular 5 cards were randomly drawn from a deck of 52 cards.
2. *Theory B*. A powerful demon intervened to ensure that the configuration A was drawn.

The essence of Bayesian analysis requires calculating the likelihood of observing X if theory A is true and calculating the likelihood of observing X if theory B is true. Your actual priors aren't particularly important, but assume that $P(A) > 0$ and $P(B) > 0$, although the probability you attach to them can be small.

The problem arises because the likelihood of the second (silly) theory is higher in the second (false) theory than in the first (true) theory. Since there are 2,598,960 different 5-card hands that can result:

$$P[X|A] = 1/2,598,960$$
$$P[X|B] = 1.$$

Regardless of your prior beliefs about A or B, whatever you believed before, equation (3.9) instructs you to increase the "weight" you give to the demon hypothesis! (Of course, your posterior density might assign little weight to B, but our interest is merely in the fact that the "experiment" induces you to give more weight than you did before to B.) If we continued drawing 5-card hands, and continued to elaborate our demon hypothesis after the fact, we could in principle move you even closer to believing that hypothesis!

If that example strikes you as fanciful, consider a more familiar example, usually called the "optional stopping" problem. To fix ideas, imagine being interested in whether some normally distributed variable (with a known variance of 1) has a mean of zero or otherwise.

1. Take a sample of size 100 and do the usual non-Bayesian hypothesis test in the manner suggested by Kmenta earlier. In this case compute $z = \frac{\sum_{i=1}^{N} X_i}{\sqrt{N}}$.
2. Continue sampling until $|z| \geq k_{0.05}$ or $N = 1000$, whichever comes first, where $k_{0.05}$ is the appropriate 5% critical value.

As the non-Bayesian knows, the first procedure provides a far more reliable indicator that the mean is zero than the second test. With the sampling size fixed in advance, if $|z|$ turns out to be greater than the appropriate critical value, the usual conclusion is that either the null is false or "something surprising happened." Under the second DGP, the probability of Type I error is 53% (see Mayo and Kruse, 2002).

What the results of the experiment would do to a non-Bayesian's *beliefs* is a separate matter, but it is clear that he/she would find a "rejection" much more informative in the first case.

By contrast, for the Bayesian who adheres to the Likelihood Principle, both experiments provide the **same** information: the posterior probability assigned to the null hypothesis should be the same, regardless of which experiment is performed. The Bayesian is free to ignore the "intentions" of the experimenter (that is, the DGP), presumably "locked up" in the mind of the experimenter. Confronted with the evidence consistent with the usual rejection of the null for the non-Bayesian, the change in the posterior beliefs of the Bayesian would be the same under both experiments.

An interesting debate on "optional stopping" can be found in the famous Savage forum (1962, p. 70 ff.), where precisely this example is discussed. For Armitage, a non-Bayesian, the DGP is very important and a flaw of Bayesian reasoning:

> I think it is quite clear that likelihood ratios, and therefore posterior probabilities, do not depend on a stopping rule. Professor Savage, Dr Cox and Mr Lindley [participants in the forum] take this necessarily as a point in favour of the use of Bayesian methods. My own feeling goes the other way. I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then "Thou shalt be misled if thou dost not know that." If so, prior probability methods seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling. (Savage *et al.,* 1962, p. 72)

G.A. Barnard, another forum participant – who originally proposed that the two experiments should be the same (Barnard, 1947a, 1947b) and introduced the notion to Savage – expressed the view that the appropriate mode of inference would depend on whether the problem was really a matter of choosing among a finite set of well-defined alternatives (in which case ignoring the DGP was appropriate) or whether the alternatives could not be so clearly spelled out (in which case ignoring the DGP was not appropriate.)[37]

3.5.4 What probabilities aren't – the non-Bayesian view

In a phrase, a Bayesian is more congenial to the notion that probabilities generated in the course of hypothesis testing represent the "personal probability that some claim is true or not," while such probabilities are merely devices that help "guide inductive behavior by assessing the usefulness of an experiment in revealing an 'error.'"[38] One problem sometimes cited by Bayesians is that non-Bayesians don't understand what "probability" means. To put it succinctly, a "*p*-value" is *not*:

- "The probability of the null hypothesis.
- The probability that you will make a Type I error if you reject the null hypothesis.
- The probability that the observed data occurred by chance."(Goodman, 2004)

The usual set-up begins with a "null hypothesis" and an "alternative hypothesis." Hypotheses can be simple or composite: an example of a simple hypothesis is "The population mean of a binomially distributed random variable is 0.5." That is, we can completely characterize the distribution of the random variable under the hypothesis. A "composite" hypothesis is a hypothesis that does not completely characterize the distribution of the random variable. An example of such a hypothesis is "The population mean of a binomially distributed variable is greater than 0.5." In addition to a set of "maintained hypotheses" ("the experimental apparatus is working correctly"), the next step is specifying a *test statistic.* In the usual hypothesis testing procedure, the distribution of this test statistic under the null hypothesis is known.

There are many ways to demonstrate that the probabilities that are used in hypothesis testing do not represent the probability that some hypothesis is true. A distinction that is sometimes made is "before trial" and "after trial" views of power and size. The following example comes from Hacking (1965).

Consider two hypotheses, a null ($\mathcal{H}_0$) and an alternative ($\mathcal{H}_1$), which are the only two possible states of the world. Let E_1, E_2, E_3, E_4 be the four possible outcomes and let the following be true about the world:

	$P(E_1)$	$P(E_2)$	$P(E_3)$	$P(E_4)$
$\mathcal{H}_0$:	0	0.01	0.01	0.98
$\mathcal{H}_1$:	0.01	0.01	0.97	0.01

We are interested in two tests, R and S, and specifically the power and size of the tests. Let the size of a test be the probability of incorrectly rejecting the null when it is true, and let the power of the test be 1 less the probability of Type II error (not rejecting $\mathcal{H}_0$ when it is false). For tests of a given size, more powerful tests are "better." The caveat about "a given size" is necessary since we can always minimize size by deciding on a rule that always rejects.

		Before trial	
		Size	*Power*
Test R	Reject $\mathcal{H}_0$ if and only if E_3 occurs	0.01	0.97
Test S	Reject $\mathcal{H}_1$ if and only if E_1 or E_2 occurs	0.01	0.02

If one takes a naive view of "power" and "size" of tests, the example is problematic. The size of both tests are the same, but test R is much more powerful – much less likely to fail to reject the null when it is false. *Before the trial*, we would surely pick test R.

What about *after* the trial? Consider the case when E_1 occurs. In that case test R instructs us to "accept" the null when *after* the trial we *know* with complete certainty that the null is false. The standard "evasion" of the problem for non–Bayesians is to observe (as Hacking, 1965, and Mayo, 1979, observe), that this is not a test that would usually be countenanced since there exist uniformly more powerful tests than R. This evasion, however, does not get to the heart of the problem.

3.5.5 What should "tests" do?

The previous discussion has attempted to be clear about why the "probabilities" of the usual hypothesis testing procedures should **not** be conflated with the "probability that the hypothesis is true."

What, then, is the "heart of the problem"? One argument, now associated with Mayo (1996), is that hypothesis tests should be used to put propositions to "severe" tests. The purpose of the probabilities for the non-Bayesian is to ascertain, as much as one can, how reliable specific procedures are at detecting errors in one's beliefs.

What is a severe test? In C.S. Peirce's words:

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely ...in order to subject them to the test of experiment. The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true. When the hypothesis has sustained a testing as severe as the present state of our knowledge ... renders imperative, it will be admitted provisionally ... subject of course to reconsideration. (Peirce, 1958, 7.182 and 7.231, as cited in Mayo, 1996)

Perhaps no better account can be given than Peirce's quotation. A nice quick gloss of a slightly more formal version of this idea is given in Mayo (2003):

Hypothesis H passes a severe test T with **x** *if*:

(i) **x** agrees or "fits" H (for a suitable notion of fit).
(ii) with very high probability, test T would have produced a result that fits H less well than **x**, if H were false or incorrect.

Mayo (1996) gives a nice example of why error probabilities *of themselves* are not enough, and why specification of an "appropriate" test statistic is a key ingredient. Mayo's example involves testing whether the probability of heads is 0.35 ($\mathcal{H}_0$) against the alternative that it is 0.10 ($\mathcal{H}_1$). It is an "artificial" example, but doesn't suffer the defect of the previous example – namely that the test is not the best in its class.

Suppose it is agreed that four coins will be tossed and that the most powerful test of size 0.1935 will be chosen. The following table shows the likelihood of observing various outcomes in advance of the experiment:

# Heads	*0*	*1*	*2*	*3*	*4*
$P(\mathcal{H}_0\|\cdot)$	0.1785	0.3845	0.3105	0.1115	0.0150
$P(\mathcal{H}_1\|\cdot)$	0.6561	0.2916	0.0486	0.0036	0.0001

Consider the following two tests:

Test 1	Reject $\mathcal{H}_0$	$\iff h = 0, 4$	Size = 0.1935	Power = $1 - 0.3438$
Test 2	Reject $\mathcal{H}_0$	$\iff h = 0$	Size = 0.1785	Power = $1 - 0.3439$

Given the set-up, the most powerful test of size 0.1935 is Test 1 – it is (slightly) more powerful than Test 2. But preferring Test 1 clearly doesn't make sense: if one sees all heads, it is surely more likely that $\mathcal{H}_0$ is true, yet Test 1 instructs you to reject. Mayo's solution is to observe that Test 1 fails to use an *appropriate* test statistic – one that measures how well the data "fits" the hypothesis. Even though one is searching for tests of size 0.1935 or better with the most "power," one chooses Test 1 at the cost of a nonsensical test statistic. The *usual* sort of test statistic might be the fraction of heads (F) less 0.35. Such a statistic has the property of punishing the hypothesis in a sensible way.

In this case, the test statistic takes on the following values:

# Heads	$F - 0.35$
0	-0.35
1	-0.10
2	0.15
3	0.40
4	0.65

In this account, Test 2 corresponds to the decision rule "Reject if $F - 0.35 < -0.1$" and the outcomes are now ordered by their departure from the null (in the direction of the alternative). The use of an appropriate sense of "fit" serves to show that the probabilities *per se* are not important – they don't directly correspond to a measure of belief. Rather, they are one step in assessing how good the test is at revealing an "error" (Mayo, 2003). The theory doesn't tell you in most non-trivial cases, however, how to generate a sensible test statistic – that depends on context.

While this example is admittedly superficial, it helps explain why, in constructing a good experiment, the importance of other (possibly not well defined) alternatives cannot be ignored. How *severe* a test is is always relative to some other possible alternatives. Suppose we collect data on unaided eyesight and the use of corrective glasses or contact lenses. If one proposed to "test" the theory that eye glass wearing **caused** unaided eyesight to get worse and found a "significant" rejection of the null of no correlation in favor of the alternative that the correlation was negative the "p-value" might be small but it would fail to be a **severe** test against the hypothesis that people with poor uncorrected vision are more likely to wear eye glasses.

3.5.6 Randomization and severity

One place where Bayesians and non-Bayesians differ is on the usefulness of randomization. Here, we can only introduce the problem.

Consider a case where the true state of the world can be characterized simply by the following:

$$y = \beta_0 + \beta_1 T + \beta_2 X + \epsilon, \tag{3.10}$$

where, for simplicity, the β are unknown parameters, the X are things that "cause" y and are observed, ϵ are things that "cause" y but are not observed, and $T = 1$(received treatment).

For the non-Bayesian, one of the benefits of randomization is that the X variables available are usually very inadequate. Also ϵ is some convolution of omitted variables and functional form misspecification: it is not generally plausible to make a statement like "ϵ follows the normal distribution," although statements like that are often found in the literature. Hence, though one could write down a "likelihood," it isn't necessary for the non-Bayesian.

A caricature might make this clear: it is **not** the case that "on the first day, God created y and made it a linear deterministic function of T and X; on the second day, in order to make work for econometricians, God appended a normally distributed error term with mean 0."

Indeed, in a randomized controlled trial (RCT), when the experimenter can intervene and assign T randomly, the "model" the experimenter estimates is often much less complicated:

$$y \quad = \quad \beta_0 + \beta_1 T + \epsilon. \tag{3.11}$$

For purposes of estimation one *could* write down a normal likelihood for this model:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y_i - \beta_0 - \beta_1 T_i}{2\sigma^2}\right). \tag{3.12}$$

With this likelihood, one could then specify prior beliefs about the fixed parameters β_0 and β_1, stipulate the form of heterokedasticity (that is, that the variance of ϵ was a constant for all observations, or model the heteroskedasticity), and so on. After seeing the data a Bayesian could update his/her beliefs about the values of these two parameters. Note that, in this formulation, there appears to be nothing special about the likelihood to distinguish it from any other comparison of means – nothing tells us, for example, that T was assigned randomly.

Nonetheless, writing down the likelihood seems a bit bizarre for the non-Bayesian. For example, if the treatment, T, was a nicotine patch and y was some outcome like "quit smoking successfully," no one thinks that only the patch matters and nothing else that can be observed matters – clearly the price of cigarettes, social norms, and so on, play a role. Indeed, available covariates are usually not used except to "test" the validity of the design. Specifically, in repeated samples:

$$E[\overline{y}_1] - E[\overline{y}_0] = \beta_1 \tag{3.13}$$

$$E[\overline{X}_1] - E[\overline{X}_0] = 0 \tag{3.14}$$

$$E[\epsilon_1] - E[\epsilon_0] = 0, \tag{3.15}$$

where $\bar{y}_1 \equiv \frac{1}{N_1}\sum_{i;T=1} y_i$, $\bar{y}_0 \equiv \frac{1}{N_0}\sum_{i;T=0} y_i$, the subscript 1 refers to the treatment group, the subscript 0 refers to the control group, and so on.

Taken literally, (3.14) suggests that, on average, in repeated samples, the mean for *any* pre-treatment variable should be the same. An auxiliary implication is that if one ran the regression but also included pre-treatment variables, the estimate of the effect of the treatment should not change. If it does change substantially, this is evidence against the design, and a cause for concern.

To fix ideas, suppose the treatment under consideration is ECMO (which we considered in section 3.1.1) and suppose a standard randomization scheme was employed on a large sample of children. A standard procedure is to report the averages for several variables. Table 3.1 is a hypothetical table.

Table 3.1 Hypothetical RCT on the efficacy of ECMO pre-treatment values of key variables (standard errors in parentheses)

Pre-treatment variable	*Treatment*	*Control*
Birth weight (grams)	3.26	3.21
	(0.22)	(0.23)
Age (days)	52	54
	(13)	(14)

Usually researchers report whether there are any "significant" differences between the treatment and control group means. The intended purpose is to ensure that the two groups satisfy a *ceteris paribus* condition: in ways we can observe, are the two groups roughly the same? – this is sometimes referred to as "balance." If sample sizes are large enough, more frequently than not the values in the two columns will not be "significantly" different. It serves as a "check" that the randomization achieved its intended purpose.

What variables should be included in this "check"? Presumably such a list does not include hair color, although this, in principle, should be balanced as well. The usual rule is to consider "pre-treatment variables which are predictive of the outcome." These may or may not be part of a proper "model" of infant death, but are there to assure one that, if there is a large difference in the groups after treatment, the researcher will not mistakenly attribute to the treatment what was really a failure of the *ceteris paribus* condition.

Bayesians frequently point to a flaw in this argument:

> My doubts were first crystallized in the summer of 1952 by Sir Ronald Fisher. "What would you do," I had asked, "if, drawing a Latin Square at random for an experiment, you happened to draw a Knut Vik square?" Sir Ronald said he thought we would draw again and that, ideally, a theory explicitly excluding regular squares should be developed ...

> The possibility of accidentally drawing a Knut Vik square or accidentally putting just the junior rabbits into the control group and the senior ones into the experimental group illustrates a flaw in the usual . . . argument that sees randomization as injecting "objective" or gambling-device probabilities into the problem of inference. (Savage *et al.,* 1962, p. 88)

Savage's example of having all the young rabbits in the control group and all the older rabbits in the treatment group is perhaps more recognizable than the distinction between Latin squares and Knut Vik squares, which come from classical agricultural experimentation. It is also related to the problem of random or pseudo-random number generation. If generating a sequence of random 0s and 1s, for instance, by chance (if only infrequently), some of these sequences will be undesirable – for example, if drawing a sequence of 1,000 numbers, it is possible that one draws all zeroes or all ones.

In the context of the RCT, the analogous problem, loosely based on our ECMO example, is described in Table 3.2. Assignment to the two groups was randomized but "bad luck" happened and the control group was comprised of the lightest birthweight babies (and viewed by the doctors as usually the least healthy) who were, on average, potentially in need of ECMO at much older ages (again, viewed by the doctors as an indicator of general frailty).

Table 3.2 Bad luck in a hypothetical RCT on the efficacy of ECMO pre-treatment values of key variables (standard errors in parentheses)

Pre-treatment variable	*Treatment*	*Control*
Birth weight (grams)	3.26	2.1
	(0.22)	(0.23)
Age (days)	52	140
	(13)	(14)

In this example, the problem is that the treatment and control groups are not "balanced." The treatment babies are (before treatment) healthier on average than the control babies. The typical non-Bayesian would generally find the numbers in Table 3.2 evidence against the validity of the design.[39]

For Bayesians, this suggests that the logic of randomization is flawed. If "balance" is the primary reason for randomization, why not deliberately divide into groups which look similar (and would "pass" a balancing test) without randomization *per se*. How does introducing uncertainty into treatment assignment help? Indeed, to some Bayesians, all it can do is lower the value of the experiment. From Berry and Kadane (1997):

> Suppose a decision maker has two decisions available, d_1 and d_2. These two decisions have current (perhaps posterior to certain data collection) expected

> utilities $U(d_1)$ and $U(d_2)$ respectively. Then a randomized decision, taking d_1 with probability λ and d_2 with probability $1 - \lambda$ would have expected utility $\lambda U(d_1) + (1 - \lambda)U(d_2)$. If the randomization is non-trivial, i.e. if $0 < \lambda < 1$, then randomization could be optimal only when $U(d_1) = U(d_2)$, and even then a non-randomized decision would be as good.

Savage goes further: "It has been puzzling to understand, why, if random choices can be advantageous in *setting up* an experiment, they cannot also be advantageous in its analysis."

There is much more to say: for instance, it may be useful to think about this class of problems in terms of the severity concept that we introduced earlier. However, it may be more instructive to consider two examples from real research. In one, I identify the problem as lack of severe testing. In the second, I identify the problem that the world is a "complicated place": assertions that were felt to be well-grounded by numerous studies seem less so in the face of a well-designed experiment.

3.6 Case study 1: "medication overuse headache"

In section 3.1.1 I briefly mentioned the case of ECMO – a treatment for infants with persistent pulmonary hypertension whose success was initially uncertain, but retrospectively seems of great benefit. Here I would like to consider a potentially "mirror-image" case: a treatment is being administered that, in my view, is potentially quite harmful. Also, I would argue, the literature is of unbelievably low quality. I locate the problem with the theory in the fact that, instead of behaving like Mayo's "error statistician" or engaging in "Peircean severe testing," the researchers began with a prior belief and then set about "updating" it. It should be noted that none of the studies involving this topic used "Bayesian statistics."[40] Rather, the question is "Is there enough evidence to proceed with the expert consensus or is more 'severe' testing necessary?"

This case is particularly useful because, as with many problems in medicine and social sciences (and elsewhere), it involves a problem of dubious ontology (is there "really" such a thing as medication overuse headache (MOH)?) as well as the problem of "new hypotheses" that "accommodate" the evidence instead of having a theory held in advance that "predicted" the evidence (much like our "demon" example in section 3.5.3).

A road map for what follows is:

1. During a period of time when the field was considered a "backwater" a diagnosis of MOH was developed. This theory argued that people with chronic severe headache pain caused their pain by taking pain medication "too frequently" and that, if they merely stopped taking the medication, their pain condition would improve.
2. The evidence for this theory was that patients who agreed to stop their pain medication had higher rates of improvement than those who didn't. These

studies typically ignore serious selection bias due to non-random attrition and regression to the mean.
3. In one of the few published critiques of the theory, it was noted that millions of users of analgesics, for reasons other than headache, do not develop migraine. In response, the theory evolved to state that only those individuals "predisposed" to migraine get MOH.
4. When a definition of the diagnosis that required improvement in pain after cessation of the offending medication was proposed it was strongly criticized. The definition was revised to empty it of potentially refutable content.

Every challenge to the theory that pain medication causes pain has been met by "accommodating" the evidence. Rather than reject the theory, at every turn the theory has accommodated the new evidence by making it more difficult to test. Furthermore, there is a complete absence of "severe testing."

3.6.1 What is medication overuse headache? Nosology and dubious ontology

The essence of "medication overuse headache" as a term for a certain class of chronic headache pain[42] is the idea that the patient *causes* his/her pain by taking pain (or other headache) medication in excess of arbitrary norms (set by researchers in the area) of appropriate use. The "offending" medication, as it is often referred to, can be any of a very diverse set with very different effects and mechanisms of action. These include ergotamine, caffeine, morphine, sumatriptan, and many other drugs. Opioids (morphine and related medications) are generally thought to be more of a problem than the other medications (Saper and Lake, 2006a).[42] Obermann and Katsarava (2007) cite a global prevalence rate of 1% and describe it as the "third most frequent headache type after tension-type headaches and migraine."[43]

> There are two ways to account for this phenomenon. The obvious one is that these people take chronic daily analgesics because they have chronic daily headaches. This is the explanation embraced by our patients and, until recently, by most physicians [who are not headache specialists]. (Edmeads, 1990)[44]

I believe this case study is illuminating because it suggests that the problem is *not* one of failing to view probability as epistemic, but is because researchers in the area have systematically *not* confronted their long-held views with severe testing.

3.6.2 Some salient background

3.6.2.1 Early history

In a recent review of the subject, Obermann and Katsarava (2007) date the first clear identification of MOH to a 1951 study, without a control group or randomization, which described 52 patients who took daily amounts of ergotamine and improved after "the ergotamine was withdrawn. A recommendation of the first withdrawal program followed and was introduced in 1963." The view that

medication overuse was a *cause* of migraine pain "became well-established" in the early 1980s (Capobianco *et al.,* 2001). It was first officially defined by the International Headache Society (1988) – the international association of neurologists with a specialty in headache – as "drug induced headache" in the International Classification of Headache Disorders (ICHD-1) (Obermann and Katsarava, 2007).

The view that medication use *caused* head pain was developed during a period of time when it was widely held that "migraine was a disorder of neurotic women" (Silberstein, 2004).[45]

3.6.2.2 The evidence

While a complete review of the evidence is not possible, let me take one representative example: Mathew *et al.* (1990).[46] Figure 3.5 is a modified version of a table from Mathew *et al.* The title of the table is also from the original article. Patients were assigned[47] to different treatment groups and their progress was observed.[48]

As Mathew *et al.* report, the data in the figure were based only on those patients who remained in the study – 90% in the group which continued to receive medication but only 50% in the group which had the medication withdrawn. With slight variations, Mathew *et al.*'s conclusions have become standard.[49] As far as I have

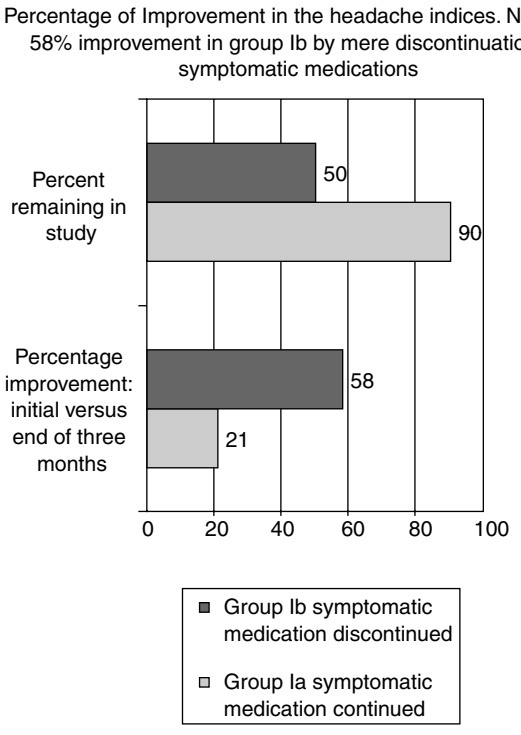


Figure 3.5 The evidence in a nutshell

been able to determine, the literature has not concerned itself with the problem of non-random attrition. One might expect that a patient who fails to improve after stopping the offending medication might be more likely to drop out than one who has improved (for some reason possibly unrelated to treatment). One possible reason for improvement might be mere "regression to the mean." For evidence that strongly suggests this is a problem see Whitney and Von Korff (1992).

3.6.2.3 First criticism

In one of the first (and extremely rare[50]) criticisms of research in this area, Fisher (1988) observed:

> As I understand it, analgesics beget more headache by making more brain serotonin available, which paradoxically increases the pain. My question is whether it holds only for headache or for other pains as well? In our arthritic clinic aspirin was used in doses of 8 to 14 tablets a day for about 15 years. I asked physicians who attended that clinic in those years whether they had ever noticed [the development of headache pain] . . . in any of the patients and they never had. Also 3 to 5 million people in the United States are taking aspirin daily to prevent arterial thrombosis. Should we expect headache . . . under these circumstances?[51]

As Fisher understood, the answer to his questions was "no" and advocated a randomized trial on the effect of withdrawal from headache medications where the control group would be subject to sham double-blind withdrawal. One can think of this as a proposal for a "severe test."

The response in the literature was to maintain that the theory was, in the main, correct and to merely amend the theory to accommodate the troubling fact highlighted by Fisher.[52] A typical amendment stated that "'analgesic abuse headache' may be restricted to those patients who are already headache sufferers [and that] individuals with . . . migraine, are predisposed to developing chronic daily headache in association with regular use of analgesic" (Bahra *et al.,* 2003).

3.6.3 Redefining MOH to avoid a severe test

The process of defining MOH provides a clear example of researchers avoiding a severe test. A useful place to start is the *initial* International Classification of Headache Disorders – 2 (ICHD-2).[53] The initial ICHD-2 definition of "medication overuse headache" is displayed in Table 3.3 (Headache Classification Committee of the International Headache Society, 2006).[54]

A key aspect of the definition is criterion C: the patient's *decision* to continue using analgesics at more than the approved rate.[55] This was immediately recognized to be a problem: existing standards of treatment for other forms of migraine, such as "menstrual migraine," *required* the use of analgesics at a rate which could then (inappropriately) be described as MOH.[56]

Another important aspect of the definition of MOH that was the subject of great dispute was item D – the requirement that, after removing the patient from the

Table 3.3 Initial 2004 International Headache Society classification criteria for analgesic-overuse headache

A. Headache present on ≥ 15 days/month with at least one of the following characteristics and fulfilling criteria C and D: 1 bilateral 2 pressing/tightening (non-pulsating) quality 3 mild or moderate intensity
B. Intake of simple analgesics on ≥ 15 days/month for > 3 months
C. Headache has developed or markedly worsened during analgesic overuse
D. Headache resolves or reverts to its previous pattern within 2 months after discontinuation of analgesics

Table 3.4 Revised ICHD-2 criteria for MOH

A. Headache present on ≥ 15 days per month
B. Regular overuse for > 3 months of one or more acute/symptomatic treatment drugs as defined under sub forms of 8.2 1. Ergotamine, triptans, opioids, or combination analgesic medication on ≥ 10 days/month on a regular basis for ≥ 3 months 2. Simple analgesics or any combination of ergotamine, triptans, analgesic, opioids on ≥ 15 days/month on a regular basis for ≥ 3 months without overuse of any single class alone
C. Headache has developed or markedly worsened during medication overuse

offending medication, the patient would improve. As reported in the literature, the problem was that such a requirement vitiated using MOH as a "diagnosis" in the traditional sense: "the problem is that medication overuse headache cannot be diagnosed until the overuse has been discontinued and the patient has been shown to improve. This means that when patients have it, it cannot be diagnosed. It can be diagnosed only after the patient does not have it any more."

After a meeting of experts in Copenhagen, this offending section (D of Table 3.3) – requiring improvement after going off the "causal" medications – was quickly removed to produce a revised version (Table 3.4). Whatever their intent, however, this redefinition seemed to make a MOH diagnosis impossible to refute.[57] Indeed, it was immediately noted that "the revision [to the definition of MOH] has eliminated the need to prove that the disorder is caused by drugs, that is, the headache improves after cessation of medication overuse" (Ferrari *et al.,* 2008). Although they suggested that "probable MOH" be introduced, their main focus was that sub-forms of MOH be defined for different types of medications, with opioids singled out as particularly problematic.[58]

The case of opioids is especially interesting since it is generally believed that opioid-related MOH is more worrisome and it has been argued that "sustained opioid therapy should rarely be administered to headache patients" (Saper and Lake, 2006b). This case is also useful since it might be falsely assumed that individuals doing research in this area (and supporting the idea of MOH) are incapable of, or not disposed to, putting hypotheses to severe testing. As noted previously,

researchers in this area routinely make no adjustment of any sort for the high rates of attrition in studies looking at chronic headache pain.

An illustrative exception to non-severe testing involves, not a test of MOH, but rather a study of the efficacy of sustained opioid therapy – opioids being considered a particularly pernicious cause of MOH (Saper and Lake, 2006a, 2006b). Although Saper *et al.* (2004) had no control group, the researchers treated individuals who dropped out for *any* reason, died for non-opioid related reasons, were suspected of "cheating" (using more opioids than allowed by the doctors), and so on, as *treatment failures*. This also included some patients who reported a substantial improvement but were considered to have "failed" to satisfy the *researchers'* definition of a significant reduction in functional impairment. In defining treatment failure more broadly, the researchers were essentially using a "worst case" bound.[59] While the use of "worst case bounds" is infrequent (or nonexistent) in the MOH literature, the argument for doing so has validity: it is entirely consistent with the notion of "severe testing." Indeed, leading researchers in MOH are aware of the potential value of such bounds. Saper and Lake (2006b), for example, harshly criticize a meta-analysis of RCTs on the efficacy of opioids for non-cancer pain for failure to adhere to "intent-to-treat" principles. In this instance, this meant treating as failures those individuals who began opioid treatment but then stopped for *any* reason.[60]

The severity of the tests to which opioid efficacy has been confronted is in sharp contrast to extant studies of MOH (sometimes by the same researchers), where a failure of a patient to reduce his medications is not treated as a failure of MOH therapy. Indeed, where attrition rates of 40% or higher are common, were the literature to treat those who were unwilling or unable to abstain from the offending medication as failures of "MOH therapy," it would appear that few, if any, of the studies in Zed *et al.* (1999) that purport to provide evidence favorable to the existence of MOH would continue to do so. Indeed, although plagued by non-random attrition and written by advocates of MOH, it has been observed that patients with MOH who "lapse" and re-establish medication overuse have higher measured "quality of life" on average than those with MOH who do not lapse (Pini *et al.*, 2001).

It might fairly be argued that an intelligent Bayesian might not have moved his/her posterior much in light of the foregoing discussion. Moreover, it is certainly the case that no formal Bayesian analysis has been employed. At least superficially, the "usual" statistical analysis was employed. What this literature *doesn't* do, however, is:

1. test the theory in such a way that the observed result would be unlikely if the obvious alternative (the one "favored by patients") was true – that it is chronic pain that causes use of pain relieving medication, not the other way around
2. employ the "usual" techniques to make tests more "severe" – the failure to use worst-case bounds, for instance, to deal with the problem of non-random attrition
3. react to each threat to the theory as potential reason to abandon the theory. Instead, the reaction of researchers has been continued modification of the theory until it is no longer capable of being refuted by evidence.

3.7 Case study 2: "union wage premium"

I would now like to consider an econometrically sophisticated literature – the literature on union wage effects. Two comprehensive and influential surveys are Lewis (1963, 1986). In these works, Lewis literally cites hundreds of studies attempting to estimate the causal effect of union status on wages.[61] See also the useful discussion in Heckman (1990).

3.7.1 Early history

The idea that labor unions might raise wages is one of the oldest debates in economics and was one of the earliest motivating examples for the famous "supply and demand" cross in a study by Jenkin (1870) (see Humphrey, 1992, for a short history). Ironically, although Jenkin (1868) concluded that the supply and demand analysis wasn't particularly relevant for explaining the consequences of union wage-setting, subsequent neoclassical theorizing in the main focused on the simple model depicted in Figure 3.6, where W is the real wage, L is the quantity of labor, D is the employer demand curve, S_c is the supply curve without unionization and S_u is the supply curve with unionization. Until Lewis' influential survey, opinions diverged between those who believed that unions could rarely control the supply of labor, such as Milton Friedman, and those that thought they could and therefore acted to create unemployment, such as Paul Samuelson (see Friedman, 1950).

3.7.2 A battery of severe tests

The analysis of union wage effects has become more sophisticated with the advent of large micro-datasets, but let me highlight some of the comparisons researchers

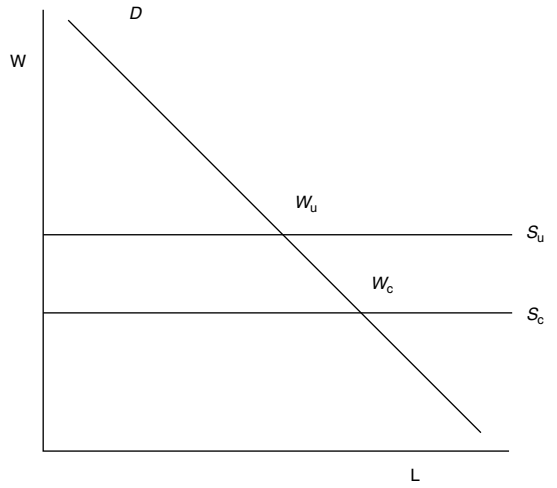


Figure 3.6 Union wage effects in the neoclassical model

have employed to analyze the question:

- Ashenfelter (1978) constructs control groups based on industry, race, and worker type (that is, craftsmen, operatives, laborers).
- Freeman (1984) compares wage rates for the same individual at *different points in time*. At one point in time the worker is in a unionized job; at a different point in time the worker is in a non-unionized job.
- Lemieux (1998) compares wage rates for the same individual who holds *two jobs*, one of which is unionized, the other which is not.
- Krashinsky (2004) compares wage rates of *identical twins*, one who is unionized and one who is not.
- Card (1992) constructs control groups, based on observable characteristics, which tend to receive the same wage in the non-union sector, as well as controlling for differences in permanent characteristics (that is, person-specific fixed effects).
- DiNardo and Lemieux (1997) and Card *et al.* (2003) compare US and Canadian workers, exploiting differential timing in the decline of unionization in the two countries.

Depending on the precise context, the union wage effect as measured in these studies ranges from 5% to 45%, with the vast majority of studies being at the higher end. All of the aforementioned studies adopt a distinctly non-Bayesian approach to the econometric analysis. The variety of research designs was not motivated by an attempt to "refine" posterior beliefs, but to put the hypothesis that unions raise wages to the most severe test possible with existing data. Each of the papers described above tried to "rule out" other explanations for the difference in union and non-union wages. (Perhaps this is why the posterior distribution of estimates of the union wage effect are as tight as they are – a survey of labor economists found remarkable unanimity on the average size of the effect (Fuchs *et al.,* 1998). The posterior mode of the economists surveyed was that unions raised wages 15% relative to similar non-union workers.)

What is also useful about this example is that there exists at least one Bayesian analysis, Chib and Hamilton (2002), which helpfully contrasts some unsophisticated non-Bayesian estimates from a small sample of workers. These non-Bayesian estimates vary from about 16% to 25%. If one treats these as "average treatment effects on the treated (ATOT)," these estimates are similar to their Bayesian posterior distributions.[62]

To put it a bit too simply, the basic empirical model has long been some variant of the following:

$$\begin{aligned} \log w_i &= X\beta_0 + \alpha_i + \epsilon_0 \quad \text{if } U_i = 0 \\ &= X\beta_1 + \psi\alpha_i + \epsilon_1 \quad \text{if } U_i = 1 \\ P(U_i = 1) &= F(Z\gamma), \end{aligned}$$

where U_i is the union status indicator for worker i, w_i is the wage of worker i, β and γ are parameters – the first two differ depending on whether the worker is unionized or not – X and Z are observed covariates, the ϵ are unobserved terms, and ψ is the ratio of the return to unobserved time invariant individual specific characteristics in the union to the non-union sector. The function $F(\cdot)$ is some type of cumulative density function and the elements of Z may overlap with X.

The Bayesian analysis of Chib and Hamilton (2002) takes a variant of the above model and is focused on how the effect of unions on wages varies across *individuals*. As has long been recognized, however, to a great extent unionization in the US occurs at the *establishment* level (this is in contrast to unionization in Europe, which frequently adheres to most workers in an industry). As Krashinsky (2004) observes, this has meant that the aforementioned empirical work has been unable to rule out the possibility of a "firm or enterprise specific fixed effect": a worker's union status could merely be a marker, for instance, for the profitability or generosity of the employer.[63] Put in other words, the list of "*ceteris paribus*" conditions considered in previous research did not include "working at the same firm."

As Freeman and Kleiner (1990) observe about the estimates of union wage effects with individual data, the "treatment effect" of most interest comes from an experiment on "firms" and not on "individuals" *per se*:

> While it is common to think of selectivity bias in estimating the union wage effect in terms of the difference between the union premium conditional on the observed union (and nonunion) sample and the differential that would result from random organization of a set of workers or establishments, we do not believe that this is the most useful way to express the problem. What is relevant is not what unionization would do to a randomly chosen establishment but rather what it would do to establishments with a reasonable chance of being unionized – to firms close to the margin of being organized rather than to the average nonunion establishment.

DiNardo and Lee (2004) use a regression discontinuity design, which, in their context, provides a very good approximation to an RCT of the sort discussed in the quote from Freeman and Kleiner (1990). Like previous work in this area, one of the motivating ideas was to put the hypothesis "do unions raise wages" to a more **severe** test, one that would allow for, among other things, a firm-specific effect.

This was possible in this research design since it used data on "firms." The research design essentially focused on comparing firms where the union "barely won" to those who "barely lost."

We can only be brief here, but the experiment is a "regression discontinuity design" based on an aspect of (US) labor law. Workers most often become unionized as the result of a highly regulated secret ballot. If more than 50% of the workers vote for the union, the workers win collective bargaining rights. If 50% or fewer do so, the workers do not win the right to collective bargaining. By comparing outcomes for employers at *firms* where unions barely won the election (for example, by one vote) with those where the unions barely lost, one comes close to the idealized RCT.

The test is severe against the hypothesis that unobserved differences in the *firms* that are unionized versus those that are not unionized can explain the different wages, and soon, of unionized workers.

It is rather easy to display the data from regression discontinuity designs. Figure 3.7 plots an idealized version of the key displays: in each, the average value of some outcome, where these averages are computed for different values of the vote share. The figure on the left corresponds to the case where unionization has an effect on the outcome in question. In the figure on the right (and one that resembles the figures in DiNardo and Lee, 2004), there is no detectable effect of unionization.

Figure 3.8 plots an idealized version of the key displays that correspond to ensuring the validity of the research design or "balance": in each the average value of some pre-treatment outcome (in the study by DiNardo and Lee this included firm size and measures of the health of the firm) are plotted for different values of the vote share. The graph on the left corresponds to the good case: firms in establishments where the union barely lost the election look the same as those where the union barely won. This corresponds to what was found in DiNardo and Lee. The figure on the right corresponds to a situation which is evidence against the design: firms in establishments where the union barely lost look much different than firms where the union barely won. In this case, the *ceteris paribus* conditions would seem to be violated.

To summarize the results of the study, the authors find (perhaps surprisingly given the huge literature documenting significant union wage effects) **no** effect of

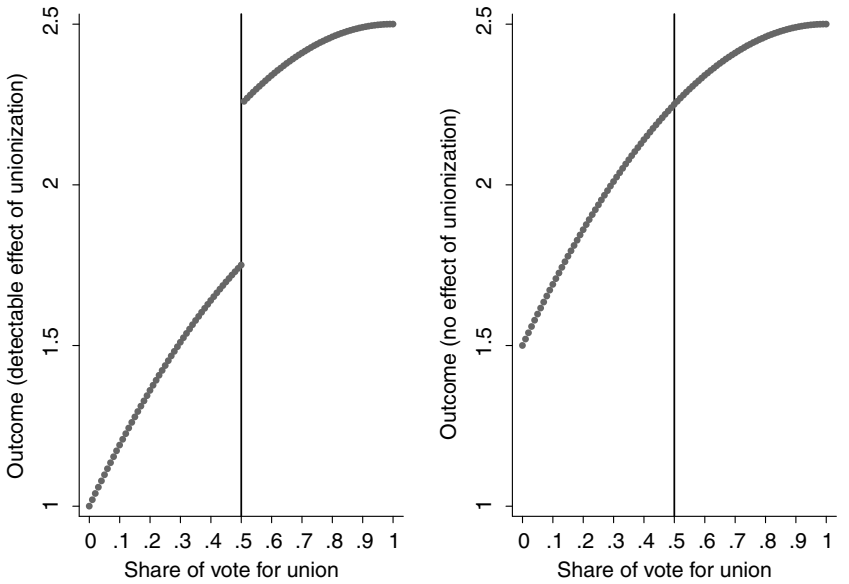


Figure 3.7 Two types of findings in a regression discontinuity design

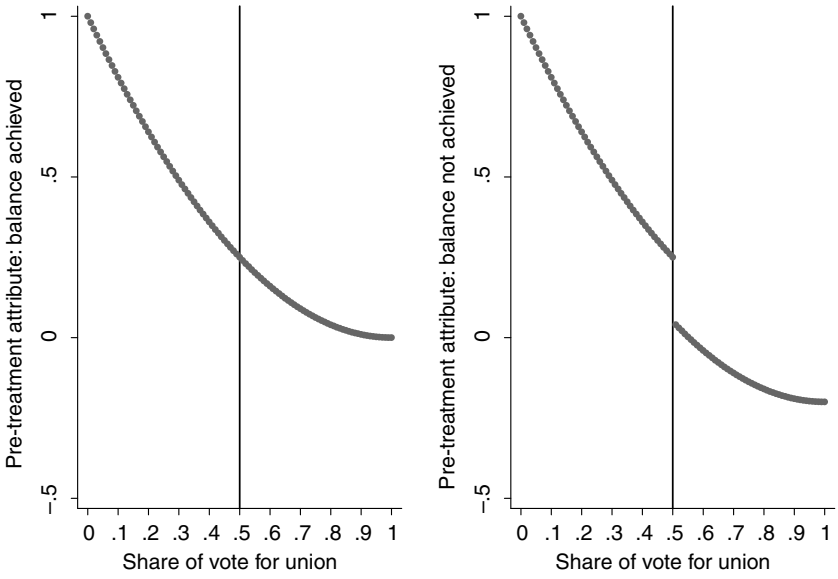


Figure 3.8 Evidence for or against "balance" in a regression discontinuity design

unionization on the myriad of outcomes they examine, such as wages, enterprise solvency, productivity, and so on. Limitations of scope prevent elaborating in more detail, but the study by DiNardo and Lee points to an important problem with any "non-severe" test of a hypothesis – Bayesian or non-Bayesian. If one takes the results from DiNardo and Lee seriously, it is hard to see how the problem could even be *addressed* with individual data – irrespective of whether Bayesian or non-Bayesian statistics were employed.

3.8 Concluding remarks

What I have written has only scratched the surface of longstanding disagreements. For any suggestion of dissent with any "Bayesian" views discussed in this chapter, there exists volumes of counter-arguments. Likewise, the debates among those who do not employ Bayesian methods are no less voluminous. I, myself, don't have a single "theory of inference" to which I adhere.

As I have sought, for reasons of clarity, to highlight the *differences* between Bayesian and non-Bayesian perspectives, I risk overstating them. It seems fitting, therefore, to conclude by illustrating that one can often find "non-Bayesian features" in Bayesian work and "Bayesian features" in non-Bayesian work.

3.8.1 Bayesian doesn't have to mean "not severe"

The idea that only non-Bayesians look for "severe tests," or "try to learn from errors," is not correct. One nice example comes from a recent careful study by Kline

and Tobias (2008), which is interested in estimating the "effect" of the Body Mass Index (BMI) on earnings.

Their point of departure is a two-equation model:

$$y_i = f(s_i) + x_i\beta + \epsilon_i \tag{3.16}$$

$$s_i = z_i\theta + u_i, \tag{3.17}$$

where y_i is log average hourly wages, x is a vector of demographic characteristics (schooling, experience, and so on) thought to have an effect on wages, s_i is the BMI of an individual, and $f(\cdot)$ is some continuous function of s which the authors introduce to allow for the reasonable possibility that, if BMI has an effect on wages, it is not necessarily linear.[64]

The most obvious possible problem is "confounding" – the relationship we observe between BMI and wages might merely represent the influence of other omitted factors that are correlated with BMI: in their model this is represented by a correlation between ϵ and u.[65] One such confounder they consider is "preferences for long-term investments, which we mean to represent characteristics that simultaneously impact decisions affecting both health and human capital accumulation."

One solution to this confounding problem is the identification of an "instrumental variable" that provides "exogenous" variation in BMI (that is, a variable which is correlated with BMI but not correlated with the unobserved determinants of wages). The authors discuss two possible instrumental variables for BMI – mother's BMI and father's BMI – and argue for their validity in several ways, including references to other literature.

In the case where $f(s)$ is linear in s, usual non-Bayesian practice is two-stage least squares or the method of instrumental variables. One test which sometimes seems to capture the notion of a"severe" test of the hypothesis that the instrumental variables are valid is an "overidentification test." Specifically, if both instrumental variables are valid, the estimated effect of BMI should be similar whether mother's BMI is used alone as an instrumental variable, father's BMI is used alone, or both are used (Newey, 1985). If the test rejects, it is unclear how to proceed, but as the authors note:

> From a theoretical perspective, however, it seems reasonable that the BMIs of the parents are either jointly valid as instruments, or jointly invalid, thus potentially calling into question what is actually learned from this procedure. On the empirical side, however, the correlation between parental BMI was found to be reasonably small (around .16), suggesting that something can be learned from this exercise, and that its implementation is not obviously redundant or "circular."

The authors' observation that the correlation between parental BMI is small seems, to me, to suggest the importance of "severity." Had the correlation been much higher, one might have been tempted to conclude that the proposed test was "obviously redundant" or "circular."

Consequently, Kline and Tobias (2008) provide an "informal test" of the validity of the "exclusion restriction" by asking whether, after using *one* instrumental variable, a model which excludes the other instrumental variable from the "structural equation" (equation (3.16) above) is well supported. Indeed, they calculate the Bayes factor associated with the hypothesis that the effect of father's BMI on wages is zero while maintaining the validity of mother's BMI as an instrumental variable and find strong support for that hypothesis. They also find strong support for the reverse.

While this does not exhaust the "specification testing" performed in the study, it does indicate an attempt to put a hypothesis to the severest test possible. Interestingly, although the study is clearly a "Bayesian" analysis, the authors found it useful to conduct such a test in an "informal" way – without attempting to "shoehorn" the specification testing into a complete Bayesian analysis.[66]

3.8.2 Non-Bayesian doesn't have to mean "severe"

My personal view is that statistical theory is often useful for situations in which we are attempting to describe something that looks like a "chance set-up." How one might go from information gleaned in such situations to draw inferences about other *different* situations, however, is not at all obvious. Some Bayesians might argue that one merely needs to formulate a prior, impose a "window on the world" (a.k.a. a likelihood) and then use Bayes' rule to revise our posterior probability. I am obviously uncomfortable with such a view and find LeCam's summary to the point: "The only precept or theory which seems relevant is the following: 'Do the best you can'. This may be taxing for the old noodle, but even the authority of Aristotle is not an acceptable substitute" (LeCam, 1977). This view also comports well with C.S. Peirce's classic description of a "severe test" I discussed earlier.

However, even at this level of vagueness and generality, it is worth observing that such views are not shared by non-Bayesians, or if they are, there is no common vision of what is meant by severe testing. Except for the most committed Bayesians, nothing in statistical theory tells you how to "infer the truth of various propositions." As I have argued elsewhere (DiNardo, 2007), often the types of theories economists seem interested in are so vague that it is often impossible to know what, in principle, would constitute "evidence" even in an "ideal" situation.

Certainly it is the case that non-Bayesian researchers are frequently unwilling to use statistical tools to change their views about *some* assessments. For one clear example, compare Glaeser and Luttmer (1997) and Glaeser and Luttmer (2003). The latter paper is a revised version of the former paper. The paper "develops a framework to empirically test for misallocation. The methodology compares consumption patterns for demographic subgroups in rent-controlled and free-market places. [They] find that in New York City, which is rent-controlled, an economically and statistically significant fraction of apartments appears to be misallocated across demographic subgroups"(Glaeser and Luttmer, 2003, p. 1027).

A significant difference between the two papers is that the latter, Glaeser and Luttmer (*ibid.*), includes an interesting falsification test (not included in Glaeser

and Luttmer, 1997). If the methodology works as intended, then when performing the same analysis on data from cities without rent control – they consider Chicago, Illinois, and Hartford, Connecticut – they should consistently estimate no welfare loss due to rent control. Contrary to such a presumption, however, for both cities they estimate large amounts of apartment misallocation (although these estimates are smaller than their estimate for New York City) that are statistically quite precise. Indeed, they acknowledge that "strictly interpreted, the results reject the identifying assumptions. In both cities, the procedure finds statistically significant misallocation" (Glaeser and Littmer, 2003, p. 1044). Nonetheless, while there are differences between the two versions of the paper, it is not clear whether such a rejection played any role in changing the inferences they draw. Indeed, they argue: "While this is disturbing, the large difference between our New York results and the results for these placebo cities suggests that even though our identifying assumptions may not exactly be true, the failure of the assumptions is unlikely to fully account for the observed misallocation in New York" (*ibid.*, p. 1044). What does it mean to say a set of assumptions fails to "fully account for the observed misallocation" and why should the results be viewed as "disturbing" (if indeed they should be)? In such a case, I think it is fair to say that we should have little confidence that their proposed *methodology* has a "truth preserving virtue." Note, however, this is only weakly related to one's views about the merits of rent control in New York City.

Much of the variation among non-Bayesians in their reaction to such statistical information seems to involve the "primacy" of certain types of (economic) *models*. Very roughly speaking, one can point to "a design-based approach" which focuses on creating or finding situations which resemble "chance set-ups" and where an analysis of the DGP proceeds separately from a single, specific, highly articulated, theoretical economic model. Historically, this approach has been associated with an emphasis on such issues as pre-specified analysis, "serious" specification testing, replicability, avoiding "confounding," identification, and so on.[67]

By contrast, one can also identify at least one strand of so-called "structural approaches" where there is little or no distinction between a DGP and a highly articulated "theoretical economic model."[68] An archetypal example of this approach, perhaps, is the multinomial logit of McFadden (1974) in which the consumer choice model – utility function, specification of heterogeneity in tastes, and so on – delivers a complete DGP in the form of a likelihood function. A feature of such an approach is that, in principle, once the model has been estimated, one can study "counterfactual policy simulations" or "experiments" which may have never been performed but can be described within the model.

This line of research gave birth to further developments which have yielded a wide variety of attitudes toward what might be called "severe testing." At one extreme, some researchers, such as Edward Prescott, apparently "completely reject econometrics as a useful scientific tool. Instead [Prescott] promotes *calibration* as the preferred method for 'uncovering' the unknown parameters of structural models and for evaluating and comparing their ability to fit the data" (Rust, 2007, p. 4).

While not rejecting the usefulness of statistics outright, Keane argues:

> that determinations of the usefulness of …"well-executed" structural models – both as interpreters of existing data and vehicles for predicting the impact of policy interventions or changes in the forcing variables – should rest primarily on how well the model performs in validation exercises. By this I mean: (1) Does the model do a reasonable job of fitting important dimensions of the historical data on which it was fit? (2) Does the model do a reasonable job at out-of-sample prediction – especially when used to predict the impact of policy changes that alter the economic environment in some fundamental way from that which generated the data used in estimation?
>
> My use of the word "reasonable" here is unapologetically vague. *I believe that whether a model passes these criteria is an intrinsically subjective judgment, for which formal statistical testing provides little guidance. This perspective is consistent with how other sciences treat validation*. (Keane, 2007, p. 32; emphasis added)

Indeed, Keane provides an illuminating illustration of this view by discussing an example of estimating the parameters of a life-cycle human capital investment model. After describing how the simplest version of the model fails to fit the data, he goes on to explain:

> by adding a number of extra features that are not essential to the model, but that seem reasonable (like costs of returning to school, age effects in tastes for schooling, measurement error in wages, and so on), we were able to achieve what we regard as an excellent fit to the key quantitative features of the data – although formal statistical tests still rejected the hypothesis that the model is the "true" data generating process (DGP). Despite these problems, there is nothing to indicate that the profession might be ready to drop the human capital investment model as a framework for explaining school and work choices over the life-cycle. (*ibid.,* p. 33)

As one might expect, Keane does not set forth a specific context in which one might find estimates of such a model "useful," or to what extent, if any, the inferences drawn from such an approach should influence the choices we make or what we advocate to others. Surely the model with or without amendments can't be "reasonable" for *all* contexts.

The "metastatistical" question is "How much confidence should one have in a judgment supported by such an approach?" The answer, to say the least, is not obvious.

Notes

1. The original author(s) of the quote are unknown. "The very model of the anonymous aphorism" (Koenker, 2007).
2. The number of Bayesian discussions are too numerous to list; nearly every book by a Bayesian has some discussion of metastatistics. A few books I found helpful: Berger and

Wolpert (1988), deFinetti (1974), Earman (1992), Good (1983), Howson and Urbach (1993), Joyce (1999), Keynes (1921), Savage (1972). Economists in particular may find Poirier (1995) useful for it's comparative approach, as well as Zellner (1984). Non-Bayesian discussions are not nearly as numerous but there are still many useful ones. Hacking (1965, 2001) are excellent introductions, as are Mayo (1996) and Venn (1888). Mayo (1996) has helped inspire a large literature trying, among other things, to provide a "philosophy of experiment." Useful articles with a broad focus include Freedman (1995) and LeCam (1977). The former includes some nice examples where economists and sociologists come off rather badly.

Occasionally all sides agree to get together and sometimes even agree to discuss issues. The famous "Savage Forum" (Savage *et al.,* 1962) is a nice introduction to a lot of the issues. Kyburg and Thalos (2003) has a nice collection of different approaches.

3. Even the term "introductory" is not mine. Hacking (1983) wrote: "Introductory topics should be clear enough and serious enough to engage a mind to whom they are new, and also abrasive enough to strike sparks off those who have been thinking about these things for years."
4. Paneth and Wallenstein (1985) observe, for example, that the survival rate among the 34 children who were considered for the trial, but did not enter because of a failure to meet one of the threshold criteria, was 100%.
5. It is also helpful to observe that the "prior" view of most ophthalmologists was that supplemental oxygen therapy was not a potential cause of Retrolental Fibroplasia (RLF) (now referred to as Retinopathy of Prematurity). From the *British Journal of Ophthalmology* (1974): "In the early days of research into the cause of RLF it was not uncommon at any meeting where oxygen was suggested as the cause, for an indignant ophthalmologist to rise from the floor and report a typical case where to his certain knowledge no supplemental oxygen was given. He would then sink back convinced that he had delivered the coup de grace to the oxygen theory. Equally challenging were those who claimed to have seen the condition in full-term infants, which seemed to deny any special vulnerability of growing retinal vessels. Although we now know these claims to have been valid, at the time they were stumbling blocks to the early acceptance of the vital importance of prematurity and oxygen."

   Much like the case of ECMO, the debate continues, as does the need for randomized controlled trials. Also, like ECMO, the debate has moved to more subtle questions, for example, about the appropriate threshold for starting oxygen in very low birth-weight children (Askie and Win, 2003; Silverman, 2004; Davis *et al.,* 2004; Hansmann, 2004; Shah, 2005; Vanderveen *et al.,* 2006).
6. See Bartlett *et al.* (1985), Wei and Durham (1978) and Zelen (1969) for a complete description of the variant of the "randomized play-the-winner" statistical method used.
7. See the several comments in *Statistical Science* **4**(4), 1989, and Ware's rejoinder in that issue. See Bartlett (2005) for a review of some of the history by one of the surgeons. The ethical issues don't end there; see also Couzin (2004): "Some companies seek out Berry Consultants [a small company founded by Bayesian advocate Donald Berry and his son] in the wild hope that a drug or device that's performed poorly in traditional trials can somehow undergo a Bayesian resurrection. (Such a 'rescue analysis' is rarely a possibility, both Berrys agree.)"
8. Indeed, while ECMO is used much more liberally today, *who* should get ECMO is a subject of considerable controversy (Allan *et al.,* 2007; Lequier, 2004; Thourani *et al.,* 2006). ECMO is now frequently employed but is still considered risky: "ECMO can have dangerous side-effects. The large catheters inserted in the baby's neck can provide a fertile field for infection, resulting in fatal sepsis" (Groopman, 2007). See *ibid.* for a case study where ECMO was begun, but then stopped because it was the "wrong treatment."

9. As it turns out, not even this sentiment is original. From Bickel and Lehmann (2001): "A chemist, Wilson (1952), [considering some issues in inference] pleads eloquently that 'There is a great need for further work on the subject of scientific inference. To be fruitful it should be carried out by critical original minds who are not only well-versed in philosophy but also familiar with the way scientists actually work (and not just with the way some of them say they work).' Wilson concludes pessimistically: 'Unfortunately the practical nonexistence of such people almost suggests that the qualities of mind required by a good philosopher and those needed by a working scientist are incompatible.' "
10. Of course, a non-Bayesian would feel that $\frac{1}{2}$ is a perfectly good estimator of the variance if you can't know which machine produced the measure.
11. Apparently, many examples of this specific type of inference can be avoided if one is a "conditional frequentist" (Poirier, 1995, p. 344).
12. The subtitle of LeCam's remarks "Toward Stating a Problem in the Doctrine of Chances" in part was an ironic twist on the title of Bayes' 1763 classic, "Toward Solving a Problem in the Doctrine of Chances" (Bayes, 1958).
13. A term frequently employed instead of "metastatistics" is the philosophy of "induction," and there is even debate on whether it is meaningful to talk about inductive inference. See Neyman (1957) and LeCam (1977), as well as Hacking (2001) and Mayo (1982).
14. LeCam's Basic Principle Zero (LeCam, 1990) was also intended to apply "in particular to the principles and recommendations listed below and should be kept in mind any time one encounters a problem worth studying." LeCam's principles seem quite sensible to me, and capture a lot of what I think non-Bayesians have in the back of their minds, including problems with the use of asymptotic approximations:

    1. Have clear in your mind what it is that you want to estimate.
    2. Try to ascertain in some way what precision you need (or can get) and what you are going to do with the estimate when you get it.
    3. Before venturing an estimate, check that the rationale which led you to it is compatible with the data you have.
    4. If satisfied that everything is in order, try first a crude but reliable procedure to locate the general area in which your parameters lie.
    5. Having localized yourself by (4), refine the estimate using some of your theoretical assumptions, being careful all the while not to undo what you did in (4).
    6. Never trust an estimate which is thrown out of whack if you suppress a single observation.
    7. If you need to use asymptotic arguments, do not forget to let your number of observations tend to infinity.
    8. J . Bertrand said it this way: "Give me four parameters and I shall describe an elephant; with five, it will wave its trunk."
15. A nice and more complete discussion can be found in Hacking (2001). Much of what follows is an abbreviated version of Hacking's discussion.
16. The Cretaceous/Tertiary Boundary is the boundary between the Cretaceous period and the Tertiary period. The Cretaceous period is the last period of the Mesozoic Era, which ended with the sudden extinction of the dinosaurs *inter alia.*
17. Even here, there is a possible non-Bayesian version of characterizing the probability: if we could re-run the world 100,000 times, in about 90% of cases an asteroid of size necessary to lead to mass extinction of the dinosaurs occurs. Perhaps ironically, on this precise question Bottke *et al.* (2007, p. 52) seem to arrive at their conclusion this way: "Using these [estimated] impact rates as input for a Monte Carlo code, we find there is a $\leq 10\%$ chance that the K/T impactor was derived from the background and a $\geq 90\%$ chance it came from the BAF [Baptistina Asteroid family]. Accordingly, we predict that the most

likely cause of the K/T mass extinction event was a collision between the Earth and a large fragment from the Baptistina asteroid shower."

18. In this regard, it is notable that there is a considerable body of non-Bayesian decision theory, the "Neyman–Pearson" framework being the best known. What is frequently referred to as the "Neyman–Pearson" statistical framework, however, is rarely *explicitly* invoked in most micro-empirical research, even though discussions about the "power" and "size" of tests are sometimes themselves the subject of debate. See, for example, McCloskey (1985), McCloskey and Ziliak (1996) and Hoover and Siegler (2008a, 2008b) for one debate on the subject.
19. There are many subtleties about the distinctions between beliefs and actions that I am ignoring. For instance, in describing Pascal's thesis, Joyce (1999, p. 21) – a "Bayesian" – is "careful to formulate [the thesis] as a *norm of rational desire* that governs the fair pricing of risky wagers [those that obey conventional axioms of probability]." In doing so he is explicit that he is making a statement about *desires* and *not actions* (*ibid.*, p. 19). "The old guard still insists that the concept of a fair price can only be understood in terms of behavioral dispositions, but it has become clear that the theoretical costs far outweigh benefits."
20. Ragnar Frisch's remarks, which accompanied Allias' article, suggest a fairly heated debate (emphasis added): "The problem discussed in Professor Allais' paper is of an extremely subtle sort and it seems to be difficult to reach a general agreement on the main points at issue. I had a vivid impression of these difficulties at the Paris colloquium in May, 1952. One evening when a small number of the prominent contributors to this field of study found themselves gathered around a table under the most pleasant exterior circumstances, it even proved to be quite a bit of a task to clear up in a satisfactory way misunderstandings in the course of the conversation. The version of Professor Allais' paper, which is now published in *ECONOMETRICA* emerged after many informal exchanges of views, including work done by editorial referees. Hardly anything more is now to be gained by a continuation of such procedures. *The paper is therefore now published as it stands on the author's responsibility. The editor is convinced that the paper will be a most valuable means of preventing inbreeding of thoughts in this important field. – R.F."*
21. By simple rearranging of terms, $B \succcurlyeq A$ yields:

$$0.11U(500,000) < 0.10U(2,500,000) + 0.01U(0),$$

and from $C \succcurlyeq D$ we get:

$$0.10U(2,500,000) + 0.01U(0) < 0.11U(500,000).$$

Hence, a contradiction.
22. For example, although games of chance greatly antedate anything resembling modern notions of probability – "someone with only modest knowledge of probability mathematics could have won himself the whole of Gaul in a week" – anything like our modern notions of probability did not "emerge permanently in discourse" until 1660 (see Hacking, 1975, 1990). Perhaps not surprisingly, the history of probability is the subject of much debate as well. For one criticism of Hacking's account see Garber and Zabell (1979).
23. Even at this point a fuller treatment would include a discussion of the problem of "logical omniscience." All I can do is cite a statement from Savage (1967): "For example, a person required to risk money on a remote digit of π would have to compute that digit, in order to comply fully with the theory [of personal probability], though this would really be wasteful if the cost of computation were more than the prize involved. For the postulates of the theory imply that you should behave in accordance with the logical implication of all that you know. Is it possible to improve the theory in this respect, making allowance within it for the cost of thinking, or would that entail paradox, as I am inclined to believe

but unable to demonstrate?" (As cited in Hacking, 1967. The published versions seems to have omitted some of the text.) See Hacking (1967) for a very useful discussion of the issue.

24. For instance, although it is easy to see how to partition a set of *events*, it is not always possible to see how to partition a set of *propositions*.
25. See Einstein (1920) for a thought-provoking discussion of Euclidean geometry as mathematical statements versus Euclidean geometry as statements about things in the world. As to Feller's observation about alternatives to Newton's law of attractions, see Cartwright (1984) for a provocative discussion of how even "the laws of physics lie" and physicists often fruitfully use different and mutually inconsistent models that makes a related point.
26. For a marvelous introductory exposition see Ch. 21 of Hacking (2001). For a more complete discussion that includes a bit more mathematical formalism and may be congenial to economists, see 2.1 of Poirier (1995).
27. There is much debate about the utility of *defining* probability as the limiting behavior of a sequence. This debate is intimately related to the debate about commending an estimator because, in repeated applications and in the long run, it would do well. The canonical problem is the "single case" exception (Hacking, 2001, Ch. 22) in which we are asked to consider a situation where, for example, there are two decks of cards: one, "the redder pack," has 25 red cards and 1 black card. The other "blacker pack" has 25 black cards and 1 red card. You are presented with two gambles. In the first gamble, you "win" if a red card is drawn randomly from the "redder" pack and lose otherwise ($P(\text{Win}) = 25/26$). In the second gamble, you "win" if a red card is drawn randomly from the "blacker" pack ($P(\text{Win}) = 1/26$). If you "win" you will be transported to "eternal felicity" and, if you lose, you will be "consigned to everlasting woe." As Hacking (and C.S. Peirce) and most people would choose the first gamble and hope, but *not* because of the long run; if we are wrong, there will be no comfort from the fact that we *would have been right most of the time*! Peirce's "evasion of the problem of induction" is to argue that we should not limit ourselves to merely "individualistic" considerations. "[Our interests] must not stop at our own fate, but must embrace the whole community. This community, again, must not be limited but extend to all races of beings with whom we can come in to immediate or mediate intellectual relation. It must reach, however vaguely, beyond this geological epoch, beyond all bounds. He would not sacrifice his own to save the whole world is, as it seems to me, illogical in all his inferences collectively. Logic is rooted in the social principle" (Peirce, 1878b, pp. 610–11).
28. Such a condition rules out, for example, the deterministic series $(H, T, H, T \ldots H, T)$ discussed above.
29. See Gillies (2000) for a nice discussion.
30. It is thus easier to understand Poirier's emphasis in the quotation above on whether the probability is "in nature" or "in themselves": that "to the extent that individuals agree on a class of events, they share an objective frequency. The objectivity, however, is in themselves, not in nature."
31. Again I have ignored "logical" probabilities, which are a class of epistemic probabilities which incorporate the notion of evidence. In this view, a probability is a "rational degree-of-belief" about a proposition or a measure of the degree of "credibility" of a proposition.
32. For three mutually exclusive analyses of what Bayes meant, and whether or not he succeeded in proving what he set out to establish, see Hacking (1965, Ch. 12). On whether Bayes' understanding is consistent with subsequent Bayesian interpreters (beginning with the rediscovery by Laplace, 1795), see Stigler (1982).
33. We have omitted one detail in this exposition, which is that the expression we are required to evaluate is:

$$\frac{\mathcal{L}(\theta|N,h)f(\theta)}{\int_0^1 \mathcal{L}(\theta|N,h)f(\theta)}.$$

In our previous notation, the denominator corresponds to $P(B)$ or $\sum_{j=1}^{k} P(B|A_j)P(A_j)$. One "nice" feature of the beta distribution is that it serves as the "natural conjugate prior" for the binomial distribution. Loosely speaking, the functional form of the prior and the likelihood are the same, which means one can treat the denominator as an *integrating constant* and it does not have to be computed directly.

34. The posterior mode can be rewritten as a weighted average of the sample mean and a prior mean, with the role of the prior vanishing as the number of actual sample observations grows large.
35. There are many variants of the following example. This particular variant is slightly adapted from Sober (2002).
36. Poirier (1995) provides a useful parody of the non-Bayesian view as he depicts a satistician's constantly evolving inference changing as he discovers more about the intent of the investigator.
37. Savage, before becoming familiar with the arguments in Barnard (1947a, 1947b), viewed the DGP as relevant ("I then thought it was a scandal that anyone in the profession could advance an idea – [that the DGP was irrelevant] – so patently wrong"). By the time of the forum he had come around to exactly the opposite point of view – "I [can] scarcely believe that people resist the idea [that the DGP was irrelevant] that is so patently right."
38. It should be noted than many Bayesians would argue that hypothesis testing *per se* is not a terribly sensible framework. They would also probably argue, nonetheless, that hypothesis tests are best interpreted in a Bayesian way.
39. There are many solutions to the problem of "inadmissible" samples in practice (unbalanced samples; see Jones, 1958, for example). One could merely conduct two experiments with more homogeneous samples. That is, one could conduct an experiment on low birth-weight babies and a separate experiment on high birth-weight babies. Sometimes block randomization is employed: the children might be sub-divided into groups according to their "healthiness" and the randomization might be performed separately within blocks.
40. See Zed *et al.* (1999) and Headache Classification Committee of the International Headache Society (2004) for extensive bibliographies.
41. The nosology of headache is elaborate and I can only coarsely define two types of headache here. According to the National Headache Foundation (http://www.headaches.org/consumer/tension_type.html, accessed December 10, 2007), "Tension-type headache is a nonspecific headache, which is not vascular or migrainous, and is not related to organic disease. The most common form of headache, it may be related to muscle tightening in the back of the neck and/or scalp [and is] characterized as dull, aching and non-pulsating pain [that] affect[s] both sides of the head.

    Symptoms … may include:

    - Muscles between head and neck contract
    - A tightening band-like sensation around the neck and/or head which is a 'vice-like' ache
    - Pain primarily occurs in the forehead, temples or the back of head and/or neck."

    Migraine headaches are most commonly associated with severe unilateral head pain, often accompanied by nausea and vomiting, photophobia (fear of light) and phonophobia (fear of sound), that can last from a few hours to several days. In some fraction of migraine patients the head pain is preceded or accompanied by visual disturbances called auras.
42. In the US, opioids were the standard of care as late as the nineteenth century until they were supplanted by aspirin in the early twentieth century at roughly the same time as over-the-counter use of such medications was made illegal (Meldrum, 2003). Moreover,

outside of the MOH literature it is generally viewed that opioids are under-prescribed because of (sometimes irrational) fears of promoting addiction, censure by police, and so on (Lipman, 2004). "The history of opioid use (or nonuse) in neuropathic pain is instructive. The natural reluctance to prescribe opioids to patients with neuropathic pain of benign cause was, for many years, reinforced by the received wisdom that opioids were ineffective in neuropathic pain [such as headache], based on weak evidence. It took many years before this 'truth' was questioned. Reexamination in the later 1980s was followed by controlled studies that clearly substantiated an important analgesic action of morphine and fentanyl and, later, other opioids in neuropathic pain" (Scadding, 2004).

43. Medication overuse headache has gone by several different names, including *analgesic rebound*, *ergotamine rebound*, *medication induced headache*, *transformed migraine*, *chronic migraine*, *daily headache*, *drug-induced headache*, *painkiller headache*, *medication-misuse headache*, *analgesic-dependent headache* (Obermann and Katsarava, 2007; Silberstein *et al.*, 1994, and so on).
44. As is widely recognized, severe chronic daily migraine occurred before the use of offending medications was common or even possible. The most widely cited example comes from the important neurologist Thomas Willis (1683), who recorded his treatment of Viscountess Anne (Finche) Conway in the seventeenth century.
45. Even current scholars in the field are negative about early developments in the field of headache. "Prior to the 1980s, the field of headache was rarely influenced by what would be generally accepted as scholarly, credible research" (Saper, 2005).
46. In describing the work as "representative," however, the view among experts in the field is considerably more favorable. Ward (2008) describes it as "his favorite article" in a recent review. Mathew (2008) responds by noting: "The impact of this article on the American and European headache communities was substantial. Until then, the Europeans had not appeared to appreciate the clinical significance of medication overuse or the existence of chronic daily headache ... One enduring fact continues to disappoint me. In spite of the extensive effort made to emphasize the importance of medication overuse in managing the headache population, many practitioners – including neurologists – continue to overprescribe symptomatic medications, thereby condemning their patients to treatment failure." For a more recent systematic review, see Zed *et al.* (1999).
47. No randomization appears to be involved. The patients were merely "grouped."
48. The measurement of "improvement" is not clear, but it appears to have been asymmetric. Improvement was measured as a percentage change in a headache index if the patient improved, and was given a value of zero if the patient did not improve.
49. One definition of MOH is:

    1. Occurs in a patient with a primary headache disorder who uses symptomatic or immediate relief medications very frequently (daily), often in excessive quantities.
    2. Tolerance to symptomatic medications develop and headaches become worse on continuing the treatment.
    3. The patient may show symptoms of withdrawal on discontinuing the medication, with increased headache lasting for a variable period of time, as long as three to four weeks.
    4. Headache ultimately improves after stopping the offending medications even though the primary headache disorder needs continuing prophylactic treatment.

50. The only other criticism I have been able to locate are letters to the editors (Gupta, 2004a, 2004b, 2004c).
51. Using language evocative of severe testing, Fisher remarked that "Claude Bernard, in speaking of an hypothesis, said that it is not sufficient to merely gather all the facts that support it but even more importantly, one must go out of one's way to find every means of refuting it."

52. While Fisher's challenge has never been even approximately met, the question of whether headaches arise *de novo* from analgesic headache has been investigated and substantiates Fisher's claim. It has been conceded that Fisher was correct (Bahra *et al.,* 2003; Lance *et al*., 1988).
53. For a brief history of the ICHD, see Gladstone and Dodick (2004).
54. Boes and Capobianco (2005) and Ferrari *et al*. (2007) describe some of the tangled history as well. It should be noted that the history is a matter of some dispute.
55. The rate of analgesic use is *usually* defined in terms of treatment days per month, such that treatment occurs at least two or three days each week, with intake of the drug on at least ten days per month for at least three months.
56. From Schuster (2004): "The definition does not apply to headache in women who take medications for five or six consecutive days for menstrually associated migraines but are treatment-free the rest of the month, acknowledged Fred D. Sheftell, MD, who participated in updating the classification. He said that if a woman took medications just four other days of the month, she would inappropriately meet the 10-days-a-month rule. For that reason, she must also be taking the drugs at least two to three days each week to meet the criteria."
57. Among the reasons given was the fact that "patients could become chronic due to medication overuse, but this effect might be permanent. In other words, it may not be reversible after discontinuation of medication overuse. Finally, a system whereby medication overuse headache became a default diagnosis in all patients with medication overuse would encourage doctors all over the world to do the right thing, namely, to take patients off medication overuse as the first step in a treatment plan."
58. See, for example, Saper and Lake (2006a) for a proposal to distinguish opioid using MOH patients from the remaining "less complicated" cases.
59. See Horowitz and Manski (1998) for a detailed discussion of such bounds. It should be noted that where such bounds are used, common practice is to report both "best case" *and* "worst case" bounds.
60. Indeed, the notion of "intent-to-treat" can be seen as part of an attempt to test a hypothesis **severely** and not a notion that is an inevitable consequence of adopting "frequentist" probability notions. See Hollis and Campbell (1999) for a discussion.
61. He referred to the union premium in wages between otherwise identical workers as the wage "gap" to distinguish it from what might obtain in a world without unions.
62. The distinction between ATOT and other estimands is important since it isn't particularly meaningful to consider the effect of union status on, say, the CEO of a large multinational, to take a stark example (US law, for example, prohibits this possibility). The paper, unfortunately, takes a naive approach to characterizes the treatment heterogeneity: in considering the variation in the effect of union status, it characterizes it by the estimated probability of being unionized. That is, the effect of unionization is allowed to vary across workers whose demographic characteristics put them at the same "risk" of being unionized. This conflates the treatment effect for workers with extremely low levels of observed human capital (who typically have very low probabilities of being unionized) with the treatment effects for those who can't be unionized (bosses) or those with high levels of education who are generally hostile to unionization. For an arguably much more sensible characterization of the heterogeneity in treatment effects, see Card (1992), for example. Card (*ibid*.) also deals with the problem of measurement error in union status, which is ignored in the empirical example in Chib and Hamilton (2002) but has long been an important issue in non-Bayesian analyses: see Freeman (1984), Jakubson (1991), or Card (1992) for three examples.
63. This possibility wasn't ignored, however. The problem was the lack of data. See, for example, Abowd and Farber (1982), Freeman and Kleiner (1990, 1999), who take the possibility quite seriously.

64. Although this is an important focus of their paper, it is one that I will not focus on since my interests in using this example lie elsewhere. To quote from their paper: "again, it is important to recognize that many applied studies in the treatment-response literature, and to our knowledge all of those that have been conducted on this specific topic, assume the relationship between the treatment variable and the outcome variable is linear (i.e., $f(s) = \alpha_0 + \alpha_1 s$), and define the slope of this function as the causal effect of interest. The assumption of linearity is likely made on computational considerations, as IV [instrumental variables] is simple to use in this context."
65. Again because my interests lie elsewhere, one might not wish to stipulate that it is possible to talk clearly about a "causal effect" of BMI on wages, because such an effect seems to presuppose that, in whatever manner we "manipulate" an individual's BMI, we would expect that the "effect" of BMI on wages would be the same. However, it is possible to imagine that the causal arrow runs *from* BMI *to* wages because (1) high BMI is equivalent to "bad health" and "bad health" lowers wages, or (2) high BMI is equivalent to "unattractive," and employers discriminate against those who are "unattractive" for reasons possibly unrelated to "productivity." If the latter were true, a "successful" but "unhealthy diet" that lowered BMI would *raise* wages; if the former were true, such a diet would *lower* wages. See DiNardo (2007) for a discussion.
66. It is also interesting to observe that this Bayesian "overidentification test" is arguably better-suited to "severity" than recent non-Bayesian interpretations of such overidentification tests. In the linear instrumental variables model, for example, the failure of the overidentification test has been recently reinterpreted *not* as a rejection of the premises of the estimated model but as evidence of "treatment effect" heterogeneity (Angrist, 2004). In this context, one possible cause (though not the only possible cause) of "treatment effect heterogeneity" is that the true relationship between BMI and wages is, say, quadratic but the investigator specifies a linear relationship. In such a case, one could no longer ensure that the estimated relationship would be invariant to the choice of instrumental variables even if the instrumental variables were "valid." The informal test proposed by Kline and Tobias (2008), however, easily accommodates such a situation since it allows $f(s)$ to be nonlinear without exhausting any overidentification.
67. Some of this discussion draws from a brief discussion in unpublished lecture notes by David Card, although for reasons of focus and brevity my account is not the same (Card, 2007).
68. It should not be surprising that the term "structural model" encompasses a wide array of activities which have very different emphases, including – to take just one example – classic studies of demand systems, and so on (see, for example, Deaton and Muellbauer, 1980). Moreover, some work involving "structural estimation" occurs in studies that also involve a design-based approach. For a simple illustration see DiNardo and Lemieux, (1992, 2001). Consequently, I use the phrase "single strand" advisedly.

References

Abowd, J. and H. Farber (1982) Job queues and the union status of workers. *Industrial and Labor Relations Review* **35**.

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole americaine. *Econometrica* **21**, 503–46.

Allan, C.K., R.R. Thiagarajan, P.J. del Nido, S.J. Roth, M.C. Almodovar and P.C. Laussen (2007) Indication for initiation of mechanical circulatory support impacts survival of infants with shunted single-ventricle circulation supported with extracorporeal membrane oxygenation. *Journal of Thoracic and Cardiovascular Surgery* **133**, 660–7.

Angrist, J. (2004) Treatment effect heterogeneity in theory and practice. *Economic Journal* **114**, 52–83.

Ashenfelter, O. (1978) Union relative wage effects: new evidence and a survey of their implications for wage inflation. In R. Stone and W. Peterson (eds.), *Economic Contributions to Public Policy*, pp. 31–63. New York: St. Martin's Press.

Askie, L.M. and T. Win (2003) The use of oxygen in neonatal medicine: half a century of uncertainty. *Neoreviews* **4**, e340–8.

Bahra, A., M. Walsh, S. Menon and P.J. Goadsby (2003) Does chronic daily headache arise de novo in association with regular use of analgesics? *Headache* **43**, 179–90.

Barnard, G.A. (1947a) The meaning of a significance test. *Biometrika* **34**, 179–82.

Barnard, G.A. (1947b) A review of *Sequential Analysis* by Abraham Wald. *Journal of the American Statistical Association* **42**, 658–69.

Barnouw, J. (1979) A review of *The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Eighteenth Century Studies* **12**, 438–43.

Bartlett, R.H. (2005) Extracorporeal life support: history and new directions. *American Society for Artificial Internal Organs Journal* **51**, 487–9.

Bartlett, R.H., D.W. Roloff, R.G. Cornell, A.F. Andrews, P.W. Dillon and J.B. Zwischenberger (1985) Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76**, 479–87.

Bayes, T. (1958) An essay towards solving a problem in the doctrine of chances. *Biometrika* **45**, 296–315 by the late Reverend Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S.

Berger, J.O. and R.L. Wolpert (1988) *The Likelihood Principle* (second edition). Hayward, Calif.: Institute of Mathematical Statistics.

Berry, D.A. (1989) [Investigating therapies of potentially great benefit: ECMO]: Comment: ethics and ECMO. *Statistical Science* **4**, 337–40.

Berry, D.A. (2006) Bayesian clinical trials. *Nature Reviews Drug Discovery* **5**, 27–36.

Berry, S.M. and J.B. Kadane (1997) Optimal bayesian randomization. *Journal of the Royal Statistical Society, Part B* **59**, 813–19.

Bickel, P.J. and E.L. Lehmann (2001) Frequentist inference. In N.J. Smelser and P.B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences* pp. 5789–96. Oxford: Pergamon.

Boes, C.J. and D.J. Capobianco (2005) Chronic migraine and medication-overuse headache through the ages. *Cephalalgia* **25**, 378–90.

Bottke, W.F., D. Vokrouhlický and D. Nesvorný (2007) An asteroid breakup 160 million years ago as the probable source of the K/T impactor. *Nature* **449**, 48–53.

British Journal of Ophthalmology (1974) Editorial: retrolental fibroplasia (rlf) unrelated to oxygen therapy. *British Journal of Ophthalmology* **58**, 487–9.

Capobianco, D.J., J.W. Swanson and D.W. Dodick (2001) Medication-induced (analgesic rebound) headache: historical aspects and initial descriptions of the North American experience. *Headache* **41**, 500–2.

Card, D. (1992) The effect of unions on the distribution of wages: redistribution or relabelling? NBER Working Paper 4195. Cambridge, Mass.: National Bureau of Economic Research.

Card, D. (1996) The effect of unions on the structure of wages: a longitudinal analysis. *Econometrica* **64**, 957–79.

Card, D. (2007) Lecture notes for Topics in Labor Economics. Cambridge: Mass.: Department of Economics, Harvard University.

Card, D., T. Lemieux and C.W. Riddell (2003) Unionization and wage inequality: a comparative study of the U.S., U.K. and Canada. NBER Working Paper 9473. Cambridge, Mass.: National Bureau of Economic Research.

Cartwright, N. (1984) *How the Laws of Physics Lie*. New York: Oxford University Press.

Chew, S.H. (1983) A generalization of the quasilinear mean with application to the measurement of income inequality and decision-theory resolving the Allais paradox. *Econometrica* **51**, 1065–92.

Chib, S. and B.H. Hamilton (2002) Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.
Couzin, J. (2004) The new math of clinical trials. *Science* **303**, 784–6.
Davis, P.G., A. Tan, C. O'Donnell and A. Schulze (2004) Resuscitation of newborn infants with 100% oxygen or air: a systematic review and meta-analysis. *Lancet* **364**, 1329–33.
Deaton, A. and J. Muellbauer (1980) *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.
deFinetti, B. (1974) *The Theory of Probability* (two volumes). New York: John Wiley.
DiNardo, J. (2007) Interesting questions in freakonomics. *Journal of Economic Literature* **45**, 973–1000.
DiNardo, J. and D.S. Lee (2004) Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics* **119**, 1383–441.
DiNardo, J. and T. Lemieux (1992) Alcohol, marijuana, and American youth: the unintended consequences of government regulation. NBER Working Paper 4212. Cambridge, Mass.: National Bureau of Economic Research.
DiNardo, J. and T. Lemieux (1997) Diverging male wage inequality in the United States and Canada, 1981–1988: do institutions explain the difference? *Industrial and Labor Relations Review* **50**(4), 629–51.
DiNardo, J. and T. Lemieux (2001) Alcohol, marijuana, and American youth: the unintended consequences of government regulation. *Journal of Health Economics* **20**, 991–1010.
Earman, J. (1992) *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*. Cambridge, Mass.: MIT Press.
Edmeads, J. (1990) Analgesic-induced headaches: an unrecognized epidemic. *Headache* **30**, 614–15.
Edwards, W., H Lindman and L.J. Savage (1963) Bayesian statistical inference for psychological research. *Psychological Review* **70**(3).
Einstein, A. (1920) *Relativity: The Special and General Theory*, trans. Robert W. Lawson. New York: Henry Holt and Company.
Feller, W. (1950) *An Introduction to Probability Theory and Its Applications*, *Volume 1*. New York: John Wiley.
Ferrari, A., C. Coccia and E. Sternieri (2008) Past, present, and future prospects of medication-overuse headache classification. *Headache* **48**(7), 1096–102.
Fisher, C.M. (1978) Analgesic rebound headache refuted. *Headache* **28**, 666.
Food and Drug Administration, US Department of Health and Human Services (2006) Guidance for the use of Bayesian statistics in medical device clinical trials – draft guidance for industry and FDA staff. Accessed Center for Devices and Radiological Health, Division of Biostatistics, Office of Surveillance and Biometrics.
Freedman, D.A. (1995) Some issues in the foundation of statistics. *Foundations of Science* **1**, 19–39.
Freeman, R. (1984) Longitudinal analysis of the effects of trade unions. *Journal of Labor Economics* **2**, 1–26.
Freeman, R.B. and M. Kleiner (1990) The impact of new unionization on wages and working conditions. *Journal of Labor Economics* **8**, S8–25.
Freeman, R.B. and M.M. Kleiner (1999) Do unions make enterprises insolvent? *Industrial and Labor Relations Review* **52**, 510–27.
Friedman, M. (1950) Some comments on the significance of labor unions for economic policy. In D.M. Wright (ed.), *The Impact of the Union: Eight Economic Theorists Evaluate the Labor Union Movement* New York: Harcourt, Brace and Company, and Institute on the Structure of the Labor Market, American University, Washington DC.
Fuchs, V.R., A.B. Krueger and J.M. Poterba (1998) Economists' views about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature* **36**, 1387–425.

Garber, D. and S. Zabell (1979) On the emergence of probability. *Archive for the History of the Exact Sciences* **21**, 33–53, communicated by C. Truesdell.
Gillies, D. (2000) *Philosophical Theories of Probability. Philosophical Issues in Science*. London: Routledge.
Gladstone, J.P. and D.W. Dodick (2004) From hemicrania lunaris to hemicrania continua: an overview of the revised international classification of headache disorders. *Headache* **44**, 692–705.
Glaeser, E.L. and E.F.P. Luttmer (1997) The misallocation of housing under rent control. NBER Working Paper 6220. Cambridge, Mass: National Bureau of Economic Research.
Glaeser, E.L. and E.F.P. Luttmer (2003) The misallocation of housing under rent control. *American Economic Review* **93**, 1027–46.
Good, I.J. (1971) 46656 varities of Bayesians. *American Statistician* **25**, 62–3.
Good, I.J. (1983) *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.
Good, I.J. and R.A. Gaskins (1971) Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–77.
Goodman, S.N. (1999) Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine* **130**, 995–1004.
Goodman, S. (2004) Basic Bayes – I. Technical report. Johns Hopkins University National Institutes of Health, Bethesda, Maryland, from a Workshop sponsored by the Food and Drug Administration and Johns Hopkins University on "Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision-Making?"
Groopman, J. (2007) *How Doctors Think*. Boston: Houghton Mifflin.
Gupta, V.K. (2004a) Chronic daily headache with analgesic overuse: epidemiology and impact on quality of life. *Neurology* **63**, 1341.
Gupta, V.K. (2004b) Classification of primary headaches: pathophysiology versus nosology? *British Medical Journal*, Letter to the Editor.
Gupta, V.K. (2004c) De novo headache and analgesic consumption: pathophysiological insights from nosologic complexity? *Headache* **44**, 375.
Hacking, I. (1965) *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Hacking, I. (1967) Slightly more realistic personal probability. *Philosophy of Science* **34**, 311–25.
Hacking, I. (1975) *The Emergence of Probability: A Philisophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.
Hacking, I. (1983) *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.
Hacking, I. (1990) *The Taming of Chance*. Number 17 in Ideas in Context. Cambridge: Cambridge University Press.
Hacking, I. (2001) *An Intoduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.
Hansmann, G. (2004) Neonatal resuscitation on air: it is time to turn down the oxygen tanks [corrected]. *Lancet* **364**, 1293–94.
Headache Classification Committee of the International Headache Society (2004) The international classification of headache disorders: second edition. *Cephalalgia* **24**, 9–160.
Headache Classification Committee of the International Headache Society (2006) New appendix criteria open for a broader concept of chronic migraine. *Cephalalgia* **26**, 742.
Heckman, J.J. (1990) Varieties of selection bias. *American Economic Review* **80**, 313–18.
Hollis, S. and F. Campbell (1999) What is meant by intention to treat analysis? Survey of published randomized controlled trials. *British Medical Journal* **319**, 670–4.
Hoover, K.D. and M.V. Siegler (2008a) Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology* **15**, 1–37.
Hoover, K.D. and M.V. Siegler (2008b) The rhetoric of signifying nothing: a rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology* **15**, 57–68.

Horowitz, J.L. and C.F. Manski (1998) Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics* **84**, 37–58.

Howson, C. (1997) A logic of induction. *Philosophy of Science* **64**, 268–90.

Howson, C. and P. Urbach (1993) *Scientific Reasoning: The Bayesian Approach* (second edition). Chicago and La Salle, Ill.: Open Court.

Humphrey, T. (1992) Marshallian cross diagrams and their uses before Alfred Marshall: the origins of supply and demand geometry. *Federal Bank of Richmond Economic Review*, March–April, 3–23.

International Headache Society (1988) Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. Headache Classification Committee of the International Headache Society. *Cephalagia* **8** (Supplement 7), 1–96.

Jakubson, G. (1991) Estimation and testing of the union wage effect using panel data. *Review of Economic Studies* **58**, 971–91.

Jaynes, E.T. (1976) Confidence intervals vs Bayesian intervals. In W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, *Volume II*. Holland: Reidel Publishing Co.

Jenkin, F. (1868) Trade-unions: how far legitimate? In S.C. Colvin and J.A. Ewing (eds.), *Papers, Literary, Scientific, &c by the late Fleeming Jenkin, Volume 2*. London: Longmans, Green & Co., originally published in the *North British Review*, March 1868.

Jenkin, F. (1870) The graphic representation of the laws of supply and demand, and their application to labour. In S.C. Colvin and J.A. Ewing (eds.), *Papers, Literary, Scientific, &c by the late Fleeming Jenkin* Volume 2. London: Longmans, Green & Co.

Jones, H.L. (1958) Inadmissible samples and confidence limits. *Journal of the American Statistical Association* **53**, 482–90.

Joyce, J.M. (1999) *The Foundations of Causal Decision Theory*. Cambridge, Mass.: MIT Press.

Keane, M.P. (2007) Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, http://gemini.econ.umd.edu/jrust/research/JE_Keynote_7.pdf.

Keynes, J.M. (1921) *A Treatise on Probability*. London: Macmillan and Co., Limited.

Kline, B. and J. Tobias (2008) The wages of BMI: Bayesian analysis of a skewed treatment-response model with nonparametric endogeneity. *Journal of Applied Econometrics*. Forthcoming.

Kmenta, J. (2000) *Elements of Econometrics* (second edition). Ann Arbor, Mich.: University of Michigan Press.

Koenker, R. (2007) Personal communication.

Krashinsky, H.A. (2004) Do marital status and computer usage really change the wage structure? *Journal of Human Resources* **3**, 774–91.

Kyburg Jr., H.E. and M. Thalos (eds.) (2003) *Probability is the Very Guide to Life: The Philosophical Uses of Chance*. Chicago and La Salle, Ill.: Open Court.

Lance, F., C. Parkes and M. Wilkinson (1988) Does analgesic abuse cause headaches de novo? *Headache* **28**, 61–2.

Laplace, P.S. (1795) *Essai Philosophique sur les Probabilités* sixth edition, translated as "Philosophical Essay on Probabilities" from the sixth French edition by Frederick Wilson Truscott and Frederick Lincoln Emory and first US edition published in 1902. New York: John Wiley & Sons .

LeCam, L. (1977) A note on metastatistics or "an essay toward stating a problem in the doctrine of chances." *Synthese* **36**, 133–60.

LeCam, L. (1990) Maximum likelihood: an introduction. *International Statistical Review* **58**, 153–71.

Lemieux, T. (1998) Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and non-random selection. *Journal of Labor Economics* **16**, 261–91.

Lequier, L. (2004) Extracorporeal life support in pediatric and neonatal critical care: a review. *Journal of Intensive Care Medicine* **19**, 243–58.
Lewis, H.G. (1963) *Unionism and relative wages in the United States*. Chicago: University of Chicago Press.
Lewis, H.G. (1986) Union relative wage effects. In O. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics, Volume 2*, Ch. 20, pp. 1139–82. Amsterdam: North-Holland.
Lindley, D.V. (1971) The estimation of many parameters. In V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*, pp. 435–447. Toronto: Holt, Rinehart and Winston.
Lindley, D.V. (2000) The philosophy of statistics. *The Statistician* **49**, 293–337.
Lipman, A.G. (2004) Does opiophobia exist among pain specialists? *Journal of Pain and Palliative Care Pharmacotherapy* **18**, 1–5.
Mathew, N.T. (2008) Response: drug-induced refractory headache. *Headache* **48**, 729.
Mathew, N.T., R. Kurman and F. Perez (1990) Drug induced refractory headache – clinical features and management. *Headache* **30**, 634–38.
Mayo, D.G. (1979) Testing statistical testing. In J.C. Pitt (ed.), *Philosophy in Economics*, Volume 16 of *The University of Western Ontario Series in Philosophy of Science*, pp. 175–203. Dordrecht, Holland: D. Reidel Publishing Company. Papers Deriving from and Related to a Workshop on Testability and Explanation in Economics held at Virginia Polytechnic Institute and State University, 1979.
Mayo, D.G. (1982) On after-trial criticisms of Neyman–Pearson theory of statistics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* **1**, 145–58.
Mayo, D.G. (1996) *Error and the Growth of Experimental Knowledge: Science and its Conceptual Foundations*. Chicago: University of Chicago Press.
Mayo, D.G. (2003) Severe testing as a guide for inductive reasoning. In H.E. Kyburg, Jr. and M. Thalos (eds.), *Probability is the Very Guide of Life: The Philosophical Uses of Chance,* pp. 89–117. Chicago and La Salle, Ill.: Open Court.
Mayo, D.G. and M. Kruse (2002) Principles of inference and their consequences. In D. Corfield and J. Williamson (eds.), *Foundations of Bayesianism*, Volume 24 of *Applied Logic*. Kluwer Academic Publishers.
Mayo, D.G. and A. Spanos (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science* **57**, 323–57.
McCloskey, D. (1985) The loss function has been mislaid: the rhetoric of significance tests. *American Economic Review* Supplement **75**, 201–5.
McCloskey, D. and S. Ziliak (1996) The standard error of regressions. *Journal of Economic Literature* **34**, 97–114.
McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers of Econometrics, Volume 3*, Ch. 4, pp. 105–42. New York and London: Academic Press.
Meldrum, M.L. (2003) A capsule history of pain management. *Journal of the American Medical Association* **290**, 2470–5.
Newey, W. (1985) Generalized method of moments specification tests. *Journal of Econometrics* **29**, 229–56.
Neyman, J. (1957) Inductive behavior as a basic concept of philosophy of science. *Revue d'Institute International de Statistique* **25**, 7–22.
Obermann, M. and Z. Katsarava (2007) Management of medication-overuse headache. *Expert Review of Neurotherapeutics* **7**, 1145–55.
Olesen, J., M.-G. Bousser, H.-C. Diener, D. Dodick, M. First, P. Goadsby, H. Göbel, M. Lainez, J. Lance, R. Lipton, G. Nappi, F. Sakai, J. Schoenen, S. Silberstein and T. Steiner (2006) New appendix criteria open for a broader concept of chronic migraine. *Cephalalgia* **26**, 742.
Paneth, N. and S. Wallenstein (1985) Extracorporeal membrane oxygenation and the play the winner rule. *Pediatrics* **76**, 622–3.
Peirce, C.S. (1878a) The doctrine of chances. *Popular Science Monthly* **12**, 604–15.

Peirce, C.S. (1878b) The probability of induction. *Popular Science Monthly* **12**, 705–18.
Peirce, C.S. (1958) Collected papers. In A. Burks (ed.), *Collected Papers, Volumes 7–8*. Cambridge, Mass.: Harvard University Press.
Pini, L.-A., A. Cicero and M. Sandrini (2001) Long-term follow-up of patients treated for chronic headache with analgesic overuse. *Cephalalgia* **21**, 878–83.
Poirier, D.J. (1995) *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, Mass.: MIT Press.
Rust, J. (2007) Comments on Michael Keane's "Structural vs. atheoretic approaches to econometrics." *Journal of Econometrics*. Forthcoming.
Saper, J.R. (2005) Editorial to the guidlines for trials of behavioral treatments for recurrent headache. *Headache* **45** (Supplement 2), S90–1.
Saper, J.R. and A.E. Lake (2006a) Medication overuse headache: type i and type ii. *Cephalalgia* **26**, 1262.
Saper, J.R. and A.E. Lake (2006b) Sustained opioid therapy should rarely be administered to headache patients: clinical observations, literature review and proposed guidelines *Headache Currents* **3**, 67–70.
Saper, J.R., A.E. Lake, R.L. Hamel, T.E. Lutz, B. Branca, D.B. Sims and M.M. Kroll (2004) Daily scheduled opioids for intractable head pain: long-term observations of a treatment program. *Neurology* **62**, 1687–94.
Savage, L.J. (1967) Difficulties in the theory of personal probability. *Philosophy of Science* **34**, 305–10.
Savage, L.J. (1972) *The Foundations of Statistics*. New York: Dover Publications. Revised and expanded version of the original 1954 work.
Savage, L.J., M. Bartlett, G.A. Barnard, D.R. Cox, E.S. Pearson, C.A.B. Smith, *et al.* (1962) The foundations of statistical inference: a discussion. In G.A. Barnard and D.R. Cox (eds.), *The Foundations of Statistical Inference: A Discussion.* Methuen's Monographs on Applied Probability and Statistics. London and Colchester: Spottiswoode Ballantyne & Co. Ltd. (A Discussion Opened by L.J. Savage at the Joint Statistics Seminar of Birbeck and Imperial Colleges. Discussants also include H. Ruben, I.J. Good, D.V. Lindley, P. Armitage, C.B. Winsten, R. Syski, E.D. Van Rest and G.M. Jenkins.)
Scadding, J.W. (2004) Treatment of neuropathic pain: historical aspects. *Pain Medicine* **5** (Supplement 1), S3–8.
Schuster, L. (2004) Revised guidelines for medication overuse headache. *Neurology Reviews* **12**.
Shah, P.S. (2005) Resuscitation of newborn infants. *Lancet* **365**, 651–2; author reply, 652–3.
Silberstein, S. (2004) Introduction: Aching heads. In J. Kempner (ed.), *Aching Heads, Making Medicine: Gender and Legitimacy in Headache*. Philadelphia.
Silberstein, S.D., R.B. Lipton, S. Solomon and N.T. Mathew (1994) Classification of daily and near-daily headaches: proposed revisions to the IHS criteria. *Headache* **34**, 1–7.
Silverman, W.A. (1980) *Retrolental Fibroplasia. A Modern Parable* London: Grune and Stratton.
Silverman, W.A. (2004) A cautionary tale about supplemental oxygen: the albatross of neonatal medicine. *Pediatrics* **113**, 394–96.
Simon, J. (1997) The philosophy and practice of resampling statistics, http://www.juliansimon.com/writings/Resampling_Philosophy/, accessed May 1, 2008.
Sober, E. (2002) Bayesianism: its scope and limits. In R. Swinburne (ed.), *Proceedings of the British Academy, Volume 113,* pp. 21–38. Oxford: Oxford University Press and the British Academy.
Society, I.H. (1988) Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalagia* **8**, 1–96.
Stigler, S.M. (1982) Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society, Series A* (General) **145**, 250–8.
Thourani, V.H., P.M. Kirshbom, K.R. Kanter, J. Simsic, B.E. Kogon, S. Wagoner, F. Dykes, J. Fortenberry and J.M. Forbess (2006) Venoarterial extracorporeal membrane oxygenation (VA-ECMO) in pediatric cardiac support. *Annals of Thoracic Surgery* **82**, 138–45.

Urbach, P. (1985) Randomization and the design of experiments. *Philosophy of Science* **52**, 256–73.

Vanderveen, D.K., T.A. Mansfield and E.C. Eichenwald (2006) Lower oxygen saturation alarm limits decrease the severity of retinopathy of prematurity. *Journal of the American Association for Pediatric Ophthalmology and Strabismus* **10**, 445–8.

Venn, J. (1888) *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, With Especial Reference to its Logical Bearings and its Application to Moral and Social Science*. London and New York: Macmillan, third edition (first edition, 1866, second edition, greatly expanded, 1876, third edition, 1888).

von Mises, R. (1957) *Probability, Statistics, and Truth*. London and New York: Allen and Unwin Ltd. and the Macmillan Company.

Ward, T.N. (2008) My favorite article: drug-induced refractory headache. *Headache* **48**, 728–9.

Ware, J.H. (1989) Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**, 298–306.

Wei, L.J. and S. Durham (1978) The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* **73**, 840–3.

Whitney, C.W. and M. Von Korff (1992) Regression to the mean in treated versus untreated chronic pain. *Pain* **50**, 281–5.

Willis, T. (1683) *Two discourses concerning the soul of brutes which is that of the vital and sensitive of man. The first is physiological, shewing the nature, parts, powers, and affections of the same. The other is pathological, which unfolds the diseases which affect it and its primary seat; to wit, the brain and nervous stock, and treats of their cures: with copper cuts. By Thomas Willis, doctor in physick, professor of natural philosophy in Oxford, and also one of the Royal Society, and of the renewned college of physicians in London. Englished by S. Pordage, student in physick*. Englished by S. Pordage, student in physick. Imprint: London: printed for Thomas Dring at the Harrow near Chancery-Lane End in Fleetstreet Ch. Harper at the Flower-de-Luce against St. Dunstan's Church in Fleet-street, and John Leigh at Stationers-Hall, 1683.

Wilson, E.B. (1952) *An Introduction to Scientific Research*. New York: McGraw-Hill.

Zed, P.J., P.S. Loewen and G. Robinson (1999) Medication-induced headache: overview and systematic review of therapeutic approaches. *Annals of Pharmacotherapy* **33**, 61–72.

Zelen, M. (1969) Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* **64**, 131–46.

Zellner, A. (1984) Causality and econometrics. In *Basic Issues in Econometrics*, Ch. 1, pp. 35–74. Chicago and London: University of Chicago Press.

This page intentionally left blank

Part II
Forecasting

This page intentionally left blank

4

Forecast Combination and Encompassing

Michael P. Clements and David I. Harvey

Abstract

Forecast combination is often found to improve forecast accuracy. This chapter considers different types of forecast combination and tests of forecast encompassing. The latter indicate when a combination is more accurate than an individual forecast *ex post*, in a range of circumstances: when the forecasts themselves are the objects of interest; when the forecasts are derived from models with unknown parameters; and when the forecast models are nested. We consider forecast encompassing tests which are framed in terms of the model's estimated parameters and recognize that parameter estimation uncertainty affects forecast accuracy, as well as conditonal tests of encompassing. We also look at the conditions under which forecast encompassing can be established irrespective of the form of the loss function.

4.1 Introduction

In this chapter we consider the closely related topics of forecast combination and tests of forecast encompassing. A test for forecast encompassing is just one of the many tests of predictive ability that might be considered. For example, one might test whether a set of forecasts are unbiased, or efficient (e.g., Mincer and Zarnowitz, 1969; Figlewski and Wachtel, 1981; Zarnowitz, 1985; Keane and Runkle, 1990) or as accurate given some loss function as a rival set of forecasts. Assuming a squared-error loss function, the approach of Granger and Newbold (1977) (sometimes known as the Morgan–Granger–Newbold test in recognition of Morgan, 1939) is one such test of whether differences between rival forecasts can

be attributed to sampling variability, or whether the differences are statistically significant once sampling variability has been taken into account. If we assume that the forecast errors are zero mean, normally distributed and serially uncorrelated (implying one-step-ahead forecasts), the Morgan–Granger–Newbold test is uniformly most powerful unbiased. A test of equal accuracy that dispenses with the restrictive assumptions that underpin this test, including that of squared-error loss, is due to Diebold and Mariano (1995). Although assessing the relative accuracy of forecasts is fundamental to forecast evaluation, from a practical or operational perspective the most useful way of doing this is not to test the null of equal accuracy but to test whether one set of forecasts encompasses the rival set. A set of forecasts is said to encompass a rival set if the rival set of forecasts do not contribute to a statistically significant reduction in forecast loss when used in combination with the original set of forecasts. Forecast encompassing is due to Chong and Hendry (1986) and is an application of the principle of encompassing (see, e.g., Mizon and Richard, 1986; Hendry and Richard, 1989) to the evaluation of forecasts, although it is formally equivalent to the earlier notion of conditional efficiency of Nelson (1972) and Granger and Newbold (1973).

The importance of testing for forecast encompassing (as opposed to equal accuracy) is that forecast combination is often found to improve forecast accuracy. That is, a linear combination of two or more forecasts may often yield more accurate forecasts than using a single forecast. If one is prepared to take a combination of forecasts rather than requiring that only one is selected, then it matters little whether one forecast is more or less accurate than another. Regression-based tests of forecast encompassing are a way of testing whether *ex post* a linear combination of forecasts results in a statistically significant reduction in (say) mean squared forecast error relative to using either individual forecast. Such tests can also be used as an indicator of when combinations might be useful *ex ante*. If neither set of forecasts encompasses the other, then subsequent forecasts should be constructed as a combination of those of the two individual sets.

If the goal is forecasting (as opposed to some other econometric endeavor, such as building a model that describes or quantifies the relationships between economic variables), then we suspect it would seldom be the case that the forecaster would be unwilling to take a combination of the available forecasts and would instead insist on selecting the single best forecasting model, or set of forecasts. Granger and Jeon (2004) present forecast combination as an example of "thick modeling," whereby the investigator pools the values of estimates of interest (parameter estimates, impulse responses, or forecasts, etc.) from a number of alternative specifications rather than seeking to select a single specification. Forecast combination has a long history: Granger and Jeon (2004) advocate the principle of thick modeling more generally. The principle of encompassing would suggest forecast encompassing be used to improve the forecasting model in the sense of re-specifying the model in an attempt to match the process that generated the data as closely as possible. However, it has been established that, even if this goal were attainable, it would not necessarily be beneficial from a forecasting perspective when there are breaks (see, e.g., Clements and Hendry, 2006).

There are a number of explanations as to why forecast combination works. Perhaps the most common is the portfolio diversification argument that underpinned the original analysis of Bates and Granger (1969), as recently discussed by Granger and Jeon (2004) and Timmermann (2006), amongst others. The idea is simply that the individual forecasts are each based on partial, and incompletely overlapping, information sets, as might be the case if they reflect private information, for example. The degree of overlap in the information sets is key, as is apparent in the discussion of Bates and Granger (1969) in section 4.2. An explanation stressed by Hendry and Clements (2004) is that of forecasts based on misspecified models when there are structural breaks and, as noted by Timmermann (2006, p. 138), a number of other papers discuss the roles of model misspecification and structural breaks. Hendry and Clements (2004) and Timmermann (2006) discuss a number of other reasons that would justify pooling.

Our survey is the latest of a number of recent reviews of the large literatures on the topics of forecast combination and on forecast encompassing. These include Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002) and Timmermann (2006). One of the key ways in which it differs from the others is the emphasis on the testing of forecast encompassing alongside the treatment of forecast combination. We are also able to include some of the important developments that have only recently found their way into the literature. For expositional convenience we focus on two forecasts, but in general more than two forecasts may be combined, and the notion of forecast encompassing can be generalized to the case of multiple forecasts (see Harvey and Newbold, 2000).

The plan of the rest of the chapter is as follows. Section 4.2 outlines the historical development of forecast combination and encompassing, and fills in some of the details. Section 4.3 describes the key developments when the forecasts are based on models and one wishes to compare the forecasting models. As the forecasts are generated from models in which the unknown parameters are replaced with estimates, an allowance is made for the true values having been replaced by random variables when the forecasts are compared. Section 4.4 considers forecasting from nested models, which is not covered by the analysis in section 4.3 and requires a separate treatment. Section 4.5 describes forecast encompassing tests within a framework of conditional testing of predictive ability, and where the emphasis shifts to testing forecasting *methods* rather than *models*. Thus far, we have maintained an assumption of symmetry of the loss function: the implications of dispensing with this assumption form the material of section 4.6. Section 4.7 offers some concluding remarks.

4.2 Historical development

4.2.1 Forecast combination

The notion of combining different forecasts of the same quantity in order to improve predictive accuracy was first proposed by Bates and Granger (1969). Suppose we have available two h-steps-ahead forecasts, f_{1t} and f_{2t}, of the quantity y_t. In this section, in line with the early literature on forecast combination (and

encompassing), we take the forecasts as given, and do not consider issues related to the method employed in obtaining the predictions, for example those of model estimation uncertainty and nesting considered in later sections. Assuming the forecasts to be unbiased, i.e., that the forecast errors $e_{it} = y_t - f_{it}$ $(i = 1, 2)$ have zero mean, Bates and Granger (1969) suggest the use of a combined forecast, f_{ct}, of the following form:

$$f_{ct} = (1 - \lambda) f_{1t} + \lambda f_{2t}. \tag{4.1}$$

When $0 \leq \lambda \leq 1$, f_{ct} comprises a simple weighted average of the two individual forecasts.

The optimal choice for the weighting parameter λ depends on the relative accuracy of the individual forecasts f_{1t} and f_{2t}, and can easily be obtained for a given loss, or cost of error, function. By far the most commonly assumed cost of error function in the literature is that of squared error loss, with forecast accuracy determined by the mean squared forecast error (MSFE) measure. Denoting the forecast error associated with f_{ct} by $\varepsilon_t = y_t - f_{ct}$, we obtain the following expression for the MSFE of the combined forecast:

$$E(\varepsilon_t^2) = (1 - \lambda)^2 \sigma_1^2 + \lambda^2 \sigma_2^2 + 2\lambda(1 - \lambda)\rho\sigma_1\sigma_2, \tag{4.2}$$

where σ_1^2 and σ_2^2 denote, respectively, the MSFEs of f_{1t} and f_{2t}, and ρ denotes the correlation between the forecast errors e_{1t} and e_{2t}. The optimal combination weight associated with a squared error loss function is then derived by choosing λ to minimize (4.2), i.e.:

$$\lambda_{opt} = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{4.3}$$

The expected squared error associated with the optimal combination weight λ_{opt} is given by:

$$E\big(\varepsilon_t^2(\lambda_{opt})\big) = \frac{\sigma_1^2\sigma_2^2(1 - \rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2},$$

where, of necessity, $E\big(\varepsilon_t^2(\lambda_{opt})\big) \leq \min\left\{\sigma_1^2, \sigma_2^2\right\}$. Suppose that f_1 and f_2 are equally accurate, i.e., $\sigma_1^2 = \sigma_2^2$. Then:

$$E\big(\varepsilon_t^2(\lambda_{opt})\big) = \frac{1}{2}\sigma_1^2(1 + \rho).$$

Given that $|\rho| \leq 1$, $\frac{1}{2}\sigma_1^2(1 + \rho) < \sigma_1^2$ for all values of ρ other than $\rho = 1$. So there are diversification gains when the forecasts are equally accurate unless the forecasts are perfectly correlated.

In practice the optimal weight parameter, and its constituent parameters ρ, σ_1^2 and σ_2^2, will not be known. However, given time series data on past forecasts and actuals, these quantities can be estimated, resulting in a sample analogue of the population weight parameter (4.3). Denoting the time series of past h-steps-ahead

forecast errors by e_{1t}, e_{2t}, $t = 1, \ldots, n$, the obvious estimator is given by:

$$\hat{\lambda}_{opt} = \frac{\sum_{t=1}^{n} e_{1t}^2 - \sum_{t=1}^{n} e_{1t}e_{2t}}{\sum_{t=1}^{n} e_{1t}^2 + \sum_{t=1}^{n} e_{2t}^2 - 2\sum_{t=1}^{n} e_{1t}e_{2t}}. \tag{4.4}$$

This estimated weight can then be used in the future to produce out-of-sample combined forecasts. The above estimator could equally be obtained from ordinary least squares estimation of the regression:

$$e_{1t} = \lambda(e_{1t} - e_{2t}) + \varepsilon_t, \tag{4.5}$$

or, equivalently:

$$y_t = (1 - \lambda)f_{1t} + \lambda f_{2t} + \varepsilon_t, \tag{4.6}$$

which is essentially a rearrangement of (4.1) with the error of the combined forecast, ε_t, now interpreted as a regression error.

The above analysis assumes that the individual forecasts are unbiased. However, it may well be the case that the individual forecasts are biased, as has been observed empirically for some macroeconomic forecasts (see, e.g., Zarnowitz and Braun, 1993; Stekler, 2002; Harvey and Newbold, 2003). In order to allow for bias in the forecasts, an intercept can be added to the regression (4.5) or (4.6):

$$e_{1t} = \alpha + \lambda(e_{1t} - e_{2t}) + \varepsilon_t,$$

which ensures that the implied combination:

$$f_{ct} = \alpha + (1 - \lambda)f_{1t} + \lambda f_{2t}, \tag{4.7}$$

is unbiased.

In addition to the possibility of biased forecasts, forecasts may also be inefficient in the sense of Mincer and Zarnowitz (1969). A generic forecast f_t is said to be Mincer–Zarnowitz efficient if $\alpha = 0$ and $\beta = 1$ in a regression $y_t = \alpha + \beta f_t + \varepsilon_t$, which implies that the forecast and forecast error are uncorrelated (see, e.g., Clements and Hendry, 1998, Ch. 3, for a discussion). If the individual forecasts are inefficient, the appropriate generalization of the combined forecast involves relaxing the implicit assumption that the combination weights sum to one. This results in an efficient combined forecast, with the more general formulation advocated by Granger and Ramanathan (1984):

$$f_{ct} = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t}. \tag{4.8}$$

The weights are now obtained from the corresponding regression:

$$y_t = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} + \varepsilon_t. \tag{4.9}$$

Clearly (4.7) and (4.1) are special cases of (4.8), where the restrictions $\beta_1 + \beta_2 = 1$, and $\alpha = 0, \beta_1 + \beta_2 = 1$ are imposed, respectively. Note that when the actuals and forecasts are non-stationary integrated processes, (4.9) should be specified using actual and predicted changes, rather than levels.

The above methods of combining forecasts are simple and straightforward to implement, but the literature also contains many extensions to these simple approaches, the relatively early contributions of which are summarized by Clemen (1989). For example, Diebold (1988) and Coulson and Robins (1993) argue that, when estimating weights based on (4.9), account should be taken of the likely autocorrelation present in ε_t, either by allowing for ARMA residuals or including y_{t-1} as an additional regressor. The possibility of time-varying combination weights could also be entertained, reflecting the potentially evolving behavior of the process governing the actuals, and also of the individual forecasters. At a simple level, this might involve using recent data only to estimate the combination weights, while more sophisticated approaches are proposed by Diebold and Pauly (1987), LeSage and Magura (1992), and Deutsch, Granger and Teräsvirta (1994).

Relatively sophisticated methods of forecast combination inherently entail a greater data requirement, and are therefore most applicable when a reasonably long history of forecast performance is available. When, as is very often the case, only small samples of historical data exist, sampling variability plays a significant role in the estimation of the combination weights. This can temper the gains that could be realized relative to when the weights are known, potentially even giving rise to forecast combinations that are less accurate than simpler approaches that do not require combination weight estimation. For example, many authors (see, e.g., Makridakis and Winkler, 1983; Stock and Watson, 1999; Fildes and Ord, 2002) have found that simple averaging of individual forecasts very often outperforms more elaborate combination techniques, while Harvey and Newbold (2005) demonstrate that situations exist where the optimal weight on, say, f_{2t} is non-zero, but sampling variability affects the weight estimates to the extent that the resulting combination has a larger MSFE than that associated with just f_{1t} alone. The Bayesian combination methods of, *inter alia*, Clemen and Winkler (1986), Diebold and Pauly (1990) and Min and Zellner (1993) provide a means of formally estimating the combination weights, while mitigating the effects of sampling variability by shrinking the weights towards some prior mean. The widely observed robust performance of simple averages of forecasts motivates a prior of equal weights in this setting.

Another extension to combining forecasts is to allow nonlinear combination methods. Such methods may be useful when relatively large samples of forecasts are available, and/or when the nature of the forecasts suggests methods other than linear combination. Given a large number of forecasts, an attractive way of considering nonlinear combination schemes is via Artificial Neural Networks (ANNs), as ANNs are able to approximate large classes of nonlinear functions. Donaldson and Kamstra (1996) use single hidden-layer ANNs to combine forecasts of the volatility of daily stock returns from GARCH and moving-average variance models, and compare the results to traditional linear combination. Specifically, the ANNs are of the form:

$$f_{ct} = \alpha + \sum_{j=1}^{k} \beta_j f_{jt} + \sum_{i=1}^{p} \delta_i \Psi\left(z_t \gamma_i\right),$$

where:

$$\Psi\left(z_t\gamma_i\right) = \left(1 + \exp\left[-\left(\gamma_{0i} + \gamma_{1i}z_{1t} + \gamma_{2i}z_{2t}\right)\right]\right)^{-1},$$

and $z_{it} = s_y^{-1}\left(f_{it} - \bar{y}\right)$, $i = 1, 2$, i.e., the f_{it} are normalized by the in-sample mean and standard deviation of y_t. When $p = 0$ and $k = 2$, the ANN specializes to standard linear combination of the two sets of forecasts. Donaldson and Kamstra (1996) allow values of p up to $p = 3$, and find values of 1 or 2 are typically selected by their cross-validation procedure. The combinations are estimated by choosing the γ_i as random drawings from a $U(-1, 1)$ distribution, and then estimation of α, β, δ can be carried out by ordinary least squares (OLS).

Finally, for forecasts other than point forecasts, other forms of combination are used. For example, experts' subjective probability distributions are often combined using the logarithmic opinion pool, or LoOP (see, e.g., Genest and Zidek, 1986; Clemen and Winkler, 1999), which for discrete probability distributions is given by:

$$f^j = \frac{\prod_{i=1}^N \left(f_i^j\right)^{\beta_i}}{\sum_{j=1}^M \prod_{i=1}^N \left(f_i^j\right)^{\beta_i}} = \frac{\exp\left(\sum_{i=1}^N \beta_i \log f_i^j\right)}{\sum_{j=1}^M \exp\left(\sum_{i=1}^N \beta_i \log f_i^j\right)},$$

where f_i^j is individual i's probability of "class c_j," where there are M classes. The denominator is a scaling factor, and typically $\sum_{i=1}^N \beta_i = 1$. Clements and Harvey (2007) consider a form of LoOP for probability forecasts ($M = 2$) and two rival forecasts ($N = 2$), and Kamstra and Kennedy (1998) (henceforth, KK) suggest a form of combination for probability forecasts that involves the combining of log odds ratios by logit regressions. The KK combination of f_{1t} and f_{2t} is:

$$\begin{aligned} f_{ct} &= \frac{\exp\left[\alpha + \beta_1 \ln\left(\frac{f_{1t}}{1-f_{1t}}\right) + \beta_2 \ln\left(\frac{f_{2t}}{1-f_{2t}}\right)\right]}{1 + \exp\left[\alpha + \beta_1 \ln\left(\frac{f_{1t}}{1-f_{1t}}\right) + \beta_2 \ln\left(\frac{f_{2t}}{1-f_{2t}}\right)\right]} \\ &= \frac{\exp(\alpha)\left(\frac{f_{1t}}{1-f_{1t}}\right)^{\beta_1}\left(\frac{f_{2t}}{1-f_{2t}}\right)^{\beta_2}}{1 + \exp(\alpha)\left(\frac{f_{1t}}{1-f_{1t}}\right)^{\beta_1}\left(\frac{f_{2t}}{1-f_{2t}}\right)^{\beta_2}}, \end{aligned} \tag{4.10}$$

where β_1 and β_2 are the maximum likelihood estimates of the slope coefficients from a logit regression of y_t on a constant, $\ln\left(\frac{f_{1t}}{1-f_{1t}}\right)$ and $\ln\left(\frac{f_{2t}}{1-f_{2t}}\right)$. Clements and Harvey (2007) show that the KK combination is optimal when the data-generating process has the form:

$$\begin{aligned} y_t &= 1\left(\frac{\exp(\delta_1 X_{1t} + \delta_2 X_{2t})}{1 + \exp(\delta_1 X_{1t} + \delta_2 X_{2t})} > v_t\right) \\ f_{1t} &= \frac{\exp(\theta_{11} X_{1t})}{1 + \exp(\theta_{11} X_{1t})} \\ f_{2t} &= \frac{\exp(\theta_{12} X_{2t})}{1 + \exp(\theta_{12} X_{2t})}, \end{aligned} \tag{4.11}$$

with X_{1t} and X_{2t} scalar explanatory variables underpinning the two forecasts, although KK combination is a computationally attractive method of combining forecasts whilst ensuring $f_t \in (0, 1)$ more generally.

4.2.2 Forecast encompassing

The concept of forecast encompassing relates to whether or not one forecast encapsulates all the useful predictive information contained in a second forecast. Formally, using a squared error loss function as above, f_{1t} is said to encompass f_{2t} if, in a linear combination of the two forecasts, f_{2t} optimally receives zero weight, so that combining f_{1t} with f_{2t} does not lead to a reduction in the MSFE. Thus, using the simplest Bates and Granger (1969) form of linear combination, f_{1t} encompasses f_{2t} if the optimal value of λ in (4.1) is zero. This concept was originally proposed by Nelson (1972) and Granger and Newbold (1973), with f_{1t} referred to as being conditionally efficient with respect to f_{2t}; the terminology has subsequently been modified by Chong and Hendry (1986) to that of forecast encompassing, adopting the language of the model encompassing literature (see, *inter alia*, Mizon, 1984; Mizon and Richard, 1986).

Chong and Hendry (1986) focus on the fact that, if f_{1t} encompasses f_{2t}, the forecast errors of the encompassing forecast, e_{1t}, should be uncorrelated with the encompassed forecast f_{2t}. This obtains since e_{1t} should be uncorrelated with information available at the time of the forecast, while, on the other hand, correlation between e_{1t} and f_{2t} would imply that the accuracy of f_{1t} could be improved by linear combination with f_{2t}. This approach results in an alternative definition of forecast encompassing, namely that f_{1t} encompasses f_{2t} if the optimal value of λ is zero in a regression:

$$e_{1t} = \lambda f_{2t} + \varepsilon_t.$$

The two definitions of forecast encompassing presented thus far implicitly assume that the forecasts are unbiased and efficient. To account for potential forecast bias and inefficiency, alternative forecast encompassing specifications can also be derived by defining encompassing as f_{2t} receiving zero optimal weight in a more general forecast combination such as (4.7) or (4.8). Both approaches have been proposed in the literature, with Fair and Shiller (1989, 1990) using (4.8), allowing for the forecasts to be both biased and inefficient, and Andrews, Minford and Riley (1996) advocating use of (4.7), an in-between case allowing for forecast bias while retaining the assumption that the combination weights sum to one.

The alternative forecast encompassing definitions can be summarized and given a regression interpretation as follows. Beginning with the most general formulation, the Fair and Shiller (1989) definition of f_{1t} encompassing f_{2t} equates to $\beta_2 = 0$ in the regression:

$$\text{FE(1):} \qquad y_t = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} + \varepsilon_t.$$

Relative to this specification, the Nelson (1972) and Granger and Newbold (1973) approach imposes the restrictions $\alpha = 0$ and $\beta_1 + \beta_2 = 1$, with encompassing defined by $\lambda = 0$ in the regression:

$$\text{FE(2):} \qquad e_{1t} = \lambda(e_{1t} - e_{2t}) + \varepsilon_t.$$

The Chong and Hendry (1986) definition can be obtained by assuming f_{1t} in FE(1) to be efficient, i.e., imposing $\alpha = 0$ and $\beta_1 = 1$, with encompassing then defined by $\lambda = 0$ in the regression:

$$\text{FE(3):} \qquad e_{1t} = \lambda f_{2t} + \varepsilon_t.$$

The FE(2) and FE(3) cases can also be modified to allow for forecast bias by relaxing the $\alpha = 0$ assumption, yielding the Andrews, Minford and Riley (1996) regression:

$$\text{FE(2}'\text{):} \qquad e_{1t} = \alpha + \lambda(e_{1t} - e_{2t}) + \varepsilon_t,$$

and a modified Chong–Hendry regression:

$$\text{FE(3}'\text{):} \qquad e_{1t} = \alpha + \lambda f_{2t} + \varepsilon_t.$$

In the remainder of this chapter, we focus on the more commonly used definitions FE(1), FE(2) and FE(3).

It is clear that if the restrictions imposed by FE(2) and FE(3) are not satisfied, the three definitions of encompassing are not equivalent, and the forecast f_{1t} may encompass f_{2t} according to one definition, but not another. FE(1) is the most general approach of the three, and allows an analysis of the forecast encompassing hypothesis without the additional requirements that the individual forecasts are unbiased and efficient. Note that if the optimal value of β_2 in FE(1) is zero, it does not follow that the optimal forecast is simply f_{1t}: the correct inference is that a *linear function of* f_{1t}, i.e., $\alpha + \beta_1 f_{1t}$, cannot be improved (in terms of MSFE) through combination with f_{2t}. If the restrictions underlying FE(2) and FE(3) do hold, tests based on these approaches should be more powerful.

When the actuals and forecasts are integrated time series processes, FE(1) should be implemented using actual and predicted changes rather than the levels of the series, as otherwise the test statistics will not have their standard distributions. When the data (and forecasts) are integrated, the FE(3) approach can be problematic, as noted by Ericsson (1992). Because a forecast f_{1t} would be expected to be cointegrated with y_t, the resulting forecast error e_{1t} is integrated of order zero. The regression FE(3) is then unbalanced in the sense that the dependent and explanatory variables have differing orders of integration, and Phillips (1995) shows that the resulting forecast encompassing tests have an asymptotic size of one.

Tests of the null hypothesis of forecast encompassing can be conducted using any of the above definitions. In terms of the more general FE(1) definition, the null and alternative can be expressed as:

$$\begin{aligned} H_0: &\quad \beta_2 = 0 \quad (f_{1t} \text{ encompasses } f_{2t}) \\ H_1: &\quad \beta_2 > 0 \quad (f_{1t} \text{ does not encompass } f_{2t}). \end{aligned}$$

The alternative hypothesis is often chosen to be one-sided, to rule out the possibility of negative combination weights. Note, however, that negative weights can arise: from (4.3) it is apparent that the weight on f_{2t} in the Bates–Granger combination will be negative if $\sigma_2\rho > \sigma_1$ (and the weight on f_{1t} will exceed unity). For

FE(2) and FE(3), the null and alternative hypotheses take the same form as above but with the parameter λ replacing β_2.

The most straightforward approach to testing would be to simply estimate the relevant regression, FE(1), FE(2) or FE(3), by OLS, and then perform a standard t-test of the null that $\beta_2 = 0$ or $\lambda = 0$. However, as Harvey, Leybourne and Newbold (1998) show in the context of the FE(2) regression, such an approach is not robust to properties of the forecast errors that one might expect to encounter in practice. First, an implicit assumption of the regression error ε_t being identically and independently distributed (i.i.d.) is not plausible for forecasts at horizons greater than one, since even optimal forecasts in this setting would be expected to have errors that follow a moving-average process of order $h-1$. Second, in some applications it is likely that the forecast errors are non-normally distributed (for example, Harvey and Newbold, 2003, find substantial evidence of non-normality in US macroeconomic forecast errors). Non-normality in the errors induces conditional heteroskedasticity in the regression FE(2), resulting in oversized tests if conventional t-tests are applied.

Harvey, Leybourne and Newbold (1998) consider two ways of obtaining asymptotically correctly-sized tests in these situations. One is to continue with a regression-based t-test, but using standard errors that are robust to heteroskedasticity and autocorrelation. Specifically, assuming the forecast errors are at most lag $(h-1)$ dependent (in line with forecast optimality), they propose the use of a rectangular lag window for long-run variance estimation, as in Diebold and Mariano (1995). This approach yields the test statistic:

$$R_1 = \frac{\hat{\lambda}}{\sqrt{\dfrac{\sum_{j=-(h-1)}^{h-1} \sum_{t=|j|+1}^{n} (e_{1t} - e_{2t})\hat{\varepsilon}_t (e_{1,t-|j|} - e_{2,t-|j|})\hat{\varepsilon}_{t-|j|}}{\left[\sum_{t=1}^{n} (e_{1t} - e_{2t})^2\right]^2}}},$$

where $\hat{\lambda}$ is the least squares estimate of λ in the FE(2) regression. Alternatively, one could impose the information that $\lambda = 0$ under the null, replacing $\hat{\varepsilon}_t$ with e_{1t} in the variance estimator:

$$R_2 = \frac{\hat{\lambda}}{\sqrt{\dfrac{\sum_{j=-(h-1)}^{h-1} \sum_{t=|j|+1}^{n} (e_{1t} - e_{2t})e_{1t}(e_{1,t-|j|} - e_{2,t-|j|})e_{1,t-|j|}}{\left[\sum_{t=1}^{n} (e_{1t} - e_{2t})^2\right]^2}}}.$$

Under standard assumptions about the forecast errors, both R_1 and R_2 are asymptotically standard normally distributed.

The second approach proposed by Harvey, Leybourne and Newbold (1998) observes that the null hypothesis under the FE(2) specification requires:

$$E[e_{1t}(e_{1t} - e_{2t})] = 0.$$

This motivates testing for forecast encompassing via a test for whether the series $d_t = e_{1t}(e_{1t} - e_{2t})$ has zero mean, along the lines of Diebold and Mariano (1995),

who presented such an approach for testing the null of equal forecast accuracy. Under standard assumptions:

$$\sqrt{n}[\bar{d} - E(d_t)] \Rightarrow N(0, S), \tag{4.12}$$

where S denotes the long-run variance of d_t, giving rise to the statistic:

$$DM = \frac{n\bar{d}}{\sqrt{\sum_{j=-(h-1)}^{h-1} \sum_{t=|j|+1}^{n} (d_t - \bar{d})(d_{t-|j|} - \bar{d})}}, \tag{4.13}$$

where $\bar{d} = n^{-1} \sum_{t=1}^{n} d_t$ and the implied estimator of S uses a rectangular lag window as above. This statistic has an asymptotic standard normal distribution under the null of forecast encompassing, and is robust to the aforementioned forecast error properties of autocorrelation and non-normality. Harvey, Leybourne and Newbold (1998) propose a small modification of this test which has improved finite sample properties, drawing on work by Harvey, Leybourne and Newbold (1997). The modified statistic is:

$$MDM = n^{-1/2}[n + 1 - 2h + n^{-1}h(h-1)]^{1/2} DM, \tag{4.14}$$

and the recommendation is to use critical values from the t_{n-1} distribution rather than those from the limiting standard normal. Simulation results in Harvey, Leybourne and Newbold (1998) show that *MDM* has better finite-sample size properties than the regression-based variants R_1 and R_2, although some loss in size-adjusted power relative to R_1 is observed for small samples.

In addition to forecast errors being autocorrelated (for $h > 1$) and possibly non-normally distributed, it may also be the case that the errors exhibit autoregressive conditional heteroskedasticity (ARCH – see, e.g., Engle, 1982), with the squared forecast errors following a dependent sequence. Intuitively, this describes the situation where, if a variable proves difficult to forecast in one period, it is likely to prove difficult to forecast in the next period as well. In such circumstances, and again in the context of FE(2), Harvey, Leybourne and Newbold (1999) show that the forecast encompassing tests suffer asymptotic size distortions, rejecting the forecast encompassing null too frequently. They also propose a simple modification which largely overcomes the size problem; this involves computing *MDM* as above, but replacing h in (4.13) and (4.14) with $\lfloor 0.5n^{1/3} \rfloor + h$, where $\lfloor . \rfloor$ denotes integer part. This modification should be employed whenever ARCH in the forecast errors is suspected, or is detected through prior testing.

Although Harvey, Leybourne and Newbold (1998) propose the *MDM* test (4.14) in the context of FE(2), the test can also be applied using the forecast encompassing specifications FE(1) and FE(3), the only difference being the specification of d_t. For FE(1), application of the Frisch–Waugh theorem shows that β_2 is identical to that in the regression $\eta_{1t} = \beta_2 \eta_{2t} + \nu_t$, where η_{1t} and η_{2t} denote the errors from regressions of y_t and f_{2t}, respectively, on a constant and f_{1t}. This allows us to write the null hypothesis as:

$$E(\eta_{1t}\eta_{2t}) = 0,$$

with the corresponding specification for d_t being $d_t = \eta_{1t}\eta_{2t}$; in practice, η_{1t} and η_{2t} can be replaced with their residual counterparts $\hat{\eta}_{1t}$ and $\hat{\eta}_{2t}$, respectively. For FE(3), testing can proceed by setting $d_t = e_{1t}f_{2t}$.[1] Variants of R_1 and R_2 can also be constructed for FE(1) and FE(3) in a straightforward manner; see, for example, Newbold and Harvey (2002) for FE(1).

Tests for forecast encompassing can also be devised for the ANN combination. This would require the computationally more burdensome estimation of $\theta = \left(\alpha, \beta_1, \ldots \beta_k, \gamma_1, \ldots \gamma_p, \delta_1, \ldots \delta_p\right)$ by NLS, rather than choosing the γ_i as random draws from a $U(-1, 1)$ distribution: i.e., minimizing $Q_n(\theta) = \sum_{t=1}^{n}\left[y_t - f_t(\theta)\right]^2$ to give $\hat{\theta}_n$. We can then use the results that, under general conditions, $\hat{\theta}_n$ converges to θ^*, where:

$$\theta^* = \arg\min_{\theta} E\left[y_t - f_t(\theta)\right]^2,$$

and that $\sqrt{n}\left(\hat{\theta} - \theta^*\right) \Rightarrow N\left(0, \Omega_\theta\right)$, where Ω_θ can be consistently estimated, to conduct inference. The null hypothesis that f_{1t} encompasses f_{2t} based on the ANN combination can be constructed as a test of the joint significance of all the parameters related to f_{2t}, i.e., $\beta_2 = \gamma_{21} = \ldots = \gamma_{2p} = 0$: see White (1989), Kuan and White (1994) and the discussion by Franses and van Dijk (2000, pp. 230–2) for details.

4.3 Model-based forecasts

The analysis and forecast encompassing tests considered in the previous section treat the forecasts as given. However, in many practical applications forecasts are obtained using estimated regression models, and the impact of estimation uncertainty on the encompassing tests then needs to be examined if we wish to assess the predictive ability of the underlying models. West and McCracken (1998) and West (2001) study the impact of estimation uncertainty for the forecast encompassing specifications FE(3) and FE(2) respectively, drawing on the work of West (1996), although the general results are equally applicable to FE(1). Suppose, by way of a simple example, that the forecasts f_{1t} and f_{2t} are generated using the non-nested regression models:

$$\begin{aligned} \text{Model 1:} \quad & y_t = \theta_1 X_{1t} + e_{1t} \\ \text{Model 2:} \quad & y_t = \theta_2 X_{2t} + e_{2t}, \end{aligned}$$

where the scalar regressors X_{1t} and X_{2t} are assumed to be stationary and well behaved, and where $E(e_{1t}X_{1t}) = E(e_{2t}X_{2t}) = 0$. Given estimates of the model parameters ($\hat{\theta}_{1t}$ and $\hat{\theta}_{2t}$) using data prior to time t, the corresponding forecasts can then be constructed as:

$$\begin{aligned} \hat{f}_{1t} &= \hat{\theta}_{1t} X_{1t} \\ \hat{f}_{2t} &= \hat{\theta}_{2t} X_{2t}. \end{aligned}$$

As in the previous section, assume that a record of n past forecasts and actuals are available for evaluation, denoting the corresponding forecast errors by $\hat{e}_{1t}$ and $\hat{e}_{2t}$. The n forecasts can be derived using model parameter estimates obtained in one of three main ways. First, a *fixed* estimation scheme involves a one-off estimation of $\hat{\theta}_{1t}$ and $\hat{\theta}_{2t}$ using data from, say, $t = 1, \ldots, R$, and then using that same set of estimates to produce n forecasts from $t = R + h$ to $R + n + h - 1$. Second, a *recursive* estimation scheme might be adopted, where the sample used for estimation uses all available information at each point, increasing by one observation per period, i.e., the models are first estimated over $t = 1, \ldots, R$ to produce forecasts for $t = R + h$, then the model parameters are re-estimated over $t = 1, \ldots, R + 1$ to give forecasts for $t = R + 1 + h$, etc. Finally, a *rolling* scheme uses a moving window of R observations to estimate the models, so that recent data is included, but more distant observations discarded, i.e., the initial estimation sample is again $t = 1, \ldots, R$ for forecasts of the period $t = R + h$, then $t = 2, \ldots, R + 1$ for use in forecasts for $t = R + 1 + h$, etc.

The encompassing tests of the previous section must now be constructed using $\hat{d}_t$, defined as d_t for the appropriate forecast encompassing specification FE(1), FE(2) or FE(3), but based on the quantities $\hat{f}_{1t}$, $\hat{f}_{2t}$, $\hat{e}_{1t}$ and $\hat{e}_{2t}$, which embody the estimated parameters $\hat{\theta}_{1t}$ and $\hat{\theta}_{2t}$. The results of West and McCracken (1998) and West (2001) show that the additional uncertainty implicit in $\hat{d}_t$ affects the asymptotic variance of $\bar{\hat{d}}$. We now have:

$$\sqrt{n}[\bar{\hat{d}} - E(d_t)] \Rightarrow N(0, \Omega), \tag{4.15}$$

where $\bar{\hat{d}} = n^{-1} \sum_{t=R+h}^{R+n+h-1} \hat{d}_t$ and:

$$\Omega = S + \delta_{dg}(DBS'_{dg} + S_{dg}B'D') + \delta_{gg}DBS_{gg}B'D', \tag{4.16}$$

with S denoting the long run variance of d_t as before, and:

$$\begin{aligned}
D &= E\left[\begin{array}{cc} \partial d_t/\partial\theta_1 & \partial d_t/\partial\theta_2 \end{array} \right] \\
B &= \left[\begin{array}{cc} [E(X_{1t}^2)]^{-1} & 0 \\ 0 & [E(X_{2t}^2)]^{-1} \end{array} \right] \\
S_{gg} &= \textstyle\sum_{j=-\infty}^{\infty} E(g_t g'_{t-j}), \qquad g_t = \left[\begin{array}{c} e_{1t}X_{1t} \\ e_{2t}X_{2t} \end{array} \right] \\
S_{dg} &= \textstyle\sum_{j=-\infty}^{\infty} E\{[d_t - E(d_t)]g'_{t-j}\},
\end{aligned}$$

and where the parameters δ_{dg} and δ_{gg} are given in the following table:

Estimation scheme	δ_{dg}	δ_{gg}
Fixed	0	π
Recursive	$1-\pi^{-1}\ln(1+\pi)$	$2\left[1-\pi^{-1}\ln(1+\pi)\right]$
Rolling, $\pi \leq 1$	$\pi/2$	$\pi-\pi^2/3$
Rolling, $\pi > 1$	$1-(2\pi)^{-1}$	$1-(3\pi)^{-1}$

with $\pi = \lim_{R,n\to\infty}(n/R), 0 \leq \pi < \infty$. Note that $BS_{gg}B'$ in the last term of (4.16) defines the asymptotic ($R \to \infty$) variance-covariance matrix of the estimator of the parameter vector $\theta = [\theta_1, \theta_2]'$, denoted V_θ. The above results can also be generalized beyond the example considered of scalar linear regression models estimated by least squares, provided the models continue to be non-nested. For results pertaining to multiple regressors in a linear framework, and also more general models and estimation techniques, see West and McCracken (1998) and West (2001). Essentially, (4.15) and (4.16) continue to hold, but involve more general representations for the constituent components of Ω.

Comparing (4.12) and (4.15)–(4.16), it can be seen that the uncertainty involved through estimation of the model parameters gives rise to additional terms in the asymptotic variance of $\bar{\hat{d}}$. To see how this arises, consider the FE(2) *MDM* test for the simple example above, assuming the forecasts have been obtained via the fixed estimation scheme, so that $\hat{\theta}_{it} = \hat{\theta}_i = \sum_{t=1}^{R} y_t X_{it} / \sum_{t=1}^{R} X_{it}^2$, $t = R+h, \ldots, R+n+h-1$, $i = 1, 2$. In this case, the expression (4.16) simplifies to $\Omega = S + \pi D V_\theta D'$ with $D = [E(e_{2t}X_{1t}), E(e_{1t}X_{2t})]$. Now the forecast errors can be written as $\hat{e}_{it} = e_{it} - (\hat{\theta}_i - \theta_i)X_{it}$, $i = 1, 2$, resulting in the decomposition:

$$\bar{\hat{d}} = \bar{d} + n^{-1}\sum_{t=R+h}^{R+n+h-1}\Big[(\hat{\theta}_1-\theta_1)e_{2t}X_{1t} + (\hat{\theta}_2-\theta_2)e_{1t}X_{2t} + (\hat{\theta}_1-\theta_1)^2X_{1t}^2 - (\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)X_{1t}X_{2t} - 2(\hat{\theta}_1-\theta_1)e_{1t}X_{1t}\Big].$$

It then follows that:

$$\begin{aligned} V\left[\sqrt{n}[\bar{\hat{d}} - E(d_t)]\right] &= V\left[\sqrt{n}[\bar{d} - E(d_t)]\right] + nE\Big\{n^{-2}\Big[(\hat{\theta}_1-\theta_1)^2\Big(\sum_{t=R+h}^{R+n+h-1} e_{2t}X_{1t}\Big)^2 \\ &\quad + (\hat{\theta}_2-\theta_2)^2\Big(\sum_{t=R+h}^{R+n+h-1} e_{1t}X_{2t}\Big)^2 \\ &\quad + 2(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)\sum_{t=R+h}^{R+n+h-1} e_{2t}X_{1t}\sum_{t=R+h}^{R+n+h-1} e_{1t}X_{2t}\Big]\Big\} + o_p(1) \\ &= V\left[\sqrt{n}[\bar{d} - E(d_t)]\right] + (n/R)E\Big\{[R^{1/2}(\hat{\theta}_1-\theta_1)]^2\Big(n^{-1}\sum_{t=R+h}^{R+n+h-1} e_{2t}X_{1t}\Big)^2 \\ &\quad + [R^{1/2}(\hat{\theta}_2-\theta_2)]^2\Big(n^{-1}\sum_{t=R+h}^{R+n+h-1} e_{1t}X_{2t}\Big)^2 \end{aligned}$$

$$+2R^{1/2}(\hat{\theta}_1-\theta_1)R^{1/2}(\hat{\theta}_2-\theta_2)(n^{-1}\textstyle\sum_{t=R+h}^{R+n+h-1} e_{2t}X_{1t})(n^{-1}\sum_{t=R+h}^{R+n+h-1} e_{1t}X_{2t})\Big\}$$

$$+ o_p(1)$$

$$\Rightarrow S + \pi\{V_{\theta,11}[E(e_{2t}X_{1t})]^2 + V_{\theta,22}[E(e_{1t}X_{2t})]^2 + 2V_{\theta,12}E(e_{2t}X_{1t})E(e_{1t}X_{2t})\}$$

$$= S + \pi D V_\theta D',$$

where $V_{\theta,ij}$ denotes the (i,j) element of V_θ.

If forecast encompassing tests are conducted without taking account of the additional terms present in Ω, i.e., by simply using $\hat{d}_t$ in place of d_t in the tests in the previous section, and using the usual implicit long run variance estimator that only estimates S, then asymptotic size distortions are generally obtained. For example, simulations by West (2001), using the example considered above with $\theta_1 = 1$, $\theta_2 = 0$ and $(e_{1t}, X_{1t}, X_{2t})'$ i.i.d. normal with variance-covariance matrix $diag(1,1,2)$, show that the FE(2) *MDM* test of the previous section, run at the nominal 5% significance level against a two-sided alternative, has empirical size around 25% for large n when $n/R = 2$.

In order to obtain asymptotically correctly-sized tests in general, the variance estimators implicit in the forecast encompassing tests must be modified so as to consistently estimate Ω. Consistent estimation of Ω can be obtained by estimating the constituent quantities S, S_{dg}, S_{gg}, B and D using their natural sample counterparts, and using n/R in place of π for determining δ_{dg} and δ_{gg}. To illustrate, consider again the FE(2) *MDM* test for the simple example above, assuming the forecasts have been obtained via the fixed estimation scheme. A further simplification of Ω is possible, since under the encompassing null hypothesis, e_{1t} cannot be predicted by model 2, so $E(e_{1t}X_{2t}) = 0$, yielding $D = [E(e_{2t}X_{1t}), 0]$ and:

$$\Omega = S + \pi V_{\theta,11}[E(e_{2t}X_{1t})]^2.$$

A consistent estimator is:

$$\hat{\Omega} = \hat{S} + (n/R)\hat{V}_{\theta,11}\left[n^{-1}\sum\nolimits_{t=R+h}^{R+n+h-1}\hat{e}_{2t}X_{1t}\right]^2,$$

where:

$$\hat{S} = [n+1-2h+n^{-1}h(h-1)]^{-1}\sum\nolimits_{j=-(h-1)}^{h-1}\sum\nolimits_{t=|j|+R+h}^{R+n+h-1}(\hat{d}_t-\overline{\hat{d}})(\hat{d}_{t-|j|}-\overline{\hat{d}}),$$

and $\hat{V}_{\theta,11}$ is the heteroskedasticity consistent estimator of the asymptotic variance of $\hat{\theta}_1$:

$$\hat{V}_{\theta,11} = \frac{R\sum_{t=1}^{R}\hat{e}_{1t}^2X_{1t}^2}{\left(\sum_{t=1}^{R}X_{1t}^2\right)^2}.$$

The resulting *MDM* statistic is then given by:

$$MDM = \frac{n^{1/2}\overline{\hat{d}}}{\sqrt{\hat{\Omega}}}.$$

There are, however, a number of special cases where the forecast encompassing tests do not require correction for model parameter estimation uncertainty. First, if $\pi = 0$, then the parameters δ_{dg} and δ_{gg} are zero regardless of the model estimation scheme, ensuring that Ω in (4.16) reduces to S. Thus if the ratio n/R is very small, a case could be made for abstracting from the issue of model estimation uncertainty and proceeding with the unadjusted tests of the previous section. Intuitively, this arises because, if R is very large compared to n, the model parameter estimation uncertainty becomes relatively insignificant compared to the uncertainty that would be present in the testing problem even if the model parameters were known.

A second case where model estimation corrections are not required is when the two models are linear, estimated by least squares, with each involving just a single regressor, as in the example above, and the forecast encompassing approach is FE(1). This case is considered by Clements and Harvey (2006). As noted in the previous section, for the FE(1) *MDM* test, $d_t = \eta_{1t}\eta_{2t}$, with η_{1t} and η_{2t} the errors from regressions of y_t and f_{2t}, respectively, on a constant and f_{1t}. Hence, letting $C(.,.)$ denote a covariance:

$$d_t = \left\{[y_t - E(y_t)] - \frac{C(y_t, f_{1t})}{V(f_{1t})}[f_{1t} - E(f_{1t})]\right\} \times \left\{[f_{2t} - E(f_{2t})] - \frac{C(f_{1t}, f_{2t})}{V(f_{1t})}[f_{1t} - E(f_{1t})]\right\}.$$

For forecasts from linear single regressor models, results such as $E\left(f_{it}\right) = \theta_i E\left(X_{it}\right)$, $i = 1, 2$, are obtained, so that d_t can be written as:

$$d_t = \theta_2 \left\{[y_t - E(y_t)] - \frac{C(y_t, X_{1t})}{V(X_{1t})}[X_{1t} - E(X_{1t})]\right\} \times \left\{[X_{2t} - E(X_{2t})] - \frac{C(X_{1t}, X_{2t})}{V(X_{1t})}[X_{1t} - E(X_{1t})]\right\}.$$

Clearly, $\partial d_t/\partial\theta_1 = 0$, and:

$$E\left(\frac{\partial d_t}{\partial \theta_2}\right) = C(y_t, X_{2t}) - \frac{C(X_{1t}, X_{2t})C(y_t, X_{1t})}{V(X_{1t})}. \tag{4.17}$$

The null hypothesis of $\beta_2 = 0$ in the FE(1) regression implies that:

$$V(f_{1t})C(y_t, f_{2t}) - C(f_{1t}, f_{2t})C(y_t, f_{1t}) = 0.$$

Substituting for f_{1t} and f_{2t} in this expression, and dividing both sides by $\theta_1^2\theta_2$ (noting that $\theta_1 \neq 0$, $\theta_2 \neq 0$), the right-hand side of (4.17) equals zero. So $D = (0, 0)$ under the null, and estimation uncertainty is irrelevant asymptotically for FE(1), contrasting with the findings using the *MDM* variants of the FE(2) and FE(3) tests.

Finally, West and McCracken (1998) show that while tests based on the FE(3) specification do require adjustment for estimation uncertainty, an alternative to directly estimating Ω exists when the models are linear. Consider augmenting

the FE(3) regression with the regressors from Model 1, i.e., in the above example, replacing the FE(3) regression with:

$$\hat{e}_{1t} = \lambda \hat{f}_{2t} + \gamma X_{1t} + \varepsilon_t.$$

West and McCracken (1998) show that tests of $\lambda = 0$ from this augmented regression only require standard autocorrelation and heteroskedasticity robust variance estimators, such as those of the previous section, and do not need estimators that explicitly account for the model parameter estimation uncertainty. In terms of an *MDM*-type testing approach, this augmented version of FE(3) can be executed by computing (4.14) with $\hat{d}_t = \hat{u}_{1t}\hat{u}_{2t}$ in place of d_t, where $\hat{u}_{1t}$ and $\hat{u}_{2t}$ denote the residuals from regressions of $\hat{e}_{1t}$ and $\hat{f}_{2t}$, respectively, on X_{1t}.

4.4 Nested model comparisons

The results of the previous section apply in the case where the rival forecasting models are non-nested. However, it is also common in forecast evaluation exercises for the forecasts under consideration to be obtained from models that are, in fact, nested. The primary situation where this arises is when forecast encompassing tests are employed to help determine whether a particular variable is useful for prediction, by testing whether a forecast based on a model including that variable as a regressor is encompassed by a forecast from the same model with that variable excluded. In such situations, the forecasts are asymptotically equivalent under the encompassing null hypothesis, and this affects the usual asymptotic results derived under a non-nested model assumption.

Clark and McCracken (2001) examine the asymptotic properties of FE(2)-based forecast encompassing tests for the special case of one-step-ahead forecasts ($h = 1$), when the models are nested, linear and estimated by OLS. Consider the following nested models:

$$\text{Model 1: } y_t = X_{1t}'\theta_{11} + e_{1t}$$
$$\text{Model 2: } y_t = X_{1t}'\theta_{21} + X_{2t}'\theta_{22} + e_{2t},$$

where the vectors X_{1t} and X_{2t} contain k_1 and k_2 regressors, respectively. The corresponding forecasts are denoted by:

$$\hat{f}_{1t} = X_{1t}'\hat{\theta}_{11t}$$
$$\hat{f}_{2t} = X_{1t}'\hat{\theta}_{21t} + X_{2t}'\hat{\theta}_{22t},$$

where the parameter vectors are first estimated from the above models, using data prior to time t. Under the null hypothesis that f_{1t} encompasses f_{2t}, Model 2 contains k_2 redundant variables (those in X_{2t}), and the population forecasts f_{1t} and f_{2t} are identical. Under the conditions outlined by Clark and McCracken (2001), which include conditionally homoskedastic forecast errors, the asymptotic null distribution of the FE(2) *MDM* statistic (4.14) (with $\hat{d}_t$ replacing d_t), for the general

case $0 < \pi < \infty$ (with $\pi = \lim_{R,n\to\infty}(n/R)$ as before), is given by:

$$MDM \Rightarrow \frac{\Gamma_1}{\sqrt{\Gamma_2}}, \tag{4.18}$$

where the terms Γ_1 and Γ_2 depend on the estimation scheme as follows:

Estimation scheme	Γ_1	Γ_2
Fixed	$\lambda^{-1}[W(1) - W(\lambda)]'W(\lambda)$	$\pi\lambda^{-1}W(\lambda)'W(\lambda)$
Recursive	$\int_\lambda^1 r^{-1}W(r)'dW(r)$	$\int_\lambda^1 r^{-2}W(r)'W(r)dr$
Rolling	$\lambda^{-1}\int_\lambda^1 [W(r) - W(r-\lambda)]'dW(r)$	$\lambda^{-2}\int_\lambda^1 [W(r) - W(r-\lambda)]'$ $[W(r) - W(r-\lambda)]dr$

with $\lambda = (1+\pi)^{-1}$ and $W(r)$ a $(k_2 \times 1)$ vector standard Brownian motion. In the case of the fixed estimation scheme, $(\Gamma_2)^{-1/2}\Gamma_1 \sim N(0,1)$, so that standard critical values can be employed. However, for the recursive and rolling estimation schemes, the forecast encompassing statistic no longer has a standard limit distribution under the null; critical values for these non-standard distributions are provided for a range of values of k_2 and π by Clark and McCracken (2000, 2001).

The above results assume the presence of model estimation uncertainty, with $0 < \pi < \infty$. If, on the other hand, $\pi = 0$, Clark and McCracken (2001) show that the *MDM* statistic is again standard normally distributed in the limit under the null hypothesis. Thus, if the ratio n/R is very small, standard normal critical values may be employed.

In the more general case where $h > 1$ and conditionally heteroskedastic forecast errors are permitted, the above results no longer hold in general. Clark and McCracken (2005) analyze this situation for predictions from nested linear models, where the forecasts are obtained using direct multi-step methods (see, e.g., Bhansali, 2002; Marcellino, Stock and Watson, 2006), as opposed to forecasts obtained by iterated one-step methods. They find that, for $0 < \pi < \infty$, FE(2) *MDM*-type test statistics do not have pivotal asymptotic null distributions, instead depending on nuisance parameters that vary with the second moments of the forecast errors, the model regressors, and the orthogonality conditions implicit in the OLS model estimations.

Two exceptions exist where FE(2) *MDM*-type forecast encompassing tests do possess pivotal limit distributions under the null for $h > 1$. First, if $k_2 = 1$, then the nuisance parameters vanish and the limit distribution of the test statistics reduces to that for $h = 1$, as given by (4.18) above. Second, if $\pi = 0$, the test statistics are standard normally distributed. Aside from these exceptions, however, no nuisance parameter-free asymptotic distributions exist from which critical values can be obtained. In such cases, critical values must instead be generated by simulation or bootstrap methods. Clark and McCracken (2005) outline a method for simulating the asymptotic critical values using estimates of the nuisance parameters, and also an algorithm for obtaining critical values via a parametric bootstrap;

simulation evidence suggests that the latter approach yields considerably better finite sample size control.

Finally, Clark and McCracken (2001, 2005) also propose a new FE(2)-based encompassing test for use with nested linear model forecasts, which is shown to be more powerful than the corresponding *MDM*-type tests. The test statistic is given by:

$$ENC\text{-}F = \frac{n\bar{\hat{d}}}{n^{-1}\sum_{t=R+h}^{R+n+h-1}\hat{e}_{2t}^2}.$$

When $h = 1$ and the forecast errors are conditionally homoskedastic, the statistic has the following limit distribution for $0 < \pi < \infty$ under the null:

$$ENC\text{-}F \Rightarrow \Gamma_1,$$

and critical values from this distribution are provided by Clark and McCracken (2000, 2001) for the fixed, recursive and rolling estimation schemes. When $h > 1$, the *ENC-F* statistic does not have a pivotal asymptotic distribution, even when $k_2 = 1$; in this more general case of multi-step prediction, therefore, critical values must be obtained by bootstrapping. When $\pi = 0$, *ENC-F* is degenerate, and needs to be rescaled by $(R/n)^{1/2}$ to obtain a limit distribution under the null.

4.5 Conditional tests of forecast encompassing

Hitherto, we have considered tests of forecast encompassing that are based on the notion of unconditional expected loss. Giacomini and White (2006) present a general framework for out-of-sample predictive ability testing which is characterized by the formulation of tests (such as tests for forecast encompassing) based on conditional expected loss. Tests of forecast encompassing based on unconditional expected loss indicate whether f_{1t} encompasses f_{2t} on average, i.e., over the whole sample, whereas a conditional evaluation would indicate that f_{1t} encompasses f_{2t} if it were not possible to predict whether the combination of f_{1t} and f_{2t} would outperform f_{1t} based on information known at $t - 1$. The approach of Giacomini and White (2006) also differs from the standard approach to testing for predictive ability in that it compares forecasting *methods* rather than forecasting *models*. Following the seminal contribution of West (1996), the underlying aim is to compare the forecast performance of the models *in population*. Although forecasts are derived from models with estimated parameters, hypotheses concerning predictive ability are framed in terms of forecasts based on the population values of the model parameters, necessitating an allowance for the impact of estimation uncertainty, as discussed in section 4.3. Instead, the approach of Giacomini and White (2006) compares the forecast performance of the methods, where the method comprises the method of estimation and the number of observations to include in the estimation window, in addition to the specification of the model. Estimation uncertainty is thus a key feature of the forecasting method and affects forecast performance.

An implication of evaluating methods rather than models is that it may be optimal to combine forecasts from the data-generating process with those from other models. This situation is ruled out when models are compared. Clements and Hendry (1998, Ch. 10) provide the following illustration based on an AR(1) process, $y_t = \psi y_{t-1} + v_t$, where $v_t \sim$ i.i.d. $N(0, \sigma_v^2)$ and $|\psi| < 1$. Then the h-step-ahead conditional MSFE, assuming an in-sample size of R observations, is:

$$E[\hat{e}_{R+h}^2 \mid y_R] = \frac{\sigma_v^2\left(1-\psi^{2h}\right)}{\left(1-\psi^2\right)} + E[\left(\psi^h - \hat{\psi}^h\right)^2]y_R^2, \tag{4.19}$$

where $\hat{\psi}$ is the OLS estimator of the unknown parameter ψ. The first term is the contribution of future disturbances, and the second is due to parameter uncertainty. Using the asymptotic formula in Baillie (1979, equation 1.6, p. 676):

$$E\left[\hat{e}_{R+h}^2 \mid y_R\right] = \frac{\sigma_v^2\left(1-\psi^{2h}\right)}{\left(1-\psi^2\right)} + h^2\psi^{2(h-1)}R^{-1}\left(1-\psi^2\right)y_R^2. \tag{4.20}$$

Chong and Hendry (1986) note that $h^2\psi^{2(h-1)}$ in the second term of (4.20) has a maximum at $h = -1/\ln\psi$, and so is not monotonic. It is straightforward to show that (4.20) will exceed the unconditional variance of the process ($\sigma_y^2 = E\left(y_t^2\right) = \sigma_v^2\left(1-\psi^2\right)^{-1}$) when:

$$\frac{y_R^2}{\sigma_y^2} > \frac{R\psi^2}{h^2\left(1-\psi^2\right)},$$

where the unconditional variance can be viewed as the expected squared forecast error of a forecast of zero (the unconditional mean), $\bar{y} = 0$. This establishes that the data-generating forecast can be beaten in terms of (squared-error loss) accuracy when there is estimation uncertainty. Moreover, consider the combined forecast:

$$\tilde{y}_{R+h} = \beta\bar{y} + (1-\beta)\hat{y}_{R+h} = (1-\beta)\hat{y}_{R+h}, \tag{4.21}$$

where $0 \leq \beta \leq 1$, with h-step-ahead forecast error:

$$\tilde{e}_{R+h} = y_{R+h} - \tilde{y}_{R+h} = \beta\bar{e}_{R+h} + (1-\beta)\hat{e}_{R+h}, \tag{4.22}$$

where $\hat{e}_{R+h} \equiv y_{R+h} - \hat{y}_{R+h}$ and $\bar{e}_{R+h} \equiv y_{R+h} - \bar{y} = y_{R+h}$. Minimizing $E\left[\tilde{e}_{R+h}^2 \mid y_R\right]$ with respect to β yields:

$$\beta_h^* = \left(1 + \frac{R\psi^2}{h^2\left(1-\psi^2\right)}\right)^{-1}. \tag{4.23}$$

Clements and Hendry (1998) compare the performance of $\hat{y}_{R+h}$, $\bar{y}$ and $\tilde{y}_{R+h}$ in terms of the unconditional MSFE, and establish that there are gains to forecast combination.

The key differences between Giacomini and White (2006) and the approach of Diebold and Mariano (1995) and West (1996) (DMW) to testing predictive ability

become apparent by contrasting their null hypotheses for a general test of equal forecast accuracy for a general loss function $L(.)$. The DMW null is that:

$$H_0: \quad E\left[L\left(Y_t, f_{1t}\left(\beta_1^*\right)\right) - L\left(Y_t, f_{2t}\left(\beta_2^*\right)\right)\right] = 0,$$

so that L is defined over the random variable Y_t and the one-step forecast based on information up to $t-1$ from Model i with population parameter vector β_i^*, denoted $f_{it}\left(\beta_i^*\right)$. The expectation of the loss differential is unconditional. By contrast, the Giacomini and White (2006) null is:

$$H_0: \quad E\left[L\left(Y_t, f_{1t}\left(\hat{\beta}_1\right)\right) - L\left(Y_t, f_{2t}\left(\hat{\beta}_2\right)\right) \mid \digamma_{t-1}\right] = 0, \tag{4.24}$$

so that loss depends on the estimates, and the expectation is with respect to an information set $\digamma_{t-1}$.

In terms of testing for forecast encompassing with standard squared-error loss (i.e., $L\left(Y_t, f_{it}\left(\beta_i^*\right)\right) = e_{it}^2$, where $e_{it} = Y_t - f_{it}^*$), the DMW null becomes:

$$H_0: \quad E(d_t) = 0,$$

where $d_t = e_{1t}\left(e_{1t} - e_{2t}\right)$ for the null that Model 1 encompasses Model 2 using FE(2). Assuming β_1^* and β_2^* are known, under standard conditions (see Diebold and Mariano, 1995; Harvey, Leybourne and Newbold, 1998; West and McCracken, 1998) we obtain (4.12), namely:

$$\sqrt{n}[\bar{d} - E(d_t)] \Rightarrow N(0, S).$$

When β_1^* and β_2^* are not known and the forecasts are based on $\hat{\beta}_1$ and $\hat{\beta}_2$, the variance of the limiting normal distribution will in general include an additional term for the effect of estimation uncertainty, as described in section 4.3.

To test for forecast encompassing using (4.24), let $\hat{d}_t = \hat{e}_{1t}\left(\hat{e}_{1t} - \hat{e}_{2t}\right)$, where $\hat{e}_{it} = y_t - f_{it}(\hat{\beta}_i)$ to make explicit the use of parameter estimates. Then $E(\hat{d}_t \mid \digamma_{t-1}) = 0$ is equivalent to $E(h_{t-1}\hat{d}_t) = 0$ when $\digamma_{t-1} = \mathcal{F}_{t-1}$ (where $\mathcal{F}_t$ is the information available at time t) and h_{t-1} is a $\mathcal{F}_{t-1}$-measurable function of dimension q. Standard asymptotic normality arguments then give rise to the (one-step) Conditional Forecast Encompassing Test (see Giacomini and White, 2006, Theorem 1, p. 1553):

$$T_n^h = n\bar{Z}_n'\hat{\Omega}_n^{-1}\bar{Z}_n,$$

where $\bar{Z}_n = n^{-1}\sum_{t=1}^n Z_t$, $Z_t = h_{t-1}\hat{d}_t$, and $\hat{\Omega}_n$ is the standard variance estimator, $\hat{\Omega}_n = n^{-1}\sum_{t=1}^n Z_t Z_t'$. Under the null:

$$T_n^h \Rightarrow \chi_q^2,$$

as $n \to \infty$. The sequences of forecasts are based on rolling estimation windows of fixed size to ensure non-vanishing parameter estimation uncertainty as the sample of forecasts (n) goes to infinity. (This aspect is suppressed in the notation for convenience.) The choice of h_{t-1} is crucial, in that the test will have no power if $E(\hat{d}_t \mid \digamma_{t-1}) \neq 0$ for some elements of $\digamma_{t-1}$, but an injudicious choice of h_{t-1}

leads to those elements not being included in h_{t-1} : h_{t-1} should include variables that are thought likely to distinguish between the two forecasting methods: such variables are likely to include indicators of past performance, as well as context-specific variables, such as business cycle indicators if it is thought that the relative performance of the two sets of forecasts may vary systematically with the business cycle. Further details of conditional forecast encompassing tests are provided in the discussion of the application of these methods to quantile forecasts in the following section.

4.5.1 Quantile forecasts

Giacomini and Komunjer (2005) present an application of the general approach of Giacomini and White (2006) to forecasting tests for quantile forecasts. The conditional aspect of their approach can be brought to the fore by considering the tests of correct conditional and unconditional coverage of Christoffersen (1998). According to Christoffersen (1998), a set of quantile forecasts is efficient with respect to an information set (denoted Ω_t) if $E\left(\alpha - 1\left(Y_t - \hat{q}_t < 0\right)|\Omega_{t-1}\right) = 0$, where $\hat{q}_t$ is a forecast of $Q_{t,\alpha}$, the α-quantile of the distribution of Y_t conditional on $\mathcal{F}_{t-1}$, namely $Q_{t,\alpha} \equiv F_t^{-1}(\alpha)$, with F_t the conditional distribution function of Y_t. $1(.)$ is the indicator function that takes the value one when the argument is true and zero otherwise. If we define $I_t \equiv 1\left(Y_t - \hat{q}_t < 0\right)$, then the condition for conditional efficiency can be written more succinctly as $E\left(\alpha - I_t|\Omega_{t-1}\right) = 0$. Testing whether this holds is a test of correct conditional coverage, because it requires both that (i) on average over the sample ($t = 1, \ldots, n$) the probability of an exceedence is not significantly different from α, and (ii) that there is no systematic relationship between these exceedences and any variables in the agent's information set at the time the forecast was made. The first requirement is that of correct *unconditional* coverage, often termed a test for unbiasedness, as it is based on whether the sample proportion of exceedences (say, $\hat{\pi} = n^{-1}\sum_{t=1}^{n} I_t$) is significantly different from the nominal proportion α. The null hypothesis is that $E\left(\alpha - I_t\right) = 0$ versus $E\left(\alpha - I_t\right) \neq 0$, and the standard likelihood ratio test is:

$$LR = -2\left[n_0 \ln\left(\frac{1-\alpha}{1-\hat{\pi}}\right) + n_1 \ln\frac{\alpha}{\hat{\pi}}\right] \overset{asy}{\sim} \chi_1^2,$$

where $n_1 = n\hat{\pi}$ and $n_0 = n - n_1$. Tests for correct unconditional coverage, or bias, can also be found in Granger, White and Kamstra (1989), Baillie and Bollerslev (1992) and McNees (1995).

The second requirement can be tested by restricting the information set to past values of I_t, namely $\Omega_{t-1} = \left\{I_{t-1}, I_{t-2}, \ldots\right\}$. The suggestion of Christoffersen (1998) is to test whether $E\left(\alpha - I_t|\Omega_{t-1}\right) = 0$ by testing whether $\left\{I_t\right\}$ follows a binary first-order Markov chain. If the transition probabilities are defined as:

$$\pi_{ij} = \Pr(I_t = j|I_{t-1} = i),$$

where $i, j = \{0, 1\}$, a lack of a systematic relationship between I_t and Ω_{t-1} (here $\Omega_{t-1} = I_{t-1}$) requires that $\pi_{0j} = \pi_{1j}$, $j = \{0, 1\}$, which gives rise to a simple likelihood ratio test. Note that this test does not consider unbiasedness mentioned

under (i). Granger, White and Kamstra (1989, note c to Table 1, p. 91) suggest using a contingency table approach to test the conditional aspect, based on whether the number of occurrences of (say) zeros followed by zeros is consistent with there being no association between the occurrence of a zero in one period and the occurrence of a zero in the following period. Clements and Taylor (2003) suggest a regression-based approach that facilitates the inclusion of variables besides lagged values of $\{I_t\}$ in the information set.

The tests of Christoffersen (1998) illustrate the distinction between conditional and unconditional tests in the context of the evaluation of a single sequence of forecasts. From a conceptual point of view, the approach of Giacomini and Komunjer (2005) can be viewed as replacing the single quantile forecast $\hat{q}_t$ by a combination of two (or more) quantile forecasts, $\theta'\hat{q}_t$, where $\theta = (\theta_1, \theta_2)'$ and $\hat{q}_t = (\hat{q}_{1t}, \hat{q}_{2t})'$, followed by the development of tests of conditional and unconditional forecast encompassing based on the estimated weights $\hat{\theta}$. Their treatment follows Giacomini and White (2006) (although complications arise due to the discontinuous nature of the moment conditions on which the generalized method of moments (GMM) estimation of θ is based).

The conditional α-quantile of Y_t, Q_t, is the optimal forecast for a "tick" or "check" loss function:

$$L_\alpha(e_t) = (\alpha - 1(e_t < 0))\, e_t,$$

where $e_t = y_t - \hat{q}_t$, so that $L(.)$ is used as the basis for assessing whether combinations of forecasts reduce loss. A straightforward application of the definition of forecast encompassing to quantile forecasts gives the following definition of conditional quantile forecast encompassing based on Giacomini and Komunjer (2005, Definition 1, p. 418): $\hat{q}_{1t}$ encompasses $\hat{q}_{2t}$ at time t if:

$$E_{t-1}\left[L_\alpha\left(Y_t - \hat{q}_{1t}\right)\right] = E_{t-1}\left[L_\alpha\left(Y_t - \left(\theta_{1t}^*\hat{q}_{1t} + \theta_{2t}^*\hat{q}_{2t}\right)\right)\right],$$

where $E_{t-1}(.) \equiv E(. \mid \mathcal{F}_{t-1})$ and where $\theta^* = \left(\theta_1^*, \theta_2^*\right)'$ are the optimal weights in that they minimize tick loss:

$$\left(\theta_1^*, \theta_2^*\right) \equiv \underset{(\theta_1,\theta_2)\in\Theta}{\arg\min} E_{t-1}\left[L_\alpha\left(Y_t - \left(\theta_{1t}\hat{q}_{1t} + \theta_{2t}\hat{q}_{2t}\right)\right)\right].$$

Thus the optimal weights are $\left(\theta_1^*, \theta_2^*\right) = (1, 0)$, so assigning a zero weight to $\hat{q}_{2t}$.

Giacomini and Komunjer (2005, Lemma 1, p. 419) show that θ^* satisfies the first-order condition:

$$E_{t-1}\left[\alpha - 1\left(Y_t - \theta^*\hat{q}_t < 0\right)\right] = 0, \tag{4.25}$$

which is the correct conditional coverage condition of Christoffersen (1998). These moment conditions are used to estimate the optimal weights by GMM and to test for forecast encompassing (that $\hat{q}_{1t}$ encompasses $\hat{q}_{2t}$): $\left(\theta_1^*, \theta_2^*\right) = (1, 0)$. The conditional moment conditions (4.25) are replaced by:

$$E\left[\left(\alpha - 1\left(Y_t - \theta^*\hat{q}_t < 0\right)\right) W_{t-1}^*\right] = 0,$$

where W_t^* is an $\mathcal{F}_t$-measurable function. W_{t-1} plays the same role as h_{t-1} in the Conditional Forecast Encompassing Test of Giacomini and White (2006) and the same issues relate to its selection. The sample moment function is given by:

$$g_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left[\alpha - 1\left(y_t - \theta'\hat{\boldsymbol{q}}_t < 0\right)\right]w_{t-1}^*, \tag{4.26}$$

and the GMM estimator of θ^*, denoted by $\hat{\theta}_n$, is the solution of:

$$\min_{(\theta_1,\theta_2)\in\Theta}\left[g_n(\theta)\right]'\hat{S}_n^{-1}\left[g_n(\theta)\right], \tag{4.27}$$

where:

$$\hat{S}_n = \frac{1}{n}\sum_{t=1}^{n}\left[\alpha - 1\left(y_t - \theta_t'\hat{q}_t < 0\right)\right]^2 w_{t-1}^* w_{t-1}^{*\prime}, \tag{4.28}$$

and $\hat{\theta}_n$ is obtained by solving (4.27) and (4.28) iteratively, starting with (4.27) and $\hat{S}_n = \mathrm{I}$. This gives $\hat{\theta}_n\left(\hat{S}_n = \mathrm{I}\right)$, after which we obtain an updated estimate of $\hat{S}_n$ from (4.28), etc.

Giacomini and Komunjer (2005, Propositions 1 and 2, p. 420) establish the consistency and asymptotic normality of $\hat{\theta}_n$. Specifically, under some conditions:

$$\left(\gamma' S^{-1}\gamma\right)^{-1/2}\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \Rightarrow N(0,1),$$

where $\gamma \equiv -E[f_t(\theta^{*\prime}\mathrm{q}_t)W_{t-1}^*\mathrm{q}_t']$, $S = E[g(\theta^*; Y_t, W_{t-1}^*)g(\theta^*; Y_t, W_{t-1}^*)']$, and f_t is the conditional density of Y_t. Given the asymptotic distribution of $\hat{\theta}_n$, the CQFE (Conditional Quantile Forecast Encompassing) test that $\hat{q}_{1t}$ encompasses $\hat{q}_{2t}$ is given by:

$$ENC_n = n\left(\hat{\theta}_n' - (1,0)\right)\hat{\Omega}_n^{-1}\left(\hat{\theta}_n' - (1,0)\right)',$$

where $\hat{\Omega}_n^{-1}$ is a consistent estimate of $\Omega = \left(\gamma' S^{-1}\gamma\right)^{-1}$. Under the null, $ENC_n \Rightarrow \chi_2^2$ as $n \to \infty$. The estimate for S is given by (4.28), and that for γ is given by Giacomini and Komunjer (2005).

4.6 Loss functions and forecast combination

Ever since the early work on forecast combination of Bates and Granger (1969) and Granger and Ramanathan (1984), combination weights have generally been chosen to minimize a symmetric, squared-error loss function, and the empirical forecast performance of the combination has typically been assessed by squared-error loss. This reflects the widespread use of squared-error loss in the forecast evaluation literature. For example, the "regression method" of Granger and Ramanathan (1984) estimates by OLS an equation such as:

$$y_t = \beta_1 f_{1t} + \beta_2 f_{2t} + e_t,$$

so that the combination weights are selected to minimize $\sum_t e_t^2$, the sum of squares of the forecast error for the combined forecast. Similarly, the "variance-covariance approach" of Bates and Granger (1969) selects the weights to minimize the variance of the forecast error of the combination of forecasts. Nevertheless, a number of papers have allowed for asymmetric loss, and have considered how the properties of optimal forecasts change once we dispense with the assumption of symmetric loss, as well as providing tests of rationality once we allow forecasters to have asymmetric loss functions.[2]

A key paper that investigates forecast combination in the context of asymmetric loss is Elliott and Timmermann (2004). They show that, for general loss functions and forecast error distributions, the optimal combination weights depend on higher-order moments of the forecast error distribution, such as the skew. However, under certain restrictions on the form of the forecast error distribution, they establish an *invariance* result, whereby the optimal combination weights on the individual forecasts are identical to the squared-error loss weights for almost all loss functions and that only the value of the constant term in the combination will differ. The value of the constant is chosen to generate the optimal amount of bias in the combination given the degree of asymmetry of the loss function. Their invariance result holds when the marginal distribution of the forecast errors depends only on the first two moments of the forecast errors, which holds when the joint distribution of the actual and forecasts $(y_t, \mathrm{f}_t)'$ is elliptically symmetric (which includes the multivariate normal and t-distributions: see Elliott and Timmermann, 2004, Proposition 2, p. 53).

Suppose:

$$E\begin{pmatrix} y_t \\ \mathrm{f}_t \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu \end{pmatrix}, \quad C\begin{pmatrix} y_t \\ \mathrm{f}_t \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{21}' \\ \sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then the forecast combination error is:

$$e_t = y_t - \beta_0 - \beta' \mathrm{f}_t,$$

with moments:

$$\mu_e = \mu_y - \beta_0 - \beta' \mu \tag{4.29}$$

$$\sigma_e^2 = \sigma_y^2 + \beta' \Sigma_{22} \beta - 2\beta' \sigma_{21}. \tag{4.30}$$

The decision maker selects (β_0, β) according to:

$$\min_{\beta_0, \beta} \int L(e_t)\, dF(e_t),$$

i.e., to minimize $E[L(e_t)]$. Under elliptical symmetry we can write $E[L(e_t)] = g\left(\mu_e, \sigma_e^2\right)$. From (4.29) and (4.30), only μ_e depends on β_0. Thus the first-order condition for minimizing $E[L(e_t)]$ with respect to β_0 is:

$$\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \beta_0} = \frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \mu_e} \frac{\partial \mu_e}{\partial \beta_0} = 0.$$

As $\frac{\partial \mu_e}{\partial \beta_0} = -1$, the optimal value for β_0, β_0^*, solves $\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \mu_e} = 0$. β_0^* depends on $L(.)$, and is set to generate the optimal amount of bias (μ_e^*) given the form of $L(.)$. For squared-error loss, $L\left(e_t\right) = e_t^2$, $E\left[L\left(e_t\right)\right] = \mu_e^2 + \sigma_e^2$, and $\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \beta_0} = -2\mu_e$, so that the optimal amount of bias is, of course, zero ($\mu_e^* = 0$).

Consider the first-order condition with respect to β:

$$\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \beta} = \frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \sigma_e^2} \frac{\partial \sigma_e^2}{\partial \beta} = 0.$$

Provided $\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \sigma_e^2} \neq 0$, $\frac{\partial g\left(\mu_e, \sigma_e^2\right)}{\partial \beta} = 0$ implies that $\frac{\partial \sigma_e^2}{\partial \beta} = 0$, so from (4.30), $2\Sigma_{22}\beta^* = 2\sigma_{21}$ and $\beta^* = \Sigma_{22}^{-1}\sigma_{21}$ irrespective of the form of $L(.)$, matching the expression for squared-error loss.

As Elliott and Timmermann (2004) remark, if an element of β^* is zero under squared-error loss, then the corresponding forecast will also receive zero weight under any other loss function, assuming that the stated properties of the forecast error distribution hold. In the general case, it will not be possible to set up a general forecast encompassing test that does not depend on the form of the loss function.

4.7 Conclusions

We have discussed the different types of standard linear forecast combination that are commonly applied in the literature and the related tests of forecast encompassing. The tests of forecast encompassing depend upon whether the forecasts are generated by models with unknown parameters and on whether the underlying aim is to compare the forecasts themselves or the models on which they are based. There is also an important distinction to be drawn between conditional and unconditional tests.

More sophisticated forms of combination are reviewed, including nonlinear forms of combination that might be useful when large numbers of forecasts are available, and types of combination that might be preferable when the forecasts are density or probability forecasts. For the most part, forecast accuracy is assessed by the standard squared-error loss, although under certain conditions on the data-generating process forecast encompassing is invariant to the form of the loss function.

Notes

1. For the specification FE(2′) one would use $d_t = (e_{1t} - \bar{e}_1)[(e_{1t} - \bar{e}_1) - (e_{2t} - \bar{e}_2)]$, and for FE(3′), $d_t = (e_{1t} - \bar{e}_1)(f_{2t} - \bar{f}_2)$.
2. These include, *inter alia*, Granger (1969), Zellner (1986), Christoffersen and Diebold (1997), Clements (1997), Elliott, Komunjer and Timmermann (2005), Patton and Timmermann (2007) and Clements (2008).

References

Andrews, M.J., A.P.L. Minford and J. Riley (1996) On comparing macroeconomic forecasts using forecast encompassing tests. *Oxford Bulletin of Economics and Statistics* **58**, 279–305.

Baillie, R.T. (1979) The asymptotic mean squared error of multistep prediction from the regression model with autoregressive errors. *Journal of the American Statistical Association* **74**, 175–84.

Baillie, R.T. and T. Bollerslev (1992) Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics* **52**, 91–113.

Bates, J.M. and C.W.J. Granger (1969) The combination of forecasts. *Operations Research Quarterly* **20**, 451–68. Reprinted in T.C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Bhansali, R.J. (2002) Multi-step forecasting. In M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, pp. 206–21: Oxford: Blackwell.

Chong, Y.Y. and D.F. Hendry (1986) Econometric evaluation of linear macro-economic models. *Review of Economic Studies* **53**, 671–90. Reprinted in C.W.J. Granger (ed.), *Modelling Economic Series*. Oxford: Clarendon Press, 1990.

Christoffersen, P.F. (1998) Evaluating interval forecasts. *International Economic Review* **39**, 841–62.

Christoffersen, P.F. and F.X. Diebold (1997) Optimal prediction under asymmetric loss. *Econometric Theory* **13**, 808–17.

Clark, T. E. and M.W. McCracken (2000) Not-for-publication appendix to "Tests of equal forecast accuracy and encompassing for nested models." Manuscript, Federal Reserve Bank of Kansas City.

Clark, T.E. and M.W. McCracken (2001) Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105**, 85–110.

Clark, T.E. and M.W. McCracken (2005) Evaluating direct multi-step forecasts. *Econometric Reviews* **24**, 369–404.

Clemen, R.T. (1989) Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* **5**, 559–83. Reprinted in T.C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Clemen, R.T. and R.L. Winkler (1986) Combining economic forecasts. *Journal of Business and Economic Statistics* **4**, 39–46.

Clemen, R.T. and R.L. Winkler (1999) Combining probability distributions from experts in risk analysis. *Risk Analysis* **19**, 187–203.

Clements, M.P. (1997) Evaluating the rationality of fixed-event forecasts. *Journal of Forecasting* **16**, 225–39.

Clements, M.P. (2008) Internal consistency of survey respondents' forecasts: evidence based on the Survey of Professional Forecasters. In J.L. Castle and N. Shephard (eds.), *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Clements, M.P. and D.I. Harvey (2006) Forecast encompassing tests and probability forecasts. Working Paper, Department of Economics, University of Warwick.

Clements, M.P. and D.I. Harvey (2007) Combining probability forecasts. Working Paper, Department of Economics, University of Warwick.

Clements, M.P. and D.F. Hendry (1998) *Forecasting Economic Time Series*. Cambridge: Cambridge University Press. The Marshall Lectures on Economic Forecasting.

Clements, M.P. and D.F. Hendry (2006) Forecasting with breaks. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 605–57: Amsterdam: Elsevier, North-Holland.

Clements, M.P. and N. Taylor (2003) Evaluating prediction intervals for high-frequency data. *Journal of Applied Econometrics* **18**, 445–56.

Coulson, N.F. and R.P. Robins (1993) Forecast combination in a dynamic setting. *Journal of Forecasting* **12**, 63–8.

Deutsch, M., C.W.J. Granger and T. Teräsvirta (1994) The combination of forecasts using changing weights. *International Journal of Forecasting* **10**, 47–57.

Diebold, F.X. (1988) Serial correlation and the combination of forecasts. *Journal of Business & Economic Statistics* **6**, 105–11.

Diebold, F.X. and J.A. Lopez (1996) Forecast evaluation and combination. In G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics*, *Volume 14*, pp. 241–68. Amsterdam: North-Holland.

Diebold, F.X. and R.S. Mariano (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253–63. Reprinted in T. C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Diebold, F.X. and R. Pauly (1987) Structural change and the combination of forecasts. *Journal of Forecasting* **6**, 21–40.

Diebold, F.X. and R. Pauly (1990) The use of prior information in forecast combination. *International Journal of Forecasting* **6**, 503–8.

Donaldson, R.G. and M. Kamstra (1996). Forecast combining with neural networks. *Journal of Forecasting* **15**, 49–61.

Elliott, G., I. Komunjer and A. Timmermann (2005) Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies* **72**, 1107–25.

Elliott, G. and A. Timmermann (2004) Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics* **122**, 47–79.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

Ericsson, N.R. (1992) Parameter constancy, mean square forecast errors, and measuring forecast performance: an exposition, extensions, and illustration. *Journal of Policy Modeling* **14**, 465–95.

Fair, R.C. and R.J. Shiller (1989) The informational content of *ex ante* forecasts. *Review of Economics and Statistics* **71**, 325–31.

Fair, R.C. and R.J. Shiller (1990) Comparing information in forecasts from econometric models. *American Economic Review* **80**, 39–50.

Figlewski, S. and P. Wachtel (1981) The formation of inflationary expectations. *Review of Economics and Statistics* **63**, 1–10.

Fildes, R. and K. Ord (2002) Forecasting competitions – their role in improving forecasting practice and research. In M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, pp. 322–53. Oxford: Blackwell.

Franses, P.H. and D.J. van Dijk (2000) *Non-linear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Genest, C. and J.V. Zidek (1986) Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* **1**, 114–48.

Giacomini, R. and I. Komunjer (2005) Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics* **23**, 416–431.

Giacomini, R. and H. White (2006) Tests of conditional predictive ability. *Econometrica* **74**, 1545–78.

Granger, C.W.J. (1969) Prediction with a generalized cost of error function. *Operations Research Quarterly* **20**, 199–207.

Granger, C.W.J. and Y. Jeon (2004) Thick modeling. *Economic Modelling* **21**, 323–43.

Granger, C.W.J. and P. Newbold (1973) Some comments on the evaluation of economic forecasts. *Applied Economics* **5**, 35–47. Reprinted in T.C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Granger, C.W.J. and P. Newbold (1977) *Forecasting Economic Time Series*. New York: Academic Press.

Granger, C.W.J. and R. Ramanathan (1984) Improved methods of combining forecasts. *Journal of Forecasting* **3**, 197–204.

Granger, C.W.J., H. White and M. Kamstra (1989) Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics* **40**, 87–96.

Harvey, D.I., S.J. Leybourne and P. Newbold (1997) Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**, 281–91.

Harvey, D.I., S.J. Leybourne and P. Newbold (1998) Tests for forecast encompassing. *Journal of Business and Economic Statistics* **16**, 254–9. Reprinted in T.C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Harvey, D.I., S.J. Leybourne and P. Newbold (1999) Forecast evaluation tests in the presence of ARCH. *Journal of Forecasting* **18**, 435–45.

Harvey, D.I. and P. Newbold (2000) Tests for multiple forecast encompassing. *Journal of Applied Econometrics* **15**, 471–82.

Harvey, D.I. and P. Newbold (2003) The non-normality of some macroeconomic forecast errors. *International Journal of Forecasting* **19**, 635–53.

Harvey, D.I. and P. Newbold (2005) Forecast encompassing and parameter estimation. *Oxford Bulletin of Economics and Statistics* **67**, Supplement, 815–36.

Hendry, D.F. and M.P. Clements (2004) Pooling of forecasts. *Econometrics Journal* **7**, 1–31.

Hendry, D.F. and J.-F. Richard (1989) Recent developments in the theory of encompassing. In B. Cornet and H. Tulkens (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, Mass.: MIT Press. Reprinted in J. Campos, N.R. Ericsson and D.F. Hendry (eds.), *General to Specific Modelling*. Cheltenham: Edward Elgar, 2005.

Kamstra, M. and P. Kennedy (1998) Combining qualitative forecasts using logit. *International Journal of Forecasting* **14**, 83–93.

Keane, M.P. and D.L. Runkle (1990) Testing the rationality of price forecasts: new evidence from panel data. *American Economic Review* **80**, 714–35.

Kuan, C.-M. and H. White (1994) Artificial neural networks: an econometric perspective. *Econometric Reviews* **13**, 1–143.

LeSage, J.P. and M. Magura (1992) A mixture-model approach to combining forecasts. *Journal of Business and Economic Statistics* **10**, 445–52.

Makridakis, S. and R.L. Winkler (1983) Averages of forecasts: some empirical results. *Management Science* **29**, 987–96.

Marcellino, M., J.H. Stock and M.W. Watson (2006) A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* **135**, 499–526.

McNees, S.K. (1995) Forecast uncertainty: can it be measured? Discussion paper, Federal Reserve Bank of New York.

Min, C. and A. Zellner (1993) Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* **56**, 89–118.

Mincer, J. and V. Zarnowitz (1969) The evaluation of economic forecasts. In J. Mincer (ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Mizon, G.E. (1984) The encompassing approach in econometrics. In D.F Hendry and K.F. Wallis (eds.), *Econometrics and Quantitative Economics*, pp. 135–72. Oxford: Blackwell.

Mizon, G.E. and J.-F. Richard (1986) The encompassing principle and its application to non-nested hypothesis tests. *Econometrica* **54**, 657–78.

Morgan, W.A. (1939) A test for significance of the difference between the two variances in a sample form a normal bivariate population. *Biometrika* **31**, 13–19.

Nelson, C.R. (1972) The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review* **62**, 902–17. Reprinted in T.C. Mills (ed.), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar, 1999.

Newbold, P. and D.I. Harvey (2002) Forecasting combination and encompassing. In M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, pp. 268–83: Oxford: Blackwell.

Patton, A.J. and A. Timmermann (2007) Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* **102**, 1172–84.

Phillips, P.C.B. (1995) Spurious regression in forecast-encompassing tests. *Econometric Theory* **11**, 1188–90.

Stekler, H.O. (2002) The rationality and efficiency of individuals' forecasts. In M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, pp. 222–40. Oxford: Blackwell.

Stock, J.H. and M.W. Watson (1999) A comparison of linear and nonlinear models for forecasting macroeconomic time series. In R.F. Engle and H. White (eds.), *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Timmermann, A. (2006) Forecast combinations. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 135–96. Amsterdam: Elsevier, North-Holland.

West, K.D. (1996) Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–84.

West, K.D. (2001) Tests for forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics* **19**, 29–33.

West, K.D. and M.W. McCracken (1998) Regression-based tests of predictive ability. *International Economic Review* **39**, 817–40.

White, H. (1989) Some asymptotic results for learning in single-layer feedforward network models. *Journal of the American Statistical Association* **84**, 1003–13.

Zarnowitz, V. (1985) Rational expectations and macroeconomic forecasts. *Journal of Business and Economic Statistics* **3**, 293–311.

Zarnowitz, V. and P. Braun (1993) Twenty-two years of the NBER-ASA quarterly economic outlook surveys: aspects and comparisons of forecasting performance. In J. Stock and M. Watson (eds.), *Business Cycles, Indicators, and Forecasting*, pp. 11–84. Chicago: University of Chicago Press and NBER.

Zellner, A. (1986) Biased predictors, rationality and the evaluation of forecasts. *Economics Letters* **21**, 45–48.

5
Recent Developments in Density Forecasting

Stephen G. Hall and James Mitchell

Abstract

With the growing recognition that point forecasts, the traditional focus, are better seen as the central points of ranges of uncertainty, in recent years increased emphasis has been given to density forecasts. This chapter reviews these recent developments, with a focus on the production and use of density forecasts in macroeconomics. Particular attention is paid to the evaluation and combination of density forecasts.

5.1 Introduction

Forecasts of the future values of economic variables are used widely in decision making. For example, in many countries inflation forecasts are now central to the setting of monetary policy since monetary policy works with a lag (e.g., see Svensson, 2005, for a review). But it has become increasingly well understood that it is not a question of this forecast proving to be "right" and that forecast proving to be "wrong." Point forecasts, the traditional focus, are better seen as the central points of ranges of uncertainty. A forecast of, say, 2% must mean that people should not be surprised if actual inflation turns out to be a little larger or smaller than that. Moreover, at a time of heightened economic uncertainty, they should not be very surprised if it turns out to be much larger or smaller. Consequently, to provide a complete description of the uncertainty associated with the point forecast many professional forecasters now publish density forecasts, or more popularly "fan charts." A famous example is the Bank of England's fan chart (see Britton, Fisher and Whitley, 1998). Importantly, as this chapter reviews, just as point forecasts are commonly evaluated using the subsequent outturn, so the reliability of uncertainty forecasts can be evaluated.

More formally, density forecasts of inflation provide an estimate of the probability distribution of its possible future values. In contrast to interval forecasts, which give the probability that the outcome will fall within a stated interval, such as inflation falling within its target range, density forecasts provide a complete description of the uncertainty associated with a forecast; they can thus be seen to provide information on all possible intervals.

In conjunction with the increased use of density forecasts by professional forecasters and central banks, the academic literature has also devoted increased emphasis to density forecasting, with the *Journal of Forecasting* devoting a special issue to it in 2000 and the *Handbook of Economic Forecasting*, published in 2006, containing a chapter surveying methods for predictive density evaluation. This chapter, with a macroeconomic focus, reviews several aspects of these recent developments, breaking them down into four areas which, in turn, are considered in separate sections of the chapter:

1. The importance of density forecasts
2. The production of density forecasts
3. The evaluation of density forecasts
4. The combination of density forecasts.

In so doing we extend the coverage, and update in the light of recent research, previous surveys and textbooks, including Tay and Wallis (2000), Clements (2005), Timmermann (2006) and Wallis (2008). The principal extensions in terms of coverage are the last two sections. In particular we focus on how to choose the weights when combining density forecasts. Reflecting the infancy of this material, considerable applied and theoretical work remains to be done to establish a consensus about how best this should be achieved. Nevertheless, this chapter attempts

to bring together current wisdom and indicates areas where more research would be welcome. Our discussion also draws out, particularly in section 5.3, which has much relevance for how forecasting is conducted in practice by professional forecasters, the distinction between model-based and subjective density forecasts, and we also consider their reconciliation. We do not discuss presentation issues, which we defer to Tay and Wallis (2000) and Wallis (2007).

We confine attention to univariate density forecasts. There is a smaller but growing literature on multivariate density forecasting, some of it drawing on recent applications of copula functions in economics (e.g., Patton, 2006), where the copula characterizes the dependence between the density forecasts. Diebold, Gunther and Tay (1998) and Diebold, Hahn and Tay (1999) show that the principle behind the evaluation of univariate density forecasts, discussed in section 5.4 below, generalizes to the multivariate case (see also Clements and Smith, 2000, 2002). Adolfson, Linde and Villani (2007) use a multivariate scoring rule to compare density forecasts of the Euro-area from vector autoregressive (VAR) and dynamic stochastic general equilibrium (DSGE) models. Barrell, Hall and Hurst (2006) and Mitchell (2007a) consider how bivariate density forecasts for inflation and output growth facilitate the evaluation of policy rules simultaneously with respect to their performance against the inflation target and any output growth target that the policy makers may also have in mind.

5.2 The importance of density forecasts

Periodically, and perhaps especially at times of heightened uncertainty, one hears the argument that it is time to jettison economic forecasts given their unreliability. But, as discussed above, in fact we should not be surprised by the unreliability of point forecasts – indeed, the unreliability of point forecasts is itself a useful indication of uncertainty. In a loose sense, ignoring for now moments higher than the second, what is important is the ability of the point forecast, relative to its variance, to track the outturn. More generally, it is important to provide a quantitative indication of the uncertainty associated with a point forecast, along with the balance of risks (skewness) on the upside and downside and the probability of extreme events (fat tails or kurtosis). This is achieved by publishing a density forecast. Importantly, the density forecast gives any users of the forecast an indication, in advance, of the health risks associated with its use.

Although it is a truism to say that density forecasts cannot capture unknowable uncertainty (Knightian uncertainty) and only capture "risk" (knowable uncertainty), the distinction introduced by Knight (1921), these "risk" assessments can be evaluated *ex post*. Indeed, they should be assessed on a regular and ongoing basis. There is no reason to expect, especially at times of structural change, that the density forecast correctly captures uncertainty. Forecasters' statements about the underlying uncertainty may be, and indeed often are, unreliable. When forecasters expect them to be unreliable the variance of the conditional variance forecast need not equal zero. Therefore evaluation tests, reviewed in section 5.4, have been developed to test, essentially, whether on average over a given sample a forecaster's

assessment of "risk" was correct. When it is correct their "risk" forecast might be said to have captured "true" uncertainty.

These sort of historical evaluations of fan charts, complementing the traditional and widespread practice of evaluating the trackrecord of point forecasts, are beginning to be carried out routinely by forecasters (for an appraisal of UK inflation density forecasts by the Bank of England and/or the National Institute of Economic and Social Research (NIESR), see Clements, 2004; Wallis, 2004; Mitchell, 2005; Elder *et al.*, 2005). The results provide an indication as to whether, albeit historically, a series of fan charts was reliable. Just as measures of point forecast accuracy indicate, to a degree, our confidence in point forecasts, these tests provide an indication of our confidence in fan charts.

5.2.1 Forecasting under general loss functions

What really matters is how forecasts affect decisions. The "better" forecasts are those that deliver "better" decisions. On this basis it is argued that the appropriate way of evaluating forecasts is not to use some arbitrary statistical loss function, but the appropriate economic loss function (see Granger, 1969; Granger and Pesaran, 2000). Only when the forecast user has a symmetric, quadratic loss function, and the constraints (if relevant) are linear, is it correct to focus on the point forecast alone. This is what the textbook's call "certainty equivalence" (for further discussion and a proof, see Ljungqvist and Sargent, 2000, pp. 57–9). In the more general case, the degree of uncertainty matters. Publishing a point forecast alone is not sufficient; users are not indifferent to the degree of uncertainty about the point forecast.[1] They will not then make decisions as if they were certain. Uncertainty is expected to attenuate their response or reaction to the point forecast (see Brainard, 1967). For more recent discussion in the context of policy makers' reactions to real-time output gap estimates, which are known to be unreliable (Orphanides and van Norden, 2002), although this is not surprising (as discussed above: see also Mitchell, 2007b), see Swanson (2004).

The importance of publishing density forecasts then follows from the fact that we tend, in reality, not to know users' loss functions. Central banks do not quantify, explicitly at least, their loss functions, but we should not expect these (unknown to us) functions to be quadratic. For example, we should expect the range of uncertainty to matter to the Federal Reserve since it probably does not care equally about inflation above and below the zero bound. The central bank has then to be what Svensson (2001) calls a *distribution* forecast targeter.

When the forecast user's loss function is asymmetric, such that positive and negative forecasting errors have differing costs, the user's "optimal" forecast need not equal the conditional mean (e.g., Zellner, 1986; see Pesaran and Weale, 2006, for a survey). Working out the optimal forecast can be complex, but if it is assumed that the conditional distribution of $y_t \mid \Omega_{t-h}$ is normal such that $y_t \mid \Omega_{t-h} \sim N(E(y_t \mid \Omega_{t-h}), V(y_t \mid \Omega_{t-h}))$ and the loss function is modeled via the Linex loss function, an analytical solution can be derived. Under these conditions the optimal or minimum loss point forecast, $\widehat{y}_{t|t-h}$, is no longer equal to the conditional mean

but equals:

$$\widehat{y}_{t|t-h} = E(y_t \mid \Omega_{t-h}) + \frac{\phi V(y_t \mid \Omega_{t-h})}{2}, \tag{5.1}$$

where ϕ is the asymmetry parameter in the Linex function. It reflects differing costs to over and under prediction. This means that it can be "rational" for the user to focus on what are effectively biased point (conditional mean) forecasts.

The trend towards forecasters publishing density forecasts is also explained by the obvious advantages they bring when communicating with the public and reminding them that the forecasters themselves expect the point forecasts to be "wrong." Indeed, interest may lie in the dispersion or tails of the density itself; e.g., inflation targets often focus the attention of monetary authorities to the probability of future inflation falling within some predefined target range, while users of growth forecasts may be concerned about the probability of recession. These probability event forecasts can readily be extracted from the density forecast. In addition, ranking forecasting models according to their point forecasting performance alone can be misleading. For example, Clements *et al.* (2003) find that the failure to find empirical support out-of-sample for nonlinear business cycle forecasts may be explained by the traditional focus on point forecasts and their root mean squared error (RMSE). They argue that nonlinear models may do better at forecasting the higher moments that are captured by density forecasts.

5.3 The production of density forecasts

In general, forecasts can be produced in a wide variety of ways, ranging from complete model-based approaches to pure judgmental approaches, sometimes referred to as Delphic forecasts; indeed, almost any combination of model and judgment is possible. In the conventional point forecasting world, it is probably fair to say that almost all forecasts which are made by policy or commercial institutes involve a considerable degree of judgment, although there is, of course, a considerable academic literature on pure model-based forecasts. When we consider density forecasting, a similar range of formal and informal techniques are used, although it is probably fair to say that, given the greater complexity of a density forecast, there should be more reliance on formal model-based information.

There is not a widespread, long history of regular published density forecasts in macroeconomics. One of the longest continuously published series is the Survey of Professional Forecasters (SPF), which is now conducted by the Federal Reserve Bank of Philadelphia and was originally started in 1968 by the American Statistical Association and the National Bureau of Economic Research.

Nevertheless, there is a tradition, which has been maintained to the present day, of publishing (unbalanced) panel data sets of competing point forecasts. For example, in the UK each month since January 1987 Her Majesty's Treasury (HMT) has collected together, in its publication *Forecasts for the UK Economy: A Comparison of Independent Forecasts*, the point forecasts of (as of December 2007) 43 independent City and non-City forecasters. Disagreement among forecasters (as measured by the variance of competing point forecasts at a given point in time) has then

been used as a proxy for true uncertainty. As the SPF also offers a direct measure of uncertainty, since forecasters are asked to report not just their point but density forecasts, it has provided the opportunity, not possible with the HMT dataset, to test the reliability of disagreement as a measure of uncertainty (see Zarnowitz and Lambros, 1987; Bomberger, 1996; Giordani and Söderlind, 2003. Boero, Smith and Wallis, 2008, introduce a new source of survey data for the UK, the Bank of England's Survey of External Forecasters, which also facilitates a comparison of disagreement and uncertainty).

Macroeconomic forecasters have also studied the density of their forecasts over a long period, although they have not typically published, on a regular and ongoing basis, density forecasts as such. The reason for this is, partly, that for a long time it was felt that density forecasts were too sophisticated for the public to understand and partly that, as the models being used made the assumption that the parameters and the error terms had constant covariance structures, the overall density function would not vary from one period to another except with respect to its conditional mean. There was therefore relatively little interest in publishing the same error bands over and over again. However, it is certainly true that modelers and forecasters in the 1970s and 1980s were calculating the uncertainty surrounding their forecasts and, occasionally at least, publishing it. For example, the London Business School, one of the UK's leading forecasters at the time, began regularly publishing the average absolute errors for its forecasts in October 1983. Some early work in this area includes Schink (1971), Bianchi and Calzolari (1980) and Fair (1980). Fair (1984) surveyed a range of stochastic simulation techniques which were being used to calculate the density functions for large nonlinear forecasting models. Hall (1986), and later Blake (1996), reported studies of the density of the NIESR's forecasts, again using extensive stochastic simulations.

These later model-based studies contrast strongly with the SPF, which was purely judgmentally based. This introduces an important theme into this section, which is the issue of combining judgment with formal model-based analysis. Another and related theme is the production of density forecasts where the density itself changes over time. It is only when the whole density changes in a significant way through time that it is worth going to the lengths of publishing a regular full density forecast.

5.3.1 Sources of uncertainty

A forecast is usually subject to a range of types of uncertainty, which we can begin to categorize by considering the following simple decomposition.[2] Let Y be the actual outcome of an event and let $\widehat{Y}$ be the forecast of that event. Then we may decompose the forecast error into a number of components:

$$Y - \widehat{Y} = (Y - Y^1) + (Y^1 - Y^2) + (Y^2 - Y^3) + (Y^3 - Y^4) + (Y^4 - \widehat{Y}), \qquad (5.2)$$

where $(Y - Y^1)$ is the contribution to the total error coming from the model's error term, $(Y^1 - Y^2)$ is the contribution coming from the uncertain parameters, $(Y^2 - Y^3)$ is the contribution coming from misspecified functional form, $(Y^3 - Y^4)$

is the contribution coming from incorrect exogenous assumptions, and $(Y^4 - \widehat{Y})$ is the contribution to the complete forecast error coming from the judgment imposed on the forecast from outside the model. This is a useful way of categorizing the total error composition; however, numerically the order in which this decomposition is carried out can affect the numerical evaluation of these components (see Hall and Henry, 1988). It is also worth noting that the variances which lie behind these components and which make up the complete density of the forecast may be either time varying or constant over time, and there will generally be non-zero covariances between these components.

Part of the usefulness of this decomposition is to emphasize the range of sources of uncertainty. Virtually no formal model-based analysis can deal with all of these and so we may justify the use of judgment, which will be discussed below, at least partly on the grounds of this failure on the part of the formal analysis.

5.3.2 Model-based densities

If we now turn to density forecasts produced by a range of models, a natural starting place is a conventional VAR, as this may be thought to capture the basic properties of most standard forecasting models, including large macroeconomic forecasting models. Indeed, the linearized solution of DSGE models, the workhorse of modern macroeconomics (see Woodford, 2003) and from which density forecasts may readily be constructed using simulation methods, can, under certain conditions, be approximated by a (restricted) finite-order VAR (e.g., see Pesaran and Smith, 2006; Ravenna, 2007).

Thus, consider a VAR of the form:

$$Y_t = B(L)Y_{t-1} + \varepsilon_t, \tag{5.3}$$

$(t = 1, \ldots, T)$, where Y_t is a vector of N variables, $B(L)$ is a suitably dimensioned matrix lag polynomial of estimated parameters with fixed covariance matrix and ε_t is an $N \times 1$ vector of residuals with constant covariance matrix. Given the constant covariance assumption for both the parameters and residuals, the density of a forecast of Y_{t+h} $(h = 1, \ldots, H)$ will, in general, vary only with the initial values Y_{t-1}. If the errors in the VAR process are normally distributed then the density function of the VAR forecast h steps ahead is also normally distributed with a covariance structure which can be approximated analytically. Lutkepohl (1991, p. 87) provides an approximate analytical expression for the conditional variance equal to the approximate mean squared error of the forecast with parameter uncertainty. Allowance can also be made for the uncertain parameters of the VAR or non-normality by using either Monte Carlo methods or bootstrap techniques (see Garratt *et al.*, 2006, Ch. 7). For a Bayesian approach, see Zellner (1971, pp. 233–6). In most practical settings, the value of Y_{t-1} will not vary sufficiently to produce large variations in the shape of the density of Y_{t+h}, and so for most practical purposes we can assume that the density is constant across time except for the mean of the distribution. Hence there is little interest in regularly publishing full details of the density forecast from this type of model, as the only element of the density which would change substantially through time is the mean.

Given this obvious limitation of the standard model, when a researcher is interested in producing a regular model-based density forecast there is an obvious need to base it on a model which has a richer structure that allows some interesting time variation in the shape of the density. The earliest and most obvious model to be developed which allowed for this possibility is the ARCH process of Engle (1982), and the associated family of GARCH models which has grown from it. A basic GARCH(1,1) model has the following general form:

$$Y_t = B(L)Y_{t-1} + \varepsilon_t \tag{5.4}$$

$$\varepsilon_t = \omega_t h_t;\ \omega_t \sim N(0,1) \tag{5.5}$$

$$h_t^2 = \alpha_0 + \alpha_1 h_{t-1}^2 + \alpha^2 \varepsilon_{t-1}^2. \tag{5.6}$$

Given the time variation in the variance of the error term, the complete density forecast for Y_{t+h} will also exhibit time variation. A wide range of variants of this basic model have grown up (for an extensive survey, see Bollerslev, Engle and Nelson, 1994), which allow not only for time variation in the variance but also for asymmetry and non-normality. The GARCH family of models has had an enormous impact on econometric modeling and forecasting, especially in the area of finance, but for the purposes being considered here it does have a number of limitations. The first and most obvious is that there are practical difficulties in modeling large systems of equations with GARCH-like structures. While there are a few extensions of the GARCH approach to systems (e.g., Engle and Kroner, 1995), these extensions are not very practical for large systems (beyond five or six variables). The only approaches within the GARCH framework which may be extended to substantial systems are the Orthogonal GARCH model and the Dynamic Conditional Correlation model of Engle (2002), and even these are not easily applied in the context of forecasting a large system of equations. The GARCH structure also imposes a parametric form on the way the density of the forecast variable evolves which may not always be reasonable.

As a result of these limitations, a number of studies have emerged recently which bring together two strands of the literature which both allow fairly general time variation in forecasting models. The first of these strands introduces VAR models which allow time variation in the coefficients; this literature includes Canova (1993), Sims (1993), Stock and Watson (1996) and Cogley and Sargent (2001). The second strand allows for stochastic volatility in the error process of multivariate systems; this includes work by Harvey, Ruiz and Shephard (1994), Jacquier, Polson and Rossi (1995), Kim, Shephard and Chib (1998) and Chib, Nardari and Shephard (2006). Allowing both time variation in the parameters and an error term with stochastic volatility potentially allows considerable variation in the density forecast. Macroeconomists have found this helpful when seeking to explain the "Great Moderation," namely the apparent decline in the volatility of both inflation and output growth in the US since the mid 1980s (see Blanchard and Simon, 2001; Stock and Watson, 2002).

A good example of this approach to generating density forecasts from a complex model is Cogley, Morosov and Sargent (2005), who develop a forecasting Bayesian

VAR (BVAR) model, which incorporates both drifting coefficients and stochastic volatility in the errors. Earlier work by Sims and Zha (1998) considered the use of Bayesian methods to compute fan charts from VAR models. Geweke and Whiteman (2006) review Bayesian methods for the construction of density forecasts, or *posterior* predictive densities. Adolfson, Linde and Villani (2007), in a forecasting application to the Euro-area, use Bayesian methods to produce density forecasts from both DSGE and VAR models.

Cogley, Morosov and Sargent (2005) construct a second-order VAR for RPIX inflation, the output gap and the nominal three-month treasury bill rate, denoted by the vector Y_t:

$$Y_t = X_t'\theta_t + \varepsilon_t, \tag{5.7}$$

where the vector X_t includes lags of Y_t and a constant, and $\varepsilon_t = R_t^{0.5}\xi_t$ is a vector of measurement innovations, where $\xi_t \sim N(0,1)$ and R_t is a stochastic volatility matrix, discussed below.

This model differs from a standard VAR in two important ways. The error process has a stochastic volatility structure, to be described below, and the parameters of the VAR are time-varying and follow a random walk, which is represented by the following joint prior:

$$f(\theta^T, Q) = f(\theta^T \mid Q)f(Q) = f(Q)\prod_{s=0}^{T-1} f(\theta_{s+1} \mid \theta_s, Q), \tag{5.8}$$

where $\theta^T = [\theta_1', \ldots, \theta_T']'$ represents the history of the drifting parameters, $\theta^{T+1,T+F} = [\theta_{T+1}', \ldots, \theta_{T+F}']'$ their potential future paths, and:

$$f(\theta_{t+1} \mid \theta_t, Q) \sim N(\theta_t, Q). \tag{5.9}$$

Thus the parameters are effectively random walks without drift and with a constant covariance structure.

In addition to this, a prior belief is imposed on the VAR that the roots of the lag polynomial must lie inside the unit circle, to ensure that the VAR is stable. This is done by creating a reflecting barrier using an indicator function:

$$I(\theta^T) = \prod_{s=1}^{T} I(\theta_s), \tag{5.10}$$

where the function $I(\theta_s) = 0$ when the roots of the VAR are stable and $I(\theta_s) = 1$ when they are unstable. This reflecting barrier then modifies the random walk prior so that we then have:

$$p(\theta^T, Q) \propto I(\theta^T)f(\theta^T, Q). \tag{5.11}$$

Hence the conditional prior is:

$$p(\theta^T \mid Q) = \frac{I(\theta^T)f(\theta^T \mid Q)}{\int I(\theta^T)f(\theta^T \mid Q)d\theta^T}. \tag{5.12}$$

Finally the reflecting barrier alters the transition density to give:

$$p(\theta_{t+1} \mid \theta_t, Q) \propto I(\theta_{t+1}) f(\theta_{t+1} \mid \theta_t, Q) \int I(\theta^{t+2,T}) f(\theta^{t+2,T} \mid \theta_{t+1}, Q) d\theta^{t+2,T}. \tag{5.13}$$

So far this defines a VAR with parameters which follow a random walk with the additional constraint that the parameters are not allowed to wander into a region with unstable behavior. In addition to this, Cogley, Morozov and Sargent (2005) further extend the model to include a drifting conditional variance. Following the stochastic volatility literature they define:

$$R_t = B^{-1} H_t B^{-1'}, \tag{5.14}$$

where B is lower triangular with unity along the main diagonal and H is assumed diagonal with univariate stochastic volatilities along the main diagonal which evolve as:

$$\ln h_{it} = \ln h_{it-1} + \sigma_i \eta_{it}, \tag{5.15}$$

where the η_{it} are mutually independent volatility innovations and σ_i is a free parameter.

This model then generates a very rich density function as it involves both changing parameters in the VAR and an error term which follows a time-varying stochastic distribution. In terms of a pure model-based density forecast, this type of Bayesian framework is probably as general as is currently possible. Cogley, Morosov and Sargent (2005) detail how to construct the fan chart by simulating the BVAR posterior predictive density, $p(Y^{T+1,T+F} \mid Y^T)$.

5.3.3 Subjective density forecasts

Most of the density forecasts which are produced regularly by national or international institutions are, however, constructed in a much less formal way. The SPF, mentioned in the introduction to this section, is largely judgmental. Each forecaster arrives at his or her own forecast in completely different ways; some may use models but most certainly do not. The key questions of interest here focus on each individual's view of the likely uncertainty surrounding their forecast for inflation and output growth. These individual forecasts are then presented as a set of histograms which are then averaged, using equal weights (see section 5.5), to give a mean density forecast.

The Bank of England began publishing density forecasts at the beginning of 1993. At the beginning of 1996 the Bank changed its methodology somewhat. Before the last quarter of 1995 the Bank's density forecast was implicitly normal. Since the beginning of 1996 the Bank of England has stated clearly that its density forecasts are non-normal, following a two-piece normal distribution; i.e., each side of the mode has a normal shape but they do not have the same standard deviation – hence each side does not represent half of the distribution. This distribution is arrived at as the subjective assessment by the Bank's Monetary Policy Committee (MPC), based partly on past forecast errors, partly on a range of formal models, and partly on subjective judgments regarding the asymmetry of risk in the forecast.

The NIESR began publishing regular density forecasts in 1996, although it had been publishing a mean absolute error for its forecasts of inflation since the second quarter of 1992. The NIESR, in contrast to the Bank, imposes a normal distribution around their point forecast, with the variance determined on the basis of past forecast errors. The window used to calculate this average error turns out to be quite important – there is uncertainty about the variance used to quantify the degree of uncertainty inherent in the fan chart (see Mitchell, 2005). Until 2002 the NIESR used a sample which started in 1982 to estimate the variance. From 2002 onwards it used a sample which began in 1993. Given the general fall in volatility of most economic series in the UK, this change brought about a considerable fall in the estimated size of the variance of its density forecasts, and hence we can see that the choice of this window can be very important in achieving a good forecast.

Mitchell (2005) has found that a break in the unconditional variance of the NIESR's forecast errors around 1993–94 could have been detected via recursive analysis of these forecast errors towards the end of 1996, rather than in 2002. It is therefore important to monitor historical forecast errors regularly, using statistical tests for structural breaks at an unknown point, to help select a period of history which is informative about the future. Stochastic simulation has been discussed as an alternative to historical errors for measuring the uncertainty associated with the forecast. This might be expected to deliver a better measure of uncertainty if a new policy regime (such as a new target for inflation) is adopted.

All of these forecasts can be viewed as being, basically, subjective in nature as they are not the direct result of a formal model. Even in this case the techniques discussed below to evaluate a forecast formally may still be applied, and we argue that an important stage in constructing even a subjective density forecast is an evaluation of the track record of those forecasts.

5.3.4 Combining model-based and subjective density forecasts

In the point forecasting area there has long been a common practice of combining model-based information and subjective judgment. Very few real forecasts are purely the result of a model and, similarly, most forecasters would use a formal model in one form or another to structure the forecasting procedure. It would seem reasonable, therefore, that when we come to consider density forecasts we would similarly want to consider a formal mixture of model and subjective information. One approach would be simply to form two quite separate forecasts, one subjective and one model-based, and to combine them. We discuss density forecast combinations in detail in section 5.5 below, so we will not discuss this possibility here. An alternative, formal means of combining model-based density forecasts with judgment is to adopt a Bayesian approach, with the non-data information summarized by the "prior" (see Sims and Zha, 1998). Waggoner and Zha (1999) consider how to use Bayesian methods to compute density forecasts for conditional forecasts in VAR models, which allow one to impose conditions on the likely future values of endogenous variables. Del Negro and Schorfheide (2004) use a DSGE model as a prior for a VAR and find this improves point forecasting performance as measured by RMSE. Clark and McCracken (2008) essentially impose a hard informative prior

on the steady-state of the VAR model by detrending, in particular inflation, prior to forecasting. Villani (2005) proposes methods to impose an informative prior on the steady-state, in particular the unconditional means of the model, and hence on the long-run forecasts. The priors push the long-run forecasts towards the chosen steady-state, say trend or target inflation.

We will, however, discuss a recently proposed technique which allows a model's density forecast to be altered at a second step in the light of subjective or "off-model" information. This proposal stems from recent work by Robertson, Tallman and Whiteman (2005), which is based on earlier work by Stutzer (1996) and Kitamura and Stutzer (1997).

We are interested in a density forecast for an M-dimensional vector of variables Y. In general, if we are attempting to derive a model-based density forecast based on a possibly nonlinear model it will not be possible to derive this density analytically. However, given the long history of stochastic simulation analysis referenced above, it is usually possible to approximate this density. Thus, assume we have derived a sample of N draws, denoted $\{Y_i\}$, $(i = 1, \ldots, N)$, and that we also have a set of weights $\{\pi_i\}$, $(i = 1, \ldots, N)$. It is then possible to approximate the model's density function by simply weighting together a transformation of the sample of draws. If we have a random sample from the predictive density, the weights are:

$$\pi_i = 1/N, \forall i. \tag{5.16}$$

The mean of the density forecast is:

$$\overline{Y} = \sum_{i=1}^{N} \pi_i Y_i, \tag{5.17}$$

and so on for any other moments.

This, therefore, provides a means to approximate the density forecast of the model. Now assume that, in addition to the model, we have some extra information which we wish to incorporate into the density forecast. We may think of this as a set of moment conditions which we wish the final density forecast to obey. In a very simple example we might wish to locate the mean of a variable at a particular point or we might wish to impose a certain degree of skewness. Suppose we wished the mean of the vector of variables to take some particular set of values, $\overline{g}$. In general, of course, this will not coincide with $\overline{Y}$ as:

$$\sum_{i=1}^{N} \pi_i Y_i \neq \overline{g}. \tag{5.18}$$

The idea then is to create a new set of weights π_i^* such that this restriction holds exactly. Of course, for N sufficiently large, there will generally be an infinite number of sets of weights which would satisfy this restriction, so the idea is to choose a set of weights which satisfy the restriction while, at the same time, remaining as close as possible to the original weights. This, of course, requires a definition of closeness and Robertson *et al.* (2005) establish that, under a set of regularity conditions, the appropriate measure of closeness is the Kullback–Leibler information criterion

(KLIC), which may be stated as:

$$K(\pi^* : \pi) = \sum_{i=1}^{N} \pi_i^* \log\left(\frac{\pi_i^*}{\pi_i}\right). \tag{5.19}$$

Obviously, if $\pi_i^* = \pi_i$ then the KLIC will be zero, so this is in effect measuring the distance between the two distributions. If π_i is uniform then this is sometimes termed the entropy; if it is non-uniform it is termed the relative entropy.

The idea is simply to choose a set of weights π^* which minimize $K(\pi^* : \pi)$ subject to the following set of restrictions:

$$\pi_i^* \geq 0; \ \sum_{i=1}^{N} \pi_i^* = 1; \ \sum_{i=1}^{N} \pi_i^* g(Y_i) = \overline{g}. \tag{5.20}$$

Once we have solved for the new weights we may then easily calculate the new density forecast simply by weighting together all the original draws with the new weights π^*.

Robertson, Tallman and Whiteman (2005) illustrate this technique by considering a small VAR model for the Federal Funds rate, inflation and the output gap and then imposing a set of prior restrictions on the forecast via a set of moment conditions. These include the prior view that inflation should be at its target rate of 2.5% three years in the future, various assumptions regarding the operation of a Taylor rule, and that the output gap eventually closes. They argue that there is some evidence that the restricted forecasts are an improvement over the standard VAR.

Cogley, Morosov and Sargent (2005) combine this technique with their time-varying parameter BVAR with stochastic volatility, outlined above, to generate forecasts for UK inflation which they contrast with the Bank of England's density forecast. They find that to impose the Bank of England's density forecast on the VAR forecasts requires a considerable change to the weighting vector, sometimes referred to as "twisting" the weights. They warn that the relative entropy (KLIC distance) points to "a severe twisting," which may be interpreted as saying that the Bank's density is a long way from the BVAR model and, on this basis, they recommend "a careful review of the evidence being used to twist the forecast."

5.4 The evaluation of density forecasts

In practice, forecasters make successive forecasts of the same event, so-called "fixed-event" forecasts, as well as series of forecasts of fixed length h, so-called "rolling" forecasts. There exist well established statistical techniques for the *ex post* evaluation of both fixed event and rolling point forecasts. For rolling point forecasts these are often based around the RMSE of the forecast relative to the subsequent outturn. Indeed, publication of RMSE statistics is itself a welcome indication that point forecasts are uncertain; in the absence of knowledge of the true loss function, squared-error loss has become the most commonly used function (see Lee, 2007, for a review of loss functions). The unbiasedness and efficiency of point forecasts are also tested using Mincer and Zarnowitz (1969) tests. For fixed-event point

forecasts, again under quadratic loss (see Clements, 1997, for an extension), the most common evaluation method (see Nordhaus, 1987) is to test whether revisions to successive point forecasts of the same event are independent. Clements and Hendry (1998, Ch. 3) provide a textbook discussion of these tests. Patton and Timmermann (2007) establish tests of point forecast optimality when the loss function is unknown.

In turn, the *ex post* evaluation of rolling density forecasts has begun to attract considerable attention, and there now exist established evaluation methods based on both the probability integral transforms and the logarithmic score, as we review below, although there remains some uncertainty about their implementation in practice and their relative merits (see Gneiting, Balabdaoui and Raftery, 2007).

The genesis of these evaluation tests, as indicated in Diebold *et al*. (1998), was the literature on the evaluation of interval forecasts and probability forecasts. Since these tests can also be applied to density forecasts, as a density forecast can always be reduced to an interval forecast, they also constitute a means of evaluating density forecasts. We therefore start our review of extant evaluation methods, in section 5.4.1, with interval forecasts. Interval evaluation tests also serve as the basis for tests of probability event forecasts. But since there are an infinity of possible interval forecasts implied by a given density forecast, rendering it impracticable to test all but plausible (or at least a finite set of) intervals, we then move our attention to "whole" density evaluation methods. In the ensuing discussion we distinguish between distributional (unconditional) and dependence (conditional) aspects of the evaluation tests (see Giacomini and White, 2004).

In contrast, little attention has been paid, at least explicitly, to the fixed-event aspect of density forecasts. This is despite the availability of the aforementioned tests for the efficiency of fixed-event point forecasts – the testable proposition (for weak efficiency) is that, under quadratic loss, forecast revisions should be uncorrelated with past forecast revisions.

Extending efficiency tests to the density case, say using the KLIC (discussed again in more detail below) to measure revisions to successive densities (as in Lahiri and Liu, 2006), is the subject of ongoing research and, to date, there are no established tests to review. But it does appear that KLIC revisions to successive densities do not convey any information on forecast efficiency since conditional variance forecasts, unlike conditional mean forecasts, are predictable even when the forecaster is assumed efficient (see Mitchell, 2007c). Consistent with the "fan" shape of density forecasts published by the Bank of England and others, conditional variance forecasts decline as we get closer to the event of interest. This "trend" precludes testing the efficiency of density forecasts, as with point forecasts, simply by testing the independence of revisions. But Mitchell (2007c) does note that fixed-event density forecasts can always be evaluated similarly to fixed-event point forecasts by reducing them to an event forecast. As we briefly explain in section 5.4.1, fixed-event probability event forecasts can be evaluated just like fixed-event point forecasts. Variance rationality has been examined by Batchelor and Zarkesh (2000).

5.4.1 Interval forecasts

A "good" interval forecast should, at a minimum, have correct coverage *ex post*; i.e., the outturn should fall in the interval the predicted proportion of times: for example, on 95% of occasions for a 95% confidence interval. But, as argued by Christoffersen (1998), in a time series context a "good" interval forecast should not just have correct *unconditional* coverage, but correct *conditional* coverage, so that in volatile periods the interval is wider than in less volatile periods. This means that occurrences inside the interval should not come in clusters over time. This is analogous to expecting independence of orders greater than or equal to h when evaluating a sequence of rolling optimal h-step-ahead point forecasts or optimal fixed-event point forecasts; e.g., see Clements and Hendry (1998, pp. 56–62).

More formally, define I_t as an indicator variable that takes the value 1 if the outcome falls within the interval forecast at time t, and 0 otherwise. Consider an interval forecast for coverage probability p, $0 \leq p \leq 1$. Then Christoffersen (1998) defines a set of *ex ante* forecasts as having correct *conditional* coverage, or as being "efficient" with respect to the information set (say, Ω_{t-1}), if $\mathsf{E}(I_t \mid \Omega_{t-1}) = p$. If $\Omega_{t-1} = \{I_{t-1}, I_{t-2}, \ldots\}$ then this implies $\{I_t\}$ is independent and identically distributed (i.i.d.) Bernoulli with parameter p.

Christoffersen (1998) then suggests a likelihood ratio (LR) test for correct *conditional* coverage. When $\Omega_{t-1} = \varnothing$, the empty set, the test reduces to an *unconditional* test of the null hypothesis that $\mathsf{E}(I_t) = p$. Wallis (2003) describes an asymptotically equivalent Pearson chi-squared test, with the advantage that, unlike the LR tests, its exact distribution can be derived. Wallis (2003) also extends the tests to density forecasts. The extension is based on reducing the density forecast to a k-interval forecast; Boero, Smith and Wallis (2004) explore, as a function of the size of k, the properties of the chi-squared test in small to moderate sample sizes typical to macroeconomics.

Christoffersen (1998) also suggested, and Clements and Taylor (2003) refined, regression-based tests of interval forecasts. They involve estimating:

$$I_t = \alpha + \beta \Omega_{t-1} + \varepsilon_t, \tag{5.21}$$

where the set of interval forecasts are conditionally efficient when $\alpha = p$ and $\beta = 0$, implying that, as before, $\mathsf{E}(I_t \mid \Omega_{t-1}) = p$. Similarly to Mincer–Zarnowitz regressions, these regression-based tests distinguish between conditional and unconditional objectives. The forecasts have correct unconditional coverage when $\alpha = p$, and are conditionally efficient when the forecast "errors" are uncorrelated with information available at the time the forecast was made, i.e., $\alpha = p$ and $\beta = 0$.

Equation (5.21) also serves as the basis for tests of probability event forecasts; see Clements (2004). Consider $p_{t|t-1}$ to be the probability forecast made one period ahead of an event (such as a breach of the inflation target, a) happening at time t; $p_{t|t-1} = P(y_t \geq a \mid \Omega_{t-1})$. Conditional efficiency, $\mathsf{E}\left[I_t \mid \Omega_{t-1}\right] = p_{t|t-1}$, then

implies $\lambda = 1$, $\alpha = 0$ and $\beta = 0$ in the following variant of (5.21):

$$I_t = \lambda p_{t|t-1} + \alpha + \beta \Omega_{t-1} + \varepsilon_t, \tag{5.22}$$

where the indicator variable I_t is redefined with respect to the event forecasts.

It is also worth noting, given our reference above to the evaluation of fixed-event probability event forecasts, that when probability event forecasts are conditionally efficient, which they are when the density forecast from which they are extracted is "correct" (as defined in section 5.4.2.1), we know from the law of iterated expectations that:

$$\mathsf{E}\left\{\mathsf{E}(I_t \mid \Omega_{t-h}) \mid \Omega_{t-h-1}\right\} = \mathsf{E}(I_t \mid \Omega_{t-h-1}). \tag{5.23}$$

This implies:

$$\mathsf{E}(p_{t|t-h} - p_{t|t-h-1} \mid \Omega_{t-h-1}) = 0, \tag{5.24}$$

which says that the revision to the probability event forecast is orthogonal to information available at $(t - h - 1)$, including lagged revisions to the probability event forecasts. Thus a testable proposition for (weakly) efficient fixed-event density forecasts is that revisions to probability forecasts, extracted from the density forecast, are independent. When there is a clear objective, such as a central bank keeping inflation at less than 2%, it is obvious what a to consider. However, for the density forecast to be well calibrated overall, (5.24) needs to hold for all possible a's. Since an infinity of event forecasts can be extracted from the density forecast, in an application evaluating the fixed-event aspect of the SPF density forecasts, Mitchell (2007c) evaluates both over a large number of arbitrary events and over events of specific interest, such as inflation falling in its "comfort zone" of 1–2%.

5.4.2 Rolling density forecasts

When evaluating the performance of density forecasts as a "whole," economists have tended to rely on using goodness-of-fit tests to establish whether the probability integral transforms of the forecast density with respect to the realizations of the variable are uniform or, via a transformation, normal. In contrast, others have employed scoring rules. Both evaluation criteria have proved popular, since they avoid having to estimate the true but unknown conditional density $f(y_t \mid \Psi_{t-h})$ (where the density of the random variable y_t is defined with respect to the total information set Ψ_{t-h} (where the forecasters' information set $\Omega_{t-h} \subset \Psi_{t-h}$), and only require a time series of realizations $\{y_t\}_{t=1}^{T}$.[3] We review both evaluation criteria below.

Derivatives of both evaluation criteria have also been developed in the papers referred to below when interest lies not in the "whole" density but in specific areas, such as the probability of tail events or economic events of interest, such as a (one period) recession.

5.4.2.1 Goodness-of-fit tests: in theory

Diebold, Gunther and Tay (1998) popularized the idea in economics of statistically evaluating a sample of density forecasts based on the probability integral transforms (pit's) of the realization of the variable with respect to the forecast densities. An alternative approach is based on the integrated squared difference between the density forecast and a nonparametric estimate of $f(y_t \mid \Psi_{t-h})$; see Li and Tkacz (2006). We focus on the former approach since it does not require the strict stationarity of y_t.

Diebold, Gunther and Tay (1998) proved that a sequence of estimated h-step-ahead density forecasts, $\{g(y_t \mid \Omega_{t-h})\}_{t=1}^{T}$, for the realizations of the process $\{y_t\}_{t=1}^{T}$, coincides with the (unknown) true densities $\{f(y_t \mid \Psi_{t-h})\}_{t=1}^{T}$ when the sequence of pit's, $z_{t|t-h}$, are uniform variates, where:[4]

$$z_{t|t-h} = \int_{-\infty}^{y_t} g(u \mid \Omega_{t-h})du = G(y_t \mid \Omega_{t-h}); \ (t = 1, \ldots, T). \tag{5.25}$$

Since the correct density forecast will be preferred by all users, irrespective of their loss function, testing the pit's is attractive as it offers a means of evaluating forecasts without the need to specify a loss function. This is convenient given that it is hard to define an appropriate general (economic) loss function, although it is sometimes possible: Clements (2004) provides an evaluation of the Bank of England's fan charts for inflation based on economic as well as statistical loss.

But just as a "good" interval forecast should be correctly calibrated both unconditionally and conditionally, so should a "good" density forecast. This translates into the requirement that, when $h = 1$, $z_{t|t-h}$ is not just uniform but also independently distributed. In other words, one-step-ahead density forecasts are optimal and capture all aspects of the distribution of y_t only when the $z_{t|t-1}$ are independently and uniformly distributed. When $h > 1$ we should expect serial dependence in $z_{t|t-h}$ even for correctly specified density forecasts. Again this is analogous to expecting dependence (an $MA(h-1)$ process) when evaluating a sequence of optimal rolling h-step-ahead point forecasts. There is not, however, a one-for-one relationship between the point forecast errors and $z_{t|t-h}$.

It is important, as stressed by Mitchell and Wallis (2008), to test density forecasts not just unconditionally, via a distributional test, but conditionally via a test for independence. Otherwise one does find, as in Gneiting, Balabdaoui and Raftery (2007) and motivating their advocation of scoring rules, that uniformity of the pit's is a necessary but not sufficient condition for optimal density forecasts.

5.4.2.2 Goodness-of-fit tests: in practice

Following the lead of Diebold, Gunther and Tay (1998), evaluation tests are commonly based on the difference between the empirical distribution of $z_{t|t-h}$ and the cumulative distribution function of a uniform random variable on [0,1], i.e., the 45° line. In many empirical studies, this has simply involved the application of a Kolmogorov–Smirnov or Anderson–Darling test for uniformity. For one-step-ahead forecasts this is often supplemented with a separate test for the independence of

$z_{t|t-h}$. For empirical examples and references, see Clements and Smith (2000), Clements (2004) and Hall and Mitchell (2004).

By taking the inverse normal cumulative density function (c.d.f.) transformation of $z_{t|t-h}$ to give, say, $z^*_{t|t-h}$, the test for uniformity can be considered to be equivalent to one for normality of $z^*_{t|t-h}$; see Berkowitz (2001). For Gaussian forecast densities with mean given by the point forecast, $z^*_{t|t-h}$ is simply the standardized forecast error (outturn minus point forecast divided by the standard error of the Gaussian density forecast). Testing normality is convenient as normality tests are widely seen to be more powerful than uniformity tests. However, testing is complicated by the fact that the impact of dependence on the tests for uniformity/normality is unknown, as is the impact of non-uniformity/non-normality on tests for dependence.

Consequently, various single and joint tests of uniformity/normality and independence have been employed in empirical studies.[5] These include Kolmogorov–Smirnov, Anderson–Darling and Doornik and Hansen (1994) tests for uniformity/normality, Ljung–Box tests and Lagrange multiplier (LM) tests for independence, and Hong (2002), Thompson (2002) and Berkowitz (2001) LR tests for both uniformity/normality and independence. Using Monte Carlo techniques, Noceti, Smith and Hodges (2003) found the Anderson–Darling test to have more power to detect misspecification than the Kolmogorov–Smirnov test (and related distributional tests). However, they maintained an assumption of a random sample and did not consider the effect dependence may have on the performance of the tests. Many of the popular distributional tests, such as the Kolmogorov–Smirnov and Anderson–Darling tests, are not robust to dependence, their properties having been developed under independence.

Parameter uncertainty and dependence Testing uniformity is complicated by both parameter uncertainty and possible dependence in the $z_{t|t-h}$. For a review and derivation of out-of-sample versions of the tests we consider below, see Corradi and Swanson (2006c).

Parameter uncertainty is a concern when the density forecast is model-based and depends on estimated parameters. This is because when parameters are estimated the Kolmogorov test is no longer asymptotically distribution free, meaning that critical values cannot be tabulated as they are dependent on the null hypothesis and the parameter values. Bai (2003) therefore developed a modified Kolmogorov-type test, based on a martingale transformation, which is asymptotically distribution free. While this test has power against violations of uniformity, it does not necessarily have power against violations of independence in the $z_{t|t-h}$.

This means that these Kolmogorov tests require the density forecast not just to capture the distribution of y_t correctly but to be correctly specified dynamically. Following Corradi and Swanson (2006c), let us illustrate what this means via a simple example. Let the true (conditional) density be $f(y_t \mid \Psi_{t-1}) = N(\alpha_1 y_{t-1} + \alpha_2 y_{t-2}, \sigma_2)$, but the density forecast, while normal, be misspecified in terms of its dynamics: $g(y_t \mid \Omega_{t-1}) = N(\alpha_1^* y_{t-1}, \sigma_1)$, where $\alpha_1^* \neq \alpha_1$. In this case $z_{t|t-1}$ is no longer independent but remains uniform. To test the null hypothesis that the

density forecasts are optimal with dynamic misspecification under both the null and alternative hypotheses therefore requires a test for uniformity that is robust to dependence. Use of the traditional Kolomogorov-type tests, including the Bai test, will lead to invalid inference as the critical values are invalid.

Accordingly, Hong (2002) and Corradi and Swanson (2005a) have developed uniformity tests robust to dependence. The Hong test is based on the generalized cross-spectrum; see also Hong, Li and Zhao (2004). Corradi and Swanson suggest a Kolmogorov-type test. At the expense of an assumption of strict stationarity for $\{y_t\}$ and having to use the block bootstrap, which they prove can be used to construct valid critical values, the advantage of the Corradi and Swanson test relative to Hong's is that it converges at a parametric rather than nonparametric rate. In addition, it directly accounts for parameter estimation uncertainty; Hong assumes parameter estimation error vanishes asymptotically.

Multi-step-ahead density forecasts A distinct form of dependence to that caused by dynamic misspecification can be induced in the pit's when forecasting more than one step ahead ($h > 1$). Even when the density forecasts are correctly conditionally calibrated we should expect dependence of order $(h-1)$ because consecutive observations are subject to common shocks. This complicates further the task of evaluating multi-step-ahead ($h > 1$) density forecasts since one risks confounding *good* dependence, explained by $h > 1$, with *bad* dependence, due to dynamic misspecification. Distributional tests applied to the pit's, which are designed to be robust to dependence, will ideally distinguish between *good* and *bad* dependence. Otherwise, when $h > 1$, one risks declaring incorrect density forecasts "correct," on the basis that the pit's are uniform, with the dependence in the pit's dismissed on the grounds that it is not a symptom of dynamic misspecification but attributable to $h > 1$.

Distributional tests designed to accommodate dependence of order $(h - 1)$, i.e., *good* dependence, have been considered. Most simply, as suggested by Diebold, Gunther and Tay (1998), the pit's have been partitioned into $(h-1)$ blocks for which we expect uniformity and independence when the density forecasts are conditionally well-calibrated. For further discussion see Clements and Smith (2000). Dowd (2007) compares, using simulation experiments, alternative methods of dealing with the dependence and finds that it is best to carry out tests on a bootstrapped resample of the pit's designed to be independent. This remains an active area for research, since applied studies continue to employ different evaluation methods in similar contexts.

Joint tests Another option to overcome the deleterious effects of dependence is to consider a joint test for uniformity and independence of $z_{t|t-h}$ (see Hong, 2002). Berkowitz (2001) also presents a parametric, LR, test for the null of standard normality against autoregressive alternatives. While obviously not robust under the null hypothesis to dynamic misspecification, these joint tests do, at least in principle, have power against violations of both uniformity/normality and independence.

For $h = 1$ Berkowitz (2001) proposes a three degrees-of-freedom LR test of the joint null hypothesis of a zero mean, unit variance and independent $z^*_{t|t-1}$ against

$z^*_{t|t-1}$ following an AR(1) process: $z^*_{t|t-1} = \mu + \rho z^*_{t-1|t-2} + \varepsilon_t$, where $\varepsilon_t \sim N(0, \sigma^2)$. The test statistic LR_B is computed as:

$$LR_B = -2\left[L(0,1,0) - L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})\right], \tag{5.26}$$

where $L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})$ is the value of the exact log-likelihood of a Gaussian AR(1) model (e.g., see Hamilton, 1994, p. 119). Under the null $LR_B \sim \chi^2_3$. The test can be readily generalized to higher-order AR models; squared (and higher power) lagged values of $z^*_{t|t-1}$ can also be included in the model in an attempt to pick up nonlinear dependence. The test abstracts from parameter uncertainty, but is expected to perform well in the small samples typical to macroeconomics; in contrast, the nonparametric goodness-of-fit tests discussed above rely on larger samples.

When $h > 1$, recognizing that this LR test is not designed to deal with dependence, Clements (2004) used a two degrees-of-freedom LR test, which drops the test for autocorrelation, to evaluate the Bank of England's year-ahead (h = five quarters) density forecasts. But this test still assumes independence in the construction of the likelihood function; since the likelihood function is misspecified a robust Wald or LM test might be considered instead (see White, 1982). Alternatively, Dowd (2008) suggests that the dependence be mopped up by first fitting an ARMA process to $z^*_{t|t-h}$.

A criticism of these LR tests is the maintained assumption of normality. They only have power to detect non-normality through the first two moments. Consequently, some authors, such as Clements and Smith (2000) and Hall and Mitchell (2004), have supplemented the LR test with a nonparametric normality test, such as the Doornik–Hansen test. But, as Bao, Lee and Saltoglu (2007) explain, one can still construct a Berkowitz-type LR test without maintaining the normality assumption. They let ε_t follow a more general distribution, specifically a semi-nonparametric density, which nests normality. Alternatively, Chen and Fan (2004) generalize Berkowitz (2001) by proposing the use of copula functions to design tests which have power against a wider range of alternative processes. Berkowitz also proposed a censored version of the LR test which focuses on the tails of the forecast density. Diks, Panchenko and van Dijk (2008) show that the censored LR test can be biased when it is used to *compare* alternative density forecasts, rather than just test a given model for goodness-of-fit. Promising new joint tests, using autocontours, have been developed by Gonzalez-Rivera, Senyuz and Yoldas (2007) which are robust to parameter uncertainty.

The KLIC as the loss function Despite the apparent choice over which distributional test to apply to the pit's, which explains the variety used in extant applied work, these evaluation tests can all be related to the KLIC. In particular, following Bao, Lee and Saltoglu (2007), we consider how one of the most popular tests, namely the Berkowitz (2001) LR test, can be directly related to the KLIC. The KLIC can therefore be interpreted as the loss function for density forecast evaluation (see Lee, 2007). As argued by Mitchell and Hall (2005), it offers a unifying framework

for density forecast evaluation, as well as comparison and combination, to which we turn below.

The KLIC offers a measure of distance, or more accurately "divergence," between the "true" but unknown conditional density $f(y_t \mid \Psi_{t-h})$, defined with respect to the total information set Ψ_{t-h}, and the ith conditional density forecast $g(y_t \mid \Omega_{it-h})$, defined with respect to forecaster i's information set Ω_{it-h}:

$$KLIC^i_{t|t-h} = \mathsf{E}\left[\ln\, f(y_t \mid \Psi_{t-h}) - \ln\, g(y_t \mid \Omega_{it-h})\right] = \int f(y_t \mid \Psi_{t-h}) \ln\left\{\frac{f(y_t \mid \Psi_{t-h})}{g(y_t \mid \Omega_{it-h})}\right\} dy_t. \tag{5.27}$$

$KLIC^i_{t|t-h} = 0$ if and only if $g(y_t \mid \Omega_{it-h}) = f(y_t \mid \Psi_{t-h})$. But, as explained, since $f(y_t \mid \Psi_{t-h})$ is unknown even *ex post*, typically density forecasts are evaluated by employing a goodness-of-fit test on the pit's $z_{it|t-h} = \int_{-\infty}^{y_t} g(u \mid \Omega_{it-h})du$. These amount to a test for whether $KLIC^i_{t|t-h} = 0$.

Estimates of $KLIC^i_{t|t-h}$ can be obtained when we follow Bao, Lee and Saltoglu (2007), invoke Proposition 2 of Berkowitz (2001) and note the following equivalence:

$$d^i_{t|t-h} = \ln\, f(y_t \mid \Psi_{t-h}) - \ln g(y_t \mid \Omega_{it-h}) = \ln q(z^*_{it|t-h}) - \ln \phi(z^*_{it|t-h}) = \ln h(z_{it|t-h}), \tag{5.28}$$

where $z^*_{it|t-h} = \Phi^{-1} z_{it|t-h}$, $q(.)$ is the unknown density of $z^*_{it|t-h}$, which needs to be specified, $\phi(.)$ is the standard normal density and Φ is the c.d.f. of the standard normal. Equation (5.28) offers a direct link between the Berkowitz test and the KLIC. For the nonparametric uniformity tests we see that, when $h(z_{it|t-h}) = 1$, as it does under the null of correct conditional calibration, $d^i_{t|t-h} = 0$.

Under some regularity conditions, $\mathsf{E}\left[KLIC^i_{t|t-h}\right]$ can be consistently estimated by sample ($t = 1, \ldots, T$) information:

$$\overline{KLIC}^i_{t-h} = \frac{1}{T}\sum\nolimits_{t=1}^{T} d^i_{t|t-h}, \tag{5.29}$$

where, following Berkowitz (2001), for $h = 1$, we could assume:

$$q(z^*_{it|t-1}) = \phi\left[\left(z^*_{it|t-1} - \mu - \rho z^*_{it-1|t-2}\right)/\sigma\right]/\sigma. \tag{5.30}$$

Noting that the LR test as traditionally written equals:

$$LR^i_B = 2\sum\nolimits_{t=1}^{T}\left[\ln q(z^*_{it|t-1}) - \ln \phi(z^*_{it|t-1})\right], \tag{5.31}$$

reveals that $\overline{KLIC}^i_{t-1} = LR^i_B/2T$.

More general specifications for $q(.)$, allowing ε_t to follow a more general distribution than the Gaussian, could also be specified. When $h > 1$ we should expect dependence, due to overlapping observations, and we might then

consider the two degrees-of-freedom LR test referred to above, where $q(z^*_{it|t-h}) = \phi\left[\left(z^*_{it|t-h} - \mu\right)/\sigma\right]/\sigma$.

5.4.2.3 *Scoring rules*

In contrast to evaluation based on the pit's there is a tradition, particularly within meteorology, of employing scoring rules (see Gneiting and Raftery, 2007, for a review, and Hall and Mitchell 2007; Amisano and Giacomini, 2007; Adolfson, Linde and Villani, 2007, for applications in economics). Scoring rules are (specific) loss functions that assign a numerical score based on the density forecast and the subsequent realization of the variable. They evaluate relative, but not absolute, density forecast performance. Gneiting, Balabdaoui and Raftery (2007) provide a related discussion of the *sharpness* of density forecasts, which refers to the concentration of the density forecast, and argue that, subject to correct calibration, the sharper the better.

Following the aforementioned applied papers, we restrict attention to the logarithmic scoring rule: $S(g(y_t \mid \Omega_{it-h}), y_t) = \ln g(y_t \mid \Omega_{it-h})$, where the density forecast is evaluated at the realisation of the random variable. The logarithmic scoring rule is intuitively appealing as it gives a high score to a density forecast that provides a high probability to the value y_t that materializes. It also conveniently relates to the KLIC; see (5.27). When the predictive density $g(y_t \mid \Omega_{it-h})$ is normal with mean m_{it} and variance v_{it} (defined below):

$$S(g(y_t \mid \Omega_{it-h}), y_t) = -0.5 \ln 2\pi - 0.5 \ln v_{it} - 0.5 \frac{(y_t - m_{it})^2}{v_{it}}, \tag{5.32}$$

indicating that the logarithmic score depends on the conditional forecasts for both the mean and variance. Competing density forecasts can be ranked according to the size of $S(g(y_t \mid \Omega_{it-h}), y_t)$, with higher values indicating better performance. $S(g(y_t \mid \Omega_{it-h}), y_t)$ cannot be used to test the null hypothesis H_0: $\overline{KLIC}^i_{t-h} = 0$, as this can be achieved only if the practitioner specifies $f(.)$, or $q(.)$ or $h(.)$; see (5.28).

5.4.2.4 *Comparing competing density forecasts*

The KLIC, and also $S(g(y_t \mid \Omega_{it-h}), y_t)$ given its relationship with the KLIC, can be used to compare competing density forecasts; Bao, Lee and Saltoglu (2007) developed a test for equal predictive performance. It formalizes previous attempts that visually compared alternative density forecasts according to their relative distance to, say, the uniform distribution; e.g., see Clements and Smith (2000). Bao, Lee and Saltoglu (2007) test is a direct generalization of tests of equal point forecast accuracy popularized by Diehold and Mariano (1995) (DM) and extended by West (1996) and White (2000). These tests assume some, usually a quadratic, loss function.

A test for equal density forecast accuracy of two competing (non-nested) density forecasts $g(y_t \mid \Omega_{1t-h})$ and $g(y_t \mid \Omega_{2t-h})$, both of which may be misspecified, is

then constructed based on $\left\{d_{t|t-h}\right\}_{t=1}^{T}$, where:

$$d_{t|t-h} = \left[\ln f(y_t \mid \Psi_{t-h}) - \ln g(y_t \mid \Omega_{1t-h})\right] - \left[\ln f(y_t \mid \Psi_{t-h}) - \ln g(y_t \mid \Omega_{2t-h})\right], \tag{5.33}$$

$$= \ln g(y_t \mid \Omega_{2t-h}) - \ln g(y_t \mid \Omega_{1t-h}), \tag{5.34}$$

$$= \left[\ln q(z^*_{1t|t-h}) - \ln \phi(z^*_{1t|t-h})\right] - \left[\ln q(z^*_{2t|t-h}) - \ln \phi(z^*_{2t|t-h})\right]. \tag{5.35}$$

The null hypothesis of equal accuracy is then:

$$H_0 : \mathsf{E}(d_{t|t-h}) = 0 \Rightarrow \overline{KLIC}^1_{t-h} - \overline{KLIC}^2_{t-h} = 0. \tag{5.36}$$

The sample mean $\overline{d}$ is defined as:

$$\overline{d} = \frac{1}{T}\sum_{t=1}^{T}\left[\left[\ln q(z^*_{1t|t-h}) - \ln \phi(z^*_{1t|t-h})\right] - \left[\ln q(z^*_{2t|t-h}) - \ln \phi(z^*_{2t|t-h})\right]\right]. \tag{5.37}$$

A test can be constructed since we know that $\overline{d}$, under appropriate assumptions, has the limiting distribution:

$$\sqrt{T}(\overline{d} - \mathsf{E}(d_{t|t-h})) \xrightarrow{d} \mathsf{N}(0, \widehat{v}), \tag{5.38}$$

where Bao, Lee and Saltoglu (2007), following West (1996), discuss estimators $\widehat{v}$ for the long-run asymptotic variance of $d_{t|t-h}$ allowing for parameter uncertainty, e.g., when the forecasts are model-based and the models are estimated using an expanding, not rolling (fixed length), window (see Giacomini and White, 2006). In the absence of parameter uncertainty, the test (5.38) reduces to a DM-type test: $\overline{d}/\sqrt{\frac{S_d}{T}} \xrightarrow{d} \mathsf{N}(0,1)$, where $S_d = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j$ and $\gamma_j = \mathsf{E}(d_{t|t-h}d_{t-j|t-j-h})$. As suggested by White (2000), the test of equal predictive accuracy (5.36) can readily be extended to multiple (greater than two) models.

To avoid having to postulate an unknown density $q(.)$, it is more convenient to couch the test in terms of (5.34) rather than (5.35).[6] In this case we see clearly that the test is equivalent to that proposed by Amisano and Giacomini (2007).

Amisano and Giacomini (2007), independently of Bao, Lee and Saltoglu (2007), proposed tests that can be used to compare the accuracy of density forecasts where evaluation is based on logarithmic scores, e.g., $\ln g(y_t \mid \Omega_{1t-h})$, rather than the pit's. These tests can be superficially related to the traditional Bayesian approach to comparing models using Bayes factors (e.g., see Koop, 2003). When there are no parameters to be estimated, the logarithmic Bayes factor is equal to the difference of the two models' logarithmic scores seen in (5.34) (see Gneiting and Raftery, 2007).

Related approaches of comparing density forecasts statistically have been proposed by Sarno and Valente (2004) and Corradi and Swanson (2006b). Rather than using the KLIC measure of "distance," these rely on the integrated squared difference between the forecast density and the true density (Sarno and Valente) and

the mean square error between the c.d.f. of the density forecast and the true c.d.f., integrated out over different quantiles of the c.d.f. (Corradi and Swanson). Rather than relying on the pit's or the logarithmic score, in both cases they estimate the true density or c.d.f. empirically. y_t is then required to be strictly stationary, an assumption often not supported for economic time-series.

(5.36) is an *unconditional* test for equal forecast accuracy (see Giacomini and White, 2006 (GW)). GW have developed more general *conditional* tests. These test which forecast will be more accurate at a future date, rather than, as with the unconditional tests, testing which forecast was more accurate "on average." One could, for example, then recursively select at time t the best forecasting method for $t+1$. Conditional tests can be straightforwardly implemented in our framework. The null hypothesis of equal conditional forecast accuracy (for one-step-ahead forecasts) amounts to testing $\mathsf{E}(d_{t|t-1} \mid h^*_{t-1}) = \mathsf{E}(h^*_{t-1} d_{t|t-1}) = 0$ $(t = 2, 3, \ldots)$, where h^*_t is a vector of "test functions" which we set equal to $h^*_{t-1} = (1, d_{t-1|t-2})'$. The GW test statistic GW_T can be computed as the Wald statistic:

$$GW_T = T\left(T^{-1}\sum_{t=2}^{T} h^*_{t-1} d_{t|t-1}\right)' \widehat{\Sigma}_T^{-1} \left(T^{-1}\sum_{t=2}^{T} h^*_{t-1} d_{t|t-1}\right), \qquad (5.39)$$

where $\widehat{\Sigma}_T$ is a consistent estimator for the asymptotic variance of $h^*_{t-1} d_{t|t-1}$ and $GW_T \xrightarrow{d} \chi^2_2$. GW note that a robust *HAC* estimator for this variance could be employed, as with DM-type tests, but they also explain that the sample variance is a consistent estimator when one exploits the fact that the null hypothesis implies $\left\{h^*_{t-1}, d_{t|t-1}\right\}_{t=2}^{T}$ is a martingale difference sequence. GW argue that this has the advantage of allowing the data $\{y_t\}$ to be heterogeneous and characterized by arbitrary structural breaks at unknown points. Their test is also valid for nested models.

5.5 The combination of density forecasts

Rather than select a single "best" forecast it can be felicitous to combine competing forecasts. This follows from appreciation of the fact that, although one model may be "better" than the others, we may not select it with probability one; we may not be sure that it is the best forecast. Therefore, if we considered this single forecast alone, we would be overstating its precision. We may better approximate the truth, and account for the uncertainty in model selection, by combining forecasts. Forecast combination also provides a means of reconciling subjective and model-based densities as discussed above (see also Osterholm, 2006).

Indeed, it is well recognized both theoretically and empirically that combining competing individual point forecasts of the same event can deliver more accurate forecasts, in the sense of a lower RMSE (see Bates and Granger, 1969; Stock and Watson, 2004; Timmermann, 2006). The success of combination follows from the fact that individual forecasts may be based on misspecified models, poor estimation or non-stationarities. Moreover, recent work (e.g., Hendry and Clements,

2004) has begun to explore further why point forecast combination works through analytical and Monte Carlo investigation. But, given that measures of uncertainty surrounding a point forecast enhance its usefulness, the natural next step is to consider density forecast combination. While Clements (2006) and Granger, White and Kamstra (1989) have considered, respectively, the combination of event and quantile forecasts, that inevitably involve a loss of information compared with consideration of the "whole" density, the combination of density forecasts has been relatively neglected. In fact, Clements (2003, p. 2) identified this as "an area waiting investigation."

5.5.1 Combination methods

While methods for combining point forecasts are well established and much exploited, less direct attention has been given in econometrics to the combination of density forecasts. This is also a concern in practice since many professional forecasters, particularly central banks, consult more than one forecast. Central bankers often look at what they call a "suite of models." These competing forecasts are produced using both structural macro (usually large-scale but increasingly DSGE) models and atheoretical models, as well as variants in-between. In addition, central bankers routinely add off-model information ("judgment") to model-based forecasts to produce predictive densities. The issue then again arises as to how they should internally reconcile or combine competing density forecasts of the same event to arrive at a single density which is then used to communicate policy.

However, the expert combination literature, more commonly seen in management science and risk analysis journals, has considered density forecast combination, although not evaluation as discussed in section 5.4. This literature adopts a Bayesian approach whereby competing densities are combined by a "decision maker" who views them as data that are used to update a prior distribution (for reviews see Genest and Zidek, 1986; Clemen and Winkler, 1999). There is also, as we discuss in section 5.5.4.1, a related Bayesian literature in econometrics, but only recently has this turned to the combination and evaluation of combined density forecasts. Its use for the combination of subjectively-formed density forecasts is also little discussed.

Within the expert combination literature Clemen and Winkler (1999) distinguish behavioral and mathematical approaches to combination. The behavioral approach seeks to combine experts' opinions by letting the experts interact in some manner to reach a collective opinion. This approach is not considered further since one can imagine many situations in economic forecasting, e.g., when forecasts are model-based, when it is inappropriate. By contrast, mathematical approaches combine the information across experts by using some rule or model. Work has focused on combination rules that satisfy certain properties or axioms. Two common axiomatic approaches are the "linear opinion pool" (Morris, 1974, 1977; Winkler, 1981; Lindley, 1983; Genest and McConway, 1990) and the "logarithmic opinion pool."

The "linear opinion pool" takes a weighted linear combination of the forecasters' probabilities. The combined density is then defined as the finite mixture:

$$p(y_t \mid \Omega_{t-h}) = \sum_{i=1}^{N} w_i g(y_t \mid \Omega_{it-h}), \tag{5.40}$$

where $g(y_t \mid \Omega_{it-h})$ are the h-step-ahead density forecasts of model i ($i = 1, \ldots, N$) of a random variable y_t at time t ($t = 1, \ldots, T$) conditional on the information set Ω_{it-h}, and $\Omega_{t-h} = \cup_{i=1}^{N} \{\Omega_{it-h}\}$. The set of non-negative weights, w_i, sum to unity. The restriction that each weight is positive might be relaxed (for discussion and references, see Genest and Zidek, 1986). In the finite mixture distribution the weights, the mixing proportions, are positive by construction (see Everitt and Hand, 1981). In (5.40) the weights are assumed time-invariant, $w_{it} = w_i$, since below we consider their estimation using sample averages ($t = 1, \ldots, T$). But in general, e.g., when computed on an out-of-sample (recursive) basis, they can be time-varying. (5.40) satisfies certain properties such as the "unanimity" property (if all forecasters agree on a probability then the combined probability agrees also). For further discussion, and consideration of other properties, see Genest and Zidek (1986) and Clemen and Winkler (1999).

The logarithmic opinion pool is defined as:

$$p(y_t \mid \Omega_{t-h}) = k \prod_{i=1}^{N} g(y_t \mid \Omega_{it-h})^{w_i}, \tag{5.41}$$

where k is a normalizing constant. When $w_i = (1/N)$, $p(y_t \mid \Omega_{t-h})$ is proportional to the geometric mean of the experts' distributions. In (5.41) $p(y_t \mid \Omega_{t-h})$ is, in fact, that density forecast "closest," in a KLIC sense, to each of the N competing density forecasts (see Heskes, 1998).

5.5.2 The linear opinion pool

We follow Mitchell and Hall (2005), Wallis (2005), Timmermann (2006, p. 177) and Hall and Mitchell (2007) and focus on density forecast combination via the linear opinion pool. Indeed, (5.40) offers a well understood and much exploited means of combining density forecasts. The SPF, previously the ASA-NBER survey, has essentially used it since 1968 to publish a combined density forecast of inflation, amongst other things, from the individual-level density forecasts which are supplied to it.

Inspection of (5.40) reveals that taking a weighted linear combination of the forecasters' densities can generate a combined density with characteristics quite distinct from those of the forecaster, although this will not be the case for the combination of natural-conjugate densities (Winkler, 1968). For example, if all the forecasters' densities are normal, but with different means and variances, then the combined density will be mixture normal. Mixture normal distributions can have heavier tails than normal distributions, and can therefore potentially accommodate skewness and kurtosis. Combining individual normal density forecasts may mitigate misspecification of the individual densities. As $N \to \infty$ the mixture distribution

is essentially nonparametric and can accommodate any possible distribution. For finite N the mixture distribution still offers a very flexible modeling approach.

Further characteristics of the combined density $p(y_t \mid \Omega_{t-h})$ can be drawn out by defining m_{it} and v_{it} as the mean and variance of forecast i's distribution at time t:

$$m_{it} = \int_{-\infty}^{\infty} y_t g(y_t \mid \Omega_{it-h}) dy_t \text{ and } v_{it} = \int_{-\infty}^{\infty} (y_t - m_{it})^2 g(y_t \mid \Omega_{it-h}) dy_t; (i = 1, \ldots, N).$$

Then the mean and variance of (5.40) are given by:[7]

$$\mathsf{E}\left[p(y_t \mid \Omega_{t-h})\right] = \sum_{i=1}^{N} w_i m_{it}, \tag{5.42}$$

$$\mathsf{Var}\left[p(y_t \mid \Omega_{t-h})\right] = \sum_{i=1}^{N} w_i v_{it} + \sum_{i=1}^{N} w_i \left\{m_{it} - m_t^*\right\}^2. \tag{5.43}$$

(5.43) indicates that the variance of the combined distribution equals average individual uncertainty ("within" model variance) plus disagreement ("between" model variance).[8] This result stands in contrast to that obtained when combining point forecasts, where combination using "optimal" (variance or RMSE minimizing) weights means the RMSE of the combined forecast must be equal to or less than that of the smallest individual forecast (see Bates and Granger, 1969, and, for related discussion in a regression context, Granger and Ramanathan, 1984). Density forecast combination will in general increase the combined variance. However, this increase in uncertainty need not be deleterious; when evaluated the combined density forecast may perform better than the individual density forecasts. Hall and Mitchell (2004) distinguish between combining competing forecasts of various moments of the forecast density and directly combining the individual densities themselves, as with the finite mixture density.

Focusing on the predictive accuracy of the combination, rather than the individual components, the key practical issue is to determine w_i.[9] We consider two methods in sections 5.5.3 and 5.5.4.

5.5.3 Equal weights

Most simply, equal weights, $w_i = 1/N$, have been advocated (see Hendry and Clements, 2004; Smith and Wallis, 2008). Indeed, equal weights are used by the SPF when publishing their combined density forecasts. Also based on equal weights, there are derivative combination methods which use some *ad hoc* rule, such as trimming or thick-modeling (Granger and Jeon, 2004), to eliminate the k% worst performing forecasts and then take an equal weighted average of the remaining forecasts.

As experience of combining point forecasts has taught us, irrespective of its performance in practice, use of equal weights is only one of many options. For example, one popular alternative to equal weights in the point forecast literature,

the so-called regression approach, is to tune the weights to reflect the historical performance of the competing forecasts (e.g., see Granger and Ramanathan, 1984). Choosing the weights via OLS estimation of the realizations of the variable on the competing point forecasts is "optimal," given quadratic loss; the optimal weighted combination of the point forecasts is the most "accurate" point forecast, in the sense of minimum RMSE. In the following section we consider extensions to density forecasts that, essentially, involve choosing the weights to maximize the in-sample or out-of-sample (predictive) "fit" of (5.40).

5.5.4 KLIC minimizing weights

How we measure the accuracy of forecasts is central to how we might choose to combine them. Similar to how RMSE (least squares) has been the historical basis for much analysis of point forecasts, Mitchell and Hall (2005) and Hall and Mitchell (2007) suggest that the KLIC can serve as the basis for density forecast combination, as well as evaluation and combination. The KLIC offers a unifying framework in which to consider choosing the combination weights.

The KLIC *distance* between the true density $f(y_t \mid \Psi_{t-h})$ and the combined density forecast $p(y_t \mid \Omega_{t-h})$ $(t = 1, \ldots, T)$ is defined as:

$$\begin{aligned} KLIC_{t|t-h} &= \int f(y_t \mid \Psi_{t-h}) \ln \left\{ \frac{f(y_t \mid \Psi_{t-h})}{p(y_t \mid \Omega_{t-h})} \right\} dy_t \\ &= \mathsf{E}\left[\ln f(y_t \mid \Psi_{t-h}) - \ln p(y_t \mid \Omega_{t-h}) \right]. \end{aligned} \tag{5.44}$$

The smaller this distance, the closer the density forecast to the true density. $KLIC_{t|t-h} = 0$ if and only if $f(y_t \mid \Psi_{t-h}) = p(y_t \mid \Omega_{t-h})$, which is an "average form" of the rational expectations hypothesis (see Pesaran and Weale, 2006, p. 722).

Given this loss function, Hall and Mitchell (2007) define the "optimal" combined density forecast as:

$$p^*(y_t \mid \Omega_{t-h}) = \sum_{i=1}^{N} w_i^* g(y_t \mid \Omega_{it-h}), \tag{5.45}$$

where the optimal weight vector $\mathbf{w}^* = (w_1^*, .., w_N^*)$ minimizes the KLIC distance between the combined and true density, (5.44). This minimization is achieved as follows:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{1}{T} \sum_{t=1}^{T} \ln p(y_t \mid \Omega_{t-h}), \tag{5.46}$$

where $\frac{1}{T}\sum_{t=1}^{T} \ln p(y_t \mid \Omega_{t-h})$ is the average logarithmic score of the combined density forecast over the sample $t = 1, \ldots, T$; for related discussion in terms of quasi-maximum likelihood estimation, see White (1982). For an analytical discussion of "optimal" pooling using (5.46), see Geweke and Amisano (2008). Minimizing the KLIC distance by maximizing the logarithmic score is convenient as it avoids having to postulate and estimate $f(y_t \mid \Psi_{t-h})$, which is unknown. At the expense of having to make an assumption about the form of $q(.)$, Hall and Mitchell (2007) do consider how, for those goodness-of-fit tests directly related to

the KLIC, such as the Berkowitz (2001) LR test, the KLIC can also be minimized by searching for those weights that minimize LR_B. For other goodness-of-fit tests the relevant test statistic can again be minimized, but since the direct link with the KLIC is lost, the weights that deliver this minimum cannot be interpreted as KLIC minimizing. This is because KLIC minimization using tests based on the pits is only as good as the underlying goodness-of-fit test.

This methodology for combining density forecasts is designed to try and mimic the optimal combination of point forecasts. It is motivated by the desire to obtain the most "accurate" density forecast, in a statistical sense, as measured by the KLIC.

The KLIC minimizing weights, $\mathbf{w}^*$, are the maximum likelihood (ML) estimates of the weights in (5.40). These ML estimates, requiring iteration via the EM algorithm, are given as (see Hamilton, 1994, p. 688):

$$w_i^* = \frac{1}{T} \sum_{t=1}^{T} \frac{g(y_t \mid \Omega_{it-h}) w_i}{p(y_t \mid \Omega_{t-h})}, \tag{5.47}$$

where w_i is the probability at the previous iteration that the data are generated by the ith density. This ML interpretation may be helpful to move from inspection of combination weights to tests of their statistical significance by accounting for their uncertainty using the inverse of the Hessian matrix. This might facilitate tests for "conditional efficiency" (*encompassing*) of forecast i relative to its competitors, tests which have yet to be applied to density forecasts, although a definition introduced by Clemen, Murphy and Winkler (1995) has been discussed by Timmermann (2006, p. 176).

5.5.4.1 *Bayesian Model Averaging (BMA)*

From a Bayesian perspective, the KLIC minimizing weights, (5.46), based on the logarithmic score, have some superficial similarities with a BMA approach. Geweke and Amisano (2008) explain the differences. Hoeting *et al.* (1999), Koop (2003, Ch. 11) and Geweke and Whiteman (2006) provide recent general discussions of BMA methods.

BMA offers a conceptually elegant means of dealing with model uncertainty. BMA is an application of Bayes' theorem; model uncertainty is incorporated into the theorem by treating the set of models S as an additional parameter and then integrating over S, where $S \equiv \{S_i, i = 1, \ldots, N\}$ and the models S_i are defined as continuous density functions $g(y_t \mid \Omega_{it-h})$ for the variable of interest y_t. BMA, especially approximate, methods are also feasible, unlike iterative methods such as (5.46), even for large N.

Specifically, $p(y_t \mid \Omega_{t-h})$ can be interpreted as the posterior density of y_t given "data" Ω_{t-h} and written like (5.40) as:

$$p^{\text{BMA}}(y_t \mid \Omega_{t-h}) = \sum_{i=1}^{N} w_i^{\text{BMA}} g(y_t \mid \Omega_{it-h}), \tag{5.48}$$

where $g(y_t \mid \Omega_{it-h}) = \Pr(y_t \mid S_i, \Omega_{t-h})$, and the weights w_i^{BMA} are the model's posterior probabilities. As shown by Draper (1995) and Hoeting *et al.* (1999), these

weights are given as:

$$w_i^{\text{BMA}} = \Pr(S_i \mid \Omega_{t-h}) = \frac{\Pr(\Omega_{t-h} \mid S_i)\Pr(S_i)}{\sum_{i=1}^{N} \Pr(\Omega_{t-h} \mid S_i)\Pr(S_i)}, \tag{5.49}$$

where all probabilities are implicitly conditional on the set of all models S under consideration.

The posterior probabilities, w_i^{BMA}, provide a natural means of ranking the N models, which relates to the discussion above about the comparison of alternative density forecasts. w_i^{BMA} indicate the probability that model i is the best model in a KLIC sense (see, e.g., Fernandez-Villaverde and Rubio-Ramirez, 2004).

Equal weights combination (see section 5.5.3) attaches equal (prior) weight to each model with no updating of the weights based on the "data."

A relationship between (5.47) and (5.49) is apparent when (i) $\Pr(S_i) = w_i$, and (ii) $\Pr(\Omega_{t-1} \mid S_i) = g(y_{t-1} \mid \Omega_{it-2})$, so that in both the "no parameters" and univariate case the log density of Ω_{t-1}, conditional on model i, equals the logarithmic score.[10] More generally, $\Pr(\Omega_{t-1} \mid S_i)$ is specified only up to unknown parameters (in forecasting model i) and the logarithmic *integrated* likelihood can now be viewed as the relevant scoring rule. Further, as discussed by Andersson and Karlsson (2007), when combining forecasts from different multivariate models, in-sample measures of fit based on the marginal likelihood for the system differ from measures of forecasting performance based on the logarithmic score for the variable of interest. Geweke and Amisano (2008) discuss how the properties of w_i^{BMA} differ from those of w_i^*, considering the case when the "true" model is not in the set of N models under consideration. Unlike w_i^{BMA}, w_i^* do not then necessarily tend to zero or one asymptotically.

In practice, when the density forecasts are model-based, approximate Bayesian methods based on information criteria are often used to proxy w_i^{BMA} (see Garratt *et al.*, 2003; Garratt *et al.,* 2006, Ch. 7; Kapetanios, Labhard and Price, 2008). These methods measure the fit of the models, corrected in line with their parsimony, such that:

$$w_i^{\text{BMA}} = \frac{\exp(\Delta_i)}{\sum_{i=1}^{N} \exp(\Delta_i)} \ (i = 1, \ldots, N), \tag{5.50}$$

where $\Delta_i = IC_i - \max(IC_j)$ and $IC_i = \sum_{t=h+1}^{T} \ln g(y_t \mid \Omega_{it-h}) - K_i$ is the information criterion for model i, such that $\sum_{t=h+1}^{T} \ln g(y_t \mid \Omega_{it-h})$ is the maximized value of the log-likelihood (or logarithmic score) and K_i is a penalty term for over-parameterization. Therefore $\Delta_i = 0$ for the best density and is positive for the other density forecasts; the larger Δ_i the less plausible is density i as the best density. Popular choices are to set K_i equal to the number of freely estimated parameters in model i (k_i), so that IC_i equals the Akaike criterion, or to set $K_i = (k_i/2)\ln(T)$, so that IC_i equals the Schwarz Bayesian information criterion. The Schwarz weights are asymptotically optimal when the true model lies in the set of N models under consideration; otherwise the Akaike weights are likely to perform better. w_i^{BMA} in (5.50) can be interpreted as the probability that the model is the best approximation

to the truth given the data (posterior probabilities) when attaching equal (prior) weight to each model, which a Bayesian would term non-informative priors (see Burnham and Anderson, 2002). Minimizing the Akaike criterion is approximately equivalent to minimizing the expected Kullback–Leibler distance between the true density and the estimated density; again see Burnham and Anderson (2002, Ch. 2, Ch. 6).

The combined density forecast $p^{\text{BMA}}(y_t \mid \Omega_{t-h})$ also has established optimality properties given the set of models under consideration (see Madigan and Raftery, 1994; Raftery and Zheng, 2003). The central estimate from $p^{\text{BMA}}(y_t \mid \Omega_{t-h})$ minimizes mean squared error, the prediction intervals are well calibrated and $p^{\text{BMA}}(y_t \mid \Omega_{t-h})$ maximizes the logarithmic score given $\Pr(S_i)$. On this basis the combined density cannot provide worse forecasts (in-sample, $t = 1, \ldots, T$), as evaluated by the average logarithmic score, than the best individual forecast. This follows from:

$$KLIC_{it} = \mathsf{E}\left[\ln f(y_t \mid \Psi_{t-h}) - \ln g(y_t \mid \Omega_{it-h})\right] \geq 0 \tag{5.51}$$

$$\Rightarrow \mathsf{E}(\ln p^{\text{BMA}}(y_t \mid \Omega_{t-h})) \geq \mathrm{E}(\ln g(y_t \mid \Omega_{it-h})) \Leftrightarrow KLIC_t \leq KLIC^{i}_{t|t-h}, \tag{5.52}$$

($i = 1, \ldots, N$; $t = 1, \ldots, T$). However, BMA implicitly assumes that all the models under consideration are stable. When they are not, perhaps if there are structural breaks, and when the set of models under consideration is not exhaustive (and, therefore, does not include the *true* model), non-Bayesian weights, like equal weights or w_i^*, might be more appropriate and, e.g., deliver a higher log score (see Geweke and Amisano, 2008). In the presence of unknown structural breaks, Pesaran and Timmermann (2007) proved that it can be helpful to average not just over different models, but over different estimation windows for a given model; Assenmacher-Wesche and Pesaran (2008) found equal weights performed best (i.e., delivered the lowest RMSE) in an application to the Swiss economy.

In contrast to methods designed to combine point forecasts (cf. Bates and Granger, 1969), the weights w_i^* or w_i^{BMA} do not allow explicitly for correlation between forecasts. One possibility for future research is to consider the use of copula functions to account for the dependence between the density forecasts (see Jouini and Clemen, 1996; Mitchell, 2007a).

5.5.4.2 Out-of-sample measures of fit

To protect against in-sample overfitting, predictive, rather than in-sample (likelihood-based), measures of fit have also been proposed as the basis for forecast combination (see Eklund and Karlsson, 2007; Kapetanios, Labhard and Price, 2006; Andersson and Karlsson, 2007), although these papers do not consider forecast density evaluation. However, in a specific sense, the marginal likelihood can be interpreted as a measure of out-of-sample predictive performance, as well as a measure of in-sample fit. This is because the marginal likelihood can be written, as seen, as the product of one-step-ahead predictive densities; also see Geweke and Whiteman (2006, pp. 15–17). However, it cannot be decomposed directly into the product of h-step-ahead density forecasts. Moreover, to interpret the marginal

likelihood as an out-of-sample measure of fit relies on the prior being informative (see Eklund and Karlsson, 2007). When an uninformative prior is used, as is common (see Fernandez, Ley and Steel 2001), the marginal likelihood reduces to an in-sample measure of fit. The relationship between in-sample (system) fit and the expected forecasting performance of the variable of interest is also lost in multivariate forecasting models, prompting Andersson and Karlsson (2007) to suggest use of the predictive likelihood for forecast combination since the univariate predictive density of interest can be readily simulated.

These out-of-sample measures of fit involve splitting the available sample ($t = 1, \ldots, T$) in two and measuring the fit of the models according to how well recursively computed h-step ahead forecasts perform, according to the marginal likelihood or logarithmic score, over a hold-out (or predictive) period ($t = t_0, \ldots, T$). Importantly, this means the measure of fit varies according to h. Empirically, the size of the hold-out period also matters. Theoretically, as the size of the hold-out period increases we should expect the weights w_i^{BMA} based on the predictive likelihood to select the correct model consistently (Eklund and Karlsson, 2007). But there is a trade-off when selecting the size of the hold-out period. A small value means the weights adapt quickly to change; but a larger value means the weights are better estimated. Pesaran and Timmermann (2007) have established the optimal trade-off between bias and forecast error variance in regression models subject to one or more structural breaks. Similarly when density forecasting, since economic time-series are known to exhibit structural breaks (Stock and Watson, 1996), unless the true model is in the set of models under consideration we might expect the ranking of the N models to vary over time. This might make it advantageous to consider selecting the size of the hold-out period using the data.

Out-of-sample measures of fit, based on the logarithmic score, can serve as the basis for density forecast combination whether the forecasts are model-based or subjectively formed (see Hall and Mitchell, 2007; Jore, Mitchell and Vahey, 2008). All that is required is the history of the density forecasts. Alternatively, Mitchell and Hall (2005) advocate density forecast combination using what they call 'KLIC' weights, which use the pit's rather than the logarithmic score, measured over a hold-out period, to measure fit. This involves using each model's $\overline{KLIC}_{t-h}^{i}$ value (see (5.29)) computed recursively over the out-of-sample window, to determine $-\Delta_i = \overline{KLIC}_{t-h}^{i} - \min(\overline{KLIC}_{t-h}^{i})$ in (5.50).

5.5.4.3 Empirical applications combining and evaluating density forecasts

Despite considerable experience combining point forecasts, and an extensive BMA literature, there has been little applied macroeconomic work devoted explicitly to density forecast combination and evaluation. For example, in his review of the available empirical literature on forecast combination, Timmermann (2006) focuses on point forecasts. Therefore, a consensus about if and when density forecast combination "works" in macroeconomics has yet to emerge. Nevertheless, a tentative start has been made (see Mitchell and Hall, 2005; Hall and Mitchell, 2007; Jore, Mitchell and Vahey, 2008). An early suggestion is that there can be substantial

gains in forecast accuracy when some variant of KLIC minimizing or BMA, rather than equal, weights are used. This contrasts with the conventional wisdom about point (conditional mean) forecasts, where equal weights are generally preferred. However, for other samples and sets of models, Gerard and Nimark (2008) and Kascha and Ravazzolo (2008) have found that equal weight density combinations can remain hard to beat.

Hall and Mitchell (2007) combine, based on their out-of-sample fit, the one-year-ahead density forecasts of UK inflation published in real time by the Bank of England and the NIESR; they search for the set of weights that maximize the logarithmic score, cf. (5.46). When this is undertaken over the full sample period the combined density forecast is found not to beat the best individual density forecast, the Bank's forecast. The KLIC minimizing weights are unity on the Bank and zero on the NIESR, i.e., there is no combination. This is consistent with the view that combination with an inferior forecast need not help. But, similar to when variance minimizing weights are used to combine point forecasts, combination does not make matters worse (as judged by the logarithmic score, and for point forecasts as judged by the RMSE) when the weights are selected using full-sample information (i.e., in-sample). But when the weights are chosen recursively (i.e., out-of-sample) this need not be the case. Hall and Mitchell (2007) find that an equal weighted combination produces a less accurate density forecast.

Again combining the Bank's and the NIESR's inflation densities, Mitchell and Hall (2005) consider the merits of an alternative means of deriving the combination weights. This involves using $-\Delta_i = \overline{KLIC}^i_{t-h} - \min(\overline{KLIC}^i_{t-h})$ in (5.50), with $\overline{KLIC}^i_{t-h}$ estimated from LR^i_B with two degrees-of-freedom. Mitchell and Hall (2005) then find, even in-sample, that the weighted combination performs worse than the Bank's density, as measured by both the logarithmic score and the LR^i_B test statistic. This can be attributed to the danger, discussed in Hall and Mitchell (2007), that the true density for $z^*_{it|t-h}$, essential to estimation of $\overline{KLIC}^i_{t-h}$, need not be nested by the chosen specification for $q(.)$. Parameter uncertainties in small samples, due to the estimation of, in this case, μ and σ, may also be playing a role. Reconciling and comparing the efficacy of the alternative methods of estimating w^*_i and $w_i{}^{\text{BMA}}$ remains the subject of ongoing research.

The starting point for Jore, Mitchell and Vahey (2008) is the application of Clark and McCracken (2008) to real-time US data. Clark and McCracken (2008) argue that combining real-time point forecasts from VAR models of output, prices and interest rates improves point forecast accuracy in the presence of uncertain model instabilities. Using the same real-time dataset, Jore, Mitchell and Vahey (2008) generalize their approach to consider forecast density combinations and evaluations. Whereas Clark and McCracken (2008) show that the point forecast errors from particular equal-weight pairwise averages are typically comparable to or better than benchmark univariate time series models, Jore, Mitchell and Vahey (2008) show that neither approach produces accurate real-time forecast densities for recent US data. But substantially improved predictive density accuracy is obtained when the competing density forecasts are combined on the basis of the fit of the individual

VAR model forecast densities, as measured by the logarithmic score, over the hold-out period. This weighted combination gives greater weight to models that allow for the shifts in volatilities associated with the Great Moderation. This result again contrasts with that typically found with point forecasts, where equal weighted averages are hard to beat.

5.6 Conclusion

The past decade has seen a considerable increase in the production and use of density forecasts in macroeconomics. This reflects both changes in the dynamics of the macroeconomy, with important shifts in both the level and volatility of many macroeconomic variables making it important to forecast the overall density function rather than just the conditional mean, and the development of econometric models that allow for time variation in the conditional variance, as well as the conditional mean. With this increased use of density forecasts, important new econometric challenges have arisen as macroeconomists seek recourse to a toolbox comparable to that routinely used both to produce and use point (conditional mean) forecasts. This chapter has reviewed recent additions to this toolbox, focusing on the practicality rather than rigour of the methods.

This first involved surveying some methods for the production of density forecasts. We also considered combining model-based and subjective information, by *twisting* the model-based densities to reflect prior (perhaps subjectively formed) information. We should imagine that the techniques considered could be of particular use to professional forecasters, like many central banks, who use both model-based information and judgment when forming their density forecasts. Second, we provided a practical discussion of methods for the *ex post* evaluation of density forecasts. This involved discussion of both rolling and fixed-event density forecasts. Numerous tests for both absolute and relative density forecasting performance, using both the probability integral transforms and scoring rules, were discussed, and their relationship to the Kullback–Leibler information criterion considered. But we stressed the need for further work to establish a consensus on the appropriate test(s), especially in the small-samples typical of macroeconomics and when forecasting more than one step ahead. Finally, again reflecting a common situation for many macroeconomists who forecast from a suite of models, we reviewed methods for the combination of density forecasts to overcome the uncertainty in model selection. Particular focus was given to how to choose the combination weights. In contrast to the conventional wisdom about point forecasts, where equal weights are generally preferred, recent applied work has found that the predictive accuracy of combined density forecasts improves when a greater weight is given to models that allow for the shifts in volatility which have been observed in many economies over the last 20 years.

Acknowledgments

The authors have benefited from numerous discussions with Shaun Vahey and Ken Wallis and thank John Geweke and Terry Mills for useful comments.

Notes

1. However, Granger and Machina (2006) show that, under conditions on the second derivative of the loss function, there is always some point forecast which leads to the same loss as if the decision maker had minimized loss given the density forecast.
2. See Clements and Hendry (1998, Ch. 7) for a complete taxonomy of forecast errors. See also Garratt *et al*. (2006, Ch. 7).
3. We fail to distinguish between random variables and their realizations to avoid introducing yet more notation; but the meaning should be clear from the context. When it is not, we clarify.
4. Diebold, Gunther and Tay (1998) show that the principle generalizes to the case when y_t is multivariate rather than univariate.
5. Alternatively, graphical means of exploratory data analysis are often used to examine the quality of density forecasts (see Diebold, Gunther and Tay, 1998; Diebold, Tay and Wallis, 1999).
6. This also means we do not have to worry about any additional uncertainty introduced because estimation of the loss differential series $d_{t|t-h}$ itself requires parameters (e.g., μ, ρ and σ) to be estimated.
7. Related expressions decomposing the aggregate density (5.40), based on the "law of conditional variances," are seen in Giordani and Söderlind (2003). This law states that for the random variables y_t and i : $V(y_t) = E[V(y_t|i)] + V[E(y_t|i)]$. For criticism see Wallis (2005).
8. For further discussion of the relationship, if any, between dispersion/disagreement and individual uncertainty see Bomberger (1996).
9. The individual forecasts of $g(y_t|\Omega_{it-h})$ are treated as given. Alternatively, one might estimate a finite mixture where the moments of the $g(\cdot)$ are estimated simultaneously with w_i; see Raffery *et al*. (2005) and Geweke and Amisano (2008).
10. The marginal likelihood is the product of the one-step-ahead densities: $\Pr(\Omega_T \mid S_i) = \prod_{t=1}^{T} g(y_t \mid \Omega_{it-1})$.

References

Adolfson, M., J. Linde and M. Villani (2007) Forecasting performance of an open economy DSGE model. *Econometric Reviews* **26**, 289–328.

Amisano, G. and R. Giacomini (2007) Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics* **25**, 177–90.

Andersson, M. and S. Karlsson (2007) Bayesian forecast combination for VAR models. Sveriges Riksbank Working Paper Series No. 216.

Assenmacher-Wesche, K. and M.H. Pesaran (2008) Forecasting the Swiss economy using VECX* models: an exercise in forecast combination across models and observation windows. *National Institute Economic Review* **203**, 91–108.

Bai, J. (2003) Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics* **85**, 531–49.

Bao, Y., T.-H. Lee and B. Saltoglu (2007) Comparing density forecast models. *Journal of Forecasting* **26**, 203–25. (Working Paper version available 2004.)

Barrell, R., S.G. Hall and I. Hurst (2006) Evaluating policy feedback rules using the joint density function of a stochastic model. *Economics Letters* **93**, 1–5.

Batchelor, R. and F. Zarkesh (2000) Variance rationality: evidence from inflation expectations. In F. Gardes and G. Prat (eds.), *Expectations in Goods and Financial Markets*. pp. 156–71. Cheltenham: Edward Elgar.

Bates, J.M. and C.W.J. Granger (1969) The combination of forecasts. *Operational Research Quarterly* **20**, 451–68.

Berkowitz, J. (2001) Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* **19**, 465–74.

Bianchi, C. and G. Calzolari (1980) The one period forecast error in non-linear econometric models. *International Economic Review* **21**, 201–8.
Blake, A. (1996) Forecast error bounds by stochastic simulation. *National Institute Economic Review* **156**, 72–9.
Blanchard, O.J. and J. Simon (2001) The long and large decline in U.S. output volatility. *Brookings Papers on Economic Activity* (1), 135–64.
Boero, G., J. Smith and K.F. Wallis (2004) The sensitivity of chi-squared goodness-of-fit tests to the partitioning of data. *Econometric Reviews* **23**, 341–70.
Boero, G., J. Smith and K.F. Wallis (2008) Uncertainty and disagreement in economic prediction: the Bank of England Survey of External Forecasters. *Economic Journal* **118**(530), 1107–27.
Bollerslev, T., R.F. Engle and D.B. Nelson (1994) Arch models. In R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, *Volume IV*, pp. 2959–3038. Amsterdam: Elsevier.
Bomberger, W. (1996) Disagreement as a measure of uncertainty. *Journal of Money, Credit and Banking* **28**, 381–92.
Brainard, W. (1967) Uncertainty and the effectiveness of monetary policy. *American Economic Review* **57**(2), 411–25.
Britton, E., P. Fisher and J. Whitley (1998) The inflation report projections: understanding the fan chart. *Bank of England Quarterly Bulletin* **38**, 30–7.
Burnham, K.P. and D.R. Anderson (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (second edition). New York: Springer-Verlag.
Canova, F. (1993) Modelling and forecasting exchange rates with a Bayesian time-varying coefficient model. *Journal of Economic Dynamics and Control* **17**, 233–61.
Chen, X. and Y. Fan (2004) Evaluating density forecasts via the copula approach. *Finance Research Letters* **1**, 74–84.
Chib, S., F. Nardari and N. Shephard (2006) Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics* **127**(2), 341–71.
Christoffersen, P. (1998) Evaluating interval forecasts. *International Economic Review* **39**, 841–62.
Clark, T.E. and M.W. McCracken (2008) Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*. Forthcoming. Revision of Federal Reserve Bank of Kansas City Working Paper 06–12.
Clemen, R.T., A.H. Murphy and R.L. Winkler (1995) Screening probability forecasts: contrasts between choosing and combining. *International Journal of Forecasting* **11**, 133–46.
Clemen, R. and R. Winkler (1999) Combining probability distributions from experts in risk analysis. *Risk Analysis* **19**, 187–203.
Clements, M.P. (1997) Evaluating the rationality of fixed-event forecasts. *Journal of Forecasting* **16**, 225–39.
Clements, M.P. (2003) Editorial: some possible directions for future research. *International Journal of Forecasting* **19**, 1–3.
Clements, M.P. (2004), Evaluating the Bank of England density forecasts of inflation. *Economic Journal* **114**, 844–66.
Clements, M.P. (2005) *Evaluating Econometric Forecasts of Economic and Financial Variables*. Basingstoke and London: Palgrave Macmillan.
Clements, M.P. (2006) Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts. *Empirical Economics* **31**, 49–64.
Clements, M.P., P. Franses, J. Smith and D. van Dijk (2003) On SETAR non-linearity and forecasting. *Journal of Forecasting* **22**, 359–75.
Clements, M.P. and D.F. Hendry (1998) *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.
Clements, M.P. and J. Smith (2000) Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting* **19**, 255–76.

Clements, M.P. and J. Smith (2002) Evaluating multivariate forecast densities: a comparison of two approaches. *International Journal of Forecasting* **18**, 397–407.
Clements, M.P. and N. Taylor (2003) Evaluating interval forecasts of high-frequency financial data. *Journal of Applied Econometrics* **18**, 445–56.
Cogley, T., S. Morozov and T.J. Sargent (2005) Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system. *Journal of Economic Dynamics and Control* **29**, 1893–925.
Cogley, T. and T.J. Sargent (2001) Evolving post World War II U.S. inflation dynamics. In NBER *Macroeconomics Annual*. pp. 331–73. Cambridge, Mass.: MIT Press.
Corradi, V. and N.R. Swanson (2006a) Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics* **127**, 779–806.
Corradi, V. and N.R. Swanson (2006b) Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics* **127**, 187–228.
Corradi, V. and N.R. Swanson (2006c) Predictive density evaluation. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1*, pp. 197–284. Amsterdam: North-Holland.
Del Negro, M. and F. Schorfheide (2004) Priors from general equilibrium models for VARs. *International Economic Review* **45**, 643–73.
Diebold, F.X., T. Gunther and A. Tay (1998), Evaluating density forecasts with application to financial risk management. *International Economic Review* **39**, 863–83.
Diebold, F.X., J. Hahn and A. Tay (1999) Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics* **81**, 661–73.
Diebold, F.X. and R.S. Mariano (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253–63.
Diebold, F.X., A. Tay and K.F. Wallis (1999) Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In R. Engle and H. White (eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W.J. Granger.* Oxford: Oxford University Press.
Diks, C., V. Panchenko and D. van Dijk (2008) Partial likelihood-based scoring rules for evaluating density forecasts in tails. Economics Discussion Paper No. 2008/10, University of New South Wales.
Doornik, J.A. and H. Hansen (1994) A practical test for univariate and multivariate normality. Discussion Paper, Nuffield College, Oxford.
Dowd, K. (2007) Validating multiple-period density forecasting models. *Journal of Forecasting* **26**, 251–70.
Dowd, K. (2008) GDP fan charts: an empirical evaluation. *National Institute Economic Review* **203**, 59–67.
Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* **57**, 45–97.
Eklund, J. and S. Karlsson (2007) Forecast combination and model averaging using predictive measures. *Econometric Reviews* **26**(2–4), 329–63.
Elder, R., G. Kapetanios, T. Taylor and T. Yates (2005) Assessing the MPC's fan charts. *Bank of England Quarterly Bulletin* **45**, 326–48.
Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007.
Engle, R.F. (2002) Dynamic conditional correlation: a simple class of multivariate generalised conditional heteroskedasticity models. *Journal of Business and Economic Statistics* **20**(3), 339–50.
Engle, R.F. and K. Kroner (1995) Multivariate simultaneous generalized GARCH. *Econometric Theory* **11**, 122–50.
Everitt, B.S. and Hand D.J. (1981), *Finite Mixture Distributions*. London: Chapman and Hall.

Fair, R.C. (1980) Estimating the predictive accuracy of econometric models. *International Economic Review* **21**, 355–78.
Fair, R.C. (1984) *Specification, Estimation and Analysis of Macroecometric Models*. Cambridge, Mass.: Harvard University Press.
Fernandez, C., E. Ley and M.F.J. Steel (2001) Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**(2), 381–427.
Fernandez-Villaverde, J. and J. Rubio-Ramirez (2004) Comparing dynamic equilibrium economies to data: a Bayesian approach. *Journal of Econometrics* **123**, 153–87.
Garratt, A., K. Lee, M.H. Pesaran and Y. Shin (2003) Forecast uncertainties in macroeconometric modelling: an application to the UK economy. *Journal of the American Statistical Association* **98**, 829–38.
Garratt, A., K. Lee, M. Pesaran and Y. Shin (2006) *Global and National Macroeconometric Modelling: A Long Run Structural Approach*. Oxford: Oxford University Press.
Genest, C. and K.J. McConway (1990) Allocating the weights in the linear opinion pool. *Journal of Forecasting* **9**, 53–73.
Genest, C. and J. Zidek (1986) Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* **1**, 114–35.
Gerard, H. and K. Nimark (2008) Combining multivariate density forecasts using predictive criteria. RBA Discussion Paper 2008-02.
Geweke, J. and G. Amisano (2008) Optimal prediction pools. Working Paper, Department of Economics, University of Iowa.
Geweke, J. and C. Whiteman (2006), Bayesian forecasting. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1*, pp. 3–80. Amsterdam: North-Holland.
Giacomini, R. and H. White (2006) Tests of conditional predictive ability. *Econometrica* **74**, 1545–78.
Giordani, P. and P. Söderlind (2003) Inflation forecast uncertainty. *European Economic Review* **47**, 1037–59.
Gneiting, T., F. Balabdaoui and A.E. Raftery (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69**, 243–68.
Gneiting, T. and A.E. Raftery (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–78.
Gonzalez-Rivera, G., Z. Senyuz and E. Yoldas (2007) *Autocontours: Dynamic Specification Testing*. Riverside, Calif.: Department of Economics, UC Riverside.
Granger, C.W.J. (1969) Prediction with a generalized cost of error function *Operational Research Quarterly* **20**, 199–207.
Granger, C.W.J. and Y. Jeon (2004) Thick modeling. *Economic Modelling* **21**, 323–43.
Granger, C.W.J. and M. Machina (2006) Forecasting and decision theory. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.) *Handbook of Economic Forecasting, Volume 1*, pp. 81–98. Amsterdam: North-Holland.
Granger, C.W.J. and M.H. Pesaran (2000) Economic and statistical measures of forecast accuracy. *Journal of Forecasting* **19**, 537–60.
Granger, C.W.J. and R. Ramanathan (1984) Improved methods of combining forecasts. *Journal of Forecasting* **3**, 197–204.
Granger, C.W.J., H. White and M. Kamstra (1989) Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics* **40**, 87–96.
Hall, S.G. (1986) The application of stochastic simulation techniques to the National Institute's Model 7. *Manchester School* **54**, 180–201.
Hall, S.G. and S.G.B. Henry (1988) *Macroeconomic Modelling*. Amsterdam: North-Holland.
Hall, S.G. and J. Mitchell (2004) Density forecast combination. National Institute of Economic and Social Research Discussion Paper No. 249.
Hall, S.G. and J. Mitchell (2007) Combining density forecasts. *International Journal of Forecasting* **23**, 1–13.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
Harvey, A.C., E. Ruiz and N. Shephard (1994) Multivariate stochastic variance models. *Review of Economics and Statistics* **61**, 247–64.
Hendry, D.F. and M.P. Clements (2004) Pooling of forecasts. *Econometrics Journal* **7**, 1–31.
Heskes, T. (1998) Selecting weighting factors in logarithmic opinion pools. *Advances in Neural Information Processing Systems* **10**, 266–72.
Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky (1999) Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–417.
Hong, Y. (2002) *Evaluation of out-of-sample probability density forecasts*. Cornell University Discussion Paper.
Hong, Y., H. Li and F. Zhao (2004) Out-of-sample performance of discrete-time spot interest rate models. *Journal of Business and Economic Statistics* **22**, 457–73.
Jacquier, E., N. Polson and P. Rossi (1995) Models and priors for multivariate stochastic volatility. CIRANO Working Paper 1995–18.
Jore, A.S., J. Mitchell and S.P. Vahey (2008) Combining forecast densities from VARs with uncertain instabilities. NIESR Discussion Paper No. 303.
Jouini, M.N. and R.T. Clemen (1996) Copula models for aggregating expert opinions. *Operations Research* **44**, 444–57.
Kapetanios, G., V. Labhard and S. Price (2006) Forecasting using predictive likelihood model averaging. *Economics Letters* **91**, 373–9.
Kapetanios, G., V. Labhard and S. Price (2008). Forecasting using Bayesian and information-theoretic model averaging: an application to UK inflation. *Journal of Business and Economic Statistics* **26**, 33–41.
Kascha, C. and F. Ravazzolo (2008) Combining density forecasts: some cross country evidence. Discussion Paper, Norges Bank.
Kim, S., N. Shephard and S. Chib (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economics and Statistics* **65**, 361–93.
Kitamura, Y. and M. Stutzer (1997) An information-theoretic alternative to Generalized Methods of Moments estimation. *Econometrica* **65**, 861–74.
Knight, F. (1921) *Risk, Uncertainty and Profit*, Boston, Mass.: Houghton Mifflin.
Koop, G. (2003) *Bayesian Econometrics*. Chichester: John Wiley & Sons.
Lahiri, K. and F. Liu (2006) Modelling multi-period inflation uncertainty using a panel of density forecasts. *Journal of Applied Econometrics* **21**, 1199–219.
Lee, T.-H. (2007), Loss functions in time series forecasting. In *International Encyclopedia of the Social Science* (second edition), Farmington Hills, Mich.: Macmillan Reference USA.
Li, F. and G. Tkacz (2006) A consistent bootstrap test for conditional density functions with time dependent data. *Journal of Econometrics* **127**, 863–86.
Lindley, D. (1983). Reconciliation of probability distributions. *Operations Research* **31**, 866–80.
Ljungqvist, L. and T.J. Sargent (2000) *Recursive Macroeconomic Theory*. Cambridge: Mass.: MIT Press.
Lutkepohl, H. (1991) *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
Madigan, D.M. and A.E. Raftery (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1335–46.
Mincer, J. and V. Zarnowitz (1969) The evaluation of economic forecasts. In J. Mincer (ed.), *Economic Forecasts and Expectations*. New York: NBER.
Mitchell, J. (2005) The National Institute density forecasts of inflation. *National Institute Economic Review* **193**, 60–9.
Mitchell, J. (2007a) Constructing bivariate density forecasts of inflation and output growth using copulae: modelling dependence using the Survey of Professional Forecasters. NIESR Discussion Paper No. 297.

Mitchell, J. (2007b) Density estimates for real-time eurozone output gap estimates. In G. Mazzi and G. Savio (eds.), *Growth and Cycle in the Eurozone*, pp. 310–20. Basingstoke: Palgrave Macmillan.
Mitchell, J. (2007c) Density forecast revisions and forecast efficiency. NIESR Discussion Paper No. 296.
Mitchell, J. and S.G. Hall (2005) Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR "fan" charts of inflation. *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.
Mitchell, J. and K.F. Wallis (2008) Evaluating density forecasts: is sharpness needed? NIESR Discussion Paper No. 320.
Morris, P. (1974) Decision analysis expert use. *Management Science* **20**, 1233–41.
Morris, P. (1977) Combining expert judgments: a Bayesian approach. *Management Science* **23**, 679–93.
Noceti, P., J. Smith and S. Hodges (2003) An evaluation of tests of distributional forecasts. *Journal of Forecasting* **22**, 447–55.
Nordhaus, W.D. (1987), Forecast efficiency: concepts and applications. *Review of Economics and Statistics* **69**, 667–74.
Orphanides, A. and S. van Norden (2002) The unreliability of output-gap estimates in real time. *Review of Economics and Statistics* **84**, 569–83.
Osterholm, P. (2006) Incorporating judgement in fan charts. Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2006–39.
Patton, A.J. (2006) Modelling asymmetric exchange rate dependence. *International Economic Review* 47, 527–56.
Patton, A.J. and A. Timmermann (2007) Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* **102**, 1172–84.
Pesaran, M.H. and R. Smith (2006) Macroeconomic modelling with a global perspective. *Manchester School* **74**, 24–49.
Pesaran, M.H. and A. Timmermann (2007) Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137**(1), 134–61.
Pesaran, M.H. and M.R. Weale (2006) Survey Expectations. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1* pp. 715–76, Amsterdam: North-Holland.
Raftery, A.E., T. Gneiting, F. Balabdaoui and M. Polakowski (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**, 1155–74.
Raftery, A.E. and Y. Zheng (2003) Long-run performance of Bayesian model averaging. *Journal of the American Statistical Association* **98**, 931–8.
Ravenna, F. (2007) Vector autoregressions and reduced form representations of DSGE models. *Journal of Monetary Economics* **54**, 2048–64.
Robertson, J.C., E.W. Tallman and C.H. Whiteman (2005) Forecasting using relative entropy. *Journal of Money, Credit and Banking* **37**, 383–401.
Sarno, L. and G. Valente (2004) Comparing the accuracy of density forecasts from competing models. *Journal of Forecasting* **23**, 541–57.
Schink, G.R. (1971) Small sample estimates of the variance-covariance matrix of forecast errors for large econometric models: the stochastic simulation technique. University of Pennsylvania PhD dissertation.
Sims, C. (1993) A nine variable probabilistic macroeconomic forecasting model. In J. Stock and M. Watson, (eds.), *Business Cycles, Indicators, and Forecasting. NBER Studies in Business Cycles Volume 28*, pp. 179–214.
Sims, C.A. and T. Zha (1998) Bayesian methods for dynamic multivariate models. *International Economic Review* **39**(4), 949–68.
Smith, J. and K.F. Wallis (2008) A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*. Forthcoming.

Stock, J.H. and M. Watson (2004) Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* **23**, 405–30.
Stock, J.H. and M.W. Watson (1996) Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* **14**, 11–30.
Stock, J.H. and M.W. Watson (2002) Has the business cycle changed and why? In M. Gertler and K. Rogoff (eds.), *NBER Macroeconomics Annual 2002*. Cambridge: Mass.: MIT Press.
Stutzer, M. (1996) A simple non-parametric approach to derivative security valuation. *Journal of Finance* **88**, 11–16.
Svensson, L. (2001) Price stability as a target for monetary policy: defining and maintaining price stability. In Deutsche Bundesbank (ed.), *The Monetary Transmission Process: Recent Developments and Lessons for Europe*, pp. 60–102. New York: Palgrave Macmillan.
Svensson, L.E.O. (2005) Monetary policy with judgment: forecast targeting. *International Journal of Central Banking* **1**, 1–54.
Swanson, E.T. (2004) On signal extraction and non-certainty-equivalence in optimal monetary policy rules. *Macroeconomic Dynamics* **8**, 27–50.
Tay, A.S. and K.F. Wallis (2000) Density forecasting: a survey. *Journal of Forecasting* **19**, 235–54.
Thompson, S. (2002) Evaluating the goodness of fit of conditional distributions, with an application to affine term structure models. Manuscript, Economics Department, Harvard University.
Timmermann, A. (2006) Forecast combinations. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting, Volume 1*, pp. 135–96. Amsterdam: North-Holland.
Villani, M. (2007) *Steady state priors for vector autoregressions*. Sveriges Riksbank Working Paper No. 181 (revised).
Waggoner, D.F. and T. Zha (1999) Conditional forecasts in dynamic multivariate models. *Review of Economics and Statistics* **81**, 639–51.
Wallis, K.F. (2003) Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting* **19**, 165–75.
Wallis, K.F. (2004) An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review* **189**, 64–71.
Wallis, K.F. (2005) Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics* **67**, 983–94.
Wallis, K.F. (2007) Forecast uncertainty, its representation and evaluation. In R.S. Mariano and Y.K. Tse (eds.), *Econometric Forecasting and High-Frequency Data Analysis*, pp. 1–51. Volume 13 of the Lecture Notes Series of the Institute for Mathematical Sciences, National University of Singapore. Singapore: World Scientific.
West, K.D. (1996) Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–84.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
White, H. (2000) A reality check for data snooping. *Econometrica* **68**, 1097–126.
Winkler, R. (1968) The consensus of subjective probability distributions. *Management Science* **15**, 61–75.
Winkler, R. (1981) Combining probability distributions from dependent information sources. *Management Science* **27**, 479–88.
Woodford, M. (2003) *Interest and Prices*. Princeton: Princeton University Press.
Zarnowitz, V. and L. Lambros (1987) Consensus and uncertainty in economic prediction. *Journal of Political Economy* **95**, 591–621.
Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.
Zellner, A. (1986) Biased predictors, rationality and the evaluation of forecasts. *Economics Letters* **21**, 45–48.

This page intentionally left blank

Part III
Time Series Applications

This page intentionally left blank

6
Investigating Economic Trends and Cycles

D.S.G. Pollock

Abstract

Methods are described for extracting the trend from an economic data sequence and for isolating the cycles that surround it. The latter often consist of a business cycle of variable duration and a perennial seasonal cycle.

There is no evident point in the frequency spectrum where the trend ends and the business cycle begins. Therefore, unless it can be represented by a simple analytic function, such as an exponential growth path, there is bound to be a degree of arbitrariness in the definition of the trend.

The business cycle, however defined, is liable to have an upper limit to its frequency range that falls short of the Nyquist frequency, which is the maximum observable frequency in sampled data. This must be taken into account in fitting an ARMA model to the detrended data.

6.1 Introduction

It has been traditional in economics to decompose time series – more accurately described as temporal sequences – into a variety of components, some or all of which may be present in a particular instance. The essential decomposition is a multiplicative one of the form:

$$Y(t) = T(t) \times C(t) \times S(t) \times E(t), \tag{6.1}$$

where:

$T(t)$ is the global trend,

$C(t)$ is a secular cycle, or business cycle,

$S(t)$ is the seasonal variation and

$E(t)$ is an irregular component.

Occasionally, other cycles of relatively long durations are included. Amongst these are the mysterious Kondratieff cycle, reflecting the ebb and flow of human fortunes over half a century, the Shumpeterian cycle, reflecting currents and tides of technological innovation, and the demographic cycle, reflecting the fluctuations in the procreative urges of human beings.

The factors $C(t)$, $S(t)$ and $E(t)$ in equation (6.1) serve to modulate the trend $T(t)$ by inducing fluctuations in its trajectory. They take the generic form of $X(t) = 1 + \xi(t)$, where $\xi(t)$ is a process that fluctuates about a mean of zero.

Typically, $Y(t)$ and $T(t)$ are strictly positive and, therefore, the modulating factors, which are usually deemed to act independently of each other, must also be bounded away from zero. This condition will be satisfied whenever the generic factor can be expressed in an exponential form:

$$X(t) = 1 + \xi(t) = 1 + \sum_{j=1}^{\infty} \frac{\{x(t)\}^j}{j!} = \exp\{x(t)\}. \tag{6.2}$$

In that case, it is appropriate to take logarithms of the expression (6.1) and to work with an alternative additive decomposition instead of the multiplicative one. This is:

$$y(t) = \tau(t) + c(t) + s(t) + \varepsilon(t), \tag{6.3}$$

where $y(t) = \ln Y(t)$, $\tau(t) = \ln T(t)$, $c(t) = \ln C(t)$, $s(t) = \ln S(t)$ and $\varepsilon(t) = \ln E(t)$. An additional assumption, which might be plausible, is that the components $c(t)$, $s(t)$, and $\varepsilon(t)$ have amplitudes that remain roughly constant over time.

In the absence of extraneous information that correlates them with other variables, it is impossible to distinguish the components of (6.3) perfectly, one from another, unless they occupy separate frequency bands. If their bands do overlap, then any separation of the components will be tentative and doubtful. Thus, a sequence that is deemed to represent one of the components will comprise, to some extent, elements that rightfully belong to the other components.

However, as we shall see, the components of an econometric data sequence often reside within bands of frequencies that are separated by wide dead spaces where there are no spectral elements of any significance. The possibility of definitely separating the components is greater than analysts are likely to perceive unless they work in the frequency domain.

The exception concerns the separation of the business cycle from the trend. These components are liable to be merged within a single spectral structure, and there is no uniquely appropriate way of separating them. Their separation depends upon adopting whatever convention best suits the purposes of the analysis. No such difficulties will affect the simple schematic model of the business cycle that we shall consider in the next section.

6.2 A schematic model of the business cycle

In order to extract the modulating components from the data, it is also necessary to remove the trend component from $Y(t)$. To understand what is at issue in detrending the data, it is helpful to look at a simple schematic model comprising an exponential growth trajectory $T(t) = \beta \exp\{rt\}$, with $r > 0$, that is modulated by a exponentiated cosine function $C(t) = \exp\{\gamma \cos(\omega t)\}$ to create a model for the trajectory of aggregate income:

$$Y(t) = \beta \exp\{rt + \gamma \cos(\omega t)\}. \tag{6.4}$$

The resulting business cycles, which are depicted in Figure 6.1, have an asymmetric appearance. Their contractions are of lesser duration than their expansions and they become shorter as the growth rate r increases.

Eventually, when the rate exceeds a certain value, the periods of contraction will disappear and, in place of the local minima, there will be only points of inflection. In fact, the condition for the existence of local minima is that $\omega\gamma > r$, which is to say the product of the amplitude of the cycles and their angular velocity must exceed the growth rate of the trend.

Next, we take logarithms of the data to obtain a model, represented in Figure 6.2, that has additive trend and cyclical components. This gives:

$$\ln\{Y(t)\} = y(t) = \mu + rt + \gamma \cos(\omega t), \tag{6.5}$$

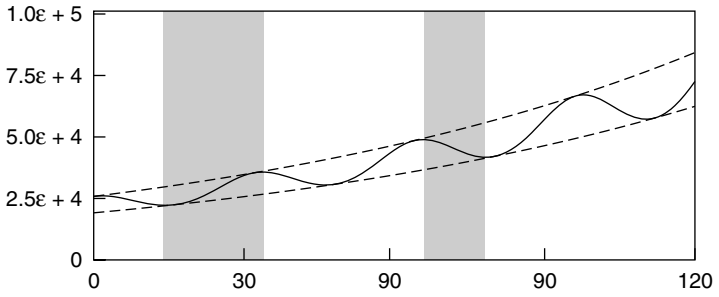


Figure 6.1 The function $Y(t) = \beta \exp\{rt + \gamma \cos(\omega t)\}$ as a model of the business cycle. Observe that, when $r > 0$, the duration of an expansion exceeds the duration of a contraction

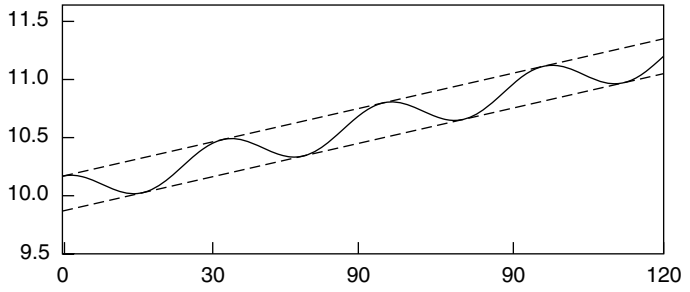


Figure 6.2 The function $\ln\{Y(t)\} = \ln\{\beta\} + rt + \gamma \cos(\omega t)$ representing the logarithmic business cycle data. The durations of the expansions and the contractions are not affected by the transformation

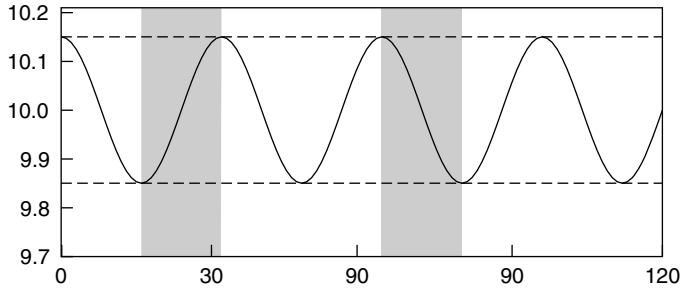


Figure 6.3 The function $\mu + \gamma \cos(\omega t)$ representing the detrended business cycle. The durations of the expansions and the contractions are equal

where $\mu = \ln\{\beta\}$. Since logs effect a monotonic transformation, there is no displacement of the local maxima and minima. However, the amplitude of the fluctuations around the trend, which has become linear in the logs, is now constant.

The final step is to create a stationary function by eliminating the trend. There are two equivalent ways of doing this in the context of the schematic model.

On the one hand, the linear trend $\xi(t) = \mu + rt$ can be subtracted from $y(t)$ to create the pure business cycle $\gamma\cos(\omega t)$. Alternatively, the function $y(t)$ can be differentiated to give $dy(t)/dt = r - \gamma\omega\sin(\omega t)$. When the latter is adjusted by subtracting the growth rate r, by dividing by ω and by displacing its phase by $-\pi/2$ radians – which entails replacing the argument t by $t - \pi/2$ – we obtain the function $\gamma\cos(\omega t)$ again. Through the process of detrending, the phases of expansion and contraction acquire equal durations, and the asymmetry of the business cycle vanishes, as is shown by Figure 6.3.

There is an enduring division of opinion, in the literature of economics, on whether we should be looking at the turning points and phase durations of the original data or at those of the detrended data. The task of finding the turning points is often a concern of analysts who wish to make international comparisons of the timing of the business cycle. There is a belief, which bears investigating, that these cycles are becoming increasingly synchronized amongst member countries of the European Union.

However, since the business cycle is a low-frequency component of the data, it is difficult to find the turning points with great accuracy. In fact, the pinnacles and pits that are declared to be the turning points often seem to be the products of whatever high-frequency components happen to remain in the data after it has been subjected to a process of seasonal adjustment.

If the objective is to compare the turning points of the cycles, then the trends should be eliminated from the data. The countries that are to be compared are liable to be growing at differing rates. From the trended data, it will appear that those with higher rates of growth have shorter recessions with delayed onsets, and this can be misleading.

The various indices of an expanding economy will also grow at diverse rates. Unless they are reduced to a common basis by eliminating their trends, their fluctuations cannot be compared easily. Amongst such indices will be the percentage rate of unemployment, which constitutes a trend-stationary sequence. It would be difficult to collate the turning points in this index with those within a rapidly growing series of aggregate income, which might not exhibit any absolute reductions in its level. A trenchant opinion to the contrary, which opposes the practice of detrending the data for the purposes of describing the business cycle, has been offered by Harding and Pagan (2002).

6.3 The methods of Fourier analysis

A means of extracting the cyclical components from a data sequence is to regress it on a set of trigonometrical functions. The relevant procedures have been described within the context of the statistical analysis of time series by numerous authors, including Bloomfield (1975), Fuller (1976) and Priestley (1989).

In the Fourier decomposition of a finite sequence $\{x_t; t = 0, 1, \ldots, T-1\}$, the T data points are expressed as a weighted sum of an equal number of trigonometrical functions of frequencies that are equally spaced in the interval $[0, \pi]$.

We define $[T/2]$ to be the integer part to $T/2$, which will be $n = T/2$, if T is even, or $(T-1)/2$, if T is odd. Then:

$$\begin{aligned} x_t &= \sum_{j=0}^{[T/2]} \left\{ \alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t) \right\} \\ &= \sum_{j=0}^{[T/2]} \rho_j \cos(\omega_j t - \theta_j). \end{aligned} \tag{6.6}$$

Here, $\rho_j^2 = \alpha_j^2 + \beta_j^2$ and $\theta_j = \tan^{-1}(\beta_j/\alpha_j)$, whilst $\alpha_j = \rho_j \cos(\theta_j)$ and $\beta_j = \rho_j \sin(\theta_j)$. The equality of (6.6) follows in view of the trigonometrical identity:

$$\cos(A - B) = \cos(A)\cos(B) + \sin(A)\sin(B). \tag{6.7}$$

The frequency $\omega_j = 2\pi j/T$ is a multiple of the fundamental frequency $\omega_1 = 2\pi/T$. The latter belongs to a sine and a cosine function that complete a single cycle in the time spanned by the data. The zero frequency ω_0 is associated with the constant function $\cos(\omega_0 t) = \cos(0) = 1$, whereas $\sin(\omega_0 t) = \sin(0) = 0$.

If $T = 2n$ is an even number, then the highest frequency is $\omega_n = \pi$; and, within (6.6), there are $\cos(\omega_n t) = \cos(\pi t) = (-1)^t$ and $\sin(\omega_n t) = \sin(\pi t) = 0$. If T is an odd number, then the highest frequency is $\pi(T-1)/T$, and there is both a sine and a cosine function at this frequency. Counting the number of non-zero functions in both cases shows that they are equal in number to the data points. Therefore, there is a one-to-one correspondence between the data points and the coefficients of the non-zero functions in the Fourier expression of (6.6).

In equation (6.6), the temporal index $t \in \{0, 1, \ldots, T-1\}$ assumes integer values. However, by allowing $t \in [0, T)$ to vary continuously, one can generate a continuous function that interpolates the T data points. This method of generating the continuous function from sampled values may be described as Fourier interpolation. It is notable that the interpolated function is analytic in the sense that it possesses derivatives of all orders.

Although the process generating the data may contain components of frequencies higher than the Nyquist frequency, these will not be detected when it is sampled regularly at unit intervals of time. In fact, the effects on the process of components of frequencies in excess of the Nyquist value will be confounded with those of frequencies that fall below it.

To demonstrate this, consider the case where the process contains a component that is a pure cosine wave of unit amplitude and zero phase and of a frequency ω that lies in the interval $\pi < \omega < 2\pi$. Let $\omega^* = 2\pi - \omega$. Then:

$$\begin{aligned} \cos(\omega t) &= \cos\left\{(2\pi - \omega^*)t\right\} \\ &= \cos(2\pi)\cos(\omega^* t) + \sin(2\pi)\sin(\omega^* t) \\ &= \cos(\omega^* t), \end{aligned} \tag{6.8}$$

which indicates that ω and ω^* are observationally indistinguishable. Here, $\omega^* < \pi$ is described as the alias of $\omega > \pi$.

Since the trigonometrical functions are mutually orthogonal, the Fourier coefficients can be obtained via a set of T simple inner-product formulae, which are in the form of ordinary univariate least squares regressions, with the values of the sine and cosine functions at the points $t = 0, 1, \ldots, T-1$ as the regressors.

Let $c_j = [c_{0,j}, \ldots, c_{T-1,j}]'$ and $s_j = [s_{0,j}, \ldots, s_{T-1,j}]'$ represent vectors of T values of the generic functions $\cos(\omega_j t)$ and $\sin(\omega_j t)$ respectively, and let $x = [x_0, \ldots, x_{T-1}]'$ be the vector of the sample data and $\iota = [1, \ldots, 1]'$ a vector of units. The "regression" formulae for the Fourier coefficients are:

$$\alpha_0 = (\iota'\iota)^{-1}\iota' x = \frac{1}{T}\sum_t x_t = \bar{x}, \tag{6.9}$$

$$\alpha_j = (c_j'c_j)^{-1}c_j' x = \frac{2}{T}\sum_t x_t \cos(\omega_j t), \tag{6.10}$$

$$\beta_j = (s_j's_j)^{-1}s_j' x = \frac{2}{T}\sum_t x_t \sin(\omega_j t), \tag{6.11}$$

$$\alpha_n = (c_n'c_n)^{-1}c_n' x = \frac{1}{T}\sum_t (-1)^t x_t. \tag{6.12}$$

However, in calculating the coefficients, it is more efficient to use the family of specialised algorithms known as fast Fourier transforms, which deliver complex-valued spectral ordinates from which the Fourier coefficients are obtained directly. (See, for example, Pollock 1999.)

The power of a sequence is the time average of its energy. It is synonymous with the mean square deviation which, in statistical terms, is its variance. The power of the sequence $x_j(t) = \rho_j \cos(\omega_j t)$ is $\rho_j^2/2$. This result can be obtained in view of the identity $\cos^2(\omega_j t) = \{1 + \cos(2\omega_j t)\}/2$, for the average of $\cos(2\omega_j t)$ over an integral number of cycles is zero. The assemblage of values $\rho_j^2/2; j = 1, 2, \ldots, [T/2]$ constitutes the power spectrum of $x(t)$, which becomes the periodogram when scaled by a factor T. Their sum equals the variance of the sequence. If $T = 2n$ is even, then:

$$\frac{1}{T}\sum_{t=0}^{T-1}(x_t - \bar{x})^2 = \frac{1}{2}\sum_{j=1}^{n-1}\rho_j^2 + \alpha_n^2. \tag{6.13}$$

Otherwise, if T is odd, then the summation runs up to $(T-1)/2$, and the term α_n^2 is missing.

The indefinite sequence $x(t) = \{x_t; t = 0, \pm 1, \pm 2, \ldots\}$, expressed in the manner of (6.6), is periodic with a period T equal to the length of the sample. It is described as the periodic extension of the sample, and it may be obtained be replicating sample elements over all preceding and succeeding intervals of T points. An alternative way of forming the periodic sequence is by wrapping the sample around a circle of circumference T. Then, the periodic sequence is generated by traveling perpetually around the circle.

6.3.1 Approximations, resampling and Fourier interpolation

By letting $t = 0, \ldots, T-1$ in equation (6.6), the data sequence $\{x_t; t = 0, \ldots, T-1\}$ is generated exactly. An approximation to the sequence may be generated by taking a partial sum comprising the terms of (6.6) that are associated with the Fourier frequencies $\omega_0, \ldots, \omega_d$, where $d < [T/2]$. It is straightforward to demonstrate that this is the best approximation, in the least squares sense, amongst all of the so-called trigonometrical polynomials of degree d that comprise the sinusoidal functions in question.

The result concerning the best approximation extends to the continuous functions that are derived by allowing t to vary continuously in the interval $[0, T)$. That is to say, the continuous function derived from the partial Fourier sum comprising frequencies no higher than $\omega_d = 2\pi d/T$ is the minimum mean square approximation to the continuous function derived from (6.6) by letting t vary continuously.

We may exclude the sine function of frequency ω_d from the Fourier sum. Then the continuous approximation is given by:

$$\begin{aligned} z(t) &= \sum_{j=0}^{d} \left\{ \alpha_j \cos\left(\frac{2\pi jt}{T}\right) \right\} + \sum_{j=1}^{d-1} \left\{ \beta_j \sin\left(\frac{2\pi jt}{T}\right) \right\} \\ &= \sum_{j=0}^{d} \left\{ \alpha_j \cos\left(\frac{2\pi j\tau}{N}\right) \right\} + \sum_{j=1}^{d-1} \left\{ \beta_j \sin\left(\frac{2\pi j\tau}{N}\right) \right\}, \end{aligned} \tag{6.14}$$

where $\tau = tN/T$ with $N = 2d$, which is the total number of the Fourier coefficients. Here, τ varies continuously in $[0, N)$, whereas t varies continuously in $[0, T)$. On the right-hand side, there is a new set of Fourier frequencies $\{2\pi j/N; j = 0, 1, \ldots, d\}$.

The N coefficients $\{\alpha_0, \alpha_1, \beta_1, \ldots, \alpha_{d-1}, \beta_{d-1}, \alpha_d\}$ bear a one-to-one correspondence with the set of N ordinates $\{z_\tau = z(\tau T/N); \tau = 0, \ldots, N-1\}$ sampled at intervals of $\pi/\omega_d = T/N$ from $z(t)$. The consequence is that $z(t)$ is fully represented by the resampled data $z_\tau; \tau = 0, \ldots, N-1$, from which it may be derived by Fourier interpolation.

The result concerning the optimality of the approximation is a weak one; for it is possible that the preponderance of the variance of the data will be explained by sinusoids at frequencies that lie outside the range $[\omega_0, \ldots, \omega_d]$. The matter can be judged with reference to the periodogram of the data sequence, which constitutes a frequency-specific analysis of variance.

Example Figure 6.4 represents the logarithms of the data on quarterly real household expenditure in the UK for the period 1956–2005, through which a linear function had been interpolated so as to pass through the midst of the data points of the first and the final years.

This interpolation is designed to minimize any disjunction that might otherwise occur where the ends of the data sequence meet when it is mapped onto the

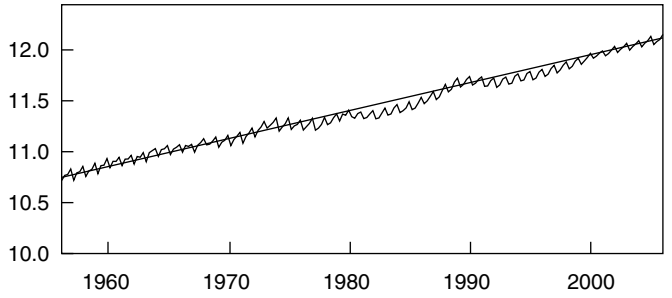


Figure 6.4 The quarterly sequence of the logarithms of household expenditure in the UK for the years 1956–2005, together with an interpolated linear trend

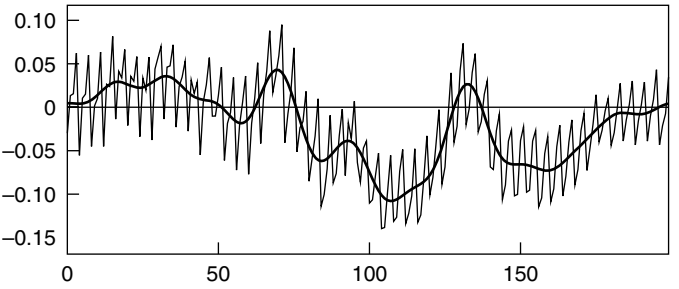


Figure 6.5 The residual deviations of the logarithmic expenditure data from the linear trend of Figure 6.4. The interpolated line, which represents the business cycle, has been synthesized from the Fourier ordinates in the frequency interval $[0, \pi/8]$

circumference of a circle. A trend line fitted by ordinary least squares regression would have a lesser gradient, which would raise the final years above the line. This would be a reflection of the relative prosperity of the times.

The residual deviations of the expenditure data from the trend line of Figure 6.4 are represented in Figure 6.5, and their periodogram is shown in Figure 6.6. Within this periodogram, the spectral structure extending from zero frequency up to $\pi/8$ belongs to the business cycle. The prominent spikes located at the frequency $\pi/2$ and at the limiting Nyquist frequency of π are the property of the seasonal fluctuations. Elsewhere in the periodogram, there are wide dead spaces, which are punctuated by the spectral traces of minor elements of noise.

The slowly varying continuous function interpolated through the deviations of Figure 6.5 has been created by combining a set of sine and cosine functions of increasing frequencies in the manner of (6.14), with the frequencies extending no further than $\omega_d = \pi/8$, and by letting t vary continuously in the interval $[0, T)$. This is a representation of the business cycle as it affects household expenditure. Observe that, since it is analytic, the turning points of this function can be determined via its first derivative.

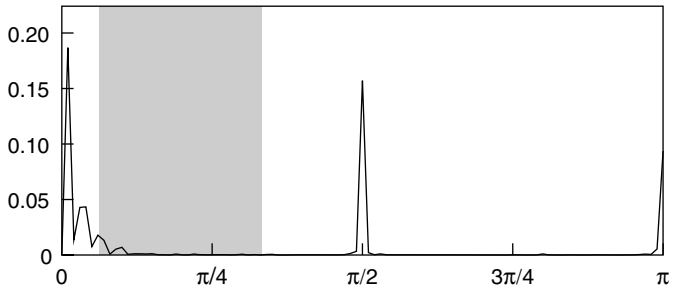


Figure 6.6 The periodogram of the residual sequence of Figure 6.5. A band, with a lower bound of $\pi/16$ radians and an upper bound of $\pi/3$ radians, is masking the periodogram

6.3.2 Complex exponentials

In dealing with the mathematics of the Fourier transform, it is common to use complex exponential functions in place of sines and cosines. This makes the expressions more concise. According to Euler's equations, these are:

$$\cos(\omega_j t) = \frac{1}{2}(e^{\mathrm{i}\omega_j t} + e^{-\mathrm{i}\omega_j t}) \quad \text{and} \quad \sin(\omega_j t) = \frac{-\mathrm{i}}{2}(e^{\mathrm{i}\omega_j t} - e^{-\mathrm{i}\omega_j t}), \tag{6.15}$$

where $\mathrm{i} = \sqrt{-1}$. Therefore, equation (6.6) can be expressed as:

$$x_t = \alpha_0 + \sum_{j=1}^{[T/2]} \frac{\alpha_j + \mathrm{i}\beta_j}{2} e^{-\mathrm{i}\omega_j t} + \sum_{j=1}^{[T/2]} \frac{\alpha_j - \mathrm{i}\beta_j}{2} e^{\mathrm{i}\omega_j t}, \tag{6.16}$$

which can be written concisely as:

$$x_t = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t}, \tag{6.17}$$

where:

$$\xi_0 = \alpha_0, \quad \xi_j = \frac{\alpha_j - \mathrm{i}\beta_j}{2} \quad \text{and} \quad \xi_{T-j} = \xi_j^* = \frac{\alpha_j + \mathrm{i}\beta_j}{2}. \tag{6.18}$$

Equation (6.17) may be described as the inverse Fourier transform. The direct transform is the mapping from the data sequence within the time domain to the sequence of Fourier ordinates in the frequency domain. The relationship between the discrete periodic function and its Fourier transform can be summarized by writing:

$$x_t = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t} \quad \longleftrightarrow \quad \xi_j = \frac{1}{T} \sum_{t=0}^{T-1} x_t e^{-\mathrm{i}\omega_j t} dt. \tag{6.19}$$

For matrix representations of these transforms, one may define:

$$\begin{aligned} U &= T^{-1/2}[\exp\{-\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \\ \bar{U} &= T^{-1/2}[\exp\{\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \end{aligned} \tag{6.20}$$

which are unitary complex matrices such that $U\bar{U} = \bar{U}U = I_T$. Then:

$$x = T^{1/2}\bar{U}\xi \quad \longleftrightarrow \quad \xi = T^{-1/2}Ux, \tag{6.21}$$

where $x = [x_0, x_1, \ldots x_{T-1}]'$ and $\xi = [\xi_0, \xi_1, \ldots \xi_{T-1}]'$ are the vectors of the data and of their Fourier ordinates respectively.

6.4 Spectral representations of a stationary process

The various equations of the Fourier analysis of a finite data sequence can also be used to describe the processes that generate the data. Thus, within the equation:

$$\begin{aligned} y_t &= \sum_{j=0}^{n} \Big\{ \alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t) \Big\} \\ &= \zeta_0 + \sum_{j=1}^{n} \Big\{ \zeta_j e^{\mathrm{i}\omega_j t} + \zeta_j^* e^{-\mathrm{i}\omega_j t} \Big\}, \end{aligned} \tag{6.22}$$

the quantities α_j, β_j can be taken to represent independent real-valued random variables, and the quantities:

$$\zeta_j = \frac{\alpha_j - \mathrm{i}\beta_j}{2} \qquad \text{and} \qquad \zeta_j^* = \frac{\alpha_j + \mathrm{i}\beta_j}{2} \tag{6.23}$$

can be regarded as complex-valued random variables.

The autocovariance of the elements y_t and y_s is given by:

$$\begin{aligned} E(y_t y_s) = \sum_{j=0}^{n}\sum_{k=0}^{n} E\Big[&\zeta_j \zeta_k e^{\mathrm{i}(\omega_j t + \omega_k s)} + \zeta_j \zeta_k^* e^{\mathrm{i}(\omega_j t - \omega_k s)} \\ &+ \zeta_j^* \zeta_k e^{\mathrm{i}(\omega_k s - \omega_j t)} + \zeta_j^* \zeta_k^* e^{-\mathrm{i}(\omega_j t + \omega_k s)} \Big]. \end{aligned} \tag{6.24}$$

The condition of stationarity requires that the covariance should be a function only of the temporal separation $|t - s|$ of y_t and y_s. For this, it is necessary that:

$$E(\zeta_j \zeta_k) = E(\zeta_j^* \zeta_k^*) = E(\zeta_j^* \zeta_k) = E(\zeta_j \zeta_k^*) = 0, \tag{6.25}$$

whenever $j \neq k$. Also, the conditions:

$$E(\zeta_j^2) = 0 \quad \text{and} \quad E(\zeta_j^{*2}) = 0, \tag{6.26}$$

must hold for all j. For (6.25) and (6.26) to hold, it is sufficient that:

$$E(\alpha_j \beta_k) = 0 \quad \text{for all} \quad j, k, \tag{6.27}$$

and that:

$$E(\alpha_j \alpha_k) = E(\beta_j \beta_k) = \begin{cases} 0, & \text{if } j \neq k; \\ \sigma_j^2, & \text{if } j = k. \end{cases} \tag{6.28}$$

An implication of the equality of the variances of α_j and β_j is that the phase angle θ_j is uniformly distributed in the interval $[-\pi, \pi]$.

Under these conditions, the autocovariance of the process at lag $\tau = t - s$ will be given by:

$$\gamma_\tau = \sum_{j=0}^{n} \sigma_j^2 \cos \omega_j \tau. \tag{6.29}$$

The variance of the process is just:

$$\gamma_0 = \sum_{j=0}^{n} \sigma_j^2, \tag{6.30}$$

which is the sum of the variances of the n individual periodic components. This is analogous to equation (6.13).

The stochastic model of equation (6.22) may be extended to encompass processes defined over the entire set of positive and negative integers as well as processes that are continuous in time. First, we may consider extending the length T of the sample indefinitely. As T and n increase, the Fourier coefficients become more numerous and more densely packed in the interval $[0, \pi]$. Also, given that the variance of the process is bounded, the variance of the individual coefficients must decrease.

To accommodate these changes, we may write $\alpha_j = dA(\omega_j)$ and $\beta_j = dB(\omega_j)$, where $A(\omega)$, $B(\omega)$ are cumulative step functions with discontinuities at the points $\{\omega_j; j = 0, \ldots, n\}$. In the limit, the summation in (6.22) is replaced by an integral, and the expression becomes:

$$\begin{aligned} y(t) &= \int_0^\pi \{\cos(\omega t) dA(\omega) + \sin(\omega t) dB)\omega)\} \\ &= \int_{-\pi}^{\pi} e^{\mathrm{i}\omega t} dZ(\omega), \end{aligned} \tag{6.31}$$

where:

$$\begin{aligned} dZ(\omega) &= \frac{1}{2}\{dA(\omega) - \mathrm{i} dB(\omega)\} \quad \text{and} \\ dZ(-\omega) &= dZ^*(\omega) = \frac{1}{2}\{dA(\omega) + \mathrm{i} dB(\omega)\}. \end{aligned} \tag{6.32}$$

Also, $y(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ stands for a doubly-infinite data sequence.

The assumptions regarding $dA(\omega)$ and $dB(\omega)$ are analogous to those regarding the random variables α_j and β_j, which are their prototypes. It is assumed that $A(\omega)$ and $B(\omega)$ represent a pair of stochastic processes of zero mean, which are indexed on the continuous parameter ω. Thus:

$$E\{dA(\omega)\} = E\{dB(\omega)\} = 0. \tag{6.33}$$

It is also assumed that the two processes are mutually uncorrelated and that non overlapping increments within each process are uncorrelated. Thus:

$$\begin{aligned} E\{dA(\omega)dB(\lambda)\} &= 0 \quad \text{for all} \quad \omega, \lambda, \\ E\{dA(\omega)dA(\lambda)\} &= 0 \quad \text{if} \quad \omega \neq \lambda, \\ E\{dB(\omega)dB(\lambda)\} &= 0 \quad \text{if} \quad \omega \neq \lambda. \end{aligned} \tag{6.34}$$

The variance of the increments is given by:

$$V\{dA(\omega)\} = V\{dB(\omega)\} = 2dF(\omega). \tag{6.35}$$

The function $F(\omega)$, which is defined provisionally over the interval $[0, \pi]$, is described as the spectral distribution function. The properties of variances imply that it is a non decreasing function of ω. In the case where the process $y(t)$ is purely random, $F(\omega)$ is a continuous differentiable function. Its derivative $f(\omega)$, which is nonnegative, is described as the spectral density function.

The domain of the functions $A(\omega)$, $B(\omega)$ may be extended from $[0, \pi]$ to $[-\pi, \pi]$ by regarding $A(\omega)$ as an even function such that $A(-\omega) = A(\omega)$ and by regarding $B(\omega)$ as an odd function such that $B(-\omega) = -B(\omega)$. Then, $dZ^*(\omega) = dZ(-\omega)$, in accordance with (6.32). From the conditions of (6.34), it follows that:

$$\begin{aligned} E\{dZ(\omega)dZ^*(\lambda)\} &= E\{dZ(\omega)dZ(-\lambda)\} = 0 \quad \text{if} \quad \omega \neq \lambda, \\ E\{dZ(\omega)dZ^*(\omega)\} &= E\{dZ(\omega)dZ(-\omega)\} = dF(\omega), \end{aligned} \tag{6.36}$$

where the domain of $F(\omega)$ is now the interval $[-\pi, \pi]$.

The sequence of the autocovariances of the process $y(t)$ may be expressed in terms of the spectrum of the process. From (6.36), it follows that the autocovariance of $y(t)$ at lag $\tau = t - s$ is given by:

$$\begin{aligned} \gamma_\tau = C(y_t, y_s) &= E\left\{ \int_\omega e^{i\omega t} dZ(\omega) \int_\lambda e^{i\lambda s} dZ(\lambda) \right\} \\ &= \int_\omega \int_\lambda e^{i\omega t} e^{i\lambda s} E\{dZ(\omega)dZ(\lambda)\} \\ &= \int_\omega e^{i\omega\tau} E\{dZ(\omega)dZ^*(\omega)\} \\ &= \int_{-\pi}^{\pi} e^{i\omega\tau} dF(\omega). \end{aligned} \tag{6.37}$$

In the case of a continuous spectral distribution function, we may write $dF(\omega) = f(\omega)d\omega$ in the final expression, where $f(\omega)$ is the spectral density function. If $f(\omega) = \sigma^2/2\pi$, then there is $\gamma_0 = \sigma^2$ and $\gamma_\tau = 0$ for all $\tau \neq 0$, which are the characteristics of a white-noise process comprising a sequence of independently and identically distributed random variables. Thus, a white-noise process has a uniform spectral density function.

The second way of extending the model is to allow the rate of sampling to increase indefinitely. In the limit, the sampled sequence becomes a continuum. Equation (6.31) will serve to represent a continuous process on the understanding that t is now a continuous variable. However, if the discrete-time process has been subject to aliasing, then the range of the frequency integral will increase as the rate of sampling increases.

Under any circumstances, it seems reasonable to postulate an upper limit to the range of the frequencies comprised by a stochastic process. However, within the conventional theory of continuous stochastic processes, it is common to consider an unbounded range of frequencies. In that case, we obtain a spectral representation of a stochastic process of the form:

$$y(t) = \int_{-\infty}^{\infty} e^{i\omega t} dZ(\omega). \tag{6.38}$$

This representation is capable, nevertheless, of subsuming a process that is limited in frequency. If the bandwidth of $Z(\omega)$ is indeed unbounded, then (6.38) becomes the spectral representation of a process comprising a continuous succession of infinitesimal impacts, which generates a trajectory that is everywhere continuous but nowhere differentiable.

Example Figure 6.7 shows the spectral density function of an autoregressive moving average ARMA(2, 2) process $y(t)$, described by the equation $\alpha(z)y(z) = \mu(z)\varepsilon(z)$, where $\alpha(z)$ and $\mu(z)$ are quadratic polynomials and $y(z)$ and $\varepsilon(z)$ are, respectively, the z-transforms of the data sequence $y(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ and of a white-noise sequence $\varepsilon(t) = \{\varepsilon_t; t = 0, \pm 1, \pm 2, \ldots\}$ of independently and identically distributed random variables.

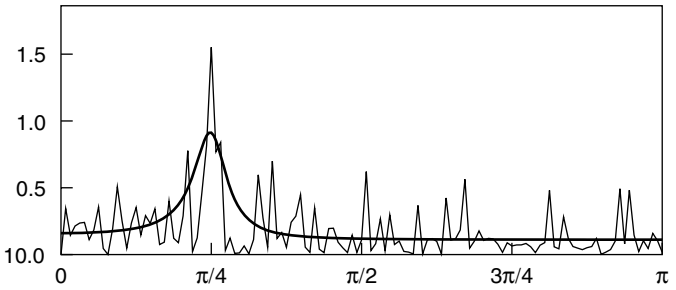


Figure 6.7 The periodogram of 256 points of a pseudo-random ARMA(2, 2) process overlaid by the spectral density function of the process

The ARMA(2, 2) process has been formed by the additive combination of a second-order autoregressive AR(2) process and an independent white-noise process. The autoregressive polynomial is $\alpha(z) = 1 + 2\rho\cos(\theta)z + \rho^2 z^2$, which has conjugate complex roots of which the polar forms are $\rho\exp\{\pm i\theta\}$. In the example, the modulus of the roots is $\rho = 0.9$ and their argument is $\theta = \pi/4$ radians.

The spectral density function attains a non-zero minimum at $\omega = \pi$. However, it is possible to decompose the ARMA(2, 2) process into an ARMA(2, 1) process and a white-noise component that has the maximum variance compatible with such a decomposition. This is a so-called canonical decomposition of the ARMA process. The moving-average polynomial of the resulting ARMA(2, 1) process is $1 + z$, which has a zero at $\omega = \pi$. By maximizing the variance of the white-noise component, an ARMA component is derived that is as smooth and as regular as possible.

Canonical decompositions are entailed in a method for extracting unobserved components from data sequences described by autoregressive integrated moving average (ARIMA) models, which will be discussed in section 6.6.3.

Figure 6.7 also shows a periodogram that has been calculated from a sample of 256 points generated by the ARMA(2, 2) process. Its volatility contrasts markedly with the smoothness of the spectrum. The periodogram has half as many ordinates as the data sequence and it inherits this volatility directly from the data. A nonparametric estimate of the spectrum may be obtained by smoothing the ordinates of the periodogram with an appropriately chosen moving average, or by subjecting the empirical autocovariances to an equivalent weighting operation before transforming them to the frequency domain.

6.4.1 The frequency-domain analysis of filtering

It is a straightforward matter to derive the spectrum of a process $y(t)$ formed by mapping the process $x(t)$ through a linear filter. If:

$$x(t) = \int_{\omega} e^{i\omega t} dZ_x(\omega), \tag{6.39}$$

then the filtered process is:

$$\begin{aligned} y(t) &= \sum_j \psi_j x(t-j) \\ &= \sum_j \psi_j \left\{ \int_{\omega} e^{i\omega(t-j)} dZ_x(\omega) \right\} \\ &= \int_{\omega} e^{i\omega t} \left(\sum_j \psi_j e^{-i\omega j} \right) dZ_x(\omega). \end{aligned} \tag{6.40}$$

On writing $\sum \psi_j e^{-i\omega j} = \psi(\omega)$, which is the frequency response function of the filter, this becomes:

$$\begin{aligned} y(t) &= \int_{\omega} e^{i\omega t} \psi(\omega) dZ_x(\omega) \\ &= \int_{\omega} e^{i\omega t} dZ_y(\omega). \end{aligned} \tag{6.41}$$

If the process $x(t)$ has a spectral density function $f_x(\omega)$, which will allow one to write $dF(\omega) = f(\omega)d\omega$ in equation (6.36), then the spectral density function $f_y(\omega)$ of the filtered process $y(t)$ will be given by:

$$\begin{aligned} f_y(\omega)d\omega &= E\{dZ_y(\omega)dZ_y^*(\omega)\} \\ &= \psi(\omega)\psi^*(\omega)E\{dZ_x(\omega)dZ_x^*(\omega)\} \\ &= |\psi(\omega)|^2 f_x(\omega)d\omega. \end{aligned} \tag{6.42}$$

The complex-valued frequency-response function $\psi(\omega)$, which characterizes the linear filter, can be written in polar form as:

$$\psi(\omega) = |\psi(\omega)|e^{-i\theta(\omega)}, \tag{6.43}$$

The function $|\psi(\omega)|$, which is described as the gain of the filter, indicates the extent to which the amplitude of the cyclical components of which $x(t)$ is composed are altered in the process of filtering.

When $x(t) = \varepsilon(t)$ is a white-noise sequence of independently and identically distributed random variables of variance σ^2, equation (6.42) gives rise to the expression $f_y(\omega) = \sigma^2|\psi(\omega)|^2 = \sigma^2\psi(\omega)\psi^*(\omega)$, which is the spectral density function of $y(t)$. Then, it is helpful to use the notation of the z-transform whereby $\psi(\omega)$ is written as $\psi(z) = \sum_j \psi_j z^j; z = e^{-i\omega}$. If we allow z to be an arbitrary complex number, then we can define the autocovariance generating function $\gamma(z) = \sum_\tau \gamma_\tau z^\tau$ wherein $\gamma_\tau = E(y_t y_{t-\tau})$. This takes the form of:

$$\gamma(z) = \sigma^2\psi(z)\psi(z^{-1}). \tag{6.44}$$

Example Figure 6.8 depicts the squared gain of the difference operator $\nabla(z) = 1 - z$, which is the curve labeled D. The squared gain of $\nabla(z)$ is obtained by setting $z = \exp\{-i\omega\}$ within $|\nabla(z)|^2 = (1 - z)(1 - z^{-1})$ to give $D(\omega) = 2 - 2\cos(\omega)$, whence $W(\omega) = D^{-1}(\omega)$ can be obtained, which is the squared gain of the summation operator. The product of $D(\omega)$ and $W(\omega)$ is the constant function $N(\omega) = 1$, which also represents the spectral density function or power spectrum of a white-noise process with a variance of $\sigma^2 = 2\pi$. Likewise, $W(\omega)$ represents the pseudo-spectrum of a first-order random walk.

This is not a well-defined spectral density function, since the random walk does not constitute a stationary process of a sort that can be defined over a doubly-infinite set of time indices. The unbounded nature of $W(\omega)$ as $\omega \to 0$ is testimony

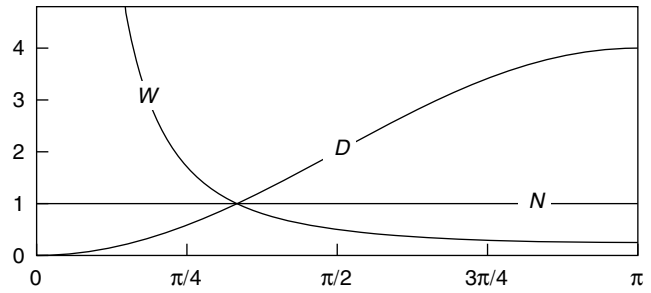


Figure 6.8 The squared gain of the difference operator, labeled D, and that of the summation operator, labeled W

to the fact that the variance of the random walk process is proportional to the time that has elapsed since its start-up. The variance will be unbounded if the start-up is in the indefinite past.

6.5 Stochastic accumulation

In the schematic model of the economy, we have envisaged business cycle fluctuations that are purely sinusoidal, and we have considered a trend that follows an exponential growth path. In a realistic depiction of an economy, both of these functions are liable to be more flexible and more variable through time.

Whereas, in some eras, a linear function, interpolated by least squares regression through the logarithms of the data, will serve as a benchmark about which to measure the cyclical economic activities, the latter usually require to be modeled by a stochastic process. It is arguable that the trend should also be modeled by a stochastic function.

A further feature of the schematic model, which is at odds with the available data, is the continuous nature of its functions. Whereas the processes that generate the data can be thought of as operating in continuous time, the sampled data are sequences of values that are indexed by dates at equal intervals. These data are liable to be modeled via discrete-time stochastic processes. Therefore, some attention needs be paid to the relationship between the discrete data and the underlying continuous process.

The theory of continuous-time stochastic models has been summarized by Bergstrom (1984, 1988), who researched the subject over a 40-year period, beginning in the mid 1960s. His posthumous contributions are to be found in Bergstrom and Nowman (2007), where the contributions of other authors are also referenced.

A linear stochastic process must have a *primum mobile* or forcing function, which is liable to be a stationary process. For the usual discrete-time processes, this is a white-noise sequence of independently and identically distributed random variables. In the theory of continuous stochastic processes, the forcing function consists, almost invariably, of the increments of a Wiener process, which is a process that has an infinite bandwidth in the frequency domain. Already, in

section 6.3, we have encountered a process with a limited bandwidth. Later, in section 6.9, we shall consider some further implications of a limited bandwidth.

The Wiener process $Z(t)$ is defined by the following conditions:

(a) $Z(0)$ is finite,
(b) $E\{Z(t)\} = 0$, for all t,
(c) $Z(t)$ is normally distributed,
(d) $dZ(s), dZ(t)$ for all $t \neq s$ are independent stationary increments,
(e) $V\{Z(t+h) - Z(t)\} = \sigma^2 h$ for $h > 0$.

The increments $dZ(s), dZ(t)$ are impulses that have a uniform power spectrum distributed over an infinite range of frequencies corresponding to the entire real line. Sampling $Z(t)$ at regular intervals to form a discrete-time white-noise process $\varepsilon(t) = Z(t+1) - Z(t)$ entails a process of aliasing, whereby the spectral power of the cumulated increments gives rise to a uniform spectrum of finite power over the frequency interval $[-\pi, \pi]$.

In general:

$$Z(t) = Z(a) + \int_a^t dZ(\tau), \tag{6.45}$$

where $Z(a)$ is a finite starting value at time a. However, if $Z(t)$ were differentiable, as some forcing functions may be, then we should have $dZ(t) = \{dZ(t)/dt\}dt$.

The simplest of stochastic differential equations is the first-order equation, which takes the form:

$$\frac{dx(t)}{dt} - \lambda x(t) = dZ(t) \quad \text{or} \quad (D - \lambda)x(t) = dZ(t). \tag{6.46}$$

Multiplying throughout by the factor $\exp\{-\lambda t\}$ gives:

$$e^{-\lambda t}Dx(t) - \lambda e^{-\lambda t}x(t) = D\{x(t)e^{-\lambda t}\} = e^{-\lambda t}dZ(t), \tag{6.47}$$

where the first equality follows from the product rule of differentiation. Integrating $D\{x(t)e^{-\lambda t}\} = e^{-\lambda t}dZ(t)$ gives:

$$x(t)e^{-\lambda t} = \int_{-\infty}^t e^{-\lambda \tau}dZ(\tau), \tag{6.48}$$

or:

$$x(t) = e^{\lambda t}\int_{-\infty}^t e^{-\lambda \tau}dZ(\tau) = \int_{-\infty}^t e^{\lambda(t-\tau)}dZ(\tau). \tag{6.49}$$

If we write $x(t) = (D - \lambda)^{-1}dZ(t)$, then we get the result that:

$$x(t) = \frac{1}{D - \lambda}dZ(t) = \int_{-\infty}^t e^{\lambda(t-\tau)}dZ(\tau), \tag{6.50}$$

from which it is manifest that the necessary and sufficient condition for stability is that $\lambda < 0$. That is to say, the root of the equation $D - \lambda = 0$, which indicates the rate of decay of the increments, must be less than zero.

The general solution of a differential equation should normally comprise a particular solution, which represents the effects of the initial conditions. However, given that their effects decay as time elapses and given that, in this case, the integral has no lower limit, no account needs to be taken of initial conditions.

When the process is observed at the integer time points $\{t = 0, \pm 1, \pm 2, \ldots\}$, it is appropriate to express it as:

$$\begin{aligned} x(t) &= e^{\lambda} \int_{-\infty}^{t-1} e^{\lambda(t-1-\tau)} dZ(\tau) + \int_{t-1}^{t} e^{\lambda(t-\tau)} dZ(\tau) \\ &= e^{\lambda} x(t-1) + \int_{t-1}^{t} e^{\lambda(t-\tau)} dZ(\tau). \end{aligned} \tag{6.51}$$

This gives rise to a discrete-time equation of the form:

$$x(t) = \phi x(t-1) + \varepsilon(t), \quad \text{or} \quad (1 - \phi L)x(t) = \varepsilon(t), \tag{6.52}$$

where:

$$\phi = e^{\lambda} \quad \text{and} \quad \varepsilon(t) = \int_{t-1}^{t} e^{\lambda(t-\tau)} dZ(\tau), \tag{6.53}$$

and where L is the lag operator, which has the effect that $Lx(t) = x(t-1)$.

The second-order equation may be expressed as follows:

$$(D^2 + \varphi_1 D + \varphi_2)x(t) = (D - \lambda_1)(D - \lambda_2)x(t) = dZ(t). \tag{6.54}$$

Using a partial-fraction expansion, this can be cast in the form of:

$$\begin{aligned} x(t) &= \frac{1}{\lambda_1 - \lambda_2} \left\{ \frac{1}{D - \lambda_1} - \frac{1}{D - \lambda_2} \right\} dZ(t) \\ &= \int_{-\infty}^{t} \left\{ \frac{e^{\lambda_1(t-\tau)} - e^{\lambda_2(t-\tau)}}{\lambda_1 - \lambda_2} \right\} dZ(\tau). \end{aligned} \tag{6.55}$$

Here, the final equality depends upon the result under (6.50). If the roots λ_1, λ_2 have real values, then the condition of stability is that $\lambda_1, \lambda_2 < 0$. If the roots are conjugate complex numbers, then the condition for stability is that they must lie in the left half of the complex plane. In that case, the trajectory of $x(t)$ will have a damped quasi-sinusoidal motion of a sort that is characteristic of the business cycle.

Equation (6.55) gives rise to a second-order difference equation. In the manner that equation (6.50) leads to equation (6.52), we get:

$$\begin{aligned} x(t) &= \frac{1}{\lambda_1 - \lambda_2} \left\{ \frac{\varepsilon_1(t)}{1 - \kappa_1 L} + \frac{\varepsilon_2(t)}{1 - \kappa_2 L} \right\} \\ &= \frac{\theta_0 + \theta_1 L}{1 + \phi_1 L + \phi_2 L} \varepsilon(t). \end{aligned} \tag{6.56}$$

Here, $(\lambda_1-\lambda_2)(1-\kappa_1 L)(1-\kappa_2 L) = 1+\phi_1 L+\phi_2 L$, and we have defined $(\theta_0+\theta_1 L)\varepsilon(t) = (1-\phi_2 L)\varepsilon_1(t) + (1-\phi_1 L)\varepsilon_2(t)$, which is a first-order moving-average process. Equation (6.56) depicts an ARMA(2, 1) process in discrete time. The correspondence between the second-order differential equation and the ARMA(2, 1) process has been discussed by Phadke and Wu (1974) and Pandit and Wu (1975).

Autoregressive models of other orders may be derived in the same manner as the second-order model by putting polynomial functions of D of the appropriate degrees in place of the quadratic function. The models can also be elaborated by applying a moving-average operator or weighting function $\rho(\tau)$ to the stochastic forcing function $dZ(t)$. This gives a forcing function in the form of:

$$\eta(t) = \int_0^q \rho(\tau) dZ(t-\tau) = \int_{t-q}^t \rho(t-\tau) dZ(\tau). \tag{6.57}$$

The consequence of this elaboration for the corresponding discrete-time ARMA model is that its moving-average parameters are no longer constrained to be functions of the autoregressive parameters alone.

In modeling a stochastic trend, it is common to adopt a first- or second-order process in which the roots are set to zeros. In that case, the stochastic increments are accumulated without decay. Therefore, it is crucial to specify the initial conditions of the process. We shall denote the process that is the m-fold integral of the incremental process $dZ(t)$ by $Z^{(m)}(t)$. Then, $Z^{(1)}(t)$ can stand for the Wiener process $Z(t)$, defined previously.

If the process has begun in the indefinite past, then there will be zero probability that its current value will be found within a finite distance from the origin. Therefore, we must impose the condition that, at any time that is at a finite distance both from the origin and from the current time, the process $Z^{(1)}(t)$ assumes a finite value. This allows us to write:

$$Z^{(1)}(t) = Z^{(1)}(t-h) + \int_{t-h}^t dZ^{(1)}(\tau), \tag{6.58}$$

where h is an arbitrary finite step in time and $a = t - h$ is a fixed point in time.

On this basis, the value of the integrated process at time t is:

$$\begin{aligned} Z^{(2)}(t) &= Z^{(2)}(t-h) + \int_{t-h}^t Z^{(1)}(\tau) d\tau \\ &= Z^{(2)}(t-h) + Z^{(1)}(t-h)h + \int_{t-h}^t (t-\tau) dZ^{(1)}(\tau). \end{aligned} \tag{6.59}$$

By proceeding through successive stages, we find that the mth integral is:

$$Z^{(m)}(t) = \sum_{k=0}^{m-1} Z^{(m-k)}(t-h)\frac{h^k}{k!} + \int_{t-h}^t \frac{(t-\tau)^{m-1}}{(m-1)!} dZ^{(1)}(\tau). \tag{6.60}$$

Here, the first term on the right-hand side is a polynomial in h, which is the distance in time from the fixed point a, whereas the second term is the m-fold integral of mean-zero stochastic increments, which constitutes a non-stationary process.

The covariance of the changes $Z^{(j)}(t) - Z^{(j)}(t-h)$ and $Z^{(k)}(t) - Z^{(k)}(t-h)$ of the jth and the kth integrated processes derived from $Z(t)$ is given by:

$$\begin{aligned} C\{z^{(j)}(t), z^{(k)}(t)\} &= \int_{s=t-h}^{t}\int_{r=t-h}^{t} \frac{(t-r)^{j-1}(t-s)^{k-1}}{j!k!} E\{dZ(r)dZ(s)\} \\ &= \sigma^2 \int_{t-h}^{t} \frac{(t-r)^{j+k-2}}{j!k!} dr = \sigma^2 \frac{h^{j+k-1}}{(j+k-1)j!k!}. \end{aligned} \tag{6.61}$$

A straightforward elaboration of the model of a stochastic trend arises when it is assumed that the expected value of the incremental process that is the forcing function has a non-zero mean. Then, $Z(t)$ is replaced by $\mu dt + dZ(t)$. This is the case of stochastic drift. If μ is relatively large, then it will make a significant contribution to the polynomial component, with the effect that the latter may become the dominant component.

6.5.1 Discrete-time representation of an integrated Wiener process

To derive the discretely sampled version of the integrated Wiener process, it may be assumed that values are sampled at regular intervals of h time units. Then, using the alternative notation of $\beta(t) = Z^{(1)}(t)$, equation (6.58) can be written as:

$$\beta(t) = \beta(t-h) + \varepsilon(t), \tag{6.62}$$

where $\varepsilon(t)$ is a white-noise process. With $\tau(t) = Z^{(2)}(t)$, equation (6.59) can be written as:

$$\tau(t) = \tau(t-h) + h\beta(t-h) + \nu(t), \tag{6.63}$$

where $\nu(t)$ is another white-noise process. Together, equations (6.62) and (6.63) constitute a so-called local linear model in which $\tau(t)$ represents the level and $\beta(t)$ represents the slope parameter. On taking the step length to be $h = 1$, the transition equation for this model is:

$$\begin{bmatrix} \tau(t) \\ \beta(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tau(t-1) \\ \beta(t-1) \end{bmatrix} + \begin{bmatrix} \nu(t) \\ \varepsilon(t) \end{bmatrix}. \tag{6.64}$$

Using the difference operator $\nabla = 1 - L$, the discrete-time processes entailed in this equation can be written as:

$$\begin{aligned} \nabla\tau(t) &= \tau(t) - \tau(t-1) = \beta(t-1) + \nu(t), \\ \nabla\beta(t) &= \beta(t) - \beta(t-1) = \varepsilon(t). \end{aligned} \tag{6.65}$$

Applying the difference operator a second time to the first of these and substituting for $\nabla\beta(t) = \varepsilon(t)$ gives:

$$\begin{aligned} \nabla^2\tau(t) &= \nabla\beta(t-1) + \nabla\nu(t) \\ &= \varepsilon(t-1) + \nu(t) - \nu(t-1). \end{aligned} \tag{6.66}$$

On the right-hand side of this equation is a sum of stationary stochastic processes, which can be expressed as an ordinary first-order moving-average process. Thus:

$$\varepsilon(t-1) + \nu(t) - \nu(t-1) = \eta(t) + \theta\eta(t-1), \tag{6.67}$$

where $\eta(t)$ is a white-noise process with $V\{\eta(t)\} = \sigma_\eta^2$. Therefore, the sampled version of the integrated Wiener process is a doubly-integrated IMA(2, 1) moving-average model.

The essential task is to find the values of the moving-average parameter θ. Thus is achieved by reference to equation (6.61), which provides the variances and covariances of the terms on the left-hand side of (6.67), from which the autocovariances of the MA process can be found. It can be shown that the variance and the autocovariance at lag 1 of this composite process are given by:

$$\gamma_0 = \frac{2\sigma_\varepsilon^2}{3} = \sigma_\eta^2(1+\theta^2) \qquad \text{and} \qquad \gamma_1 = \frac{\sigma_\varepsilon^2}{6} = \sigma_\eta^2\theta. \tag{6.68}$$

The equations must be solved for θ and σ_η^2. There are two solutions for θ, and we should take the one which fulfils the condition of invertibility: $\theta = 2 - \sqrt{3}$. (See Pollock, 1999.)

When white-noise errors of observation are superimposed upon values sampled from an integrated Wiener process at regular intervals, the resulting sequence can be described by a doubly-integrated second-order moving-average process in discrete time, which is an IMA(2, 2) process. Such a model provides the basis for the cubic smoothing spline of Reinsch (1976), which can be used to extract an estimate of the trajectory of the underlying integrated Wiener process from the noisy data. The statistical interpretation of the smoothing spline is due to Wahba (1978).

The smoothing spline interpolates cubic polynomial segments between nodes that are derived by smoothing a sequence of sampled data points. The segments are joined in such a way as to ensure that the second derivative of the spline function is continuous at the nodes. An account of the algorithm of the smoothing spline and of its derivation from the statistical model has been provided by Pollock (1999). It is shown that the means by which the nodes are obtained from the data amount to a so-called discrete-time Wiener–Kolmogorov (WK) filter.

The Wiener–Kolmogorov principle can also be used to derive the so-called Hodrick–Prescott (HP) filter, which is widely employed in macroeconomic analysis – see Hodrick and Prescott (1980, 1997). The filter, which is presented in section 6.6.2, is derived from the assumption that the process that generates the trend is a doubly-integrated discrete-time white noise. When white-noise errors are added to the sampled values of the process, the observations are once more described by an IMA(2, 2) model, and the nodes that are generated by the WK trend-extraction filter are analogous to those of the smoothing spline.

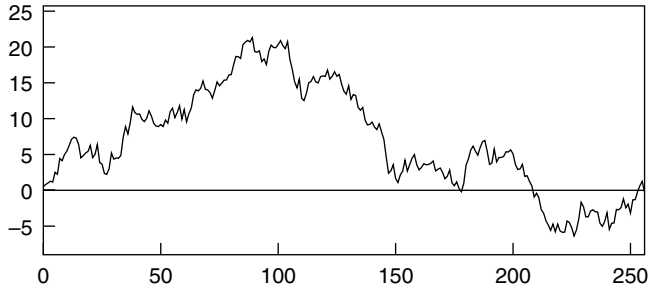


Figure 6.9 The graph of 256 observations on a simulated series generated by a random walk

The trend that is generated by the smoothing spline is an aesthetically pleasing curve, of which the smoothness belies the disjunct nature of the stochastic forcing function. That nature is more clearly revealed in the case of a model that postulates a trend that is generated by an ordinary Wiener process, as opposed to an integrated process. The discrete-time observations, which are affected by white-noise errors, are modeled by an IMA(1, 1) process, which also corresponds to the local level model that has been advocated by Harvey (1985, 1989), amongst others. The function that provides statistical estimates of the trend at the nodes and at the points between them has jointed linear segments.

It should be recognized that, if the forcing function were assumed to be bounded in frequency, then the interpolating function would be a smooth one, generated by a Fourier interpolation, that would have no discontinuities at the nodes.

In section 6.9, we shall return to the question of how best to specify the continuous-time forcing function. In the next section, we shall deal exclusively with discrete-time models, and we shall examine various ways of decomposing into its component parts a model of an aggregate process that combines the trend and the cycles.

Example A Wiener process, which is everywhere continuous but nowhere differentiable, can be represented graphically only via its sampled ordinates. If the sampling is sufficiently rapid to give a separation between adjacent points that is below the limits of visual acuity, then the sampled process, which constitutes a discrete-time random walk, will give the same visual impression as the underlying Wiener process. This is the intended effect of Figure 6.9.

Figure 6.10 depicts the trajectory of the IMA(2, 1) process that represents the sampled version of an integrated Wiener process. This is a much smoother trajectory than that of the random walk. The extra smoothness can be attributed to the effect of the summation operator, of which the squared gain has been depicted in Figure 6.8. The operator amplifies the sinusoidal elements in the lower part of the frequency range and attenuates those in the upper part.

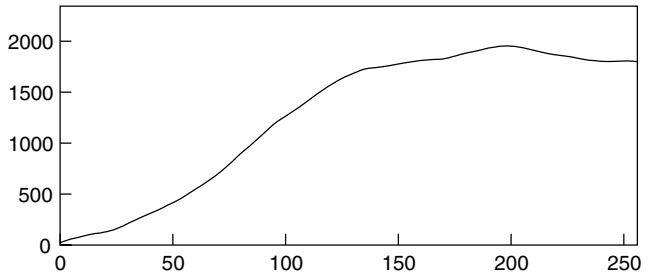


Figure 6.10 The graph of 256 observations on a simulated series generated by an IMA(2, 1) process that correspond to the sampled version of an integrated Wiener process

6.6 Decomposition of discrete-time ARIMA processes

An ARMA model can be represented by the equation:

$$\sum_{i=0}^{p} \phi_i y_{t-i} = \sum_{i=0}^{q} \theta_i \varepsilon_{t-i} \quad \text{with} \quad \phi_0 = \theta_0 = 1, \tag{6.69}$$

where t has whatever range is appropriate to the analysis. To exploit the algebra of polynomial operators, the equation can be embedded within the system:

$$\phi(z)y(z) = \theta(z)\varepsilon(z), \tag{6.70}$$

where $\varepsilon(z) = z^t\{\varepsilon_t + \varepsilon_{t-1}z^{-1} + \cdots\}$ is a z-transform of the infinite white-noise forcing function or disturbance sequence $\{\varepsilon_{t-i}; i = 0, 1, \ldots\}$ and where $y(z)$ is the z-transform of the corresponding data sequence. The embedded equation will be associated with z^t.

The polynomials $\theta(z)$ and $\phi(z)$ must have all their roots outside the unit circle to make their inverses, $\theta^{-1}(z)$ and $\phi(z)^{-1}$, amenable to power series expansions when $|z| \geq 1$. Then, it is possible to represent the system of (6.70) by the equation $y(z) = \phi^{-1}(z)\theta(z)\varepsilon(z)$.

An ARIMA process represents the accumulation of the output of an ARMA process. On defining the (backwards) difference operator $\nabla(z) = 1 - z$, the dth-order model can be represented by:

$$\nabla^d(z)\alpha(z)y(z) = \theta(z)\varepsilon(z). \tag{6.71}$$

The inverse of the difference operator is the summation operator $\nabla^{-1}(z) = \{1 + z + z^2 + \cdots\}$, and this might be used in representing the system of (6.71), alternatively, by the equation $y(z) = \nabla^{-d}(z)\alpha^{-1}(z)\theta(z)\varepsilon(z)$.

The difficulty here is that, if it is formed from an infinite number of independently and identically distributed random variables, the disturbance sequence cannot have a finite sum. For this reason, it appears that the algebra of polynomial operators cannot be applied to the analysis of non-stationary processes.

The usual recourse in the face of this problem is scrupulously to avoid the use of the cumulation operator $\nabla^{-1}(z)$ and to represent the integrated system only in the form of (6.71). This is not a wholly adequate solution to the problem since, to exploit the algebra of the operators, it is necessary to define the inverses of all of the polynomial operators. An alternative solution is to constrain the disturbance sequence to be absolutely summable, which appears to negate the assumption that it is generated by a stationary stochastic process.

The proper recourse is to replace the process of indefinite summation by a definite summation that depends upon supplying the system with initial conditions at some adjacent points in time. To show what this entails, we may consider the system of equations that is derived from (6.69) by setting $t = 0, 1, \ldots, T-1$. The set of T equations can be arrayed in a matrix format as follows:

$$\begin{bmatrix} y_0 & y_{-1} & \cdots & y_{-p} \\ y_1 & y_0 & \cdots & y_{1-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_p & y_{p-1} & \cdots & y_0 \\ \vdots & \vdots & & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p-1} \end{bmatrix} \begin{bmatrix} 1 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \varepsilon_0 & \varepsilon_{-1} & \cdots & \varepsilon_{-q} \\ \varepsilon_1 & \varepsilon_0 & \cdots & \varepsilon_{1-q} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_q & \varepsilon_{q-1} & \cdots & \varepsilon_0 \\ \vdots & \vdots & & \vdots \\ \varepsilon_{T-1} & \varepsilon_{T-2} & \cdots & \varepsilon_{T-q-1} \end{bmatrix} \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_q \end{bmatrix}. \tag{6.72}$$

Apart from the elements $y_0, y_1, \ldots, y_{T-1}$ and $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{T-1}$, which fall within the indicated period, these equations comprise the values $y_{-p}, \ldots, y_{-1}$ and $\varepsilon_{-q}, \ldots, \varepsilon_{-1}$, which are to be found in the top-right corners of the matrices, and which constitute the initial conditions at the start-up time of $t = 0$.

Each of the elements within this display can be associated with the power of z that is indicated by the value of its subscripted index. In that case, the system can be represented by equation (6.70) with the constituent polynomials defined as follows:

$$\begin{aligned} y(z) &= y_{-p}z^{-p} + \cdots + y_0 + y_1 z + \cdots + y_{T-1}z^{T-1}, \\ \varepsilon(z) &= \varepsilon_{-q}z^{-q} + \cdots + \varepsilon_0 + \varepsilon_1 z + \cdots + \varepsilon_{T-1}z^{T-1}, \\ \phi(z) &= 1 + \phi_1 z + \cdots + \phi_p z^p \quad \text{and} \\ \theta(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q. \end{aligned} \tag{6.73}$$

This scheme applies regardless of the values of the roots of the polynomial operators $\phi(z)$ and $\theta(z)$. Therefore, it can accommodate the case where $\phi(z) = \nabla^d(z)\alpha(z)$, which is that of equation (6.71). One of the virtues of this notation is that it is not burdened by an explicit representation of the initial conditions. At a later stage, in section 6.7, we shall need to represent the initial conditions explicitly.

A trend has only a tenuous existence within the context of a univariate ARIMA model of the sort represented by equation (6.71). In such a model, it amounts to nothing more that the accumulation of the fluctuations that are created by

applying a filter $\theta(z)/\alpha(z)$ to a white-noise sequence $\varepsilon(t)$ of independently and identically distributed random variables.

If the trend and the transitory motions that accompany it are due to the same motive force, which is the white-noise process, then it is difficult to draw a distinction between them. However, a distinction can be made by attributing the trend to the unit roots within $\nabla^d(z) = (1-z)^d$ and by attributing the transitory motions to the stable roots of the autoregressive operator $\alpha(z)$. This is what the decomposition of Beveridge and Nelson (1981) achieves.

Faced with the insistence that the trend and the fluctuations are due to separate sources, an obvious recourse is to attribute separate and independent ARIMA models to each of them. In that case, the aggregate data are also described by a univariate ARIMA model. Provided that their models have distinct parameters, WK filters may be used tentatively to extract the independent components from the data.

The assumption that the components originate from transformations of white-noise sequences implies that their spectra extend over the entire frequency range $[0, \pi]$. This means that they are bound to overlap substantially. In practice, the spectral structures of the components are often confined to frequency bands that are separated by wide spectral dead spaces. In that case, the separation of one component from another can be achieved in a more decisive manner than the WK filters will usually allow.

6.6.1 The Beveridge–Nelson decomposition

The Beveridge–Nelson decomposition relates to an ARIMA model with first-order integration and with a stochastic drift. This can be represented in z-transform notation by:

$$y(z) = \frac{\mu(z)}{\nabla(z)} + \frac{\theta(z)}{\alpha(z)\nabla(z)}\varepsilon(z). \tag{6.74}$$

If the system has a start-up at $t = 0$, then $\mu(z)$, which represents the drift, is the z-transform of a sequence that is constant over the integers $0, 1, \ldots, t$ and zero-valued for $t < 0$. The operator associated with $\varepsilon(z)$ has the following partial-fraction decomposition:

$$\frac{\theta(z)}{\alpha(z)\nabla(z)} = \frac{\rho(z)}{\alpha(z)} + \frac{\delta}{\nabla(z)}. \tag{6.75}$$

Multiplying both sides by $\nabla(z) = 1 - z$ and setting $z = 1$ gives $\delta = \theta(1)/\alpha(1)$, where the numerator and the denominator are just the sums of the polynomial coefficients. Substituting the result into equation (6.74) creates an additive decomposition of the form $y(z) = \tau(z) + \zeta(z)$, wherein:

$$\tau(z) = \frac{1}{\nabla(z)}\left\{\mu(z) + \delta\varepsilon(z)\right\}, \tag{6.76}$$

$$\zeta(z) = \frac{\rho(z)}{\alpha(z)}\varepsilon(z), \tag{6.77}$$

are respectively the trend component and the transitory component. This is the so-called Beveridge–Nelson decomposition.

The trend component of the Beveridge–Nelson decomposition is a first-order random walk with drift, whereas the transitory component is an ARMA process. The distinguishing feature of the decomposition is that both components have the same forcing function. It is easy to see that:

$$\tau(z) = \frac{\theta(1)}{\alpha(1)}\frac{\alpha(z)}{\theta(z)}y(z), \tag{6.78}$$

which is to say that the estimate of the trend is derived by applying an ordinary linear filter to the data sequence. The effect of the filter is to eliminate the ARMA factor from the data so as to deliver a pure random walk.

A common objection to the Beveridge–Nelson decomposition is that the resulting trend is liable to be too rough. This is a consequence of the fact that a random walk that is an accumulation of independently and identically distributed random variables comprises elements at all frequencies up to the limiting Nyquist frequency of π radians per sample period. Also, the decomposition makes no provision for the presence of seasonal fluctuations in the data. A more elaborate model can be proposed with the aim of overcoming these objections.

Consider the multiplicative seasonal ARIMA model of Box and Jenkins (1976), which can be represented by the equation:

$$\nabla^d(z)\nabla_s^D(z)y(z) = \mu(z) + \frac{\theta(z)\Theta(z^s)}{\alpha(z)A(z^s)}\varepsilon(z). \tag{6.79}$$

Here, $\alpha(z)$ and $\theta(z)$ are the autoregressive and moving-average polynomials that have appeared in equation (6.74), whereas $A(z)$ and $\Theta(z)$ are seasonal operators. Whereas $\nabla(z)$ continues to represent the ordinary difference operator, there is now a seasonal difference operator $\nabla_s(z) = 1 - z^s = (1 - z)S(z)$, which forms the differences between the data from the same season (or month) of two successive years. The factors of this operator are the ordinary difference operator and a seasonal summation operator $S(z) = 1 + z + z^2 + \cdots + z^{s-1}$. A decomposition can now be found of the form $y(z) = \tau(z) + \sigma(z) + \zeta(z)$, where:

$$\tau(z) = \frac{1}{\nabla^{d+D}}\{\mu(z) + \alpha(z)\varepsilon(z)\}, \tag{6.80}$$

$$\sigma(z) = \frac{\beta(z)}{S^D(z)}\varepsilon(z), \tag{6.81}$$

$$\zeta(z) = \frac{\gamma(z)}{\alpha(z)A(z^s)}\varepsilon(z), \tag{6.82}$$

are, respectively, the trend, the seasonal component and the transitory component. If the degree $d + D$ of the (ordinary) difference operator exceeds unity, then the trend is liable to be smoother than one generated by a first-order random walk. Also, the effect of $\alpha(z)$ might be further to attenuate the high-frequency elements of the forcing function, thereby enhancing the smoothness of the trend.

To enhance the smoothness of the trend and of the seasonal component yet further, an irregular component could be incorporated in the decomposition. The irregular elements could be extracted from the trend and the seasonal component and assigned to this additional term, which could be regarded as statistically independent of the primary forcing function $\varepsilon(t)$. However, from this point of view, it is natural to consider a model in which each of the components is driven by a statistically independent forcing function. Such a model is the basis of the Wiener–Kolmogorov methodology for signal extraction.

6.6.2 WK filtering

The modern theory of statistical signal extraction was formulated independently by Wiener (1941) and Kolmogorov (1941), who arrived at the same results in different ways. Whereas Kolmogorov took a time-domain approach to the problem, Wiener worked primarily in the frequency domain. However, the unification of the two approaches was soon achieved, and a modern account of the theory, which encompasses both, has been provided by Whittle (1983).

The purpose of a WK filter is to extract an estimate of a signal sequence $\xi(t)$ from an observable data sequence:

$$y(t) = \xi(t) + \eta(t), \tag{6.83}$$

which is afflicted by the noise $\eta(t)$. According to the classical assumptions, which we shall later amend, the signal and the noise are generated by zero-mean stationary stochastic processes that are mutually independent. Also, the assumption is made that the data constitute a doubly-infinite sequence. It follows that the autocovariance generating function of the data is the sum of the autocovariance generating functions of its two components. Thus:

$$\gamma^{yy}(z) = \gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z) \quad \text{and} \quad \gamma^{\xi\xi}(z) = \gamma^{y\xi}(z). \tag{6.84}$$

These functions are amenable to the so-called Cramér–Wold factorization, and they may be written as:

$$\gamma^{yy}(z) = \phi(z^{-1})\phi(z), \quad \gamma^{\xi\xi}(z) = \theta(z^{-1})\theta(z), \quad \gamma^{\eta\eta}(z) = \theta_\eta(z^{-1})\theta_\eta(z). \tag{6.85}$$

The estimate x_t of the signal element ξ_t is a linear combination of the elements of the data sequence:

$$x_t = \sum_j \beta_j y_{t-j}. \tag{6.86}$$

The principle of minimum mean square error estimation indicates that the estimation errors must be statistically uncorrelated with the elements of the

information set. Thus, the following condition applies for all k:

$$\begin{aligned} 0 &= E\Big\{y_{t-k}(\xi_t - x_t)\Big\} \\ &= E(y_{t-k}\xi_t) - \sum_j \beta_j E(y_{t-k}y_{t-j}) \\ &= \gamma_k^{y\xi} - \sum_j \beta_j \gamma_{k-j}^{yy}. \end{aligned} \tag{6.87}$$

The equation may be expressed, in terms of the z-transforms, as:

$$\gamma^{y\xi}(z) = \beta(z)\gamma^{yy}(z). \tag{6.88}$$

It then follows that:

$$\begin{aligned} \beta(z) &= \frac{\gamma^{y\xi}(z)}{\gamma^{yy}(z)} \\ &= \frac{\gamma^{\xi\xi}(z)}{\gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z)} = \frac{\theta(z^{-1})\theta(z)}{\phi(z^{-1})\phi(z)}. \end{aligned} \tag{6.89}$$

Now, by setting $z = \exp\{\mathrm{i}\omega\}$, one can derive the frequency-response function of the filter that is used in estimating the signal $\xi(t)$. The effect of the filter is to multiply each of the frequency elements of $y(t)$ by the fraction of its variance that is attributable to the signal. The same principle applies to the estimation of the residual component. This is obtained using the complementary filter:

$$\beta^c(z) = 1 - \beta(z) = \frac{\gamma^{\eta\eta}(z)}{\gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z)}. \tag{6.90}$$

The estimated signal component may be obtained by filtering the data in two passes according to the following equations:

$$\phi(z)q(z) = \theta(z)y(z), \quad \phi(z^{-1})x(z^{-1}) = \theta(z^{-1})q(z^{-1}). \tag{6.91}$$

The first equation relates to a process that runs forwards in time to generate the elements of an intermediate sequence, represented by the coefficients of $q(z)$. The second equation represents a process that runs backwards to deliver the estimates of the signal, represented by the coefficients of $x(z)$.

The Wiener–Kolmogorov methodology can be applied to non-stationary data with minor adaptations. A model of the processes underlying the data can be adopted that has the form:

$$\begin{aligned} \nabla^d(z)y(z) &= \nabla^d(z)\{\xi(z) + \eta(z)\} = \delta(z) + \kappa(z) \\ &= (1+z)^n\zeta(z) + (1-z)^m\varepsilon(z), \end{aligned} \tag{6.92}$$

where $\zeta(z)$ and $\varepsilon(z)$ are the z-transforms of two independent white-noise sequences $\zeta(t)$ and $\varepsilon(t)$. The condition $m \geq d$ is necessary to ensure the stationarity of $\eta(t)$,

which is obtained from $\varepsilon(t)$ by differencing $m - d$ times. Then, the filter that is applied to $y(t)$ to estimate $\xi(t)$, which is the d-fold integral of $\delta(t)$, takes the form:

$$\beta(z) = \frac{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n}{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n + \sigma_\varepsilon^2(1-z^{-1})^m(1-z)^m}, \tag{6.93}$$

regardless of the degree d of differencing that would be necessary to reduce $y(t)$ to stationarity.

Two special cases are of interest. By setting $d = m = 2$ and $n = 0$ in (6.92), a model is obtained of a second-order random walk $\xi(t)$ affected by white-noise errors of observation $\eta(t) = \varepsilon(t)$. The resulting lowpass WK filter, in the form:

$$\beta(z) = \frac{1}{1 + \lambda(1-z^{-1})^2(1-z)^2} \quad \text{with} \quad \lambda = \frac{\sigma_\eta^2}{\sigma_\delta^2}, \tag{6.94}$$

is the HP filter. The complementary highpass filter, which generates the residue, is:

$$\beta^c(z) = \frac{(1-z^{-1})^2(1-z)^2}{\lambda^{-1} + (1-z^{-1})^2(1-z)^2}. \tag{6.95}$$

Here, λ, which is described as the smoothing parameter, is the single adjustable parameter of the filter.

By setting $m = n$, a filter for estimating $\xi(t)$ is obtained that takes the form:

$$\begin{aligned}\beta(z) &= \frac{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n}{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n + \sigma_\varepsilon^2(1-z^{-1})^n(1-z)^n} \\ &= \frac{1}{1 + \lambda\left(\mathrm{i}\dfrac{1-z}{1+z}\right)^{2n}} \quad \text{with} \quad \lambda = \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2}.\end{aligned} \tag{6.96}$$

This is the formula for the Butterworth lowpass digital filter. The filter has two adjustable parameters and, therefore, is a more flexible device than the HP filter. First, there is the parameter λ. This can be expressed as:

$$\lambda = \{1/\tan(\omega_d)\}^{2n}, \tag{6.97}$$

where ω_d is the nominal cut-off point of the filter, which is the mid-point in the transition of the filter's frequency response from its pass band to its stop band. The second of the adjustable parameters is n, which denotes the order of the filter. As n increases, the transition between the pass band and the stop band becomes more abrupt.

These filters can be applied to the non-stationary data sequence $y(t)$ in the manner indicated by equation (6.91), provided that the appropriate initial conditions are supplied with which to start the recursions. However, by concentrating on the estimation of the residual sequence $\eta(t)$, which corresponds to a stationary process,

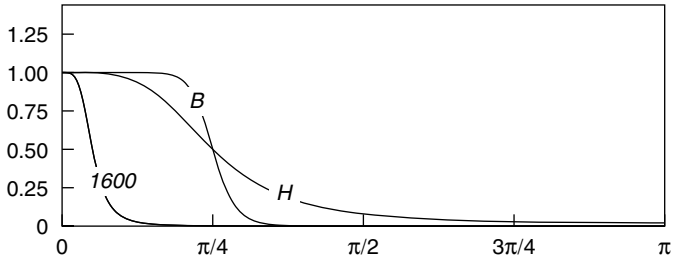


Figure 6.11 The gain of the HP filter H and of the Butterworth filter B with nominal cut-off points at $\pi/4$ radians, together with the gain of an HP filter with a smoothing parameter of 1600

it is possible to avoid the need for non-zero initial conditions. The estimate of $\eta(t)$ can then be subtracted from $y(t)$ to obtain the estimate of $\xi(t)$.

The HP filter has many antecedents. Its invention cannot reasonably be attributed to Hodrick and Prescott (1980, 1997), who cited Whittaker (1923) as one of their sources. Leser (1961) also provided a complete derivation of the filter at an earlier date. The Butterworth filter is a commonplace of electrical engineering. The digital version of the filter has been described in an econometric context by Pollock (2000) and by Gómez (2001). It has been applied to climatological data by Harvey and Mills (2003).

Example Figure 6.11 shows the gain functions of the three filters overlaid on the same diagram. The lowpass HP filter with a smoothing parameter of $\lambda = 1600$ is commonly recommended for estimating the trend in quarterly economic data. The corresponding gain function is marked in the diagram by the number 1600.

An alternative to specifying the smoothing parameter directly is to specify the frequency value ω_d for which the gain is $\beta(\omega_d) = 0.5$. For the HP filter, the correspondence between ω_d and λ is as follows:

$$\lambda = \frac{1}{4\{1 - \cos(\omega_d)\}^2} \quad \text{and} \quad \omega_d = \cos^{-1}(1 - 1\sqrt{4\lambda}). \tag{6.98}$$

The frequency ω_d corresponds to the mid-point in the transition between the pass band and the stop band of the filter. This might be described as the nominal cut-off frequency, but, in the case of the HP filter, this is a misnomer, on account of the very gradual transition of the gain. The Butterworth filter is capable of a much more rapid transition. The curve labeled B corresponds to the gain of a Butterworth filter with $n = 6$ and $\omega_d = \pi/4$.

6.6.3 Structural ARIMA models

The HP filter and the Butterworth filter are appropriate to the task of extracting the trend or the trend/cycle component from a data sequence without regard to the

structure of the residual component. More elaborate filters are available that also take account of a seasonal component.

Consider, therefore, a seasonal ARIMA model of the form:

$$y(z) = \frac{\theta(z)}{\phi(z)}\varepsilon(z) = \frac{\theta(z)}{\phi_S(z)\phi_T(z)}\varepsilon(z), \tag{6.99}$$

where $\phi_S(z)$ contains the seasonal autoregressive factors and $\phi_T(z)$ contains the non-seasonal factors.

The denominator contains both an ordinary differencing operator $\nabla^d(z)$ and a seasonal differencing operator $\nabla_s^D(z) = \nabla^D(z)S^D(z)$. The operator $\nabla_s(z) = 1 - z^s = (1-z)S(z)$ forms the differences between the data from the same season (or month) of two successive years. Its factors are the ordinary difference operator and a seasonal summation operator $S(z) = 1 + z + z^2 + \cdots + z^{s-1}$. The factorization of the seasonal operator implies that the overall degree of differencing within the ARIMA model is $d + D$. The factor $\nabla^{d+D}(z)$ is assigned to $\phi_T(z)$, whereas $S^D(z)$ belongs to $\phi_S(z)$.

On the assumption that the degree of the moving-average polynomial $\theta(z)$ is at least equal to that of the denominator polynomial $\phi(z)$, there is a partial-fraction decomposition of the autocovariance generating function of the model into three components, which correspond to the trend effect, the seasonal effect and an irregular influence. Thus:

$$\frac{\theta(z^{-1})\theta(z)}{\phi_S(z^{-1})\phi_T(z^{-1})\phi_T(z)\phi_S(z)} = \frac{Q_T(z)}{\phi_T(z^{-1})\phi_T(z)} + \frac{Q_S(z)}{\phi_S(z^{-1})\phi_S(z)} + R(z). \tag{6.100}$$

Here, the first two components on the right-hand side represent proper rational fractions, whereas the final component is an ordinary polynomial. If the degree of the moving-average polynomial is less than that of the denominator polynomial, then the irregular component is missing from the decomposition in the first instance.

To obtain the spectral density function of $y(t)$, we set $z = e^{-i\omega}$, where $\omega \in [0, \pi]$. (This function is more properly described as a pseudo-spectrum in view of the singularities occasioned by the unit roots in the denominators of the first two components.) The spectral decomposition corresponding to equation (6.100) can be written as:

$$f(\omega) = f(\omega)_T + f(\omega)_S + f(\omega)_R, \tag{6.101}$$

where $f(\omega) = \theta(e^{i\omega})\theta(e^{-i\omega})/\{\phi(e^{i\omega})\phi(e^{-i\omega})\}$.

Let $\nu_T = \min\{f(\omega)_T\}$ and $\nu_S = \min\{f(\omega)_S\}$. These correspond to the elements of white noise embedded in $f(\omega)_T$ and $f(\omega)_S$. The principle of canonical decomposition is that the white-noise elements should be reassigned to the residual component. On defining:

$$\begin{aligned} \gamma_T(z)\gamma_T(z^{-1}) &= Q_T(z) - \nu_T\phi_T(z)\phi_T(z^{-1}), \\ \gamma_S(z)\gamma_S(z^{-1}) &= Q_S(z) - \nu_S\phi_S(z)\phi_S(z^{-1}), \\ \text{and} \quad \rho(z)\rho(z^{-1}) &= R(z) + \nu_T + \nu_S, \end{aligned} \tag{6.102}$$

the canonical decomposition of the generating function can be represented by:

$$\frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})} = \frac{\gamma_T(z)\gamma_T(z^{-1})}{\phi_T(z)\phi_T(z^{-1})} + \frac{\gamma_S(z)\gamma_S(z^{-1})}{\phi_S(z)\phi_S(z^{-1})} + \rho(z)\rho(z^{-1}). \tag{6.103}$$

There are now two improper rational functions on the right-hand side, which have equal degrees in their numerators and denominators.

According to Wiener–Kolmogorov theory, the optimal signal-extraction filter for the trend component is:

$$\begin{aligned}\beta_T(z) &= \frac{\gamma_T(z)\gamma_T(z^{-1})}{\phi_T(z)\phi_T z^{-1})} \times \frac{\phi_S(z)\phi_T(z)\phi_T(z^{-1})\phi_S(z^{-1})}{\theta(z)\theta(z^{-1})} \\ &= \frac{\gamma_T(z)\gamma_T(z^{-1})\phi_S(z)\phi_S(z^{-1})}{\theta(z)\theta(z^{-1})} = \frac{C_T(z)}{\theta(z)\theta(z^{-1})}.\end{aligned} \tag{6.104}$$

This has the form of the ratio of the autocovariance generating function of the trend component to the autocovariance generating function of the process $y(t)$. This formulation presupposes a doubly-infinite data sequence, so it must be translated into a form that can be implemented with finite sequences.

The approach to the estimation of unobserved components that adopts the principle of canonical decompositions has been advocated by Hillmer and Tiao (1982) and Maravall and Pierce (1987). It has been implemented in the TRAMO–SEATS program of Gómez and Maravall (1996) and Caporello and Maravall (2004), which builds upon the work of Burman (1980).

6.6.4 The state-space form of the structural model

In the foregoing approach to modeling the components of a structural time series model, an aggregate univariate process is first estimated and then decomposed into its components. An alternative approach is to model the individual components from the start as separate entities, which are described by independent linear stochastic models.

Provision can be made for a cyclical component which is distinct from the trend component, but, if this is omitted, then the disaggregated model commonly takes the form of $y(z) = \tau(z) + \sigma(z) + \eta(z)$, where:

$$\tau(z) = \frac{(1+\alpha z)}{\nabla^2(z)}\zeta(z), \tag{6.105}$$

$$\sigma(z) = \frac{1}{S(z)}\omega(z). \tag{6.106}$$

$\tau(t)$ is the trend, $\sigma(t)$ is the seasonal component and $\eta(t)$ is the irregular noise. Here there are three independent white-noise processes driving the model, which are $\zeta(t)$, $\omega(t)$ and $\eta(t)$. The model has been described by Harvey (1989) as the basic structural model. A reason for omitting the cyclical or business-cycle component from this model is the difficulty in separating it from the trend component.

The trend process is usually depicted as the product of two processes that constitute the so-called local linear model, which has already been described in section 6.5.1:

$$\tau(t) = \tau(t-1) + \beta(t) + \nu(t), \tag{6.107}$$

$$\beta(t) = \beta(t-1) + \varepsilon(t). \tag{6.108}$$

The first of these describes the level of the trend process and the second describes its slope.

A more elaborate seasonal model is available that generates more regular cycles. A moving-average operator $M(z)$ can be included in the numerator of the expression on the right-hand side of (6.106) to give $\sigma(z) = \{M(z)/S(z)\}\omega(z)$. The autoregressive operator may be factorized as $S(z) = \prod_{j=1}^{s-1}(1 - e^{2\pi j/s})$, where s is the number of observations per annum. The complementary moving-average operator will have the form of $M(z) = \prod_{j=1}^{s-1}(1 - \rho e^{2\pi j/s})$, where $\rho < 1$ is close to unity. The zeros of the moving-average operator will serve largely to negate the effects of the poles of the autoregressive operator, except at the seasonal frequencies, where prominent spectral spikes will be found.

The basic structural model, without the elaboration of a seasonal moving-average component, can be represented in a state-space form that comprises a transition equation, which describes a first-order vector autoregressive process, and an accompanying measurement equation. For notational convenience, let $s = 4$, which corresponds to the case of quarterly observations on annual data. Then the transition equation, which gathers together equations (6.106), (6.107) and (6.108), is:

$$\begin{bmatrix} \tau(t) \\ \beta(t) \\ \sigma(t) \\ \sigma(t-1) \\ \sigma(t-2) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau(t-1) \\ \beta(t-1) \\ \sigma(t-1) \\ \sigma(t-2) \\ \sigma(t-3) \end{bmatrix} + \begin{bmatrix} \nu(t) \\ \varepsilon(t) \\ \omega(t) \\ 0 \\ 0 \end{bmatrix}. \tag{6.109}$$

This incorporates the transition equation of the non-seasonal local linear model that has been given by (6.64). The observation equation, which combines the current values of the components, is:

$$y(t) = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tau(t) \\ \beta(t) \\ \sigma(t) \\ \sigma(t-1) \\ \sigma(t-2) \end{bmatrix} + \eta(t). \tag{6.110}$$

The state-space model is amenable to the Kalman filter and the associated smoothing algorithms, which can be used in estimating the parameters of the model and in extracting estimates of the so-called unobserved components $\tau(t)$, $\sigma(t)$ and $\varepsilon(t)$. These algorithms have been described by Pollock (2003a).

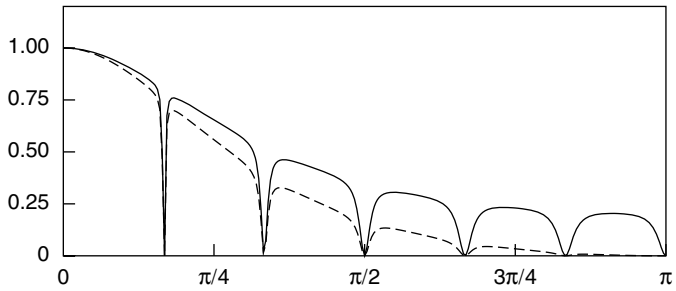


Figure 6.12 The gain function of the trend-extraction filter obtained from the STAMP program (solid line) together with that of the canonical trend-extraction filter (broken line)

Disaggregated structural time-series models have been treated at length in Harvey (1989). The methodology has been implemented in the STAMP program, which is described by Koopman *et al.* (2007). A similar approach has been pursued in a program within the Captain MATLAB Toolbox, which has been described by Pedregal, Taylor and Young (2004). A comparative analysis of the STAMP and TRAMO–SEATS programs has been provided by Pollock (2002b).

Example Figure 6.12 shows the gain of the trend extraction filter that is associated with a disaggregated structural model that has been applied to the monthly airline passenger data of Box and Jenkins (1976).

The solid line represents the gain of the ordinary filter and the broken line represents the gain of the filter that is obtained when the principle of canonical decomposition is applied to the components of the model. In that case, the white noise that is contained in the components is removed and reassigned to the residual component.

The indentations in the gain function at the seasonal frequencies $\pi j/6; j = 1, \ldots, 6$ are due to the zeros of the filter that are to be found on the circumference of the unit circle and which are effective in removing the seasonal fluctuations from the trend.

Disregarding these indentations, the gain of the filters is reduced only gradually as frequency increases. In particular, the ordinary unadjusted filter is liable to transmit a higher proportion of the high-frequency noise of the data. However, given that such high-frequency noise is largely absent from the airline passenger data, it transpires that the effect upon the estimated trend of adopting the principle of canonical decomposition is a minor one.

6.7 Finite-sample signal extraction

The classical theory of linear filtering relies heavily upon the simplifications that are afforded by the assumption that the data constitute a doubly infinite sequence. The assumption is an acceptable one in the case of finite impulse response (FIR) filters that can be realized via low-order moving-average operators. When such a

filter has only a short span, it matters little which assumptions are made about the length of the data sequence. Only at the ends of the data sequence are there liable to be problems.

The assumption of a double-infinite data sequence also sustains the theory of time-invariant infinite impulse response (IIR) rational filters, such as the Butterworth and HP filters of section 6.6.2, which correspond to moving averages of infinite order. These are not so easily applied to short sequences. Nevertheless, if the data sequence is sufficiently lengthy to allow the transient effects of the arbitrary start-up values to disappear, then such filters can be implemented successfully via bidirectional feedback procedures which comprise only a handful of recent data values. (In effect the start-up values purport to summarize the history of the infinite data sequence, insofar as it affects the IIR filter.)

In econometric applications, attention is often focused upon the most recent observations at the upper end of a short data sequence. In such cases, a theory of filtering is called for that fully recognizes the finite nature of the data sequence. Also, in cases where the data are trended, it becomes essential to supply appropriate non-zero initial conditions to the filter, and these should be the products of a finite-sample theory.

The theory that we shall expound here depends upon replacing the symbol z within the various polynomial operators by a matrix lag operator. However, it is immediately apparent that this replacement alone is insufficient for the purpose of creating adequate finite-sample filters.

To demonstrate the effects of the replacement, let $L_T = [e_1, e_2, \ldots, e_{T-1}, 0]$ be the matrix version of the lag operator, which is formed from the identity matrix $I_T = [e_0, e_1, e_2, \ldots, e_{T-1}]$ of order T by deleting the leading column and by appending a column of zeros to the end of the array. Then, the matrix of order T that corresponds to the pth difference operator $\nabla^p(z) = (1 - z)^p$ is:

$$\nabla_T^p = (I - L_T)^p. \tag{6.111}$$

We may partition this matrix so that $\nabla_T^p = [Q_*, Q]'$, where Q_*' has p rows. If y is a vector of T elements, then:

$$\nabla_T^p y = \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} y = \begin{bmatrix} g_* \\ g \end{bmatrix}, \tag{6.112}$$

and g_* is liable to be discarded, whereas g will be regarded as the vector of the pth differences of the data.

The inverse matrix is partitioned conformably to give $\nabla_T^{-p} = [S_*, S]$. It follows that:

$$\begin{bmatrix} S_* & S \end{bmatrix} \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} = S_* Q_*' + S Q' = I_T, \tag{6.113}$$

and that:

$$\begin{bmatrix} Q_*' \\ Q' \end{bmatrix} \begin{bmatrix} S_* & S \end{bmatrix} = \begin{bmatrix} Q_*'S_* & Q_*'S \\ Q'S_* & Q'S \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_{T-p} \end{bmatrix}. \tag{6.114}$$

If g_* is available, then y can be recovered from g via:

$$y = S_*g_* + Sg. \tag{6.115}$$

The lower-triangular Toeplitz matrix $\nabla_T^{-p} = [S_*, S]$ is completely characterized by its leading column. The elements of that column are the ordinates of a polynomial of degree $p-1$, of which the argument is the row index $t = 0, 1, \ldots, T-1$. Moreover, the leading p columns of the matrix ∇_T^{-p}, which constitute the submatrix S_*, provide a basis for all polynomials of degree $p-1$ that are defined on the integer points $t = 0, 1, \ldots, T-1$.

It follows that $S_*g_* = S_*Q_*'y$ contains the ordinates of a polynomial of degree $p-1$, which is interpolated through the first p elements of y, indexed by $t = 0, 1, \ldots, p-1$, and which is extrapolated over the remaining integers $t = p, p+1, \ldots, T-1$.

6.7.1 Polynomial regression and HP filtering

A polynomial that is designed to fit the data should take account of all of the observations in y. Imagine, therefore, that $y = \phi + \eta$, where ϕ contains the ordinates of a polynomial of degree $p-1$ and η is a disturbance term with $E(\eta) = 0$ and $D(\eta) = \Sigma$. Then, in forming an estimate $f = S_*r_*$ of ϕ, we should minimize the sum of squares $\eta'\Sigma^{-1}\eta$. Since the polynomial is fully determined by the elements of a starting value vector r_*, this is a matter of minimizing:

$$(y-\phi)'\Sigma^{-1}(y-\phi) = (y - S_*r_*)'\Sigma^{-1}(y - S_*r_*), \tag{6.116}$$

with respect to r_*. The resulting values are:

$$r_* = (S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}y \quad \text{and} \quad \phi = S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}y. \tag{6.117}$$

An alternative representation of the estimated polynomial is available, which avoids the inversion of Σ. This is provided by the identity:

$$\begin{aligned} P_* &= S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1} \\ &= I - \Sigma Q(Q'\Sigma Q)^{-1}Q' = I - P_Q, \end{aligned} \tag{6.118}$$

which gives two representations of the projection matrix P_*. The equality follows from the fact that, if $\text{Rank}[R, S_*] = T$ and if $S_*'\Sigma^{-1}R = 0$, then:

$$S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1} = I - R(R'\Sigma^{-1}R)^{-1}R'\Sigma^{-1}. \tag{6.119}$$

Setting $R = \Sigma Q$ gives the result. It follows that the ordinates of the polynomial fitted to the data by generalized least-squares regression can be represented by:

$$\phi = y - \Sigma Q(Q'\Sigma Q)^{-1}Q'y. \tag{6.120}$$

A more general method of curve fitting, which embeds polynomial regression as a special case, is one that involves the minimization of a combination of two sums of squares. Let x denote the vector of fitted values, which is a sequence of the ordinates of points, equally spaced in time, through which a continuous curve might be interpolated. The criterion for finding the vector is to minimize:

$$L = (y - x)'\Sigma^{-1}(y - x) + x'Q\Omega^{-1}Q'x. \tag{6.121}$$

The first term penalizes departures of the resulting curve from the data, whereas the second term imposes a penalty for a lack of smoothness in the curve.

The second term comprises $d = Q'x$, which is the vector of pth-order differences of x. The matrix Ω^{-1} serves to generalize the overall measure of the curvature of the function that has the elements of x as its sampled ordinates, and it serves to regulate the penalty, which may vary over the sample.

Differentiating L with respect to x and setting the result to zero, in accordance with the first-order conditions for a minimum, gives:

$$\begin{aligned}\Sigma^{-1}(y - x) &= Q\Omega^{-1}Q'x \\ &= Q\Omega^{-1}d.\end{aligned} \tag{6.122}$$

Multiplying the equation by $Q'\Sigma$ gives $Q'(y - x) = Q'y - d = Q'\Sigma Q\Omega^{-1}d$, whence $\Omega^{-1}d = (\Omega + Q'\Sigma Q)^{-1}Q'y$. Putting this into the equation $x = y - \Sigma Q\Omega^{-1}d$ gives:

$$x = y - \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \tag{6.123}$$

By setting $\Omega = \lambda^{-1}I$ and $\Sigma = I$ and letting Q' denote the second-order difference operator, the HP filter is obtained in the form of:

$$x = y - Q(\lambda^{-1}I + Q'Q)^{-1}Q'y. \tag{6.124}$$

This form is closely related to that of the infinite-sample filter $\beta(z) = 1 - \beta^c(z)$ which invokes equation (6.95). In the finite-sample version of the filter, the submatrix Q' of $\nabla_T^2 = (I - L_T)^2$ replaces the difference operator $(1 - z)^2$, and Q replaces $(1 - z^{-1})^2$.

If $\Omega = 0$ in (6.123), and if Q' is the matrix version of the second-difference operator, then the generalized least squares interpolator of a linear function is derived, which is subsumed under (6.120).

6.7.2 Finite-sample WK filters

To provide a statistical interpretation of the formula of (6.123), consider a data sequence $y = \xi + \eta$, where $\xi = \phi + \zeta$ is a trend component, which is the sum of a vector ϕ, containing the ordinates of a polynomial of degree p at most, and of a vector ζ from a stochastic process with p unit roots that is driven by a zero-mean forcing function. The term η stands for a vector sampled from a mean-zero stationary stochastic process which is independent of the process driving ξ such that:

$$E(\eta) = 0, \quad D(\eta) = \Sigma \quad \text{and} \quad C(\eta, \xi) = 0. \tag{6.125}$$

If Q' is the pth difference operator, then $Q'\phi = \mu\iota$, with $\iota = [1, 1, \ldots, 1]'$, will contain a constant sequence of values, which will be zeros if the degree of ϕ is less than p. Also, $Q'\zeta$ will be a vector sampled from a mean-zero stationary process. Therefore, $\delta = Q'\xi$ is from a stationary process with a constant mean. Thus, there is:

$$\begin{aligned} Q'y &= Q'\xi + Q'\eta \\ &= \delta + \kappa = g, \end{aligned} \tag{6.126}$$

where:

$$\begin{aligned} &E(\delta) = \mu\iota, \quad D(\delta) = \Omega, \\ &E(\kappa) = 0, \quad D(\kappa) = Q'\Sigma Q. \end{aligned} \tag{6.127}$$

Now consider the conditional expectation of η given $g = Q'y$, which is also its minimum mean square error estimator on the assumption that the various stochastic processes are normally distributed. This is:

$$\begin{aligned} E(\eta|g) &= E(\eta) + C(\eta, g)D^{-1}(g)\{g - E(g)\} \\ &= \Sigma Q(\Omega + Q'\Sigma Q)^{-1}\{Q'y - \mu\iota\}. \end{aligned} \tag{6.128}$$

If the vector $E(g) = \mu\iota$ is non-zero it will, nevertheless, be virtually nullified by the matrix $\Sigma Q(\Omega + Q'\Sigma Q)^{-1}$, which is a matrix version of a highpass filter. Therefore, it may be deleted from the expressions of (6.128). Next, since $\xi = y - \eta$, the estimate of the trend is $x = E(\xi|g) = y - E(\eta|g)$, which is exactly equation (6.123).

The HP filter may be derived by specializing the statistical assumptions of (6.125) and (6.127). It is assumed that:

$$D(\eta) = \Sigma = \sigma_\eta^2 I, \quad D(\delta) = \Omega = \sigma_\delta^2 I \quad \text{and} \quad \lambda = \frac{\sigma_\eta^2}{\sigma_\delta^2}. \tag{6.129}$$

Putting these details into equation (6.123) gives equation (6.124).

It is straightforward to derive the dispersion matrices that are found within the formulae for the finite-sample estimators from the corresponding autocovariance generating functions. Let $\gamma(z) = \{\gamma_0 + \gamma_1(z + z^{-1}) + \gamma_2(z^2 + z^{-2}) + \cdots\}$ denote the autocovariance generating function of a stationary stochastic process. Then, the corresponding dispersion matrix for a sample of T consecutive elements drawn from the process is:

$$\Gamma = \gamma_0 I_T + \sum_{\tau=1}^{T-1} \gamma_\tau (L_T^\tau + F_T^\tau), \tag{6.130}$$

where $F_T = L_T'$ is in place of z^{-1}. Since L_T and F_T are nilpotent of degree T, such that $L_T^q, F_T^q = 0$ when $q \geq T$, the index of summation has an upper limit of $T - 1$.

6.7.3 The polynomial component

The formula (6.123) tends to conceal the presence of polynomial components within the sequences that are generated by filtering the nonstationary data. An alternative procedure, which we have already adopted in detrending the logarithmic consumption data of the UK in the example following (6.14), is to extract a polynomial trend from the nonstationary data before applying a filter to the residual sequence, which will have the characteristics of a sequence generated by a stationary process, provided that the polynomial is of a sufficient degree.

Another procedure that can be followed requires the data to be reduced to stationarity by a process of differencing, before it is filtered. The filtered output can be reinflated thereafter to obtain estimates of the components of the non-stationary process. It transpires that, in the context of WK filtering, such a procedure produces estimates that are identical to those that are delivered by the finite-sample filter of (6.123).

To demonstrate this result, we shall assume that, within $y = \xi + \eta$, the vector ξ is generated by a stochastic process with p unit roots driven by a mean-zero white-noise process. The vector η is assumed to be from a stationary process. Therefore, the specifications of (6.125) and (6.127) remain, but we may choose to set $E(\delta) = 0$, if only to confirm that the polynomial component will arise just as surely in the absence of stochastic drift.

Let the estimates of ξ, η, $\delta = Q'\xi$ and $\kappa = Q'\eta$ be denoted by x, h, d and k respectively. Then the Wiener–Kolmogorov minimum mean square error estimates of the differenced components are:

$$E(\delta|g) = d = D(\delta)\{D(\delta) + D(\kappa)\}^{-1}g = \Omega(\Omega + Q'\Sigma Q)^{-1}Q'y, \tag{6.131}$$

$$E(\kappa|g) = k = D(\kappa)\{D(\delta) + D(\kappa)\}^{-1}g = Q'\Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \tag{6.132}$$

The estimates of ξ and η may be obtained by integrating, or re-inflating, the components of the differenced data to give

$$x = S_*d_* + Sd \quad \text{and} \quad h = S_*k_* + Sk, \tag{6.133}$$

where S_*d_* and S_*k_* are vectors of the ordinates of polynomials of degree p. For this representation, the polynomial parameters, in the form of the starting values d_* and h_*, are required.

The initial conditions in d_* should be chosen so as to ensure that the estimated trend is aligned as closely as possible with the data. The criterion is:

$$\text{Minimize} \quad (y - S_*d_* - Sd)'\Sigma^{-1}(y - S_*d_* - Sd) \quad \text{with respect to} \quad d_*. \tag{6.134}$$

The solution for the starting values is:

$$d_* = (S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}(y - Sd). \tag{6.135}$$

The equivalent starting values of k_* are obtained by minimizing the (generalized) sum of squares of the fluctuations:

$$\text{Minimize} \quad (S_*k_* + Sk)'\Sigma^{-1}(S_*k_* + Sk) \quad \text{with respect to} \quad k_*. \tag{6.136}$$

The solution is:

$$k_* = -(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}Sk. \tag{6.137}$$

The starting values k_* and d_* can be eliminated from the expressions for x and h in (6.133), which provide the estimates of the components. Thus, using expression $I - P_* = P_Q$ from (6.118), we get:

$$\begin{aligned} h &= Sk + S_*k_* \\ &= (I - P_*)Sk = P_QSk. \end{aligned} \tag{6.138}$$

Then, by using the expression for k from (6.132) together with the identity $Q'S = I_T$, we get:

$$h = \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \tag{6.139}$$

This agrees with (6.128) in the case where $\mu = 0$. The condition that $x + h = y$, which is that the sum of the estimated components equals the data vector, indicates that:

$$\begin{aligned} x &= y - h \\ &= y - \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y, \end{aligned} \tag{6.140}$$

which is equation (6.123) again.

Observe that the filter matrix $Z_\eta = \Sigma Q(\Omega + Q'\Sigma Q)^{-1}$ of (6.140), which delivers $h = Z_\eta g$, differs from the matrix $Z_\kappa = Q'Z_\eta$ of (6.132), which delivers $k = Z_\kappa g$, only in respect of the matrix difference operator Q'. The effect of omitting the operator is to remove the need for reinflating the filtered components and thus to remove the need for the starting values. These matters have been discussed at greater length by Pollock (2006).

6.8 The Fourier methods of signal extraction

If the data are generated by a stationary stochastic process, then it may be reasonable to regard them as the product of a circular process, of which the Fourier representation is readily available. There are some advantages in exploiting the Fourier representation by performing the essential filtering operations in the frequency domain – for these are usually aimed at suppressing or attenuating some of the cyclical elements of the data. It is also straightforward to provide a time-domain interpretation of the frequency domain operations, and the possibility exists of performing the equivalent operations in either domain.

The dispersion matrix of a circular stochastic process is obtained from the autocovariance generating function $\gamma(z)$ by replacing the argument z by the circulant matrix $K_T = [e_1, \ldots, e_{T-1}, e_0]$, which is formed from the identity matrix $I_T = [e_0, e_1, \ldots, e_{T-1}]$ by moving the leading column to the back of the array. In

this way, the generating function $\gamma(z)$ gives rise to the matrix:

$$\begin{aligned}\Gamma^\circ &= \gamma(K_T)\\ &= \gamma_0 I_T + \sum_{\tau=1}^{\infty} \gamma_\tau (K_T^\tau + K_T^{-\tau})\\ &= \gamma_0 I_T + \sum_{\tau=1}^{T-1} \gamma_\tau^\circ (K_T^\tau + K_T^{-\tau}).\end{aligned} \tag{6.141}$$

It can be seen from this that the circular autocovariances would be obtained by wrapping the sequence of ordinary autocovariances around a circle of circumference T and adding the overlying values. Thus:

$$\gamma_\tau^\circ = \sum_{j=0}^{\infty} \gamma_{jT+\tau}, \quad \text{with} \quad \tau = 0, \ldots, T-1. \tag{6.142}$$

Given that $\lim(\tau \to \infty)\gamma_\tau = 0$, it follows that $\gamma_\tau^\circ \to \gamma_\tau$ as $T \to \infty$, which is to say that the circular autocovariances converge to the ordinary autocovariances as the circle expands.

The circulant autocovariance matrix is amenable to a spectral factorization of the form:

$$\Omega^\circ = \gamma(K_T) = \bar{U}\gamma(D)U, \tag{6.143}$$

wherein U and $\bar{U}$ are the unitary matrices defined by (6.20) and:

$$D = \operatorname{diag}(\exp\{\mathrm{i}2\pi j/T\}; j = 0, \ldots, T-1), \tag{6.144}$$

is a diagonal matrix whose elements are the T roots of unity, which are found on the circumference of the unit circle in the complex plane. Then, $\gamma(D)$ is the diagonal matrix formed by replacing the argument z within $\gamma(z)$ by D.

The jth element of the diagonal matrix $\gamma(D)$ is:

$$\gamma(\exp\{\mathrm{i}\omega_j\}) = \gamma_0 + 2\sum_{\tau=1}^{\infty} \gamma_\tau \cos(\omega_j \tau). \tag{6.145}$$

This represents the cosine Fourier transform of the sequence of the ordinary autocovariances; it corresponds to an ordinate (scaled by 2π) sampled at the point $\omega_j = 2\pi j/T$, which is a Fourier frequency, from the spectral density function of the linear (that is, non-circular) stationary stochastic process.

The theory of circulant matrices has been described by Gray (2002) and by Pollock (2002a). Both authors provide abundant additional references.

The method of WK filtering can also be implemented using the circulant dispersion matrices that are given by:

$$\begin{aligned}&\Omega_\delta^\circ = \bar{U}\gamma_\delta(D)U, \quad \Omega_\kappa^\circ = \bar{U}\gamma_\kappa(D)U \quad \text{and}\\ &\Omega^\circ = \Omega_\delta^\circ + \Omega_\kappa^\circ = \bar{U}\{\gamma_\delta(D) + \gamma_\kappa(D)\}U,\end{aligned} \tag{6.146}$$

wherein the diagonal matrices $\gamma_\delta(D)$ and $\gamma_\kappa(D)$ contain the ordinates of the spectral density functions of the component processes. By replacing the dispersion matrices of (6.131) and (6.132) by their circulant counterparts, we derive the following formulae:

$$d = \bar{U}\gamma_\delta(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}Ug = P_\delta g, \tag{6.147}$$

$$k = \bar{U}\gamma_\kappa(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}Ug = P_\kappa g. \tag{6.148}$$

We may note that P_δ and P_κ are circulant matrices.

The filtering formulae may be implemented in the following way. First, a Fourier transform is applied to the (differenced) data vector g to give Ug, which resides in the frequency domain. Then, the elements of the transformed vector are multiplied by those of the diagonal weighting matrices $J_\delta = \gamma_\delta(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}$ and $J_\kappa = \gamma_\kappa(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}$. Finally, the products are carried back into the time domain by the inverse Fourier transform, which is represented by the matrix $\bar{U}$. (An efficient implementation of a mixed-radix fast Fourier transform, which is designed to cope with samples of arbitrary sizes, has been provided by Pollock, 1999. The usual algorithms demand a sample size of $T = 2^n$.)

An advantage of the Fourier method is that it is possible to effect a total suppression of the elements within the stop band of the desired frequency response. Also, the transition between the pass band and the stop band can be confined to the interval between adjacent Fourier frequencies, which means that it can be perfectly abrupt.

Neither of these features are available to the ordinary finite-sample WK filters. Nevertheless, it is possible to achieve both of these effects by working in the time domain. This fact is manifest in the formulae of (6.147) and (6.148) which entail the equations $d = P_\delta g$ and $k = P_\kappa g$ respectively.

In effect, a pair of wrapped filters can be applied to the data in the time domain via processes of circular convolution. If we can imagine the leading rows of the matrices P_δ and P_κ disposed around a circle of circumference T, then each of the succeeding rows is derived from its predecessor via an anticlockwise rotation through an angle of $2\pi/T$ radians.

Example It is commonly believed that, in the case of samples of finite length T, it is impossible to design a filter that will preserve completely all elements within a specified range of frequencies and that will remove all elements outside it. A filter that would achieve such an objective is described as an ideal filter. The ideal lowpass filter with a cut-off frequency of $\omega_d = 2\pi d/T$ has the following frequency response over the interval $[-\pi, \pi]$:

$$\phi(\omega) = \begin{cases} 1, & \text{if } \omega \in [-\omega_d, \omega_d], \\ 0, & \text{otherwise.} \end{cases} \tag{6.149}$$

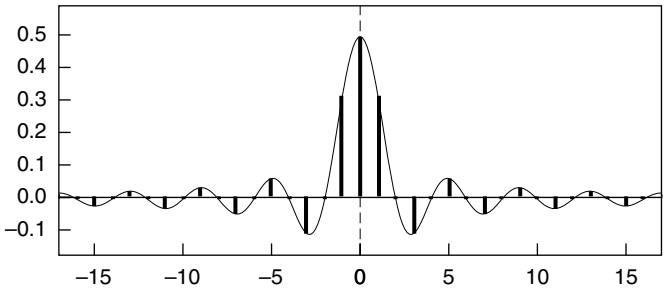


Figure 6.13 The central coefficients of the Fourier transform of the frequency response of an ideal lowpass filter with a cut-off point at $\omega = \pi/2$. The sequence of coefficients extends indefinitely in both directions. The coefficients are the sampled ordinates of a sinc function

The coefficients of the filter are given by the discrete-time sinc function, which is the (inverse) Fourier transform of the periodic frequency response function:

$$\beta_k = \frac{1}{2\pi}\int_{-\omega_d}^{\omega_d} e^{i\omega k}d\omega = \begin{cases} \dfrac{\omega_d}{\pi}, & \text{if } k = 0; \\ \dfrac{\sin(k\omega_d)}{\pi k}, & \text{if } k \neq 0. \end{cases} \tag{6.150}$$

Such a frequency response presupposes a doubly-infinite data sequence, insofar as it represents the relative amplification and attenuation of trigonometrical functions that are defined over the entire real line.

The coefficients of (6.150) form a doubly infinite sequence, of which a central part is illustrated in Figure 6.13. In order to obtain a practical filter, it seems that one must truncate the sequence, retaining only a limited number of its central elements. This truncation gives rise to a filter of which the frequency response has certain undesirable characteristics. (See Figure 6.19 for an example.)

In particular, there is a ripple effect whereby the gain of the filter fluctuates within the pass band, where it should be constant with a unit value, and within the stop band, where it should be zero-valued. Within the stop band, there is a corresponding problem of leakage whereby the truncated filter transmits elements that ought to be blocked.

However, it is clear that an ideal filter can be implemented in the frequency domain by preserving the ordinates of the Fourier transform of the data that are associated with frequencies less than ω_d and by setting all other ordinates to zero. This is a matter of applying the following set of weights to the Fourier ordinates:

$$\lambda_j = \begin{cases} 1, & \text{if } j \in \{-d, \ldots, d\}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.151}$$

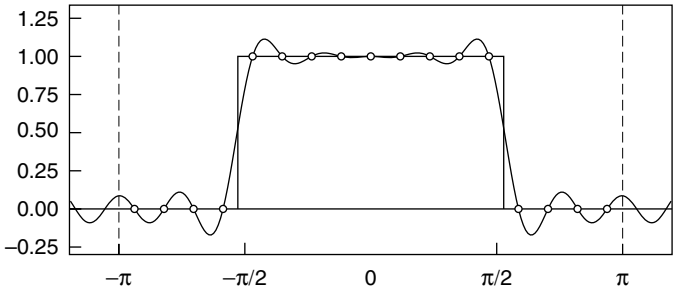


Figure 6.14 The frequency response of the 17-point wrapped filter defined over the interval $[-\pi, \pi)$. The values at the Fourier frequencies are marked by circles

By applying an inverse discrete Fourier transform to these weights, the coefficients of a circular filter are obtained, of which the values are given by:

$$\beta^{\circ}(k) = \begin{cases} \dfrac{2d+1}{T}, & \text{if k = 0,} \\ \dfrac{\sin([d+1/2]\omega_1 k)}{T\sin(\omega_1 k/2)}, & \text{for k = 1, … , [T/2],} \end{cases} \tag{6.152}$$

where $\omega_1 = 2\pi/T$. These coefficients would be obtained by wrapping coefficients of (6.150) around a circle of circumference T and adding the overlying values:

$$\beta_k^{\circ} = \sum_{j=-\infty}^{\infty} \beta_{jT+k}. \tag{6.153}$$

Applying the wrapped filter to the finite data sequence via a circular convolution is equivalent to applying the original filter to an infinite periodic extension of the data sequence.

The function of (6.152) is just an instance of the Dirichlet kernel – see Pollock (1999), for example. Figure 6.14 depicts the frequency response for this filter at the Fourier frequencies, where $\lambda_j = 0, 1$ in the case where $\omega_d = \pi/2$. It also depicts the continuous frequency response that would be the consequence of applying an ordinary filter with these coefficients to a doubly-infinite data sequence.

6.8.1 Applying the Fourier method to trended data

In an ideal application of the Fourier method, it should be possible to wrap the data sequence $y_t; t = 0, \ldots, T-1$ seamlessly around the circle, such that there is no disjunction at the point where the head of the sequence joins the tail. To achieve such an effect, it is common to taper the data so as reduce both ends to zero. To avoid corrupting the sample data, the taper can be applied to some extrapolations of the ends of the sample. However, a data sequence that follows a linear trend is not amenable to tapering, since there is liable to be a radical disjunction at the point where the head joins the tail.

The periodic extension of the linearly trended sequence, which would be generated by traveling around the circle indefinitely, has a saw-tooth profile. The corresponding spectrum or periodogram has a *one-over-f* profile that descends, as the frequency increases, in the manner of a rectangular hyperbola, from a high point that is adjacent to the zero frequency to a low point at the limiting frequency. Unless the data are adequately detrended, such a spectrum will serve to conceal all but the most prominent of the harmonic characteristics of the data.

There are two simple ways in which the data may be detrended. The first, which has been described already in section 6.7.3, is to apply the difference operator to the data as many times as are necessary to reduce them to stationarity. The components that are extracted by filtering the differenced data can be reinflated, in the manner indicated by equations (6.133)–(6.137), to obtain the components of the original data.

We denote the data by y and their differences by $g = Q'y$. The filtered sequence that underlies the trend is denoted by d and the vector of initial conditions by d_*. Then, if we set $\Sigma = I$, the relevant equations for delivering the estimate x of the trend component are:

$$x = S_*d_* + Sd \quad \text{and} \quad d_* = (S_*'S_*)^{-1}S_*'(y - Sd). \tag{6.154}$$

The detrended sequence is $h = y - x$. Underlying the detrended sequence is the filtered sequence $k = g - d$, from which the detrended data component may be obtained directly via the equations:

$$h = S_*k_* + Sk \quad \text{and} \quad k_* = -(S_*'S_*)^{-1}S_*'Sk. \tag{6.155}$$

Another way of reversing the effects of a differencing operation that has been applied to the data to reduce them to stationarity is to re-inflate the Fourier ordinates of the filtered sequence, using values from the frequency response function of the anti-differencing summation operator. Once the ordinates had been reinflated within the frequency domain, they can be transformed into the time domain to produce the filtered sequence.

This method is applicable only to components that are bounded away from the zero frequency, since the summation operator has infinite gain at zero. (See Figure 6.8.) However, if one wishes to apply a lowpass filter to the data, then one has the option of applying the complementary highpass filter and of subtracting the filtered sequence from the original data to generate the lowpass component.

The second way of detrending the data is to extract a polynomial component via an ordinary or a generalized least squares regression according to the formula of (6.120). The formula will allow greater weight to be given to the points at both ends of the sample, to ensure that the interpolated curve passes through their midst. This can be achieved by allowing Σ^{-1} to be a diagonal matrix with large values at the ends. In this way, a disjunction in the wrapped version of the residual sequence, or in its periodic extension, can be avoided.

Example Figure 6.15 shows the logarithms of the data on aggregate household expenditure in the UK for the years 1956–2005, through which a smooth trajectory

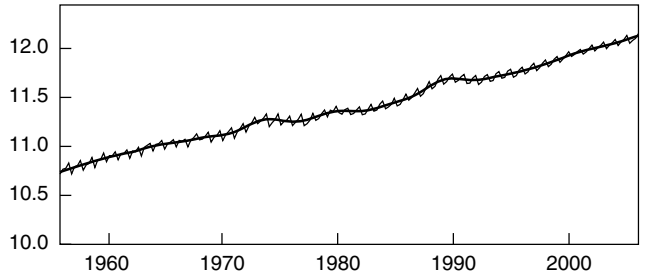


Figure 6.15 The logarithms of quarterly household expenditure in the UK, for the years 1956–2005, together with an interpolated trend

has been interpolated. This has been obtained by selecting the Fourier coefficients of the twice-differenced data that correspond to frequencies in the interval $[0, \pi/8]$. This frequency band has been chosen in the light of the periodogram of Figure 6.6, which shows that it contains an isolated spectral structure.

The sequence that has been synthesized from these coefficients has been reinflated in the manner indicated by (6.154) to produce the trajectory. The result of this procedure is a composite of the trend and the business cycle. The same trajectory of aggregate expenditure would have been obtained by adding the business cycle that is depicted in Figure 6.5 to the linear trend of Figure 6.4.

6.9 Band-limited processes

The majority of the methods that we have described for extracting the components of an econometric data sequence presuppose that the data can be described by a univariate ARIMA model. The spectral density function of an ARIMA process is supported on the entire frequency interval $[0, \pi]$, where its ordinates are strictly positive with the possible exception of a few zero-valued ordinates that constitute a set of measure zero. Such zero values will be attributable to the presence of unit roots within the moving-average operator.

It is commonly assumed that the component parts of an aggregate econometric sequence can also be described by ARIMA models. It is on this basis that the WK filters are derived. However, reference to the periodogram of Figure 6.6 and to others like it suggests that the components often reside within strictly limited frequency bands which are separated by dead spaces where the spectral ordinates are virtually zeros.

In many circumstances, the disparity between the assumptions underlying the WK filters and the nature of the data to which they are applied has no adverse effects. A lowpass filter that achieves a gradual transition from a pass band to a stop band within the region of a spectral dead space will be as effective in extracting a low-frequency trend component as is a frequency-domain filter that achieves an abrupt transition between two adjacent Fourier frequencies.

The ordinates at times $t = 0, \ldots, T-1$ of the business cycle that is represented in Figure 6.5 have been obtained by a Fourier method; but they might have been obtained by applying the Butterworth filter of order $n = 6$ and with a nominal cut-off frequency of $\omega_d = \pi/4$ radians, of which the gain is depicted in Figure 6.11. The principal advantage of the Fourier method, in this context, lies in the ease with which a continuous function can be synthesised from the Fourier coefficients.

Difficulties do arise when an attempt is made to estimate the parameters of an ARMA model from data such as those of Figure 6.5. A natural objective is to attempt to characterize the business cycle via the parameters of a fitted ARMA model. Such a model is liable to be applied to a seasonally adjusted version of the data, for which the periodogram will lack the spectral spike at the seasonal frequency of $\pi/2$ and at the harmonic frequency of π.

An AR(2) model with complex roots is the simplest of the models that might be appropriate to the purpose. The modulus of its roots should reveal the damping characteristics of the cycles, and their argument should indicate the angular velocity or, equivalently, the length, of the cycles. However, such a model will invariably deliver estimates that imply real-valued roots, which fail adequately to represent the dynamics of the business cycle. (See Pagan, 1997, for example.)

The problem of estimating the business cycle also affects the model-based approaches to econometric signal extraction, which depend upon the prior estimation of an aggregate ARIMA model or upon the estimation of ARIMA components. A business cycle component is usually missing from such models, since the estimation fails to deliver the appropriate complex roots. However, it is straightforward to include a business cycle component with a pre-specified frequency in a disaggregated structural model. (See Harvey, 1985, for example.)

To obtain parametric estimates of the business cycle, it is necessary to remove from the data all but the relevant low-frequency components. This is achieved by selecting the relevant Fourier coefficients from which the business cycle can be constituted via a Fourier synthesis in the manner of (6.14). Thereafter, it is necessary to sample the continuous function at a rate that will ensure that the Nyquist frequency π corresponds to the highest frequency that is present in the component. A successful ARMA model which represents the complex dynamics of the business cycle can be estimated from the resampled data sequence.

The Shannon–Whittaker sampling theorem indicates that the resampled data contains sufficient information to reconstitute the continuous business cycle function.

6.9.1 The Shannon–Whittaker sampling theorem

Let $x(t)$ be a square integrable continuous signal of which the Fourier transform $\xi(\omega)$ is band limited to the frequency interval $[-\omega_d, \omega_d]$. Then the signal can be recovered from its sampled ordinates provided that these are separated by a time interval of no more than π/ω_d, which is to say that the sinusoidal element of the highest frequency within the signal must take at least two sampling intervals to complete a cycle.

To demonstrate this result, we must consider the Fourier representation of a real-valued square-integrable function $x(t)$ defined over the real line. The following are the corresponding expressions for the function $x(t)$ and its Fourier transform $\xi(\omega)$:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}\omega t}\xi(\omega)d\omega \longleftrightarrow \xi(\omega) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\omega t}x(t)dt. \tag{6.156}$$

By sampling $x(t)$ at intervals of π/ω_d, a sequence:

$$\{x_\tau = x(\tau[\pi/\omega_d]); \tau = 0, \pm 1, \pm 2, \ldots\},$$

is generated. The elements of the sequence and their Fourier transform $\xi_S(\omega)$ are given by:

$$\begin{aligned} x_\tau &= \frac{1}{2\omega_d}\int_{-\omega_d}^{\omega_d} \exp\{\mathrm{i}\omega\tau[\pi/\omega_d]\}\xi_S(\omega)d\omega \\ &\longleftrightarrow \\ \xi_S(\omega) &= \sum_{\tau=-\infty}^{\infty} x_\tau \exp\{-\mathrm{i}\omega\tau[\pi/\omega_d]\}. \end{aligned} \tag{6.157}$$

Since $\xi(\omega) = \xi_S(\omega)$ is a continuous function defined on the interval $[-\omega_d, \omega_d]$, it may be regarded as a function that is periodic in frequency, with a period of $2\omega_d$. Putting the right-hand side of (6.157) into the left-hand side of (6.156), and taking the integral over $[-\omega_d, \omega_d]$ in consequence of the band-limited nature of the function $x(t)$, gives:

$$\begin{aligned} x(t) &= \frac{1}{2\pi}\int_{-\omega_d}^{\omega_d}\left\{\sum_{\tau=-\infty}^{\infty} x_\tau e^{-\mathrm{i}\omega\tau[\pi/\omega_d]}\right\}e^{\mathrm{i}\omega t}d\omega \\ &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} x_\tau \int_{-\omega_d}^{\omega_d} e^{\mathrm{i}\omega(t-[\tau\pi/\omega_d])}d\omega. \end{aligned} \tag{6.158}$$

The integral on the right-hand side is evaluated as:

$$\int_{-\omega_d}^{\omega_d} e^{\mathrm{i}\omega(t-[\tau\pi/\omega_d])}d\omega = 2\frac{\sin(t\omega_d - \tau\pi)}{t - \tau[\pi/\omega_d]}. \tag{6.159}$$

Putting this into the right-hand side of (6.158) gives:

$$x(t) = \sum_{\tau=-\infty}^{\infty} x_\tau \frac{\sin(t\omega_d - \tau\pi)}{\pi(t - \tau[\pi/\omega_d])} = \sum_{k=-\infty}^{\infty} x_\tau \phi_d(\tau - k), \tag{6.160}$$

where:

$$\phi_d(t-\tau) = \frac{\sin(t\omega_d - \tau\pi)}{\pi(t - \tau[\pi/\omega_d])}. \tag{6.161}$$

When $\tau = 0$, this becomes an ordinary sinc function that is a continuous function of t, and which is the Fourier transform of the following frequency function:

$$\phi_d(\omega) = \begin{cases} 1, & \text{if } |\omega| \in [0, \omega_d]; \\ 0, & \text{otherwise.} \end{cases} \tag{6.162}$$

When $\tau \neq 0$, it represents a sinc function that has been displaced in time by τ intervals of length π/ω_d. The set of such displaced sinc functions constitutes an orthogonal basis for all continuous functions that are band-limited to the frequency interval $[-\omega_d, \omega_d]$.

In the case of a stationary stochastic process, the sampled sequence is not square summable and, therefore, in a strict sense, this proof of the interpolation via the Nyquist–Shannon theory does not apply. However, the convergence of the interpolation formula of (6.160), when $x(\tau) = \{x_\tau; \tau = 0, \pm 1, \pm 2, \ldots\}$ is a stationary sequence, can be confirmed by considering a sum with $\tau \in [-N, N]$ for some finite integer N. The variance of the sum of discarded terms can be made arbitrarily small by increasing the value of N.

The reconstruction of a continuous function from its sampled ordinates in the manner suggested by the sampling theorem is not possible in practice, because it requires forming a weighted sum of an infinite number of sinc functions, each of which is supported on the entire real line. Nevertheless, a continuous band-limited periodic function defined on a finite interval – which corresponds to the circumference of a circle – can be reconstituted from a finite number of wrapped or periodic sinc functions, which are Dirichlet kernels by another name. However, the most practical means of reconstituting the function is by a simple Fourier synthesis of the sort described by equation (6.14).

Example The analysis of the example following (6.14) suggests that the business cycle of the detrended logarithmic consumption data fits within the frequency band $[0, \pi/8]$. If this structure can be isolated and thereafter mapped into the frequency interval $[0, \pi]$, then it will be capable of being described by an ordinary linear stochastic model of the ARMA variety. For this purpose, the spectral elements that fall outside the frequency range of the business cycle must first be removed. This operation, which constitutes an anti-alias filtering, may be carried out either in the time domain or in the frequency domain.

Given the availability of the spectral ordinates of the data, it is straightforward to operate in the frequency domain by setting the rejected ordinates to zeros. Then, a continuous low-frequency function can be synthesized from the selected ordinates. An example is provided by the interpolated function in Figure 6.5. The synthesized function can be resampled at a rate that corresponds to the maximum frequency within the spectral structure of the business cycle.

There is some advantage in fitting a trend function that is more flexible than the straight line of Figure 6.5. Therefore, a fourth degree polynomial has been fitted to

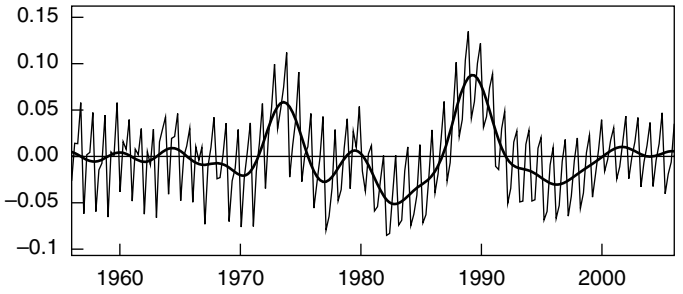


Figure 6.16 The residuals from fitting a polynomial of degree 4 to the logarithmic expenditure data. The interpolated line, which represents the business cycle, has been synthesized from the Fourier ordinates in the frequency interval $[0, \pi/8]$

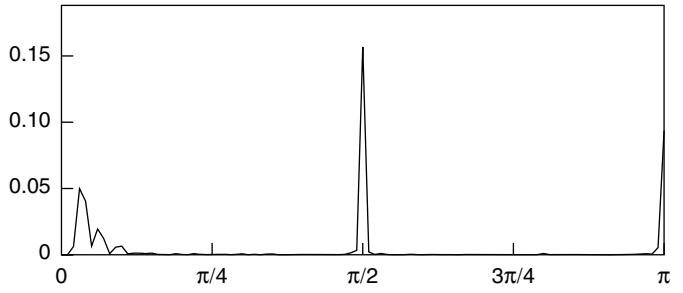


Figure 6.17 The periodogram of the data sequence of Figure 6.16

the data by a least squares regression. The effect is to remove some of the power from the Fourier ordinates adjacent to the zero frequency.

The residual sequence from this polynomial interpolation is show in Figure 6.16, together with an interpolated function that has been synthesized from the Fourier ordinates that lie in the interval $[0, \pi/8]$. This, function, which purports to represent the business cycle, is devoid of any seasonal fluctuations. Figure 6.17 displays the periodogram of the residual sequence.

After the removal of all elements of frequencies in excess of $\pi/8$ the data may be resampled at 1/8th of the original rate of observation. This simple fractional rate is a convenient one, since it implies taking one in every eight of the anti-aliased data points. In that case, there is no need to synthesize a continuous function for the purpose of resampling the data.

The periodogram of the sub-sampled anti-aliased data is show in Figure 6.18 with the parametric spectrum of an estimated AR(3) model superimposed. The periodogram represents a rescaled version of the part of the periodogram of Figure 6.17 that occupies the frequency range $[0, \pi/8]$ and it appears to be well represented by the parametric spectrum.

The continuous band-limited function of Figure 6.16 can be recovered from the sub-sample by associating to each of its elements an appropriately scaled Dirichlet

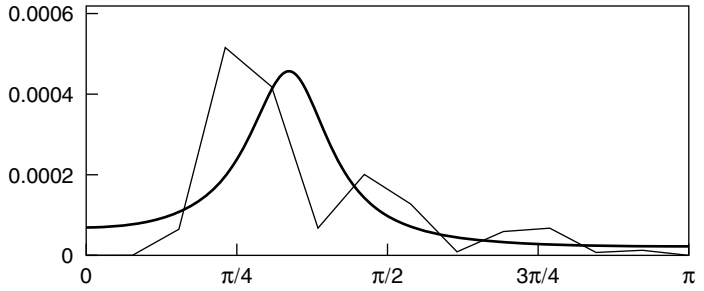


Figure 6.18 The periodogram of the sub-sampled anti-aliased data with the parametric spectrum of an estimated AR(3) model superimposed

kernel and, thereafter, by adding these kernels. This demonstrates the one-to-one correspondence that exists between the continuous function and the sub-sampled sequence. This is precisely the one-to-one correspondence that exists between the periodic function $z(t)$, synthesized by equation (6.14), and its sampled ordinates $\{z_\tau = z(\tau T/N); \tau = 0, 1, \ldots, N-1\}$.

The AR(3) model that underlies the spectral density function of Figure 6.18 provides a statistical description both of the continuous band-limited function of Figure 6.16 and of the ordinates sampled from it at the rate of one observation in eight sample periods.

6.10 Separating the trend and the cycles

The remaining issue to be discussed in this chapter is the matter of separating the trend of an economic data sequence from the cycles that surround it. This is a difficult problem. The trend and the cycles are combined within the same spectral structure and there is rarely any indication, within the periodogram, of where the trend ends and the cycles begin. In the absence of objective criteria for achieving a separation, the definition of the trend is liable to reflect the purposes of the study as well as the circumstances of the economy over the period in question.

A simple prescription that was offered by the pioneering econometrician Tintner (1940, 1953) is that the trend should contain no cyclical motions. This can be interpreted to mean that, if the trend is a differentiable function, then its first derivative should have no more than one local maximum or one local minimum. Such a function can be described as a pure trend. A polynomial function of low degree fitted to the data by least squares regression is liable to fulfill the requirement and it can provide an appropriate benchmark for measuring the cyclical variations.

An example of such a trend is the linear function of Figure 6.5, which has been applied to logarithmic data. When a quadratic function was fitted to the data by least squares regression, the result was virtually a straight line. The data are

from a period that was characterized by uninterrupted economic growth at annual rates that varied little. Therefore, the method of polynomial detrending works well.

In other eras, where there have been marked disruptions, the polynomial method is less appropriate. In order to serve as a benchmark for the ensuing periods of stability, the trend must be made to absorb the disruptions, which implies that it must have a segmented structure. In section 6.10.2 we will describe a method for achieving this.

A prescription that is to be found in the pioneering work of Burns and Mitchell (1946) is that the business cycle should be defined in terms of a limited band of frequencies. A modern interpretation of this is that the band should comprise the sinusoidal elements of the data that have cyclical durations of no more that eight years and of no less than a year and a half. Such cycles can be extracted from the data via a bandpass filter, as we will discuss below.

The definition seems arbitrary, but it might be justified by proposing that the reactions of economic agents to cycles within the frequency band differ from their reactions to cycles at other frequencies. Thus, it might be argued that their adaptations to cycles of more than eight years' duration occur mainly at a subconscious level, whereas cycles of a lesser duration incite conscious reactions.

The growth of an economy may be likened to a process of biological growth, which is affected by events that occur in the course of its evolution. Therefore, a stochastic trend based on the accumulation of random increments has been seen as an appropriate model for an economic trend. This idea has inspired the Beveridge–Nelson decomposition of an ARIMA process, which depicts the trend as an accumulation of disturbances that also give rise to accompanying fluctuations.

In practice, the Beveridge–Nelson decomposition depends upon a linear filter that is applied to the data sequence like any other filter. However, the filtered sequence that represents the trend is liable to include a substantial proportion of the high-frequency elements of the data and for that reason it may be regarded as unacceptable.

6.10.1 Bandpass filters

In an attempt to separate a business cycle component from the trend, economists have been resorting increasingly to the use of bandpass filters to implement the definition of Burns and Mitchell (1946). This appears to be in response to the fact that the structural time series methods, which use ARIMA models to represent the unobserved components, fail to isolate the business cycle.

An ideal bandpass filter that transmits all elements within the frequency range $[\alpha, \beta]$, and blocks all others, has the following frequency response:

$$\psi(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (\alpha, \beta); \\ 0, & \text{otherwise.} \end{cases} \tag{6.163}$$

The coefficients of the corresponding time-domain filter are obtained by applying an inverse Fourier transform to this response to give:

$$\begin{aligned}\psi(k) = \int_{\alpha}^{\beta} e^{\mathrm{i}k\omega} d\omega &= \frac{1}{\pi k}\{\sin(\beta k) - \sin(\alpha k)\} \\ &= \frac{2}{\pi k}\cos\{(\alpha+\beta)k/2\}\sin\{(\beta-\alpha)k/2\} \\ &= \frac{2}{\pi k}\cos(\gamma k)\sin(\delta k).\end{aligned} \tag{6.164}$$

Here, $\gamma = (\alpha+\beta)/2$ is the centre of the pass band and $\delta = (\beta-\alpha)/2$ is half its width.

The final equality, which follows from the identity $\sin(A+B) - \sin(A-B) = 2\cos A \sin B$, suggests two interpretations. On the left-hand side is the difference between the coefficients of two lowpass filters with cut-off frequencies of β and α respectively. On the right-hand side is the result of shifting a lowpass filter with a cut-off frequency of δ so that its center is moved from $\omega = 0$ to $\omega = \gamma$.

The process of frequency shifting is best understood by taking account of both positive and negative frequencies when considering the lowpass filter. Then the pass band covers the interval $(-\delta, \delta)$. To convert to the bandpass filter, two copies of the pass band are made that are shifted so that their new centers lie at $-\gamma$ and γ. In the limiting case, the copies are shifted to the centers $-\pi$ and π. There they coincide, and we have $\psi(k) = 2\cos(\pi k)\sin(\delta k)/\pi k$, which constitutes an ideal highpass filter. A bandpass filter can also be expressed as the difference of two such highpass filters.

The coefficients of (6.164) constitute an infinite sequence, which needs to be truncated to produce a practical filter. Alternatively, a wrapped or circular filter may be obtained by sampling the frequency response at a set of equally spaced points in the frequency range $[-\pi, \pi)$, equal in number to the elements of the data sequence. The wrapped filter is obtained by applying the discrete Fourier transform to the sampled ordinates and it can be applied to the data sequence by circular convolution.

The z-transform of a set of filter coefficients that are symmetric about the central point and that sum to zero incorporates the factor $(1-z)(1-z^{-1}) = -z^{-1}(1-z)^2$. This operator is effective in nullifying a linear trend and in reducing a quadratic trend to a constant. Therefore, such a filter can be applied by linear convolution to a trended data sequence in the expectation that it will produce a stationary filtered sequence.

This is one of the attractions of the truncated bandpass filter that has been proposed to economists by Baxter and King (1999). To ensure that the coefficients of the truncated filter do sum to zero, the filter can be expressed as the difference between two truncated versions of the ideal lowpass filter, of which the coefficients have been scaled so as to sum to unity.

The truncated filter has several disadvantages. In the first place, the truncation leads to the phenomenon of leakage that has already been described in section 6.8. This is illustrated by Figure 6.19. Also, a finite-order moving-average filter with

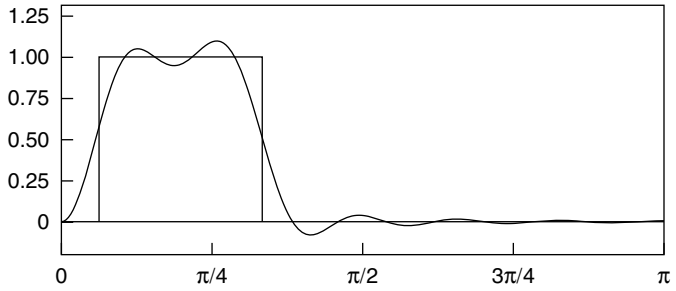


Figure 6.19 The frequency response of the truncated bandpass filter of 25 coefficients superimposed upon the ideal frequency response. The lower cut-off point is at $\pi/15$ radians (11.25°), corresponding to a period of 6 quarters, and the upper cut-off point is at $\pi/3$ radians (60°), corresponding to a period of 32 quarters

constant coefficients is incapable of reaching the ends of the sample. This problem occasions a trade-off between the accuracy of the approximation to the ideal filter, which increases with the number of coefficients, and the end-of-sample problem, which is exacerbated by increasing the span of the filter.

There are numerous ways of overcoming the end-of sample problem, including the obvious recourse of extrapolating the sample by forecasting and backcasting it with the help of an ARIMA model that purports to describe the data. Another recourse is to extend the sample by attaching its symmetric reflection to either end. However, if the data are strongly trended this will tend to increase the values at the beginning of the sample and to decrease the values at the end, relative to the values obtained via a linear extrapolation of the sample.

A circular filter should not be applied directly to a trended data sequence. When such a sequence is wrapped around a circle there is liable to be a radical disjunction where the beginning and the end of the sample are joined. The effects of this disjunction are liable to be carried into the filtered sequence in a manner that does not affect the ordinary linear filter. One way of overcoming this difficulty is to apply the circular filter to data that have been reduced to stationarity by differencing. Thereafter, the filtered differenced sequence can be cumulated to obtain an estimate of the business cycle component.

Example The filter of Baxter and King (1999) is a time-invariant moving average comprising $2q + 1$ of the central coefficients of the ideal infinite-order bandpass filter, which are symmetrically disposed around the central value. These coefficients should rescaled so that they sum to zero.

The elements of the filtered sequence are given by:

$$\begin{aligned} x_t = \phi_q y_{t-q} + \phi_{q-1} y_{t-q+1} + \cdots + \phi_1 y_{t-1} + \phi_0 y_t \\ + \phi_1 y_{t+1} + \cdots + \phi_{q-1} y_{t+q-1} + \phi_q y_{t+q}. \end{aligned} \tag{6.165}$$

Given a sample $y_0, y_1, \ldots, y_{T-1}$ of T data points, only $T - 2q$ processed values $x_q, x_{q+1}, \ldots, x_{T-q-1}$ are available, since the filter cannot reach the ends of the sample, unless some extrapolations are added to it.

To overcome this difficulty, Christiano and Fitzgerald (2003) have used a filter that comprises selections of the coefficients of the ideal filter which vary as one moves through the sample. At all times, the central coefficient of the ideal filter is aligned with the current data value. The remainder of the selection consists of the coefficients on either side that fall within the data window. Thus, the filtered values are weighted combinations of all of the sample elements.

In the case of data that might have been generated by a random-walk process, it is proposed to supplement the weighted sum by two additional terms based on the first and the final sample elements, which are the appropriate predictors of the elements of the process that fall outside the data window. In that case, the elements of the filtered sequence will be given by:

$$\begin{aligned} x_t = {} & Ay_0 + \phi_t y_0 + \cdots + \phi_1 y_{t-1} + \phi_0 y_t \\ & + \phi_1 y_{t+1} + \cdots + \phi_{T-1-t} y_{T-1} + By_{T-1}, \end{aligned} \tag{6.166}$$

where A and B are sums of the coefficients of the ideal filter that lie beyond either end of the data window. Since the filter coefficients must sum to zero, it follows that:

$$A = -\left(\frac{1}{2}\phi_0 + \phi_1 + \cdots + \phi_t\right) \quad \text{and} \quad B = -\left(\frac{1}{2}\phi_0 + \phi_1 + \cdots + \phi_{T-t-1}\right). \tag{6.167}$$

For data that appear to have been generated by a first-order random walk with a constant drift, it is appropriate to extract a linear trend before filtering the residual sequence. In fact, this has proved to be the usual practice in most circumstances.

It has been proposed to subtract from the data a linear function $f(t) = \alpha + \beta t$ interpolated through the first and the final data points, such that $\alpha = y_0$ and $\beta = (y_{T-1} - y_0)/T$. In that case, there should be $A = B = 0$. This procedure is appropriate to seasonally adjusted data. For data that manifest strong seasonal fluctuations, such as the UK expenditure data, a line can be fitted by least squares through the data points of the first and the final years. Figure 6.20 shows the effect of the application of the filter to the UK data adjusted in this manner.

Figure 6.20 can be compared with Figure 6.5 and Figure 6.16, both of which also purport to show the business cycles that affected the data in question. It is clear that the bandpass filter fails to transmit the appropriate cyclical fluctuations. An explanation for the failure can be found in Figure 6.6, which shows the periodogram of the linearly detrended data.

The highlighted band in Figure 6.6 covers the frequency interval $[\pi/16, \pi/3]$ which, according to Baxter and King (1999), is the frequency range that defines the business cycle. However, this figure indicates that only a small part of the low-frequency component falls within the interval. Therefore, it appears that the definition is at fault. In fact, the leakage that is associated with the filter does allow some of the low-frequency power of the elements that reside in the interval $[0, \pi/16]$ to pass into the filtered sequence.

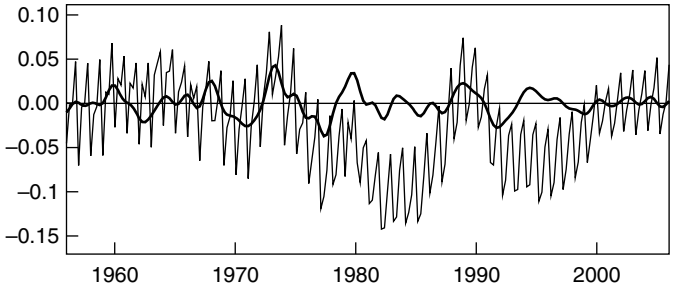


Figure 6.20 A filtered sequence obtained by applying the bandpass filter of Christiano and Fitzgerald to the logarithms of UK household expenditure

6.10.2 Flexible trends and structural breaks

Over a period of a century or so, one can expect to see occasional disturbances that disrupt the steady progress of the economy. To highlight the effects of such breaks, a firm trend function can be fitted to the data to characterize the progress of the economy broadly over the entire period. Such a trend will not be deflected by temporary disruptions, which will be seen in the residual deviations of the data from the trend.

Alternatively, it may be appropriate to absorb the breaks within the trend function. In that case, the trend will not be thrown off course for long by a break and, therefore, it should serve as a benchmark against which to measure cyclical variations when the economy resumes its normal progress. At best, the residual sequence will serve to indicate how the economy might have behaved in the absence of the break.

Numerous devices have been proposed by economists for accommodating structural breaks, which give rise to segmented trend functions. Mills (2003) has illustrated the effects of some of them by applying them to a common data sequence, which is annual UK output from 1855 to 1999. He has also provided references to an extensive literature in economics concerning structural breaks.

A common theme that unites many of the methods is their use of polynomial segments to represent the trends within sub-intervals of the data period. There is a problem of how the transition between two adjacent sub-periods should be modelled. This issue has been discussed by Granger and Teräsvirta (1993) and by Teräsvirta (1998). Others have focused on devising tests to determine the points in time when one statistical regime that describes the data should be replaced by another. Work in this area has been summarized by Perron (2006).

When a smoothing spline is used to interpolate a continuous segmented polynomial function through the data, the smoothness of the function is maintained by imposing the condition that, at the points where they join, the adjacent segments should have equal derivatives, up to some specified order.

The most common smoothing spline is that of Reinsch (1976), which is subject to the condition that the first and second derivatives of adjacent cubic segments

should be equal at the joints, which are described as the knots or the nodes. Breaks can be accommodated within such a spline by placing successive nodes in close proximity. Considerable effort has been devoted to developing algorithms that will ensure the optimal placement of the nodes. (See, for example, Luo and Wahba, 1997.)

When the abscissae of the nodes correspond to the sample dates, it is possible to increase the flexibility of the spline function by allowing local variations to occur in the smoothing parameter. The same recourse can be used to lend additional flexibility to the HP filter, which is a device that is appropriate for extracting from noisy data a trend that is generated by a discrete-time process or by a process limited in frequency to the Nyquist value.

The finite-sample version of the HP filter is provided by equation (6.124). Its generalization is provided by:

$$x = y - Q(\Lambda^{-1} + Q'Q)^{-1}Q'y, \tag{6.168}$$

where $\Lambda = \text{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_{T-3}\}$ is a diagonal matrix of smoothing parameters and Q' is the matrix of the twofold difference operator. In modifying the underlying statistical model of the HP filter, which is specified by (6.129), it is the variance σ_δ^2 of the process driving the trend that is allowed to vary, whereas the variance σ_η^2 of the process that is responsible for the errors of observation remains constant.

Setting $\Lambda^{-1} = \lambda^{-1}I$ in (6.168), which gives the smoothing parameter a globally constant value, produces the HP filter. Setting λ_t to a high value where the trend should be stiff and allowing it to take low values where the trend should be flexible will produce a device that can easily absorb structural breaks.

On the assumption that the underlying trend process is limited in frequency by the Nyquist value, it is appropriate to use the method of Fourier interpolation to create a continuous trend based on the elements of the vector x.

Example An example of a function that fails to accommodate structural breaks is provided by the polynomial of degree 4 that has been interpolated through the logarithms of 129 annual observations of the real GDP of the UK. This is shown in Figure 6.21. Figure 6.22 shows the residual sequence. In both figures, three major events can be recognized. The first is the end of World War I in 1918, which is followed by a sharp decline in GDP. The second is the recession of 1929 and the third is the end of World War II, which is also succeeded by a reduction in income. The recession has less of an impact than one might expect.

Figure 6.23 shows a trend function that has been fitted using a variable smoothing parameter. In this case, only the end-of-war breaks have been accommodated, leaving the disruptions of the 1929 recession to be expressed in the residual sequence. The effect has been achieved by attributing a greatly reduced value to the smoothing parameter in the vicinity of the post-war breaks. In the areas that are marked by shaded bands, the smoothing parameter has been given a value

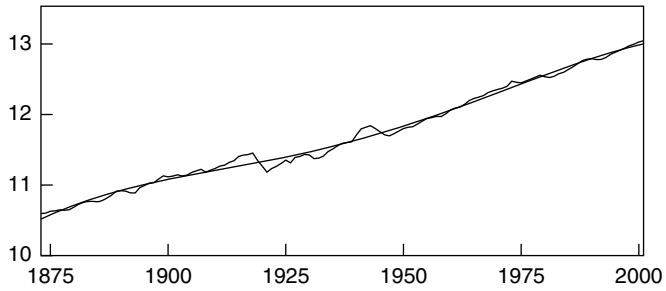


Figure 6.21 The annual series of the logarithms of real GDP in the UK, at constant prices, for the years 1873–2001. A polynomial of degree 4 has been fitted to the data by least squares regression

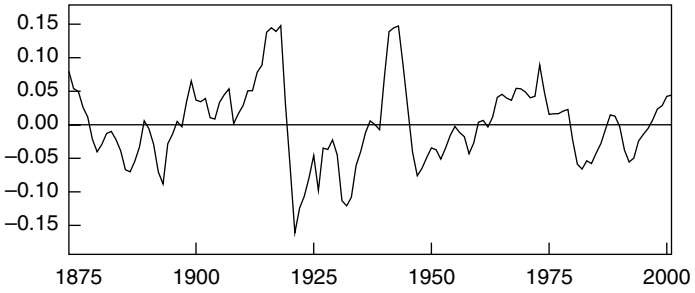


Figure 6.22 The residual obtained from fitting a polynomial of degree 4 to the logarithmic GDP data of Figure 6.21

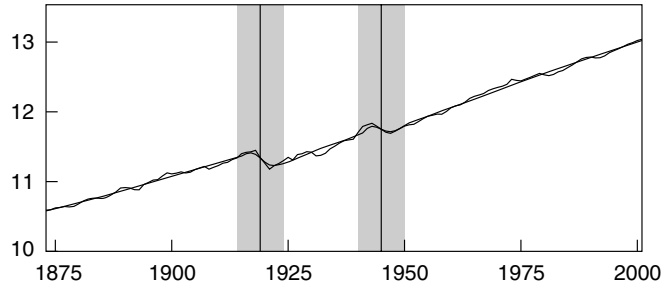


Figure 6.23 The logarithms of annual UK real GDP from 1873 to 2001 with an interpolated trend. The trend is estimated via a filter with a variable smoothing parameter

of 5. Elsewhere, it has been given a high value of 100,000, which results in trend segments that are virtually linear.

6.11 Summary and conclusions

When confronted by the wide variety of methods that are available for extracting the components of an econometric data sequence, a practitioner is liable to ask for a recommendation of the best method. In the case of business cycle analysis, there can be no unequivocal answer. The choice of an appropriate method will depend both on the nature of the data and on the purpose of the analysis. It may also depend on the aesthetic preferences of the analyst.

Nevertheless, the choice of a method ought to be made with a view to its effects in the frequency domain. Econometricians working with temporal sequences are, nowadays, paying increasing attention to the frequency aspects of their analyses, and this is where the major emphasis of the present chapter has been placed.

One of the difficulties in analyzing business cycles is that there is no unequivocal definition of what constitutes a trend. Often, a clearly defined structure that combines the trend and the cycles can be discerned within the data. An example of the successful extraction of a combination of trend and cycles that has been identified by spectral methods is provided by Figure 6.15. However, there is hardly ever a case where the data indicates a point within the frequency spectrum of this structure where the trend ends and the cycles begin.

The only unequivocal definition of the trend that might be offered is that it must have a monotonic trajectory that is devoid of cycles, which means, in practice, that it should be modeled by a polynomial of low degree. This was the practice of the generation of pioneering econometricians to which Tintner belonged.

Latterly, this approach has fallen out of favour amongst econometricians. Nowadays, they are liable to describe polynomial trends as deterministic trends, which are contrasted with stochastic trends. The latter are regarded as capable of more realistic representations of economic behavior. In particular, a stochastic trend can represent a cumulation of random events that effect the development of an economy in the course of time, in the way that the circumstances of their early lives can affect the physical statures of human beings.

Polynomial trends are an essential element within linear models of stochastic accumulation, whether they be represented in continuous time or in discrete time. Therefore, although the conceptual distinction may be a clear one, the practical distinction between a stochastic trend generated by an ARIMA process and a polynomial trend buried in noise is by no means as clear cut as, at first, it might seem to be.

The distinction becomes even more tenuous in the case of an ARIMA model that incorporates stochastic drift. Therefore, notwithstanding the recent efforts of several econometricians, it does not seem to us to be fruitful to employ statistical tests in an attempt to determine which of these alternative statistical structures actually underlies the data.

An opinion to which we adhere in this chapter is that the trend is best regarded as an analytic device, as opposed to an object that subsists within the data that might be uncovered by an appropriate technique. If the trend is to be regarded as an artificial benchmark, then its definition depends largely on what one is intending to measure.

In some cases, when the economy has had an uninterrupted progress, it is straightforward to define an appropriate benchmark. A case in point has been the UK economy over the years 1956–2005, of which the aggregate consumption is portrayed in Figures 6.4–6.6. For that period, a log-linear trend function provides a datum about which to measure the cyclical variations in consumption.

In other eras and over longer periods, where there have been substantial disruptions to the progress of the economy, the matter becomes more complicated. To highlight the major disruptions, it is appropriate to fit a polynomial of a limited degree over the entire span of the data. An example is provided by Figure 6.21. There, a fourth-degree polynomial, which adheres quite well to the data in the main, also reveals the uncommon circumstances in the periods surrounding the ends of the two world wars.

If the purpose is also to illustrate the normal workings of the economy, then it may be appropriate to fit similar polynomial trends of low degrees to the sub-periods that did not experience any disruptions. The overall result will be a segmented curve; and the issue arises of how to join the segments.

The answer that is favored in this chapter is illustrated in Figure 6.23, which shows the effect of a filter with a variable smoothing parameter. The resulting curve comprises segments that are virtually straight lines, interspersed by short segments with rapidly changing slopes.

The disjunctions that occur within the data sequence as a consequence of disruptions and breaks give rise to spectra that extend over the entire frequency range. Unless the breaks are absorbed within the trend, the residual sequence will fail to manifest the band-limited structure that we might expect to see in normal periods. Therefore, one of the criteria of a successful elimination of a break is the restoration of a band-limited spectral structure to the trend cycle component within the residual sequence.

The recognition that, at least for limited periods, the trend cycle complex is liable to be confined to a limited frequency band gives rise to further opportunities, but it also poses additional problems. The opportunities arise from the possibility of using a Fourier synthesis to create a continuous analytic function to represent the business cycle in isolation or the trend and cycle in combination.

In Figure 6.5, the business cycle has been synthesized from a limited number of the low-frequency Fourier ordinates of the linearly detrended logarithmic data. The combination of the trend and the cycle can be formed by adding the business cycle function to the linear trend of Figure 6.4. The result is shown in Figure 6.15.

The analytic nature of these functions means that they are amenable to differentiation, and their turning points are identified as the points where the first derivatives are zero-valued. This method of finding the turning points may be contrasted with the very different procedure of Bry and Boschan (1971), which had

been widely adopted by governmental statistical offices, but which often reaches doubtful conclusions.

A problem posed by band-limited processes is that they cannot easily be represented by the ARMA models that are ubiquitous in time series analysis. Such models are based on the assumption that the spectra of the processes that they represent are supported on a frequency interval that extends as far as the Nyquist frequency, which represents the limit of what is observable in sampled data.

It is often supposed that a discrete-time ARMA process is representing an underlying continuous-time process that has an unbounded frequency range. If that were the case, then the spectral density function defined over the Nyquist interval would be the product of a process of aliasing, whereby the elements of the continuous process that fall outside the Nyquist interval are attributed to frequencies that are inside.

In section 6.5, we have described a correspondence that would exist between processes that are unbounded in frequency and the discrete time models that would serve to represent them. Nevertheless, we have expressed doubts about the relevance to business cycle analysis of such unbounded processes.

In section 6.9, we have argued that processes that are limited in frequency to subintervals of the Nyquist interval, in the way that the business cycle is limited, can be resampled at a reduced rate so as to map their limited supports onto the full Nyquist interval. Thereafter, the ordinary methods of ARMA modeling can be applied to the resampled data. In that case, the Nyquist–Shannon sampling theorem indicates that there is a one-to-one correspondence between the discretely sampled process and an equivalent process in continuous time.

By these means one should be able to find an ARMA model that will capture the dynamics of the business cycle and reveal them in terms of the estimated parameters. In particular, the modulus and the arguments of the roots of the autoregressive operator should reveal the damping characteristics of the cycles and their average periods.

A modern interpretation by Baxter and King (1999) of a prescription of Burns and Mitchell (1946) is that the business cycle should be defined as a band-limited process containing cyclical elements of durations of no less than one and a half years and not exceeding eight years. This appears, at first sight, to be an unequivocal definition. However, there are difficulties in implementing it accurately. Thus, it is commonly believed that the filter that would be required to realize this definition must comprise an infinite number of coefficients, which is not practical.

In place of the infinite-order filter, a truncated approximation is commonly employed that comprises a limited number of the central coefficients. Such a filter is beset by the phenomenon of leakage, whereby the powerful low-frequency elements that would be blocked by the ideal filter find their way into the estimated business cycle. (In fact, a superior approximation is available in the form of a rational filter. See Pollock, 2003b, for example, where a rational function is employed to create a sharp lowpass filter.)

However, it has been show here that the bandpass definition can be fulfilled by selecting the appropriate ordinates of the Fourier transform of the detrended data.

The equivalent filter in the time domain is a wrapped or circular filter. Whereas such filters avoid the leakage that besets approximate bandpass filters, they deliver inappropriate estimates of the business cycle when they adhere strictly to the Baxter–King bandpass definition. Moreover, it seems that any success that the approximate bandpass filter may have in representing the business cycle must be due, in some measure, to the leakage.

The conclusion that we have reached ultimately is that, whereas it is sometimes possible to identify a trend-cycle complex within the data, there can be no definitive definition of what constitutes the trend and what, in consequence, must constitute the cyclical component. Therefore, it seems that one must be liberal in allowing any definitions that seem to fulfil their intended purposes. Even when the purpose is mistaken or unfulfilled, we should not automatically reject the resulting definition or the estimates to which it gives rise.

A Computer Program

The computer program that has been used in connection with this chapter is available at the following web address: http://www.le.ac.uk/users/dsgp1/.

References

Baxter, M. and R.G. King (1999) Measuring business cycles: approximate band-pass filters for economic time series. *Review of Economics and Statistics* **81**, 575–93.

Bergstrom, A.R. (1984) Continuous time stochastic models and issues of aggregation over time. In Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics, Volume* 2, pp. 1146–212. Amsterdam: North-Holland.

Bergstrom, A.R. (1988) The history of continuous-time econometric models. *Econometric Theory* **4**, 350–73.

Bergstrom, A.R. and K.B. Nowman (2007) *A Continuous Time Econometric Model of the United Kingdom with Stochastic Trends*. Cambridge: Cambridge University Press.

Beveridge, S. and C.R. Nelson (1981) A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* **7**, 151–72.

Bloomfield, P. (1976) *Fourier Analysis of Time Series: An Introduction*. Chichester: John Wiley and Sons.

Box, G.E.P. and G.M. Jenkins (1976) *Time Series Analysis: Forecasting and Control* (revised edition). San Francisco: Holden Day.

Bry, G. and C. Boschan, (1971) *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. New York: National Bureau of Economic Research.

Burman, J.P. (1980) Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society, Series A* **143**, 321–37.

Burns, A.M. and W.C. Mitchell (1946) *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Caporello, G. and A. Maravall (2004) *Program TSW: Revised Reference Manual*. Working Paper 0408, Servicio de Estudios, Banco de España.

Christiano, L.J. and T.J. Fitzgerald (2003) The band-pass filter. *International Economic Review*. **44**, 435–65.

Fuller, W.A. (1976) *Introduction to Statistical Time Series*. New York: John Wiley and Sons.

Gómez, V. (2001) The use of butterworth filters for trend and cycle estimation in economic time series. *Journal of Business and Economic Statistics* **19**, 365–73.

Gómez V. and A. Maravall (1996) *Programs TRAMO and SEATS – Instructions for the User* (with some updates). Working paper 9628, Servicio de Estudios, Banco de España.
Granger C.W.T. and T. Teräsvirta (1993) *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
Gray, R.M. (2002) *Toeplitz and Circulant Matrices: A Review.* Information Systems Laboratory, Department of Electrical Engineering, Stanford University, California, http://ee.stanford.edu/gray/˜toeplitz.pdf.
Harding, D. and A. Pagan (2002) Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics* **49**, 365–81.
Harvey, A.C. (1985) Trends and cycles in macroeconomic time series. *Journal of Business and Economic Statistics* **3**, 216–28.
Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.
Harvey, D.I. and T. Mills (2003) Modelling trends in central england temperatures. *Journal of Forecasting* **22**, 35–47.
Hillmer, S.C. and G.C. Tiao (1982) An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77**, 63–70.
Hodrick, R.J. and E.C. Prescott (1980) *Postwar U.S. Business Cycles: An Empirical Investigation*. Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania.
Hodrick R.J. and E.C. Prescott (1997) Postwar U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking* **29**, 1–16.
Kolmogorov, A.N. (1941) Interpolation and extrapolation. *Bulletin de l'academie des sciences de U.S.S.R.,* Ser. Math. **5**, 3–14.
Koopman, S.J., A.C. Harvey, J.A. Doornick and N. Shephard (2007) *STAMP 8.0: Structural Time Series Analyser Modeller and Predictor: The Manual*. London: Timberlake Consultants Press.
Leser, C.E.V. (1961) A simple method of trend construction. *Journal of the Royal Statistical Society, Series B* **23**, 91–107.
Luo. Z. and Grace Wahba (1997) Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–16.
Maravall, A. and D.A. Pierce (1987) A prototypical seasonal adjustment model. *Journal of Time Series Analysis* **8**, 177–93.
Mills, T.C. (2003) *Modelling Trends and Cycles in Economic Time Series*. Basingstoke: Palgrave Macmillan.
Pandit, S.M. and S.M. Wu (1975) Unique estimates of the parameters of a continuous stationary stochastic process, *Biometrika* **62**, 497–501.
Pagan, A. (1997) Towards an understanding of some business cycle characteristics. *Australian Economic Review* **30**, 1–15.
Pedregal, D.J., C.J. Taylor and P.C. Young (2004) Lancaster: Centre for Research on Environmental Systems and Statistics (CRES), Lancaster University.
Perron. P. (2006) Dealing with structural breaks. in T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory,* pp. 278–352. Basingstoke: Palgrave Macmillan.
Phadke, M.S. and S.M. Wu (1974) Modelling of continuous stochastic processes from discrete observations with applications to sunspot data. *Journal of the American Statistical Association* **69**, 325–9.
Pollock, D.S.G. (1999) *A Handbook of Time-Series Analysis, Signal Processing and Dynamics*. London: Academic Press.
Pollock, D.S.G. (2000) Trend estimation and de-trending via rational square wave filters. *Journal of Econometrics* **99**, 317–34.
Pollock, D.S.G. (2002a) Circulant matrices and time-series analysis. *International Journal of Mathematical Education in Science and Technology* **33**, 213–30.
Pollock, D.S.G (2002b) A review of TSW: the Windows version of the TRAMO-SEATS program. *Journal of Applied Econometrics* **17**, 291–9.

Pollock, D.S.G. (2003a) Sharp filters for short sequences. *Journal of Statistical Inference and Planning* **113**, 663–83.

Pollock, D.S.G. (2003b) Recursive estimation in econometrics. *Journal of Computational Statistics and Data Analysis* **44**, 37–75.

Pollock, D.S.G. (2006) Wiener–Kolmogorov filtering, frequency-selective filtering and polynomial regression. *Econometric Theory* **23**, 71–83.

Priestley, M.B., (1989), *Spectral Analysis and Time Series*. London: Academic Press.

Reinsch, C.H. (1976) Smoothing by spline functions. *Numerische Mathematik* **10**, 177–83.

Teräsvirta, T. (1998) Modelling nonlinear economic relationships with smooth transitions. in A. Ullah and D.E.A. Giles (eds.), *Handbook of Applied Economic Statistics,* pp. 507–52. New York: Marcel Dekker.

Tintner, G. (1940) *The Variate Difference Method*. Bloomington, Ind.: John Wiley and Sons.

Tintner, G. (1953) *Econometrics*. New York: John Wiley and Sons.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B* **40**, 364–72.

Whittaker, E.T. (1923) On a new method of graduation. *Proceedings of the Royal Society of Edinburgh,* **44**, 77–83.

Whittle, P. (1983) *Prediction and Regulation by Linear Least-Square Methods* (second revised edition). Oxford: Basil Blackwell.

Wiener, N. (1941) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Report on the Services Research Project DIC-6037. Published in book form in 1949 by MIT Technology Press and John Wiley and Sons, New York.

7

Economic Cycles: Asymmetries, Persistence, and Synchronization

Joe Cardinale and Larry W. Taylor

Abstract

Marking upswings and downswings for a time series $\{y_t\}$ provides insights that are not immediately obvious, but may be meaningful to academics, policy makers, and the general public. The mean and standard deviation of durations, as well as the amplitude and steepness of a given phase, yield fruitful insights about cycle asymmetries and persistence. Expansions and contractions in one series can then be compared to those in another to determine whether their respective cycles are synchronized. However, our primary focus here is on classical nonparametric methods for the analysis of duration, dating back to the seminal work of Burns and Mitchell (1946), Cutler and Ederer (1958), Bry and Boschan (1971), and Cox (1972). Although our specific application is to unemployment cycles, the ideas and techniques discussed in this chapter apply to a wide variety of micro- and macroeconometric studies.

7.1 Introduction

An event history is a longitudinal record of specified events. Examples at the individual level include dates of schooling, marriage, birth of children, job change, illness, promotion, retirement and migration. Examples at the aggregate level include dates of militarized disputes, riots, revolution, economic expansions and contractions, bull and bear markets, and massive layoffs. Regardless of the underlying activity, an event consists of a qualitative change that marks a new phase for either the individual or the collective. Consider that the beginning of a bull market marks the end of a bear market and that the beginning of a militarized interstate dispute marks the end of placid relations between states with diplomatic ties.

A pertinent issue is whether the probability of exiting a phase, or *state*, depends on its duration. For instance, is a four-hour riot more likely to end within the next hour than a one-hour riot? Does the likelihood of changing jobs decrease with the time invested in the current job? Is a young economic expansion more robust to failure than an old one? If so, we anticipate the end of an old expansion, but are surprised by the end of a young one.

The demographer sees each of the above questions answered best by a life table analysis, the biostatistician by a survival analysis, and the engineer by a reliability or failure-time analysis. The economist tends to view the individual-specific questions on job change as somehow different from the aggregate-level question on business cycles, with the question on riots fitting perhaps somewhere in between. In fact, these questions are generally handled by various sub-disciplines, as the microeconometrician examines individual employment while the macroeconometrician examines the overall business cycle. Despite the initial segregation, however, it is now understood that life table, survival, and reliability analyses are intellectually very similar, and that the analysis of business cycles can be handled in much the same way as that involving promotion and job change.[1]

7.2 Marking time

One important distinction between microeconometric and macroeconometric event studies lies in the complexity of marking time. For individual level or microeconometric events such as marriage, job change, promotion, birth and death, it is relatively easy to pinpoint when the qualitative change occurs. In contrast, for national unemployment, economic expansions and contractions, and bull and bear markets, the timing of change is less certain. Consider the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER). The committee is comprised of experts who mark time – that is, mark the *turning points* – in economic activity by consensus. They examine trends in real gross domestic product (GDP), real income, employment, industrial production and wholesale–retail sales, in conjunction with their collective reasoning that an economic contraction, or *recession*, must last more than a few months.[2]

The NBER derives its method from the graphically oriented approach of Burns and Mitchell (1946), who define the *classical cycle*. The graphical approach is based on marking turning points for several specific cycles, then aggregating that information to form a reference cycle. In formalizing the graphical approach, however, Harding and Pagan (2002) observe that Burns and Mitchell would have *preferred* to use a single series, namely GDP, had that been available to them. The single series approach seems especially appropriate for low frequency data.

Harding and Pagan (2002) thus employ a modified version of the Bry and Boschan (BB) (1971) algorithm to mark turning points for quarterly GDP observations. Their BBQ algorithm again leads to the so-called classical cycle, though defined in a more rigorous, non-qualitative manner.[3] Harding and Pagan (2006) use a similar nonparametric approach to codify the methods for deriving the NBER reference cycle. Regardless of whether single or multiple series are employed to mark time, Harding and Pagan observe that using algorithms is largely immune from deleterious compositional effects of dating committees, such as the NBER Dating Committee. Consider that Artis, Marcellino and Proietti (2004) employ a modified version of BBQ for use in their study of European business cycles.

Parametric models complement the nonparametric analysis. Pagan (1997) examines some simple linear statistical models and shows they are capable of replicating the observed phase *durations* of classical cycles in Australia, the United Kingdom and the United States. He demonstrates that realistic linear statistical models of national output have (i) deterministic trend growth, but of low magnitude, if any; (ii) near-unit-root behavior in the deterministically detrended data, if not exactly unit-root behavior; and (iii) innovations of a certain magnitude. Economic models with final equations for output that match these specifications are observationally equivalent, and this explains why King and Plosser (1994, p. 436) find it difficult to distinguish between the Klein–Goldberger model and a neoclassical real business cycle model. Despite potential identification problems, however, it is only by melding quantitative analysis with economic theory that one can hope to distill the relative importance of monetary, real, expectational, and international shocks. To do so, Harding and Pagan (2000) emphasize the production of

statistics that directly address the ability of a model to replicate business cycle characteristics.

7.2.1 Reasons for marking time

Let $y_t = \log(GDP_t)$, and consider mapping $\{y_t\}$ to a binary time series, $\{S_t\}$, with $S_t = 0$ for recessions and $S_t = 1$ for expansions. Pagan (2004) and Harding and Pagan (2007) offer the following rationale for the reader's interest in $\{S_t\}$:

1. *S_t frequently emphasizes features of y_t that are not immediately obvious.* Observing the behavior of y_t over different phases affords us a better understanding of its features.
2. *S_t may be more meaningful to decision makers than y_t.* Reaction to an economic downturn is often strong among the electorate, and it is of interest to determine whether the probability of exiting a downturn depends on how long one has been in it and whether such exit probabilities, or *hazards*, have changed in fundamental ways in recent history.
3. *S_t may be the object of interest if questions are asked about the synchronization of cycles across sectors or countries.* If S_t is derived from many underlying series, it is often more convenient to compare the representative S_t values than to compute a large number of correlations from the underlying series.
4. *S_t is generally more robust than y_t to relatively unimportant short-lived shocks.* Such shocks may substantially affect statistics based on GDP growth rates but have little impact on overall trends. In contrast, S_t emphasizes the qualitative trend, up or down.

7.2.2 Techniques for marking time

In short, mapping $\{y_t\}$ to $\{S_t\}$ yields fruitful insights about cycle asymmetries, persistence, and synchronization. Some nonparametric rules for marking time are exceptionally simple. For yearly aggregate data, Neftci (1984) and Cashin and McDermott (2002) employ the calculus rule, $S_t = 1(\Delta y_t > 0)$. For quarterly data, one often-used rule in the popular press is the extended Okun rule. The rule states that, for a recessionary phase, termination is signified by two successive quarters of positive growth, $(\Delta y_{t+1} > 0, \Delta y_{t+2} > 0)$. Similarly, for an expansionary phase, termination is signified by two successive quarters of negative growth, $(\Delta y_{t+1} < 0, \Delta y_{t+2} < 0)$. The simplicity of such rules is very attractive, and any limitations of such rules are quickly revealed by visual inspection of the observed series. In fact, regardless of the rule employed, the constructed turning points should visually coincide with those apparent in a plot of the observed time series, $\{y_t\}$.

7.2.2.1 BBQ

To locate turning points in the *level* of GDP, the BBQ algorithm first determines a potential set of local peaks and troughs. Time t is a local peak if:

$$(y_t - y_{t-2} > 0, y_t - y_{t-1} > 0, y_t - y_{t+1} > 0, y_t - y_{t+2} > 0),$$

with the inequality reversed for troughs. The algorithm ensures that peaks and troughs alternate, so that an expansion is immediately followed by a contraction, and vice versa. Finally, the algorithm considers combining phases, or creating new phases, according to a set of predetermined rules. For instance, a censoring rule for business cycles is that either a contraction or expansion must last a minimum of two quarters and complete cycles must last a minimum of five quarters.

BBQ can also mark turning points for other types of series. For example, to locate a potential peak for the *growth* cycle in GDP, replace the requirement that $y_t - y_{t-2} > 0$ with the requirement that $\Delta y_t - \Delta y_{t-2} > 0$, and so on. Pagan and Sossounov (2003) modify BBQ to factor in the *magnitude* of growth rates in financial series. Other applications of nonparametric methods to mark the turning points include Lunde and Timmermann (2004) and Ohn, Taylor and Pagan (2004), who investigate bull and bear markets; Cashin, McDermott and Scott (2002), who investigate booms and slumps in commodity markets; Eichengreen, Rose and Wyplosz (1995), who investigate exchange rate crises; and Ibbotson, Sindelar and Ritter (1994), who examine hot and cold IPO markets.

7.2.2.2 Markov chain models

Hamilton's (1989) innovative Markov chain switching-regime model can also be employed to mark time. For the parametric switching-regime model, a latent random variable, s_t^*, governs the state or *regime* with, say, $s_t^* = 0$ indicating low or negative average growth, and $s_t^* = 1$ indicating high or positive average growth. Two states, signifying negative and positive average growth rates, are adequate to mark the turning points since $\Delta y_t < 0$ indicates a downswing and $\Delta y_t > 0$ indicates an upswing in the level of GDP.

Consider a simple latent-structure model for GDP:

$$\Delta y_t - \mu_{s_t^*} = \phi(\Delta y_{t-1} - \mu_{s_{t-1}^*}) + \varepsilon_t. \tag{7.1}$$

The mean of the growth-rate process switches between "low" and "high," with $\mu_0 < \mu_1$. For each date t in the sample, Hamilton shows how to obtain an estimate of $P(s_t^* = 0|F_T)$, where F_T contains the past, or even the *complete*, sample history of growth rates, $\{\Delta y_t\}_{t=1,\ldots,T}$.

Define $S_t^* = 1[0.5 - P(s_t^* = 0|F_T)]$, so that $S_t^* = 1$ during projected high-growth phases, and $S_t^* = 0$ during low-growth phases. The observed binary series, $\{S_t^*\}$, can be used in a survival, or *duration*, analysis to determine whether the probability of remaining in a given phase, either contraction or expansion, depends on how long one has been in it. An important consideration is that Hamilton's (1989) model is such that s_t^* evolves with the probability of remaining in a given phase independent of its duration, so that contractions and expansions are assumed to be *duration independent*. Of course, this makes $\{S_t^*\}$ less than ideal to represent phases that may actually be *duration dependent*.

Although Durland and McCurdy (1994), Filardo (1994), Diebold, Lee and Weinbach (1994), Macheu and McCurdy (2000) and Jensen and Liu (2006) generalize Hamilton's assumptions in various directions, no parametric model matches

the flexibility and transparency of BBQ to mark time. In particular, for the purpose of solely *marking* the turning points in an observed series such as GDP, it is unnecessary to consider a latent-structure model with or without covariates that helps *predict* the turning points. That is, there is no need to proxy $\{S_t\}$ with $\{S_t^*\}$ for the purpose of a duration analysis.

This is not to say, however, that Markov-switching (MS) models describing fluctuations in y_t are uninformative. In fact, MS models are alternatives to the linear models emphasized by Pagan (1997).[4] Hamilton (2005) argues that linear models are incapable of replicating the cyclical pattern in key economic aggregates, and he devises a simple nonlinear model for unemployment. In particular, linear models cannot capture the fact that the unemployment rate rises more quickly than it falls over the business cycle. Although technology, the labor force, and the capital stock are all key determinants of long-run growth, the forces that contribute to a business downturn can be quite different, and they typically introduce asymmetric behavior that necessitates a nonlinear dynamic representation. Harding and Pagan (2002) also note the deficiency of linear models for replicating the *shapes* of expansions in the business cycle. The point that is often lost is that a duration analysis complements the empirical results from either linear or nonlinear models of y_t; and for the purpose of a duration analysis, it is unnecessary to specify the model for y_t.

7.2.3 Detrending the series

Likewise, there is generally no need to detrend the series to obtain $\{S_t\}$. Cooley and Prescott (1995) first remove the trend prior to marking the turning points. The trend is typically thought of as a permanent effect, and the remainder as a temporary effect. Unfortunately, confusion is likely to ensue when one attempts to separate permanent from temporary effects, because not all temporary components measure the same thing.

For example, consider decomposing aggregate output so that $y_t = P_t + z_t$, where y_t is the logarithm of GDP, P_t is the permanent effect and z_t is the temporary effect. The permanent effect captures slow-moving low-frequency movements in y_t, and the temporary effect captures the faster-moving high-frequency movements. The term P_t is typically an integrated or I(1) stochastic series, but it can just as easily be defined as some type of deterministic trend. The interpretation of z_t, either as an output gap or some function of growth rates, depends on how P_t is defined.

7.2.3.1 Output gaps versus growth rates

Consider first defining the permanent component as the deterministic trend, $P_t = a + bt$. The temporary effect is the output gap, $z_t = y_t - a - bt$, and the time trend captures steady increases in capital and labor that feed into the aggregate production function. In other words, the output gap defines the difference in actual and potential GDP. Marking time by the sign of z_t determines phases of output above or below the trend. On the other hand, if we define the perma-

nent component as $P_t = y_{t-1}$, the temporary effect is now the growth rate since $z_t = \Delta y_t = y_t - y_{t-1}$. Marking time by examining positive and negative values of z_t defines the classical business cycle.

We make a distinction here between growth cycles and gap or *deviation* cycles. For growth cycles, we seek turning points in Δy_t; there is no reason to specify a trend curve. For gap cycles, however, we seek deviations from a specified trend curve. In contrast, Zarnowitz and Ozyildirim (2006), among others, classify a gap cycle as a special type of growth cycle. For their gap analysis, Zarnowitz and Ozyildirim recommend that the trend curve be determined by the classical nonparametric phase-average-trend (PAT) algorithm of Boschan and Ebanks (1978). Zarnowitz and Ozyildirim then argue that a gap analysis is more informative than a direct growth-rate analysis in the study of national output. First, they argue that growth rates over short time spans are very erratic and must be smoothed with complex moving averages that potentially distort patterns. Second, they find that the timing of growth rates is very different from that of the corresponding level series. Of course, an alternative interpretation of the second finding is that the growth cycle is providing different information than is the classical business cycle.

7.2.3.2 Filtering procedures

The specification of any trend curve is somewhat arbitrary. However, Zarnowitz and Ozyildirim (2006) find that the PAT algorithm produces a nonlinear trend curve that smoothly transits from higher to lower growth. They also find that the trend from the PAT algorithm fits as well as a log-linear trend, the stochastic Beveridge and Nelson (1981) trend, the local linear trend of Harvey (1989), the Hodrick–Prescott (1997) trend, and Rotemberg's (1999) heuristic trend.

Intuitively, different types of trends produce different types of temporary components. For instance, King and Rebelo (1993) show that the temporary component from the Hodrick–Prescott (HP) trend is a two-sided weighted average of growth rates. However, if the data-generating process is the pure random walk, $y_t = y_{t-1} + \epsilon_t$, Harding and Pagan (2005) show that the HP temporary component is well represented by a weighted average of current and lagged values of the growth rates with slowly declining weights. In contrast, the Beveridge–Nelson temporary component for the random walk is degenerate since the permanent component is y_t.

The point is this. It is easy to think that all temporary components are measuring the same thing as long as each is a stationary process; however, this is not the case. For business cycles, the litmus test appears to be whether the turning points match well with those of the NBER. Consider that, for quarterly observations on GDP, Hamilton (1989) compares his estimated latent-structure probabilities with turning points in aggregate activity. He demonstrates that his estimate of $P(s_t^* = 0|F_T)$ is generally greater than 0.5 during recessions and less than 0.5 during expansions. On the other hand, the very flexible BBQ algorithm of Harding and Pagan (2002) also yields turning points that accord well with those of the NBER.[5]

7.3 The discrete-time hazard function

Duration data for economic cycles are invariably discrete since data are collected at discrete intervals of time, for example, weekly, monthly, quarterly, or yearly. Although data for markets such as housing or financial markets are available at short intervals, the more interesting questions about such cycles are typically best captured by intervals of at least a month. Consider also that a discrete-time duration analysis can be viewed as an approximation to a continuous-time analysis, or vice versa, and general notions about one apply to the other.[6] A discrete-time framework has the advantage of being more natural for the types of data encountered at the aggregate level; the framework is inherently semiparametric and is easy to understand and implement. Since formal statistical inference depends on the framework, it is best to adopt that of discrete-time if data are measured at long intervals.

Consider a random sub-sample of n observations $(T_1, T_2, \ldots, T_n)$ from a discrete, cumulative distribution F, such that $F(a) = 0$ for $a < 0$.[7] The probability-distribution function, or *density function*, is $f(t) = P(T = t)$, and the discrete-time *hazard function* is:

$$h(t) = P(T = t | T \geq t) = f(t)/G(t), \tag{7.2}$$

where $h(t)$ is the hazard function, and $G(t) = P(T \geq t)$ is the *survival function*. The density function, $f(t)$, gives the probability that a duration will last *exactly* t periods, the survival function, $G(t)$, gives the probability that a duration will last *at least* t periods, and the hazard function, $h(t)$, gives the conditional probability that a phase will terminate in period t, given that it has lasted t *or more* periods. If rising or falling, the hazard provides useful information about the likelihood of a change in phase. Using over 100 years of annual data, Mills (2001) finds several instances of non-constant hazards in the business cycles of 22 countries.

The hazard may also be useful in the assessment of general market conditions. For instance, Diebold and Rudebusch (1990) and Ohn, Taylor and Pagan (2004) observe that post-World War II contractions are more prone to revert to expansion than pre-World War II contractions. One explanation for this finding is that policy makers are now much better able to manage potential economic crises. A second explanation is that individuals and firms are better able to smooth shocks due to innovation and financial deregulation. On the other hand, Watson (1994) finds that, for most individual sectors of the economy, the average contraction and expansion durations for the pre-war and post-war periods are similar; and, more recently, Stock and Watson (2003) suggest that favorable market conditions in the modern era are more likely due to good luck than to good management or innovation.[8] This is especially true for recent times as there have been relatively few long-lived supply disruptions since the 1970s.

7.3.1 Hazard plots

It is generally informative to plot the hazard function, with hazard rates easily computed by the nonparametric life table method of Cutler and Ederer (1958). The

computer package LIMDEP 7.0 constructs such life tables, with an approximate, but intuitive, explanation of the procedure as follows:

- Place the contractions in ascending order by length.
- Construct the hazard rate at time "t" as the ratio of the number of contractions terminating in month "t" over the number of contractions lasting at least "t" months.

Sample information is thus used to estimate $P(T = t)/P(T \geq t)$ in a straightforward manner. A contraction that terminates is said to *exit* the sample, while those contractions lasting at least t months are said to be still *at risk*. Of course, the pool of contractions still *at risk* decreases with t. This implies that the effective sample size for estimating the hazard rates for relatively long contractions is less than for short contractions. Formal statistical tests are thus necessary to avoid spurious conclusions from inspecting the graphs alone.

7.3.2 Benchmark hazards

Nevertheless, graphs convey important information about the general shape of the hazard function. Hollander and Proschan (1975) and Hollander and Wolfe (1999) provide some important benchmarks:

1. Constant Failure Rate: CFR
2. Increasing (Decreasing) Failure Rate: IFR, DFR
3. Increasing (Decreasing) Failure Rate on Average: IFRA, DFRA
4. New Better (Worse) than Used: NBU, NWU.

Figure 7.1 depicts hazard functions from CFR, IFR, IFRA, and NBU *life distributions*.[9] Because there is a one-to-one relationship between the hazard function and the probability density function, a comparison of hazard rates is a natural way of analyzing the nature of exit probabilities, more so than a comparison of density functions. In particular, the CFR hazard is almost always given special consideration in any duration analysis.

CFR hazards correspond to the geometric density. New economic expansions are no more or less likely to terminate than mature ones. In contrast, if expansions are IFR, the hazard, or *failure,* rate is never decreasing, and our illustrated IFR hazard implies an ever more likely chance of termination, or *mortality*.

IFR hazards are not the only ones that have a tendency to rise. Although the depicted IFRA hazard has periods of decline, IFRA has the same overall upward trend. For example, militarized interstate disputes initially have a decreasing hazard rate for a short period, but then exhibit increasing hazards over most of the duration due, perhaps, to a more concentrated effort to either negotiate or impose settlement.

In contrast to IFRA, the depicted NBU hazard function rises above and then falls back to the initial value. There is no overall trend in either direction. Because the hazard function never falls below the initial value, new phases have the greatest chance of surviving an *additional* week or month. An NBU hazard may arise due to

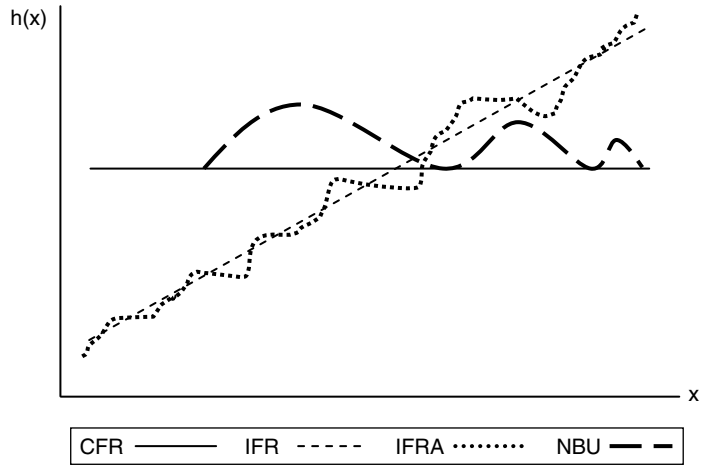


Figure 7.1 Hazard functions for various life distributions

seasonal effects such as holidays or inclement weather, and IFRA hazards are special cases of NBU hazards. A completely analogous situation holds for DFR, DFRA and NWU distributions.

The graphical approach of Cutler and Ederer (1958) is nonparametric and avoids some of the dangers of relying too heavily on parametric methods. For instance, in the emerging area of forensic economics, Bonanomi, Gaughan and Taylor (1998) use the flexible nonparametric life table method in the estimation of lost profits when the plaintiff claims lost customers due to an alleged transgression. However, in the general economics literature, Sichel (1991) advocates using parametric methods to increase the power of tests for duration dependence. In the political science literature, Bennett (1999) and Zorn (2000) observe that the parametric Weibull model is the most widely-used form. Bennett also cautions that Cox's semiparametric proportional hazards model does not allow for *precision* concerning the hazard function.

Precisely wrong results, however, are hardly helpful. Ohn, Taylor and Pagan (2004) show that the Weibull model is insufficient to capture the richness of economic contractions and expansions, and Taylor (2007) shows that the Weibull model is highly misleading for militarized interstate disputes. Attempts to apply more flexible continuous-time methods have met with mixed success. Diebold, Rudebusch and Sichel (1993), for example, apply their nonlinear exponential linear model to business cycle data that mostly duplicates the results from the Weibull model. Zuehlke (2003) applies the nonlinear model of Mudholkar, Srivastava and Kollia (1996) that allows hazards to be monotonically increasing, monotonically decreasing, U-shaped or inverted U-shaped. However, Ohn, Taylor and Pagan (2004) observe that none of these shapes adequately describes the hazard functions for pre-World War II constructions and expansions. Even for the small samples encountered in the business cycle literature, nonparametric analysis provides valuable insights that parametric analysis fails to uncover.

7.4 Testing for duration dependence

The direction of duration dependence is easily obtained from a sample of durations. First, if the mean duration equals the sample standard deviation, there is no evidence of duration dependence; that is, there is *a constant hazard*. Second, if the mean duration is greater than the sample standard deviation, there is evidence of *positive* duration dependence, or a generally increasing hazard. Finally, if the mean duration is less than the sample standard deviation, there is evidence of *negative* duration dependence, or a generally decreasing hazard.

7.4.1 The nature of duration independence

Duration independence is considered the neutral case. Long expansions have no greater chance of ending than short expansions; long bear markets have no greater chance of ending than short ones; and long housing slumps have no greater chance of ending than short slumps. The duration of the phase has no predictive power in determining the end of the phase. Because of neutrality, the constant hazard is the standard benchmark, and a graph of the hazard function is frequently employed to see if the hazard function appears roughly constant. If so, there is *duration independence,* and the hazard function does not depend on t. In other words, the null hypothesis is:

$$H_0 : h(t) = p \quad \text{for some } 0 < p < 1 \text{ and all } t > 0. \tag{7.3}$$

The density *must* be geometric for constant hazards, and the above null hypothesis is equivalent to:

$$H_0 : f(t) = P(T = t) = (1-p)^t p \text{ for } 0 \leq t \leq \infty. \tag{7.4}$$

In other words, testing for duration independence is equivalent to testing whether the durations follow the geometric density. A *direct*, or *strong-form*, test for the geometric density is the usual chi-square goodness-of-fit test employed by Ohn, Taylor and Pagan (2004).

Finally, for the geometric density, $E(T) = (1-p)/p$ and $V(T) = (1-p)/p^2$, leading to a third null hypothesis for duration dependence:

$$H_0 : V(T) - [E(T)]^2 - E(T) = 0. \tag{7.5}$$

7.4.2 Weak-form tests

Pagan (1998) and Mudambi and Taylor (1991, 1995) devise tests for duration dependence based on the *consistency relationship* as in (7.5). Such tests are called *weak-form* tests and have close links with continuous-time tests for the *exponential* density. Although the exponential density is simply the continuous-time equivalent of the discrete-time geometric density, there is one rather important difference between the two. The consistency relationship for the exponential density is $V(T) - [E(T)]^2 = 0$ rather than $V(T) - [E(T)]^2 - E(T) = 0$, and thus the choice between the discrete-time and continuous-time tests is important except when p is close to either 0 or 1, since then $V(T) \approx [E(T)]^2$.

7.4.2.1 The GMD test

Mudambi and Taylor (1995) devise a generalized method of moments (GMM) estimator based on the moment condition, $V(T) - [E(T)]^2 - E(T) - \gamma = 0$. For the geometric density, $\gamma = 0$, and they determine whether GMD = $(1/n) \sum_{i=1}^{n} (T_i - \overline{T})^2 - \overline{T}^2 - \overline{T}$ is statistically significant from zero. Once normalized, GMD is asymptotically $N(0, 1)$, and should be especially sensitive to IFRA and DFRA alternatives since it is completely analogous to a continuous-time test based on $V(T) - [E(T)]^2 = 0$ that is designed for such alternatives. Because GMD is highly skewed in finite samples, however, simulations are necessary to obtain finite-sample critical values.

7.4.2.2 The SB test

Closely related to the GMD test is the *state-based* SB test proposed by Pagan (1998), which has a number of positive attributes:

1. SB focuses directly on the conditional probabilities.
2. SB involves a regression and so is easy to explain to a nonspecialist.
3. SB can be used to examine prediction issues.
4. SB can be used to study how the exit probabilities have changed over time since the parameters can be recursively estimated.
5. SB can be easily modified to estimate discrete-time hazard functions, with or without covariates.

Although SB can be applied to whole cycles, it is best to apply the test to the separate half cycles, or phases. In other words, one should apply SB first to contractions and then to expansions. In fact, Mudambi and Taylor (1991) show that it is incompatible for expansions, contractions, and whole cycles *all* to follow a constant-hazard geometric distribution. So, if expansions and contractions are duration independent, it is certain that whole cycles are duration dependent. Likewise, if whole cycles are duration independent, there is necessarily some form of statistical dependence in the half-cycle components.

Consider a small sample of *contractions* observed at monthly intervals. In the example below, there are gaps in the time line because most months of expansion are excluded from the sample. In fact, the only included months of expansion are those that correspond to a turning point.

	Jan	Feb	Mar	Apr	...	Sep	Oct	Nov	Dec	Jan	Feb	...	Jun	Jul
S_t	0	0	0	1	...	0	0	0	0	0	1	...	0	0
d_{t-1}	0	1	2	3	...	0	1	2	3	4	5	...	0	1

$S_t = 0$ indicates a month of economic contraction, and $S_t = 1$ indicates a month of economic expansion. For instance, the first complete contraction began in January and ended in March, with April as the turning point. The string of S_t values represents two complete contractions and one incomplete, or *censored,* contraction. The durations of the contractions are $T_1 = 3, T_2 = 5$, and $T_3 = 2$. Any censored observations will invariably be at the beginnning and/or end of the string, and

such incomplete observations are dropped from the sample. For the sub-sample of expansions, we consider instead $1-S_t$, so that $(1-S_t)=1$ marks the turning point of an economic expansion.

An economic expansion or contraction is a *phase* of the business cycle. Define d_t as the number of months in a given phase. The above table demonstrates values for lagged d_t. Now drop those observations from the sample if $d_{t-1}=0$, and define m as the number of remaining S_t values. For our example, $m=8$ since we drop two observations because $d_{t-1}=0$ and we drop the last contraction due to censoring. A straightforward test for duration dependence in contractions is obtained by testing the null hypothesis, $H_0:\beta_1=0$, in the simple regression equation:

$$S_t=\beta_0+\beta_1 d_{t-1}+\epsilon_t, \tag{7.6}$$

where $E_{t-1}(\epsilon_t)=0$.

For constant hazards, $\beta_1=0$ and $\beta_0=p_{01}$, where $p_{01}=P(S_t=1|S_{t-1}=0)$. The term d_{t-1} captures autonomous changes in the hazard function. The resulting model can be written as:

$$S_t=p_{01}+\nu_t. \tag{7.7}$$

Hamilton (1994, p. 684) shows how to write such an equation if one is considering complete cycles rather than half-cycles. For the purpose of a duration analysis, however, it is only necessary to consider half-cycles: expansions or contractions, bull or bear markets, upswings or downswings.

For non-constant hazards, $\beta_1\neq 0$, and the termination probability depends on d_{t-1}, the length of time in the specified phase. Ohn, Taylor and Pagan (2004) show that the SB t-test is appropriate for testing $H_0:\beta_1=0$. We argue here the same point, but from a different approach. In particular, by the definition of the binary indicator, S_t, the duration of an expansion must be at least one month. The geometric density is thus left-censored at unity. The censored probability function is $P(T=t)=(1-p)^{t-1}p$ for $1\leq t\leq\infty$, with $E(T)=1/p$ and $V(T)=(1-p)/p^2$. Let n be the number of turning points. Then $\overline{T}=(1/n)\sum_{i=1}^{n}T_i$ is a consistent estimator for $E(T)$, $\widetilde{\sigma}^2=(1/n)\sum_{i=1}^{n}(T_i-\overline{T})^2$ is a consistent estimator for $V(T)$, and $\overline{S}$ is a consistent estimator for p.

The least squares estimator of β_1 is:[10]

$$\widehat{\beta}_1=\frac{\frac{1}{m}\sum_{t=1}^{m}(S_t-\overline{S})d_{t-1}}{\frac{1}{m}\sum_{t=1}^{m}(d_{t-1}-\overline{d})^2}, \tag{7.8}$$

where $\overline{S}=\frac{1}{m}\sum_{t=1}^{m}S_t=n/m$ and $\overline{d}=\frac{1}{m}\sum_{t=1}^{m}d_{t-1}$.

Our goal is to show that $p\lim\widehat{\beta}_1=0$ under the null hypothesis that durations follow the geometric density. To do so, consider the numerator of the least

squares estimator:

$$\frac{1}{m}\sum_{t=1}^{m}(S_t-\overline{S})d_{t-1} \tag{7.9}$$

$$=\frac{1}{m}\sum_{t=1}^{m}S_t d_{t-1}-\overline{S}\,\overline{d} \tag{7.10}$$

$$=\frac{n}{m}\overline{T}-(\frac{n}{m})^2\frac{1}{n}\sum_{i=1}^{n}\frac{(T_i+1)T_i}{2} \tag{7.11}$$

$$=\overline{S}\frac{1}{2}\{2\overline{T}-\overline{S}[\widetilde{\sigma}^2+\overline{T}^2+\overline{T}]\}. \tag{7.12}$$

Since, for the geometric distribution, $p\lim\overline{T}=1/p$ and $p\lim\widetilde{\sigma}^2=(1-p)/p^2$, it follows that:

$$p\lim\overline{S}\frac{1}{2}\{2\overline{T}-\overline{S}[\widetilde{\sigma}^2+\overline{T}^2+\overline{T}]\}=p\frac{1}{2}\left\{2\frac{1}{p}-p\left[\frac{(1-p)}{p^2}+(1/p)^2+\frac{1}{p}\right]\right\}=0. \tag{7.13}$$

An immediate implication is that $p\lim\widehat{\beta}_1=0$ for a constant-hazard function. On the other hand, just as with GMD, the distribution of SB is skewed right in finite samples, and thus it is necessary to use simulations to obtain finite-sample critical values.

In spite of their skewed distributions, GMD and SB are asymptotically pivotal; that is, asymptotically they do not depend on unknown parameters. For asymptotically pivotal statistics, the bootstrapped critical values are generally more accurate than those based on first-order asymptotic theory. Horowitz (2001) and Davidson and MacKinnon (2006) explain why it is desirable to use pivotal statistics when bootstrapping.

7.4.3 Strong-form tests

Diebold and Rudebusch (1991) and Ohn, Taylor and Pagan (2004) employ the chi-square goodness-of-fit test to determine if durations follow the geometric density. The test statistic is $\chi^2=\sum_{j=1}^{K}[(O_j-E_j)^2/E_j]$, where O_j is the observed frequency in the jth bin and E_j is the expected frequency in the jth bin. The expected frequency is derived under the null distribution, in this case the geometric density. A well-known rule of thumb is that the expected frequency, E_j, should be at least 5 for all bins (see Hoel, 1954). To be on the safe side, Ohn, Taylor and Pagan (2004) use 6 instead of 5. If one adheres to this rule-of-thumb, long-term experience suggests that χ^2 approximately follows its asymptotic chi-square distribution with K-1 degrees of freedom. Nonetheless, both Diebold and Rudebusch (1991) and Ohn, Taylor and Pagan (2004) employ simulation to obtain finite-sample critical values.

7.5 Modeling with covariates

Most researchers using aggregate-level data segment the time line to control for heterogeneity of the exit probabilities. For example, Edwards, Biscarri and

de Gracia (2003) segment the time line for Latin American and Asian countries to determine the effect of financial liberalization on stock market cycles. Models that employ covariates have not been as popular and the studies that do use them generally assume duration independence. For instance, Estrella and Mishkin (1998) employ a discrete-time analysis to examine various financial variables as predictors of US recessions, and Chin, Geweke and Miller (2000) use a similar analysis to predict turning points in the civilian unemployment rate. Conditional upon the right-hand-side variables, however, both studies assume the hazard function is independent of time.[11]

The strong assumption of duration independence in models with covariates is defensible in some circumstances. In the political science literature, Bennett (1999, p. 262) goes so far as to argue that including covariates to effectively eliminate duration dependence is a laudable goal:

> Unless we can anthropomorphize and assume that the phenomenon we are examining truly has a life of its own, then the pattern or covariation over time that we observe is somehow, somewhere, driven by a variable or set of variables that characterizes the world. If the causal factor driving duration dependence is measured and included in the model as an independent variable, then unexplained duration dependence ... may disappear.

Bennett's view aligns with the concept of *probabilistic reduction* that Spanos (1995) traces back to the biometric tradition of Galton and Pearson; see also Spanos (2006) and Hoover (2006). In the regression framework, probabilistic reduction implies that a complete theory must induce white-noise disturbances in the model. In practice, whether covariates can account for the observed duration dependence in any binary series is surely an empirical question that cannot be assumed away for the model at hand. Complete theories are ideal but rare.

Another reason that duration dependence is frequently ignored for discrete-time analysis is that many researchers apparently believe that it is not possible to incorporate such dependence. For instance, Bennett (1999, p. 259) argues that the logit model is insufficient for the analysis of duration data because "it assumes that no duration dependence exists." Below we show that a slightly modified logit model is suitable to capture autonomous changes in the discrete-time hazard as well as changes due to covariates.

Probit models are alternatives to logit models. There is no need, however, to adopt any type of latent structure for either probit or logit. In other words, it is neither necessary nor desirable to insist that there exists some type of latent variable, y_t^*, such that $S_t = 1$ if $y_t^* > 0$, and $S_t = 0$ if $y_t^* \leq 0$. Although such an assumption *is* desirable in the discrete-choice literature, where y_t^* is interpreted as a utility function, it only unnecessarily complicates a duration analysis. In fact, to mark time for economic cycles, we frequently map an *observed* series,$\{ y_t\}$, to $\{S_t\}$. Thus, the unobservables of true interest in either the logit or probit probability models are the estimable parameters controlling the termination probabilities of $\{S_t\}$.[12]

7.5.1 The logit model

Following Allison (1984), let $P(t)$ represent the discrete-time hazard function. We use $P(t)$ here rather than $h(t)$ because it is more natural to do so for logistic regression. For the logit hazard model:

$$\log(P(t)/(1 - P(t)) = a(t) + \beta_1 x_1 + \beta_2 x_2(t). \tag{7.14}$$

where $a(t)$ represents a set of dummies, one for each of the observed exit periods, that account for autonomous changes in the exit probabilities; x_1 represents a set of covariates that do not change over the course of a given contraction; and $x_2(t)$ represents a set of covariates that do change over the course of a given contraction. In other words, $a(t)$ allows for non-constant hazards, conditional upon the x values. For constant hazards, one should substitute a single intercept parameter, a, for the time-varying $a(t)$.

7.5.1.1 The LSB test

Define d_t as the number of consecutive months (that is, duration) spent in a contraction up and *through* time t, and consider a very simple model with just one $x(t)$, namely d_{t-1}, as defined for the SB test:

$$\log(P(t)/(1 - P(t)) = \beta_0 + \beta_1 d_{t-1}. \tag{7.15}$$

Drop observations from the sample if $d_{t-1} = 0$. A restriction from the assumption of duration independence is $H_0 : \beta_1 = 0$, with the test statistic computed by the corresponding asymptotic t-ratio. The constant-hazard test from logit regression, call it LSB, is obviously closely related to Pagan's regression-based SB test. The potential advantages of using SB are that it is very straightforward, the least squares algorithm is very stable, and many computer packages recursively estimate least squares (but not logit) coefficients.

Stability is an important consideration for small samples since the number of observations with $S_t = 1$ is usually small for macroeconomic data. Consider, for example, that there are only about ten post-World War II economic expansions in the American business cycle. For these expansions, the number of observations with $S_t = 0$ is large because of a low termination probability, hence the proportion of observations with $S_t = 1$ is small. With a very low proportion of observations with $S_t = 1$, small samples can be especially problematic for logistic regression.

For sufficiently large samples, however, the logit model is preferred to the linear model. For the linear model, the predicted probabilities can lie outside of the unit interval from 0 to 1, and it is well known that least squares estimators are not efficient if the dependent variable is binary. In contrast, logit probabilities are always bounded on the unit interval; and since logit estimation is based on maximum likelihood, the estimators from logit are asymptotically efficient.

7.5.1.2 A comparison with Cox's model

Thompson (1977) presents another compelling argument in favor of logit estimation. As the discrete-time intervals become smaller and smaller, the logit model

converges to Cox's (1972) continuous-time proportional hazards model. Cox's model can be written as:

$$\log h(t) = a(t) + \beta x, \tag{7.16}$$

where $h(t)$ is the continuous-time hazard rate, similar to $P(t)$, and x represents a set of covariates that do not vary over time. Like the discrete-time logit model, $a(t)$ can be any function of time. The term *proportional hazard* comes from the fact that, for any t and any two individuals i and j:

$$h_i(t)/h_j(t) = \exp(\beta x_i)/\exp(\beta x_j) = \exp(\beta(x_i - x_j)). \tag{7.17}$$

This function does not vary with time because the autonomous time-varying term, $a(t)$, cancels out. Cox's model, however, is not just limited to proportional hazards, since the model is no longer proportional if some of the x values vary with time. Some computer packages allow for time-varying covariates while others do not. Fortunately, this is not a concern to us since it is always possible to allow for time-varying covariates in logistic regression.

Cox's proportional hazards model is semiparametric because the autonomous time-varying term, $a(t)$, does not have to be specified. Although the estimators of β are asymptotically unbiased and normally distributed, they are not *fully* efficient because the exact functional form of $a(t)$ is not specified. However, Efron (1977) shows that the loss of efficiency is typically so small that it is not of practical concern. The importance of Cox's model for duration analysis is well-summarized by Allison (1984, p. 35):

> It is difficult to exaggerate the impact of Cox's work on the practical analysis of event history data. In recent years, his 1972 paper has been cited well over 100 times a year in the world scientific literature. In the judgment of many, it is unequivocally the best all-around method for estimating regression models with continuous-time data.

In practice, time is always measured in discrete intervals, although the intervals may be irregular for individual histories. Consider also that, if two or more individuals experience events at the same time, that is, if we observe a *tie*, then the model proposed by Cox (1972) is the logit model. Therefore, although some authors have argued that continuous-time methods are preferable to discrete-time methods on theoretical grounds, or have completely ignored discrete-time methods altogether, the lack of attention to the logit model seems rather unfortunate. The preference for continuous-time models is largely based on computational grounds, but this is certainly less of an issue today than it was 25 years ago!

7.5.2 Predetermined variables and unobserved heterogeneity

An important issue is the number and choice of covariates. Consider, for example, economic expansions. If the length of the prior expansion and/or contraction influences the exit probability of the current expansion, then the length of the

prior expansion and/or contraction should be included in the set of explanatory variables. The usual asymptotic t-ratio can then be employed to determine if such effects are important. Thus, with a set of appropriate explanatory variables, either predetermined or strictly exogenous, one can account for certain types of dependence that cannot be handled by simply segmenting the time line.

Unobserved, or neglected, heterogeneity is inherently a problem of omitted variables. As a practical consideration, however, the sample sizes encountered in the study of economic cycles are generally too small to consider many explanatory variables. Explanatory variables are used to control for heterogeneity in the exit probabilities. Bover, Arellano and Bentolila (2002) develop logistic discrete hazard models that can accommodate unobserved heterogeneity. In their microeconometric study of about 27,000 unemployment spells of Spanish men, they explicitly control for unemployment benefits, age, education, head of household status, and dummies for economic sector and year of unemployment. A second model is estimated with time-varying macroeconomic variables, such as the growth rate in GDP and sectoral unemployment rates, substituting for the sectoral and time dummy variables. Autonomous shift dummies, $a(t)$, are included in both models to capture flexible additive duration dependence. A dummy variable is included for each possible exit time, with time marked at monthly intervals.

Bover, Arellano and Bentolila (2002) treat the length of unemployment benefits as predetermined, though not strictly exogenous, since knowledge about future benefits can influence job choice and especially the decision to re-enter the labor market. The distinction between predetermined and strictly exogenous variables is fairly unimportant unless there is unobserved heterogeneity. In that case, predetermined variables are effectively endogenous, and it is necessary to maximize the joint mixture likelihood for the unemployment and benefit durations. This is accomplished, in part, by introducing a discrete unobserved random variable with finite support. Additional parameters are thus included to model the unobserved heterogeneity.

In the macroeconometrics literature on business cycles, however, post-war sample sizes of about ten spells preclude the possibility of estimating models that are rich in parameters. As a partial solution, segmenting the time line into pre-war and post-war is sufficient to control for omitted variables that vary across, but not within, the segments. Since any residual heterogeneity induces *negative* duration dependence, expansions and contractions may appear to be self-perpetuating even if this premise is false. On the other hand, Diebold and Rudebusch (1990) and Ohn, Taylor and Pagan (2004) find evidence of *positive*, not negative, duration dependence in the American business cycle. Thus, at least the *direction* of such duration dependence cannot be caused by residual, or neglected, heterogeneity.

7.6 The shape of cycles

Duration dependence concerns the shape of the hazard function. If the hazard function slopes upward, there is positive duration dependence; if downward, there

is negative duration dependence; if flat, there is no duration dependence. However, the shape of the hazard function is but one example of the types of shapes associated with cycles. In particular, Harding and Pagan (2002) and Pagan and Sossounov (2003) discuss the typical shapes of phases, either contractions or expansions. They address the following issues:

1. Amplitudes of phases
2. Cumulative movements within phases
3. Asymmetric behavior of phases.

After marking the turning points, the binary series, $\{S_t\}$, is employed along with the observed underlying series, $\{y_t\}$, to describe the shape of a phase. For example, during economic expansions, GDP is observed to rise quickly at first and then slows its ascent before finally reversing direction, thus marking the beginning of a contraction.

In Figure 7.2, we present a stylized economic expansion. The y axis represents log(GDP), or *amplitude*, and the x axis represents time spent in an expansionary phase, or *duration*. On the time axis, time A represents the first turning point, the trough, and time B the second turning point, the peak. The amplitude of the expansion is the vertical distance between points A and B, measuring the change in GDP from trough to peak. In this instance the amplitude is $\log(GDP_B) - \log(GDP_A)$. The hypotenuse of the triangle is a benchmark representing a constant increase in amplitude, with increases in amplitude proportional to the time spent in the expansionary phase.

Descriptive measures of interest include the average duration and average amplitude of the expansions in the sample, measures of the variability in durations and amplitudes, and a measure to show how closely growth in GDP adheres to the hypotenuse depicted in Figure 7.2. For our sample of n expansions, we observe the durations $\{T_1, T_2, \ldots, T_n\}$, and the amplitudes $\{A_1, A_2, \ldots, A_n\}$.

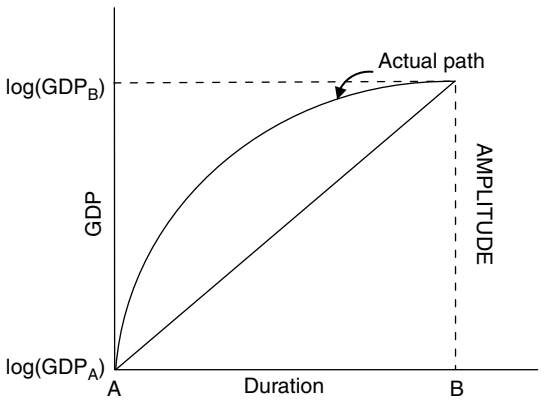


Figure 7.2 Stylized expansion phase

7.6.1 Durations

Assume that durations and amplitudes constitute random samples. If so, the $T_i's$ follow identical distributions, $T_i \backsim D(\mu_T, \sigma_T^2)$, with T_i statistically independent from T_j for $i \neq j$. Further, $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$ is a consistent estimator for μ_T and $\overline{T} \overset{a}{\backsim} N(\mu_T, \sigma_T^2/n)$. This assumes, of course, that the time line has been successfully segmented to ensure that each of the $T_i's$ follows the same distribution, that is, with no mixing of distributions.

Small sample inference, however, is particularly problematic for durations. Consider that the $T_i's$ rarely come from the normal distribution. In fact, for the discrete case with constant hazards, the $T_i's$ follow the geometric distribution, $T_i \backsim GEOM(\mu_T, \mu_T^2 - \mu_T)$, with $\mu_T = 1/p$, where p is the constant hazard. However, since the geometric distribution is considerably right-skewed, for very small samples it is generally unwise to construct confidence intervals on μ_T by employing the *t*-distribution, since normality of the durations is one of the assumptions supporting its use. Simulations by Pagan and Sossounov (2003) also suggest that $\overline{T}$ is significantly skewed in small samples.

7.6.2 Amplitudes

Amplitudes are less problematic. For the *i*th expansion, the amplitude is calculated as $A_i = y_{T_i} - y_0 = \sum_{j=1}^{T_i} \Delta y_j$, with $\Delta y_1 = y_1 - y_0$, $\Delta y_2 = y_2 - y_1$, and so on. In Figure 7.2, y_0 equals $\log(GDP_A)$ and y_{T_i} equals $\log(GDP_B)$. The amplitude of the expansion is thus the sum of the growth rates from time A to B. Although Harding and Pagan (2002) convincingly argue that GDP growth rates are not statistically independent, their sum may be approximately normally distributed by either Gordin's (1969) or Hannan's (1973) Central Limit Theorems (CLT) for stationary ergodic processes; see also White (1984, Ch. V). Thus, approximate normality for the amplitudes, A_i, is more plausible than normality for the durations, T_i, provided that expansions are sufficiently long to allow the CLT to work for each A_i.

7.6.3 Cumulative gain

Another measure of interest is the *cumulative* gain from trough to peak, that is, the area under the curve that describes the actual path of GDP. An approximation to this gain is obtained by adding together the area in rectangles of unit length and height equal to $y_j - y_0$. The approximation, however, is too large since each such rectangle overstates the actual area by approximately $(y_j - y_{j-1})/2$. Correcting for the overstatement, Harding and Pagan (2002) approximate the cumulative gain for the *i*th expansion by:

$$F_i = \sum_{j=1}^{T_i} \left[(y_j - y_0) - (y_j - y_{j-1})/2 \right] \tag{7.18}$$

$$= \left[\sum_{j=1}^{T_i} (y_j - y_0) \right] - A_i/2, \tag{7.19}$$

where $A_i = y_{T_i} - y_0$. A benchmark for F_i is the cumulative area under the hypotenuse depicted in Figure 7.2. The area of this right triangle is easily calculated as $AREA_i = (T_i \cdot A_i)/2$. Taking the difference, $F_i - AREA_i$, and then dividing by the duration of the ith expansion, one obtains:

$$E_i = (F_i - AREA_i)/T_i. \tag{7.20}$$

A positive value of E_i indicates that growth rates generally increase at a decreasing rate over the life of the expansion, and a negative value of E_i indicates that growth rates generally increase at an increasing rate over the life of the expansion. A positive value for $\overline{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$ indicates that most of the growth occurs at the beginning of the typical expansion, and a negative value of $\overline{E}$ indicates that most of the growth occurs at the end of the typical expansion. If $\overline{E} = 0$, then neither characterization is accurate; instead, the actual path for y tends to oscillate around the hypotenuse. $\overline{E}$ is thus a useful descriptive measure of the average shape, or curvature, of expansions. A similar interpretation holds for $\overline{C}$, the average shape of contractions. In the literature on business cycles, Sichel (1994) documents the rapid recovery of an expansion that leads to a positive value for $\overline{E}$. In the financial literature, Edwards, Biscarri and de Gracia (2003) observe that the excess index (equation 7.20) is particularly useful in characterizing stock market behavior.

7.7 Synchronization of cycles

The cyclic characteristics of a single series, $\{y_t\}$, are of interest because they yield insights about the underlying series. Consider also that the cyclic relationship between *two* underlying series, $\{y_{1t}\}$ and $\{y_{2t}\}$, is of like interest to both academics and policy makers. For example, do short periods of financial crisis influence the business cycle? Are cycles in foreign economies closely tied to the American business cycle? Are cycles in national unemployment related to cycles in GDP? Finally, are cycles in oil prices related to the world economic or political (dis)order? Each of these questions can be answered, in part, by examining the observed binary time series, $\{S_{1t}\}$ and $\{S_{2t}\}$, that respectively correspond to $\{y_{1t}\}$ and $\{y_{2t}\}$.[13]

7.7.1 The coincidence indicator

One way to measure the correspondence between $\{S_{1t}\}$ and $\{S_{2t}\}$ is to employ the *coincidence indicator* of Harding and Pagan (2002):

$$\widehat{I} = 1/T\sum_{t=1}^{T}[S_{1t}S_{2t} + (1 - S_{1t})(1 - S_{2t})], \tag{7.21}$$

where T is the total number of periods in the sample interval, regardless of phase. Consistent with the notation of Harding and Pagan (2002, 2006), we use T in this section to denote the sample size rather than duration. It follows that $\widehat{I}$ is the fraction of periods that $\{S_{1t}\}$ and $\{S_{2t}\}$ are synchronized. Harding and Pagan (2006) note that there is *perfect positive synchronization* between $\{S_{1t}\}$ and $\{S_{2t}\}$ if

$\widehat{I} = 1$, and there is *perfect negative synchronization* if $\widehat{I} = 0$. Beyond the literature on national output, Edwards, Biscarri and de Gracia (2003) observe that financial synchronization, or *concordance*, among Latin American countries has substantially increased after financial liberalization.

7.7.2 Correlation analysis

The sample correlation coefficient between S_1 and S_2, call it r_s, conveys similar information to $\widehat{I}$. If $r_s = 1$, there is evidence in favor of the null hypothesis, H_0 : $\rho_s = 1$, since $S_{1t} = S_{2t}$ for every paired observation in the sample. Of course, if there is a single case where $S_{1t} \neq S_{2t}$, there is reason to reject the hypothesis that the cycles are *perfectly* positively synchronized. Similar reasoning holds true for perfect negative synchronization. A formal test of $H_0 : \rho_s = 1$ is presented by Harding and Pagan (2006).

As perfect synchronization will be empirically atypical, it is still useful to compute either or both of the sample correlation coefficient and coincidence indicator to see how closely two series move in tandem. Graphing the series may also reveal an obvious translation of $\{S_{1t}\}$ that will more closely synchronize $\{S_{1t}\}$ with $\{S_{2t}\}$. For instance, consider $\{S_{3t}\} = \{S_{1,t\pm l}\}$ for some integer $l > 0$. The concordance between $\{S_{3t}\}$ and $\{S_{2t}\}$ may be considerably higher than the concordance between $\{S_{1t}\}$ and $\{S_{2t}\}$ if lagged effects are important. Alternatively, it may be that changing just a few turning points could lead to near-perfect concordance between two series. If so, sensitivity analysis is worthwhile.

7.7.2.1 Tests based on the method of moments

It is important to formally test for no synchronization $H_0 : \rho_s = 0$, since this implies that $\{S_{1t}\}$ and $\{S_{2t}\}$ are unrelated series with no common cycle. As a case in point, the business cycles of the United States and the United Kingdom could have high concordance simply because most of the time these economies are expanding, not contracting. On the other hand, whether these two economies actually *move together* is a different issue. In other words, although $\{S_{1t}\}$ and $\{S_{2t}\}$ may be highly synchronized, this does not by itself imply a common cycle.

Under classical conditions, several tests for $H_0 : \rho_s = 0$ are equivalent. For instance, we can employ:

$$t_r = r_s \sqrt{\frac{T-2}{1-r_s^2}} \overset{a}{\sim} N(0,1). \tag{7.22}$$

Numerically equivalent test statistics are the standard t-ratios for the slope coefficients in either $S_{1t} = \alpha + \beta S_{2t} + \varepsilon_{1t}$ or $S_{2t} = \gamma + \delta S_{1t} + \varepsilon_{2t}$. However, the situation is complicated by the fact that S_1 is serially correlated, as is S_2, and thus the independence assumption associated with the traditional t-test of zero-correlation is compromised. Therefore, the statistic used to test $H_0 : \rho_s = 0$ must be made robust to serial correlation and heteroskedasticity.[14] Harding and Pagan (2006) recommend that the test statistic be constructed via GMM with a robust variance estimate to account for serial correlation and heteroskedasticity.

A multivariate version of the GMM test for several *S*-series is presented by Harding and Pagan (2006), but with the GMM estimator for the bivariate case based only on the moment conditions:

$$E[S_{jt}] = \mu_j,\ j = 1, 2 \tag{7.23}$$

$$E\left[\frac{(S_{1t} - \mu_1)(S_{2t} - \mu_2)}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}} - \rho_s\right] = 0. \tag{7.24}$$

Stack the above three moment conditions into a $3x1$ vector, $h_t(\theta, S_{1t}, S_{2t})$, such that:

$$h_t(\theta, S_{1t}, S_{2t})' = \left[S_{1t} - \mu_1,\ S_{2t} - \mu_2,\ \frac{(S_{1t} - \mu_1)(S_{2t} - \mu_2)}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}} - \rho_s\right], \tag{7.25}$$

with parameter vector $\theta' = [\mu_1, \mu_2, \rho_S]$, and take the average:

$$g(\theta, \{S_{1t}, S_{2t}\}_{t=1}^T) = \frac{1}{T}\sum_{t=1}^{T} h_t(\theta, S_{1t}, S_{2t}). \tag{7.26}$$

The covariance matrix of $\sqrt{T}g(\theta, \{S_{1t}, S_{2t}\}_{t=1}^T)$ is consistently estimated by:

$$V = \Gamma_0 + \sum_{k=1}^{m}\left[1 - \frac{k}{m+1}\right][\Gamma_k + \Gamma_k'], \tag{7.27}$$

where:

$$\Gamma_k = \frac{1}{T}\sum_{t=k+1}^{T} h_t(\theta, S_{1t}, S_{2t}) h_{t-k}(\theta, S_{1t}, S_{2t})'. \tag{7.28}$$

Harding and Pagan (2006) recommend the window width, m, to be the integer part of $(T-1)^{1/3}$.

Let $\theta_0' = [\mu_1, \mu_2, 0]$ be the restricted parameter vector for $H_0 : \rho_S = 0$, with no common cycle between S_1 and S_2 under this null hypothesis. The test statistic:

$$W_{SNS} = \sqrt{T}g\left(\theta_0, \{S_{1t}, S_{2t}\}_{t=1}^T\right)' V^{-1} \sqrt{T}g\left(\theta_0, \{S_{1t}, S_{2t}\}_{t=1}^T\right), \tag{7.29}$$

follows an asymptotic chi-square distribution with one degree of freedom. Substituting sample means, $\widehat{\mu}_1$ and $\widehat{\mu}_2$, and the sample correlation, r_s, for their population counterparts does not affect the asymptotic distribution of W_{SNS}. Substituting sample moments for population moments reduces W_{SNS} to:

$$W_{SNS} = T(r_s - 0)\widehat{v}^{-1}(r_s - 0), \tag{7.30}$$

where $\widehat{v}$ is the lower right-hand element of $\widehat{V}$. An equivalent test statistic is the asymptotic t-ratio, $t_{SNS} = T^{1/2}\widehat{v}^{-1/2} r_s \overset{a}{\sim} N(0,1)$.

Closely related tests are the market-timing test of Pesaran and Timmermann (1992) and Pearson's chi-square test for independence. For instance, Artis,

Kontolemis and Osborn (1997) and Artis, Krolzig and Toro (2004) use either Pearson's test or a transformation of the concordance index to test whether the series $\{S_{1t}\}$ and $\{S_{2t}\}$ are unrelated. However, consider that the method of moments test essentially examines the moment condition implied by the covariance, that is, $E(S_1S_2) - E(S_1)E(S_2) - \sigma_s = 0$. The null hypothesis is $H_0 : \sigma_s = 0$. Observe, however, that $\sigma_s = p_{12} - p_1p_2$, where $p_{12} = P(S_1 = 1, S_2 = 1)$, $p_1 = P(S_1 = 1)$, and $p_2 = P(S_2 = 1)$. The method of moments test thus effectively determines whether S_1 and S_2 are statistically independent, $p_{12} = p_1p_2$, and this is exactly the goal of Pearson's test.

Of course, a critical assumption behind Pearson's test is that *observations* in the sample are statistically independent. This assumption clearly fails in this case since the state variables, S_1 and S_2, exhibit strong serial dependence; see, for example, Kedem (1980). Since the market timing and concordance index tests also assume that observations are statistically independent, the robust covariance matrix of Harding and Pagan's method of moments test offers an improvement since it allows for serial correlation and heteroskedasticity of unspecified type.

7.7.2.2 Regression-based tests

On the other hand, the problems induced by serial correlation and heteroskedasticity can be significantly lessened by separating contractions from expansions and by incorporating time dependency into the model. The very nature of duration dependence will most surely differ across expansions and contractions; even if both phases exhibit constant hazards, regression disturbances are generally heteroskedastic if observations on expansions are not separated from those on contractions. So, when testing for synchronization, we should separate expansions from contractions. Define the dependent binary variable so that $S_{2t} = 0$ if the contraction continues and $S_{2t} = 1$ if the contraction terminates. For expansions, consider $1 - S_{2t}$ instead of S_{2t}. Our sub-sample thus consists of strings like:

	Jan	Feb	Mar	Apr	...	Sep	Oct	Nov	Dec	Jan	Feb	...	Jun	Jul
S_{2t}	0	0	0	1	...	0	0	0	0	0	1	...	0	0
S_{1t}	0	0	1	1	...	0	1	1	1	0	0	...	0	0

In this example, for S_2 there are two complete contractions of respective length $T_1 = 3$ and $T_2 = 5$, and one incomplete contraction of length $T_3 = 2$. The number 1 signifies the beginning of a new phase, in this case an expansion. Of course, $S_{1t} = S_{2t}$ should be the predominant case if there is a high concordance. By considering the phases separately, we are able to address problems of time dependency and heterogeneity. The statistical method we propose employs logistic regression and is similar in spirit to Pagan's (1998) regression-based test. To test for dependence between S_1 and S_2, let S_2 be the dependent variable in the logistic regression:

$$\log(P(t)/(1 - P(t)) = a(t) + \beta S_{1t}. \tag{7.31}$$

The term $a(t)$ consists of a set of dummy variables, one for each possible exit period, that controls for autonomous changes in the exit probability, $P(t)$. That is, $a(t)$ accounts for the duration dependence, or serial correlation, in the series $\{S_{2t}\}$. Having removed the time dependence captured by $a(t)$, the slope coefficient,

β, captures the statistical dependence between S_2 and S_1. Our null hypothesis, $H_0 : \beta = 0$, corresponds to statistical independence, and thus to $H_0 : \sigma_s = 0$. The form of the above regression is identical for expansions, and completely analogous regressions can be employed for the alternative dependent variable, S_1. Finally, it is possible to use linear regression rather than logistic regression. For large samples, however, estimation efficiency is improved by using logistic regression.

7.8 Unemployment cycles

If the most important measure of aggregate economic well-being is output, then surely the second most important is unemployment. Although unemployment cycles are interesting to academics, policy makers, and the general public, most of the attention in the academic literature is on output cycles. Exceptions are Boldin (1994), Chin, Geweke and Miller (2000), and Hamilton (2005), whose work focuses on unemployment cycles.

Even though we anticipate high unemployment during GDP contractions and low unemployment during GDP expansions, the output and unemployment series offer separate pieces of information about the economy. Even if we use the same rules to mark turning points for unemployment and output, we do not expect perfect negative synchronization – that is, necessarily *increasing* levels of unemployment during periods of *decreasing* output, and vice versa. It is thus of interest to determine the degree of synchronization, to compare the shapes of output and unemployment cycles, and to perform a separate duration analysis for the unemployment series.

7.8.1 Cycle shapes

Our data are the logarithmic monthly Bureau of Labor Statistics (BLS) seasonally adjusted *civilian unemployment rate* series from 1948:1 through 2007:1.[15] Figure 7.3 plots the unemployment rate series and for comparison marks the NBER dated business cycle recessions. Table 7.1 presents the unemployment rate reference dates determined by the BBQ algorithm. We set the minimum phase to 9 months and the minimum cycle to 18 months. These censoring rules visually mark the turning points in unemployment much better than those employed by Harding and Pagan (2002) to mark the turning points in output – namely, a minimum phase of 6 months and a minimum cycle of 15 months. Our censoring rules also mark the turning points in unemployment much better than does the Extended Okun Rule that considers only whether there are two consecutive months opposite the prevailing phase. There are 10 post-war completed spells of contraction, or *downswings*, in the unemployment rate, lasting an average of 48 months with a standard deviation of 30 months. There are 9 post-war completed spells of expansion, or *upswings*, in the unemployment rate, lasting an average of 22 months with a standard deviation of 10 months. Similar results are obtained from using quarterly data with a minimum phase of 3 quarters and a minimum cycle of 6 quarters.

The average amplitudes of upswings and downswings are the same in magnitude. The average amplitude of downswings is -0.55 with a coefficient of variation of

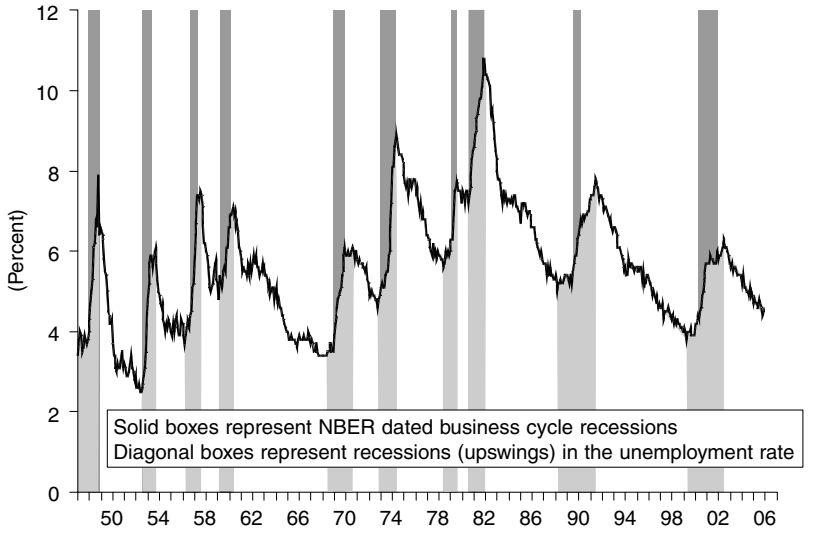


Figure 7.3 US civilian unemployment rate
Note: Unemployment data are from the US Bureau of Labor Statistics, and business cycle dates are from the National Bureau of Economic Research.

Table 7.1 Unemployment rate cycle dates: 1948–2006

Reference dates		*Duration in months*			
Peak	*Trough*	*Contraction*	*Expansion*	*Cycle*	
		Peak to trough	*Previous trough to this peak*	*Trough from previous trough*	*Peak from previous peak*
Oct 1949	June 1953	44	–	–	–
Sep 1954	Mar 1957	30	15	45	59
Jul 1958	Feb 1960	19	16	35	46
May 1961	May 1969	96	15	111	34
Aug 1971	Oct 1973	26	27	53	123
May 1975	May 1979	48	19	67	45
Jul 1980	Jul 1981	12	14	26	62
Dec 1982	Mar 1989	75	17	92	29
Jun 1992	Apr 2000	94	39	133	114
Jun 2003	Jun 2006	36	38	74	132
		Summary statistics			
	No. of phases	10 (9)	9 (10)		
	Average length	48.0 (59.1)	22.2 (10.4)		
	Standard deviation	30.3 (38.0)	10.0 (3.3)		

Note: NBER business cycle summary statistics are in parentheses. Unemployment contractions are paired with business cycle expansions, and unemployment expansions are paired with business cycle contractions.

−0.55, and the average amplitude of upswings is 0.55 with a coefficient of variation of 0.33. Since falls in unemployment appear about evenly matched with rises in unemployment, this lends some credence to the idea of a natural rate. The overall sample average unemployment rate is about 5.61%, although at times there were large deviations from the average. For example, in November 1982 the unemployment rate reached a high of 10.8%.

There are, however, significant differences between upswings and downswings. For downswings, the average cumulative movement is $\overline{F}_c = -18.43$, the average excess is $\overline{C} = -0.046$ and the coefficient of variation in the excess is −1.32. For upswings, the average cumulative movement is $\overline{F}_e = 6.40$, the average excess is $\overline{E} = 0.009$ and the coefficient of variation in the excess is 5.70. The average cumulative movement in downswings is well over twice the magnitude of the cumulative movement in upswings; this is consistent with the longer average duration of downswings. From the average excess $\overline{C}$, employment tends to fall at a decreasing rate during contractions, or downswings, and from $\overline{E}$, employment tends to rise at a decreasing rate during expansions, or upswings. The steep initial decline in unemployment during downswings is more prominent than the steep initial ascent in unemployment during upswings. In fact, from the coefficient of variation, there is much more relative variability in the excess for upswings than for downswings. Long durations in downswings, large cumulative movements, and stable excess across downswings all reflect favorably on current economic policy.

7.8.2 Synchronization with business cycles

From Figure 7.3, cycles in output and unemployment appear highly synchronized. However, even though there are about as many turning points in the unemployment cycle as there are in the business cycle (see the summary statistics in Table 7.1), the two binary series representing unemployment and output are not perfectly correlated. Let $S_{1t} = 1$ if output is rising and $S_{2t} = 1$ if unemployment is rising; also let $S_{1t} = 0$ if output is falling and $S_{2t} = 0$ if unemployment is falling. The coincidence indicator for the binary series is $\widehat{I} = 0.17$ and the correlation between them is $r_s = -0.61$. In other words, about 83% of the time rising output is associated with falling unemployment.

We can test the null hypothesis $H_0 : \rho_s = 0$ by using Harding and Pagan's (2006) method of moment test. With a bandwidth of either $m = 8$ or $m = 9$ we find that t_{SNS} is about −5.5, and thus we reject the null hypothesis that unemployment and output are statistically independent. Regression-based tests yield the same conclusion. Unemployment and output appear to be statistically dependent.

Separating upswings from downswings in unemployment provides an important insight. When measured at monthly intervals with our censoring rules, falling unemployment is coincident with rising output, though rising unemployment is not necessarily coincident with falling output. In fact, conditional on rising unemployment, output is falling only about half the time. In other words, unemployment is perfectly synchronized with output when unemployment is falling, but it is a coin toss when unemployment is rising.

7.8.3 Duration analysis

If expansions, or upswings, in the unemployment rate exhibit *positive duration dependence,* then upswings with longer maturities have a higher chance of terminating than those with shorter maturities. The hazard tends to rise with the duration of the event, and the average duration is greater than the standard deviation. In the opposite case, called *negative duration dependence,* the hazard tends to fall with the duration of the event, and the average duration is less than the standard deviation. *Duration independence* is characterized by neither of the above cases. For instance, the probability that unemployment exits an expansionary state and enters a contractionary state is constant regardless of how long the upswing has lasted.

The average duration of 22 months for upswings is more than twice the sample standard deviation of 10 months; this suggests that rises in unemployment exhibit positive duration dependence. Likewise, the average duration of 48 months for downswings is larger than the sample standard deviation of 30 months; this also suggests positive duration dependence, though such descriptive evidence is not as strong as it is for upswings. We can formally test for duration dependence by using the discrete-time SB and GMD tests from Ohn, Taylor and Pagan (2004). We subtract months from each of the observed durations to be consistent with our BBQ censoring rules.[16]

We analyze upswings and downswings separately. Under the null hypothesis of duration independence, the estimated termination probability for upswings is 0.076, but the estimated termination probability for downswings is only 0.025. However, neither SB nor GMD indicate duration dependence for either upswings or downswings. Finite-sample p-values are obtained through a parametric bootstrap algorithm, with the discrete-time geometric density corresponding to the null hypothesis of duration independence. Sensitivity analysis is performed by varying the termination probability to account for sampling variability in estimation. Since our calculated p-values are always considerably greater than 0.10, we cannot reject the null hypothesis that the probability of exiting a state of national unemployment is independent of its duration at the 10% significance level. As a further robustness check, we also employ the asymptotic LSB t-test and the continuous-time W-test from Shapiro and Wilk (1972). The distribution of W depends neither on the termination probability nor on the minimum phase, with finite-sample critical values tabulated by Shapiro and Wilk (1972). Consistent with SB and GMD, neither LSB nor W reject duration independence for either upswings or downswings in unemployment.

In comparision, Ohn, Taylor and Pagan (2004) find some evidence that post-war economic contractions exhibit positive duration dependence, though there is no such evidence for post-war expansions. However, such lack of evidence is not too surprising in light of the small number of completed cycles. For instance, post-war sample sizes are too small to employ the chi-square goodness-of-fit test for comparison with SB and GMD.

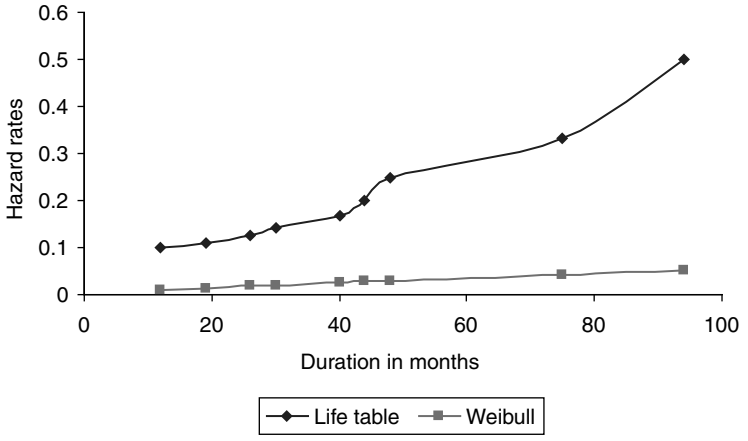


Figure 7.4 Downswings

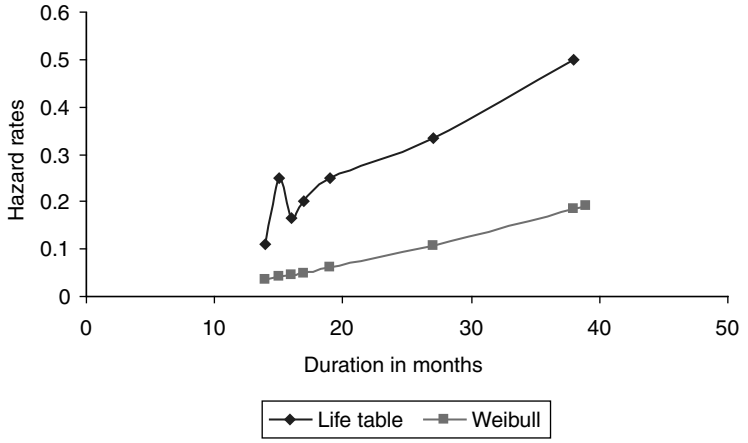


Figure 7.5 Upswings

Figures 7.4 and 7.5 present nonparametric plots of the hazard functions via the life table method of Cutler and Ederer (1958). The hazard functions for both downswings and (especially) upswings can be characterized as NBU, or perhaps even the more narrow classes, IFRA or IFR. The hazard function for upswings exhibits a local peak at 15 months, dropping at 16 months, and ascending thereafter. The rising hazard is consistent with the sharply rising hazard function for economic contractions observed by Ohn, Taylor and Pagan (2004). The hazard function for downswings exhibits strong IFR behavior with an upward trend, and this coincides well with the nonparametric hazard function for economic expansions from Ohn, Taylor and Pagan. For comparison, we also graph the parametric Weibull hazard

rates. In each case, the Weibull hazards indicate positive duration dependence. On the other hand, for upswings the parametric Weibull hazard fails to reflect the clustering of exits at 15–16 months.

We obtain further insights about the hazard functions through a regression analysis.[17] Our approach is most closely related to Estrella and Mishkin (1998) and especially that of Chin, Geweke and Miller (2000). For example, as did Chin, Geweke and Miller, we separate upswings from downswings to obtain separate estimates of the coefficients that control the hazard probabilities. As a point of contrast, our sample consists of monthly observations on unemployment from January 1948 through January 2007, whereas the sample of Chin, Geweke and Miller consists of monthly observations on unemployment from October 1949 through February 1998. We employ BBQ rules similar to those used for business cycles, whereas Chin, Geweke and Miller employ a three-month centered moving average rule subject to a threshold condition. Using either set of rules, the average upswing lasts roughly 23 months, and the average downswing lasts roughly 50 months; to be exact, for downswings we found an average of 48 months while Chin, Geweke and Miller found an average of 51 months.

Chin, Geweke and Miller employ probit estimation, closely related to our logit estimation. We favor logit estimation, however, since its use follows directly from the seminal work of Cox (1972). An additional difference is that, as is customary, we set our binary dependent variable to unity only in the month of a turning point, whereas they set the binary variable to unity in month t if a turning point occurs in months $t+1, \ldots, t+12$. Chin, Geweke and Miller construct artificial observations to control for overfitting of the model at highly leveraged values of the independent variables, but we employ a sensitivity analysis to determine whether our results are robust to dropping spells from the sample each in turn. For instance, we consider estimates from the full set of post-war contractions as well as contractions but for, say, the fourth one.

Unlike Estrella and Mishkin (1998) and Chin, Geweke and Miller (2000), we eliminate observations from the beginning of each spell to account for censoring. That is, we eliminate the first nine observations from each observed spell since the minimum phase duration allowed for either upswings or downswings is nine. As per the classical approach, we include dummy variables to allow for autonomous shifts in the hazard probabilities. In contrast, neither Estrella and Mishkin nor Chin, Geweke and Miller consider that duration dependence may be autonomous. That is, they do not allow hazard probabilities to change with the duration of the spell, independently of any change in time-varying covariates.

Harding and Pagan (2007) show that most binary time series constructed from the BBQ algorithm using NBER censoring rules are serially correlated with heteroskedastic disturbances. As is well known, either one of serial correlation or heteroskedasticity will typically lead to invalid inference. Furthermore, serial correlation in the binary time series, $\{S_t\}$, typically implies autonomous duration dependence in the spells of both contractions and expansions. Harding and Pagan account for autonomous duration dependence by directly modeling serial dependence in the state variable, S_t. Observe, however, that modeling the

serial dependence is infeasible once contractions are separated from expansions to eliminate heteroskedasticity in the disturbances. Suppose, for example, that S_t follows the second-order process:

$$S_t = \gamma_1 S_{t-1} + \gamma_2 S_{t-2} + \gamma_3 S_{t-1} S_{t-2} + u_t. \tag{7.32}$$

Having conditioned on either contractions or expansions, then $S_{t-1} = 0$ and $S_{t-2} = 0$, and it is thus infeasible to directly estimate equation 7.32. On the other hand, it is still possible to estimate the autonomous hazard function. Indeed, semi-parametric estimation of the autonomous hazard function is a very direct approach that addresses the same goal of accounting for the serially dependent nature of S_t. In other words, conditioning the constructed binary series on the states allows for estimation of the hazard function while eliminating the need to model potential serial dependence and heteroskedasticity in the state variable, S_t.

Consider first downswings with binary dependent variable S_t, such that $S_t = 1$ signifies a turning point towards rising unemployment. By considering only downswings, we eliminate one type of heteroskedasticity that occurs when upswings and downswings are considered together; see Hamilton (1989). Define the following autonomous-shift dummy variables: $D_1 = 1$ for months 1–20, inclusive, and $D_1 = 0$ otherwise; $D_2 = 1$ for months 21–30, inclusive, and $D_2 = 0$ otherwise; and $D_3 = 1$ for periods >30 and $D_3 = 0$ otherwise.[18]

Our first model is:

$$\log(P(t)/(1 - P(t)) = a_1 D_{1t} + a_2 D_{2t} + a_3 D_{3t}, \tag{7.33}$$

with estimates of the $a's$ reported in the second column of Table 7.2. Each of the coefficients is significant at the 5% level of significance. The hazard, or exit, probability is given by the formula:

$$P_i = 1/[1 + \exp(-a_i)]. \tag{7.34}$$

The estimated exit probability is about 0.019 in any month in the interval 10–20, about 0.013 in any month in the interval 21–30, and about 0.031 in any month

Table 7.2 Logit estimation: downswings in unemployment

Variables	*Equation 1*	*Equation 2*	*Equation 3*
		a(t): Autonomous shift variables	
D_{1t}	−3.922 (.0000)	−3.947 (.0000)	−4.769 (.0000)
D_{2t}	−4.331 (.0000)	−4.362 (.0000)	−5.515 (.0000)
D_{3t}	−3.434 (.0000)	−3.447 (.0000)	−5.053 (.0000)
		x: *Fixed exogenous variables*	
Lagged upswing	–	−.00025 (.8104)	–
		x(t): Changeable exogenous variables	
$(R - r)_t - (R - r)_0$	–	–	−0.726(.0181)
$CU_t - CU_{t-1}$	–	–	−0.628(.1379)

Note: p-values in parentheses.

Table 7.3 Logit estimation: upswings in unemployment

Variables	*Equation 1*	*Equation 2*	*Equation 3*
	a(t): Autonomous shift variables		
D_{1t}	−2.526 (.0000)	−1.326 (.0754)	−3.677 (.0004)
D_{2t}	−3.300 (.0012)	−0.872 (.6336)	−4.813 (.0011)
D_{3t}	−2.140 (.0042)	0.186 (.9101)	−3.543 (.0107)
	x: *Fixed exogenous variables*		
Lagged downswing	–	−0.028(.1158)	–
	x(t): Changeable exogenous variables		
$(R-r)_t - (R-r)_0$	–	–	0.338 (.3659)
$CU_t - CU_{t-1}$	–	–	2.170 (.0005)

Note: p-values in parentheses.

greater than 30. These values are very close to those obtained from the linear probability model – that is, by employing least squares with dependent variable S_t and interpreting the $a's$ from that model as probabilities. The implication is a U-shaped hazard for unemployment contractions. However, at the 5% significance level, the asymptotic likelihood ratio test from the logit model does not reject the null hypothesis of constant-hazard probabilities, $H_0 : a_1 = a_2 = a_3$.

Consider next expansions with binary dependent variable $1 - S_t$, such that $1 - S_t = 1$ signifies a turning point towards falling unemployment. For model (7.33), the estimates of the $a's$ are reported in the second column of Table 7.3. Each of the coefficients is significant at the 5% level. The estimated exit probability is about 0.074 in any month in the interval 10–20, about 0.036 in any month in the interval 21–30, and about 0.105 in any month greater than 30. Again, these values are very close to those estimated by the linear probability model. Consistent with our life table analysis, the estimated hazard function rises much more rapidly for upswings than for downswings in unemployment. However, we again fail to reject the null hypothesis of constant-hazard probabilities, $H_0 : a_1 = a_2 = a_3$, for upswings in unemployment.

Other variables may also influence the hazard probabilities. Our second equation in Table 7.2 augments the dummy variables with the duration of the immediately preceding (or lagged) upswing, and our second equation in Table 7.3 augments the dummy variables with the duration of the lagged downswing. Since neither lagged value is statistically significant, there is insufficient evidence from the logit model to conclude that the length of the current phase is influenced by the length of the preceding phase. In contrast, evidence from the linear probability model *does* suggest that the lag effect is important for upswings in unemployment. On average, the longer the preceding downswing, the shorter the current upswing.

Finally, we consider including the time-varying explanators examined by Chin, Geweke and Miller (2000). Let the symbol 0 signify the date of the last turning point prior to the date t. Then $(UR_t - UR_0)$, $(CU_t - CU_0)$, and $[(R-r)_t - (R-r)_0]$

are the respective differences in the values of unemployment (*UR*), the manufacturing capacity utilization rate (*CU*) and the spread between the monthly average of Moody's Aaa corporate bond rate and the monthly average of the 90-day treasury bill rate ($R - r$). The differences in these variables are measured from the start of the phase, t_0, up to time t. Logit or probit models can be adapted easily to incorporate such explanators that vary over time, and one purpose for including them is to account for non-stationarity of the turning points – or, in other words, duration dependence captured by covariates; see also Pesaran and Potter (1997). We also include in our regression equations the dummy variables D_1, D_2 and D_3 that account for autonomous shifts in the hazard probabilities, as well as the first differences in *UR*, *CU*, and $(R - r)$ that measure changes from time $t - 1$ to t.

Of the six time-varying explanators, only two were deemed important in our logit model: the change in the interest rate spread from time t_0 and the change in capacity utilization from time $t - 1$: $\Delta CU_t = CU_t - CU_{t-1}$. The interest rate spread typically increases about 2 percentage points over the life of an upswing, and decreases about 2 percentage points over the life of a downswing. We expect the hazard probability for downswings to be inversely related to the change in the spread from time t_0, and the hazard probability for upswings to be directly related to the change in the spread from time t_0. Inclusion of the interest rate spread corrects for drift in the hazard, though not in the same manner as do the autonomous-shift dummy variables.

From Tables 7.2 and 7.3, the *signs* on the spread coefficients are consistent with our prior reasoning. The coefficient on the spread is, however, statistically significant only for downswings. Thus, *ceteris paribus*, a large interest rate spread lowers the termination probability for a downswing in unemployment (a healthy labor market) but does not increase the termination probability for an upswing in unemployment. In other words, having controlled for autonomous shifts in the hazards, the interest rate spread has no discernable effect during unemployment upswings.

The economic interpretation of the change in capacity utilization is straightforward. An increase in utilization from time $t - 1$ should increase labor usage at time t, and a decrease in utilization should decrease labor usage. The signs on the coefficients in Tables 7.2 and 7.3 are again consistent with our prior reasoning. However, in this case the coefficient on utilization is statistically significant only for upswings in unemployment. Therefore, a decrease in capacity utilization has no discernable effects in good labor markets, but an increase marks a turnaround in bad labor markets. As a thought experiment, suppose that $D_1 = 1$ but that all other variables are set to zero in the equation for upswings. The termination probability for this relatively young upswing is only about 0.025 using the estimated coefficient for D_1 in the fourth column of Table 7.3. Consider increasing capacity utilization by 1 percentage point, say from 85% to 86%. With this change, the termination probability increases to about 0.18, a sevenfold increase in the hazard. Thus, as intuitively expected, labor fares substantially better when capital is reutilized.

At the 5% significance level, the above empirical results on the interest rate spread and capacity utilization are robust to whether we include either the

autonomous-shift dummy variables or lagged term. For with a single intercept, there is a constant hazard unless time-varying covariates actually change in value. A single-intercept probit model was chosen by Estrella and Mishkin (1998) to forecast turning points in US recessions, and by Chin, Geweke and Miller (2000) to forecast turning points in US unemployment.

Our empirical results also appear quite robust to dropping each spell, in turn, from the respective samples of either upswings or downswings. For example, we estimate the logit model with the full sample of ten downswings, as well as with the sample of nine downswings that excludes, say, the third spell. At the 5% significance level, our sensitivity analysis again indicates that the interest rate spread is statistically significant for downswings and that capacity utilization is statistically significant for upswings.

7.9 Conclusion

Duration analysis has many uses, both in academe and in industry. Consider that in a recent issue of *Business Week* (January 22, 2007) there were two duration applications in separate fields. Hugh Moore of Guerite Advisors (*Business Week*, p. 13) notes that the average amplitude in the fall of housing starts is 51% from peak to trough, and the average amplitude in the fall of housing expenditures as a percentage of GDP is 28% from peak to trough. Housing corrections, or recessions, last an average of 27 months. In the same issue (*Business Week*, p. 62), the global macro-group for Barclays Global Investors (BGI) reports devising a set of signals, or leading indicators, that predict turning points from recession to expansion in various countries. Profits are realized by buying stock and shorting bonds before the recovery is generally recognized. Using leading indicators is similar to using covariates in a duration analysis.

In this chapter we have emphasized classical, nonparametric methods in the duration analysis of unemployment cycles. The nonparametric turning point algorithm from Harding and Pagan (2002), or BBQ, is derived from the classical graphical approch of Burns and Mitchell (1946). The life table analysis follows directly from Cutler and Ederer (1958), and the logit model derives from the seminal work of Cox (1972). These classical techniques are just as relevant today as when first introduced, and have both micro- and macroeconometric applications.

Consider the many economic studies of individual job histories. Adamchik (1999), for example, uses the nonparametric Kaplan–Meier estimator and the semiparametric proportional-hazards model in her study of the effect of unemployment benefits on re-employment in Poland. In a related study, Bover, Arellano and Bentolila (2002) examine not only unemployment benefits, but also the relationship between the unemployment duration of Spanish men and the business cycle. In the latter study, they employ a logit model with autonomous shift dummies that is closely related to Cox's famous proportional-hazards model. The approach is very flexible since Cox's model is no longer proportional when explanators that vary with time are included in the model. Following the early frontier work of

Heckman and Singer (1984), Bover, Arellano and Bentolila (2002) then extend their logit model to account for unobserved heterogeneity.

In more recent macroeconometrics literature, Mudambi and Taylor (1995), Pagan (1998), and Ohn, Taylor and Pagan (2004) propose discrete-time tests for duration dependence and use bootstrap methods for finite-sample inference. Harding and Pagan (2002, 2006) propose new measures and tests for cycle asymmetries and synchronization. The practical importance of duration analysis for aggregate series is best illustrated by the March 28, 2007, testimony of Federal Reserve Chairman Ben Bernanke to the Joint Economic Committee of Congress. In response to his predecessor Alan Greenspan, who warned that the current expansion could be fizzling out, Bernanke responded:

> I would make a point, I think, which is important, which is there seems to be a sense that expansions die of old age, that after they reach a certain point, then they naturally begin to end. I don't think the evidence really supports that. If we look historically, we see that the periods of expansions have varied considerably. Some have been quite long.

Bernanke thus discounts the notion that expansions exhibit positive duration dependence. This view concurs with that of Ohn, Taylor and Pagan (2004), who fail to reject the constant-hazard assumption for post-World War II expansions but who do find statistically significant evidence of positive duration dependence in pre-World War II expansions. On the other hand, consider that the lack of support for positive duration dependence in the post-war period may be due to the small sample size. The mean duration of post-war expansions is about 50 months with a standard deviation of about 30 months. A mean larger than the standard deviation suggests positive duration dependence.

Finally, in this chapter we have stressed the advantage of a separate analysis of upswings and downswings in unemployment. Indeed, given a downswing in unemployment, aggregate output is always rising, but given an upswing in unemployment, the behavior of output is a coin toss. For a young spell of rising unemployment, an increase in capacity utilization of 1 percentage point increases by sevenfold the probability of a turning point from upswing to downswing. In contrast, for downswings the interest rate spread appears to affect the termination probability, and capacity utilization does not appear to matter.

7.10 Appendix: LIMDEP 7.0 program for jackknifing duration data

```
/*LIMDEP 7.0 PROGRAM FOR JACKKNIFING DURATION DATA, June 2007*/
  READ ; FILE = LUexp.TXT ;          ? DATA IN ASCII FORMAT
                NVAR = 6;            ? NUMBER OF VARIABLES
                NOBS = 129;          ? NUMBER OF OBSERVATIONS
                NAMES = SB,D1,D2,D3,BUS,PHASE $ ? VARIABLE NAMES
/* ADD ";TEMP = TFILE" FOR LARGE DATA SETS.
```

```
SB: SET SB=0 IF THE EXPANSION CONTINUES, AND SET SB=1 IF THE
EXPANSION TERMINATES. THUS, SB=1 SIGNIFIES THE BEGINNING OF A
CONTRACTION. CENSORING IS OBTAINED BY ELIMINATING SOME OF
THE OBSERVATIONS WITH SB=0. TO IMPOSE A MINIMUM DURATION OF ONE
MONTH, ELIMINATE THE FIRST OBSERVATION OF EACH PHASE. TO IMPOSE
A MINIMUM DURATION OF TWO MONTHS, ELIMINATE THE FIRST TWO
OBSERVATIONS OF EACH PHASE, AND SO ON.

D1,D2,D3: AUTONOMOUS CHANGE DUMMY VARIABLES.

PHASE: SET PHASE=1 FOR THE FIRST EXPANSION, PHASE = 2 FOR THE
SECOND EXPANSION, AND SO ON, UP TO PHASE=L FOR THE LAST
EXPANSION, "L".*/

LIST; SB,D1,D2,D3 $

NAMELIST; Z = D1,D2,D3 $ NAMES OF RIGHT-HAND-SIDE VARIABLES

REGRESS ; LHS = SB ; RHS = Z ; PDS=5 $ ORDINARY LEAST SQUARES
LOGIT   ; LHS = SB;  RHS = Z         $ LOGISTIC REGRESSION

CALC ; MP = MAX(PHASE) $ CALCULATE THE NUMBER OF PHASES

PROC

    SAMPLE; ALL$
    REJECT ; PHASE = I $ ELIMINATE THE ITH PHASE.

    REGRESS; LHS = SB; RHS = Z ; PDS = 5 $
    LOGIT  ; LHS = SB; RHS = Z $

ENDPROC

EXECUTE ; I = 1,MP $
STOP
```

Acknowledgments

We thank Bob Thornton and Terence Mills for useful suggestions. However, we alone are responsible for any shortcomings of the work.

Notes

1. For the reader interested in a concise discussion focused on measuring business cycles, see Harding and Pagan (2008).
2. The NBER URL is http://www.nber.org/cycles.html.
3. The Bry–Boschan (BB) program was originally designed for monthly data. James Engel, Don Harding and Mark Watson have each written or modified GAUSS programs that are similar to BB. The "Q" in the moniker BBQ stands for "quarterly" intervals even though BBQ easily accommodates either monthly or quarterly intervals. Our BBQ program is a derivative of the one written by Don Harding and is available upon request.

4. Teräsvirta (2006) provides an overview of MS and other types of univariate nonlinear time series models.
5. Our list of techniques to mark time is not exhaustive. Boldin (1994) reviews five techniques to mark time for business cycles: the NBER business cycle dating committee; GDP growth rules; the Commerce Department's Bureau of Economic Analysis (BEA) indicators; Stock and Watson's (1989, 1991) indicators; Stock and Watson's (1989, 1991) experimental business cycle indices; and a Markov-switching model for unemployment.
6. A nice introduction to continuous-time duration analysis is Greene (2006, pp. 710–12). Another good overview of duration techniques, discrete or continuous, is the chapter on transition data in the microeconometrics text by Cameron and Trivedi (2005, pp. 573–608).
7. If the length of an economic contraction is influenced by the length of the preceding expansion, *and* vice versa, the assumption of statistical independence is violated. Likewise, an economic contraction caused by an especially bad harvest could behave very differently from a contraction that occurs during the normal operation of modern market economies. These problems can be handled when modeling with covariates. Another way to help ensure homogeneity across spells of expansion and contraction is to segment the time line into distinct sampling periods with the same underlying probability distributions; see, for example, Diebold and Rudebusch (1990).
8. Watson's (1994) pre-war sampling period ranges from roughly 1860 through 1929, and his post-war sampling period ranges roughly from 1947 through 1990.
9. A realization from a so-called life distribution cannot be negative. The term "life distribution" is coined from the study of mortality, where many of these distributions were first employed.
10. Our notation for the time indices slightly abuses notation since there are gaps in the time line for a phase analysis of either contractions or expansions.
11. For instance, Chin, Geweke and Miller (2000) use covariates to capture drift in the hazard function. However, conditional on the x values, their hazard function is constant since it does not explicitly depend on the duration of the phase.
12. The transition probability for the latent state variable in the Durland and McCurdy (1994) Markov-switching model has a logit form. But $\{S_t\}$ is not the same as $\{S_t^*\}$, and there is no reason to consider the latter for the purpose of a duration analysis.
13. Vahid (2006) surveys both parametric and nonparametric methods of uncovering common cycles in multiple series. Our focus is on nonparametric methods that employ the binary variables S_1 and S_2. In contrast, to investigate synchronization in output across the G7 countries, Stock and Watson (2003) focus on the correlation between Δy_1 and Δy_2.
14. As shown by Hamilton (1989), a primary reason for the heteroskedasticity is heterogeneous transition probabilities across contractions and expansions.
15. The series ID is LNS14000000, and is available from the US Department of Labor, Bureau of Labor Statistics (http://stats.bls.gov).
16. Ohn, Taylor and Pagan (2004) refer to imposing a minimum phase jointly with the assumption of duration dependence as the *Markov hypothesis*. An alternative to subtracting the minimum phase is to incorporate the phase restriction into the econometric model. For instance, Harding and Pagan (2003) show that for a two-quarter minimum the base model that includes S_{t-1} is extended by including the variables S_{t-2} and $S_{t-1}S_{t-2}$.
17. Our LIMDEP program for the regression analysis is presented in the appendix (section 7.10).
18. Because of the limited number of post-war expansions in unemployment, it is not possible to include a dummy variable for each possible exit time. Further, the effective range for D_1 is 10–20 since it is not possible for the spell to terminate in the first nine months due to censoring.

References

Adamchik, V. (1999) The effect of unemployment benefits on the probability of re-employment in Poland. *Oxford Bulletin of Economics and Statistics* **61**, 95–108.

Allison, P.D. (1984) *Event History Analysis: Regression for Longitudinal Event Data*. London: Sage Publications, Inc.

Artis, M.J., Z.G. Kontolemis and D.R. Osborn (1997) Business cycles for G7 and European countries. *Journal of Business* **70**, 249–79.

Artis, M.J., H.M. Krolzig and J. Toro (2004) The European business cycle. *Oxford Economic Papers* **56**, 1–44.

Artis, M.J., M. Marcellino, and T. Proietti (2004) Dating business cycles: a methodological contribution with an application to the Euro area. *Oxford Bulletin of Economics and Statistics* **66**, 537–65.

Bennett, D.S. (1999) Parametric models, duration dependence, and time-varying data revisited. *American Journal of Political Science* **43**, 256–70.

Beveridge S. and C.R. Nelson (1981) A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "Business Cycle." *Journal of Monetary Economics* **7**, 151–74.

Boldin, M.D. (1994) Dating turning points in the business cycle. *Journal of Business* **67**, 97–131.

Bonanomi, L., P.A. Gaughan and L.W. Taylor (1998) A statistical methodology for measuring lost profits resulting from a loss of customers. *Journal of Forensic Economics* **11**, 103–13.

Boschan, C. and W.W. Ebanks (1978) The phase-average trend: a new way of measuring growth. In: *1978 Proceedings of the Business and Economic Statistics Section*. Washington, DC: American Statistical Association.

Bover, O., M. Arellano and S. Bentolila (2002) Unemployment duration, benefit duration, and the business cycle. *Economic Journal* **112**, 223–65.

Bry, G. and C. Boschan (1971) *Cyclical Analysis of Times Series: Selected Procedures and Computer Programs*. New York: National Bureau of Economic Research.

Burns, A.F. and W.C. Mitchell (1946) *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cashin, P. and C.J. McDermott (2002) Riding on the sheep's back: examining Australia's dependence on wool exports *Economic Record* **78**, 249–63.

Cashin, P., C.J. McDermott and A. Scott (2002) Booms and slumps in world commodity prices. *Journal of Development Economics* **69**, 277–96.

Chin, D., J. Geweke and P. Miller (2000) *Predicting Turning Points*. Washington, DC: Technical Paper Series, Congressional Budget Office.

Cooley, T.F. and E.C. Prescott (1995) Economic growth and business cycles. In T.F. Cooley (ed.), *Frontiers of Business Cycle Research,* pp. 1–38. Princeton: Princeton University Press.

Cox, D.R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34**, 187–202.

Cutler, S. and F. Ederer (1958) Maximum utilization of the life table in analyzing survival. *Journal of Chronic Disorders* **8**, 699–712.

Davidson, R. and J.G. MacKinnon (2006) Bootstrap methods in econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, pp. 812–38. New York: Palgrave Macmillan.

Diebold, F.X., J.H. Lee and G.C. Weinbach (1994) Regime switching with time-varying transition probabilities. In C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, pp. 283–302. Oxford: Oxford University Press.

Diebold, F.X. and G.D. Rudebusch (1990) A nonparametric investigation of duration dependence in the American business cycle. *Journal of Political Economy* **98**, 596–616.

Diebold, F.X. and G.D. Rudebusch (1991) Turning point prediction with the composite leading index: A real-time analysis. In K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records,* pp. 231–56. Cambridge: Cambridge University Press.

Diebold, F.X., G.D. Rudebusch and D.E. Sichel (1993) Further evidence on business cycle duration dependence. In J.H. Stock and M.W. Watson (eds.), *Business Cycles, Indicators, and Forecasting,* pp. 87–116. Chicago: University of Chicago Press for NBER.

Durland, J.M and T.H. McCurdy (1994) Duration-dependent transitions in a Markov model of US GNP growth. *Journal of Business and Economic Statistics* **12**, 279–88.

Edwards, S., J.G. Biscarri and F.P. de Gracia (2003) Stock market cycles, financial liberalization and volatility. *Journal of International Money and Finance* **22**, 925–55.

Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–65.

Eichengreen, B., A.K. Rose and C. Wyplosz (1995) Exchange rate mayhem: the antecedents and the aftermath of speculative attacks. *Economic Policy* **21**, 251–312.

Estrella, A. and F.S. Mishkin (1998) Predicting U.S. recessions: financial variables as leading indicators. *Review of Economics and Statistics* **80**, 45–61.

Filardo, A.J. (1994) Business-cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics* **12**, 299–308.

Gordin, M.I. (1969) The Central Limit Theorem for stationary processes. *Soviet Math. Dokl.* **10**, 1174–76.

Greene, W. (2006) Censored data and truncated distributions. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume I: Econometric Theory*, pp. 695–734. New York: Palgrave Macmillan.

Hamilton, J.D. (1989) A new approach to the economic analysis of non-stationary time series and the business cycle. *Econometrica* **57**, 357–84.

Hamilton, J.D. (1994) *Time Series Analysis*. Princeton: Princeton University Press.

Hamilton, J.D. (2005) What's real about the business cycle? Working Paper 11161. Cambridge, Mass.: National Bureau of Economic Research.

Hannan, E.J. (1973) Central limit theorems for time series regression. *Z.Wahrsch. Verw. Gebiete* **26**, 157–70.

Harding, D. and A.R. Pagan (2000) Knowing the cycle. In R. Backhouse and A. Salanti (eds.), *Macroeconomics in the Real World.* Oxford: Oxford University Press.

Harding, D. and A.R. Pagan (2002) Dissecting the cycle: A methodological approach. *Journal of Monetary Economics* **49**,365–81.

Harding, D. and A.R. Pagan (2003) A comparison of two business cycle dating methods. *Journal of Economic Dynamics and Control* **27**, 1681–90.

Harding, D. and A.R. Pagan (2005) A suggested framework for classifying modes of cycle research. *Journal of Applied Econometrics* **20**, 151–59.

Harding, D. and A.R. Pagan (2006) Synchronization of cycles. *Journal of Econometrics* **132**, 59–79.

Harding, D. and A.R. Pagan (2007) The econometric analysis of some constructed binary time series. Working Paper. Brisbane, Australia: National Centre for Econometric Research.

Harding, D. and A.R. Pagan (2008) Measuring business cycles. Forthcoming in S. Durlauf and L. Blume (eds.), *The New Palgrave Dictionary of Economics* (second edition).

Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* New York: Cambridge University Press.

Heckman, J. and B. Singer (1984) A method for minimizing the distributional assumptions in econometric models for duration data. *Econometrica* **52**, 271–320.

Hodrick, R.J. and E.C. Prescott (1997) Postwar U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking* **29**, 1–16.
Hoel, P.G. (1954) *Introduction to Mathematical Statistics*. New York: John Wiley and Sons.
Hollander, M. and R. Proschan (1975) Tests for the residual life. *Biometrika* **62**, 585–93.
Hollander, M. and D.A. Wolfe (1999) Life distributions and survival analysis. In M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods,* pp. 495–765. New York: John Wiley and Sons.
Hoover, K.D. (2006) The methodology of econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics Volume 1: Econometric Theory*, pp. 61–87. New York: Palgrave Macmillan.
Horowitz, J. (2001) The bootstrap in econometrics. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics, Volume 5*, pp. 3159–228. Amsterdam: Elsevier Science B.V.
Ibbotson, R.G, J.L. Sindelar and J.R. Ritter (1994) The market's problems with the pricing of initial paper. *Journal of Applied Corporate Finance* **7**, 66–74.
Jensen, M.J. and M. Liu (2006) Do long swings in the business cycle lead to strong persistence in output? *Journal of Monetary Economics*, **53**, 597–611.
Kedem, B. (1980) *Binary Time Series*. New York: Marcel Dekker.
King, R.G. and C.I. Plosser (1994) Real business cycles and the test of the Adelmans. *Journal of Monetary Economics* **33**, 405–38.
King, R.G. and S.T. Rebelo (1993) Low frequency filtering and real business cycles. *Journal of Economic Dynamics and Control* **17**, 207–31.
LIMDEP (1995) Version 7.0, Econometric Software, Inc.
Lunde, A. and A. Timmermann (2004) Duration dependence in stock prices: an analysis of bull and bear markets. *Journal of Economics and Business Statistics* **22**, 253–73.
Macheu, J. and T. McCurdy (2000) Identifying bull and bear markets. *Journal of Business and Economic Statistics* **18**, 100–12.
Mills, T.C. (2001) Business cycle asymmetry and duration dependence: an international perspective. *Journal of Applied Statistics* **28**, 713–24.
Mudambi, R. and L.W. Taylor (1991) A nonparametric investigation of duration dependence in the American business cycle: a note. *Journal of Political Economy* **99**, 654–56.
Mudambi, R. and L.W. Taylor (1995) Some nonparametric tests for duration dependence: an application to UK business cycle data. *Journal of Applied Statistics* **22**, 163–77.
Mudholkar, G.S., D.K. Srivastava and G.D. Kollia (1996) A generalization of the Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association* **91**, 1575–83.
Neftci, S.N. (1984) Are economic time series asymmetric over the business cycle? *Journal of Political Economy* **92**, 307–28.
Ohn, J., L.W. Taylor and A.R. Pagan (2004) Testing for duration dependence in economic cycles. *Econometrics Journal* **7**, 528–49.
Pagan, A.R. (1997) Policy, theory and the cycle. *Oxford Review of Economic Policy* **13**, 19–33.
Pagan, A.R. (1998) Bulls and bears: a tale of two states. Walras-Bowley Lecture. Montreal: Econometric Society.
Pagan, A.R. (2004) Some econometric analysis of constructed time series. Invited paper. Toronto: Canadian Econometric Group Meeting.
Pagan, A.R. and K. Sossounov (2003) A simple framework for analysing bull and bear markets. *Journal of Applied Econometrics* **18**, 23–46.
Pesaran, M.H. and S. Potter (1997) A floor and ceiling model of U.S. output. *Journal of Economic Dynamics and Control* **21**, pp. 661–95.
Pesaran, M.H. and A. Timmermann (1992) A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics* **10**, 461–5.
Rotemberg, J.J. (1999) A heuristic method for extracting smooth trends from economic time series. Working Paper 7439. New York: National Bureau of Economic Research.

Shapiro, S.S. and M.B. Wilk (1972) An analysis of variance test for the exponential distribution (complete samples). *Technometrics* **14**, 335–70.
Sichel, D.E. (1991) Business cycle duration dependence: a nonparametric approach. *Review of Economics and Statistics* **73**, 254–60.
Sichel, D.E. (1994) Inventories and the three phases of the business cycle. *Journal of Business and Economic Statistics* **12**, 269–77.
Spanos, A. (1995) On theory testing in econometrics: modelling with nonexperimental data. *Journal of Econometrics* **67**, 189–226.
Spanos, A. (2006) Econometrics in retrospect and prospect. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, pp. 3–58. New York: Palgrave Macmillan.
Stock, J.H. and M.W. Watson (1989) New indexes of leading and coincidental economic indicators. In O. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual – 1989*, pp. 351–94. Cambridge, Mass.: MIT Press.
Stock, J.H. and M.W. Watson (1991) A probability model of coincident economic indicators. In K Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge: Cambridge University Press.
Stock, J.H. and M.W. Watson (2003) Has the business cycle changed? Evidence and explanations. Jackson Hole, Wyoming, Federal Reserve Bank of Kansas City symposium, "Monetary Policy and Uncertainty," August, pp. 28–30.
Taylor, L.W. (2007) Estimating duration dependence in militarized interstate disputes. *Journal of Applied Statistics* **34**, pp. 423–41.
Teräsvirta, T. (2006) Univariate nonlinear time series models. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*, pp. 396–424. New York: Palgrave Macmillan.
Thompson, W.A., Jr. (1977) On the treatment of grouped observations in life studies. *Biometrics* **33**, pp. 463–70.
Vahid, F. (2006) Common cycles. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory,* pp. 610–30. New York: Palgrave Macmillan.
Watson, M.W. (1994) Business-cycle durations and postwar stabilization of the US Economy. *American Economic Review* **84**, 24–46.
White, H. (1984) *Asymptotic Theory for Econometricians*. New York: Academic Press, Inc.
Zarnowitz, V. and A. Ozyildirim (2006) Time series decomposition and measurement of business cycles, trends and growth cycles, *Journal of Monetary Economics* **53**, 1717–39.
Zorn, C.J. (2000) Modeling duration dependence. *Political Analysis* **8**, 367–80.
Zuehlke, T.W. (2003) Business cycle duration dependence reconsidered. *Journal of Business and Economic Statistics* **21**, 564–69.

8

The Long Swings Puzzle: What the Data Tell When Allowed to Speak Freely

Katarina Juselius

Abstract

The persistent movements away from long-run benchmark values in real exchange rates that are often observed in many real exchange rates during periods of currency float have been subject to much empirical and theoretical research without resolving the underlying puzzle. This chapter demonstrates how the cointegrated VAR approach of grouping together components of similar persistence can be used to uncover structures in the data that ultimately may help to explain theoretically the forces underlying such puzzling movements. The characterization of the data into components which are empirically $I(0)$, $I(1)$ and $I(2)$ is shown to be a powerful organizing principle, allowing us to structure the data into long-run, medium-run, and short-run behavior. Its main advantage is the ability to associate persistent movements away from fundamental benchmark values in one variable/relation with similar persistent movements somewhere else in the economy.

8.1 Introduction

International macroeconomics is known for a number of empirical puzzles, the most notable among them being the "PPP (purchasing power parity) puzzle," which is closely related to the "long swings puzzle" and the "exchange disconnect puzzle" (Rogoff, 1996). These puzzles are all related to the pronounced persistence away from equilibrium states that have been observed in many real exchange rates during periods of currency float. Among these, the Dmk–$ rate in the post-Bretton Woods period is one of the more extreme cases.

One important purpose of this chapter is to demonstrate how the cointegrated vector autoregressive (CVAR) approach (Johansen, 1995; Juselius, 2006) can be used to uncover structures in the data that ultimately may help to explain theoretically the forces underlying such persistent movements in the data. The CVAR approach starts from a general unrestricted VAR model that gives a good characterization of the raw data. It then tests down until a parsimonious representation of the data with as much economic content as possible has been achieved. When properly applied, the CVAR is able to extract valuable information about the dynamics of the pulling and pushing forces in the data without distorting this information. This entails the identification of stationary relationships between non-stationary variables, interpretable as long-run equilibrium states, and the dynamic adjustment of the system to deviations from these states. It also entails the identification of the transitory and permanent shocks that have affected the variables and the short- and long-run impacts of these shocks.

For the results to be reliable, the statistical properties of the model have, however, to be taken seriously. This implies adequately controlling for reforms, interventions, regime changes, etc., that are often part of the data-generating mechanism. The reunification of East and West Germany is an example of such an important event. The approach also entails the untying of any transformation of the variables, such as the real exchange rate transformation, imposed from the outset on the data. Such transformations, common in empirical economics, often seriously distort signals in the data that otherwise might help to uncover precisely those empirical regularities which give a clue to the underlying reasons for the puzzling behavior.

The weight of the empirical analysis is on characterizing data within the broad framework of a theory model. To facilitate the interpretation of the empirical results, the chapter argues that it is essential first to translate the underlying assumptions of the theoretical model into hypotheses on the pulling and pushing forces of the VAR model (Juselius and Johansen, 2006; Juselius, 2006; Juselius and Franchi, 2007). A careful formulation of such a scenario is indispensable for

being able to structure and interpret the empirical results so that empirical regularities either supporting or rejecting the theoretical assumptions become visible. In particular, the latter are valuable as they should ultimately lead to empirically more relevant theory models. Thus, to some extent, the CVAR approach switches the role of theory and statistical analysis in the sense of rejecting the privileging of *a priori* economic theory over empirical evidence. In the language of the CVAR approach, empirical evidence is the pushing force and economic theory is adjusting (Hoover, Johansen and Juselius, 2008).

The approach will be illustrated with an empirical analysis of the long swings in real exchange rates based on German and US prices and the Dmk–$ rate over the period 1975:09–1998:12. Using the above decomposition into pulling and pushing forces, the empirical analysis identifies a number of "structured" (rather than stylized) facts describing important empirical regularities underlying the long swings puzzle. These provide clues suggesting where to dig deeper (see Hoover, 2006) to gain an empirically more relevant understanding of the puzzling behavior in the goods and foreign exchange markets.

To structure the data as efficiently as possible, this chapter argues that the order of integration, rather than being regarded as a structural parameter, should be considered an empirical approximation, measuring the degree of persistent behavior in a variable or a relation. Organizing the data into directions where they are *empirically* $I(0)$, $I(1)$ or $I(2)$ is not the same as claiming they are *structurally* $I(0)$, $I(1)$ or $I(2)$. In the first case, some implications of the statistical theory of integrated processes are likely to work very well, such as inference on structures; others are likely to work less well, such as inference on the long-run values towards which the process converges when all the errors have been switched off. The focus of this chapter is on structure rather than long-run values (Johansen, 2005).

The statistical analysis suggested that the two prices (and possibly even the nominal exchange rate) were empirically $I(2)$. Thus another important aim of this chapter is to discuss the $I(2)$ model, how it relates to the $I(1)$ model, and what can be gained by interpreting the empirical reality within the rich structure of the $I(2)$ model. Because the $I(2)$ model is also more complex, the analysis is first done within the $I(1)$ model, emphasizing those signals in the results suggesting data are $I(2)$. Though most of the $I(1)$ results can be found in the $I(2)$ model, the chapter demonstrates that the $I(2)$ results are more precise and that the $I(2)$ structure allows for a far richer interpretation.

The exposition of the chapter is as follows. Section 8.2 defines the $I(1)$ and $I(2)$ models as parameter restrictions on the unrestricted VAR. Section 8.3 introduces the persistent features of the real exchange rate data for the German–US case and discusses how they can be formulated as the pulling and pushing forces of a CVAR model. Section 8.4 discusses under which conditions $I(2)$ data can be modeled with the $I(1)$ model, why it works, and how the interpretation of the results has to be modified. Section 8.5 presents the empirical $I(1)$ analysis of prices and nominal exchange rates inclusive of specification testing and estimation of the long-run structure. Section 8.6 gives a brief account of the $I(2)$ model and discusses at some

length the specification of the deterministic components. Section 8.7 discusses an estimation procedure based on maximum likelihood and shows how the $I(2)$ structure can be linked to the $I(1)$ model. Section 8.8 provides hypothetical scenarios for the real exchange rate data. Section 8.9 presents the empirical results of the pulling and pushing forces structured by the $I(2)$ model, summarizes the puzzling facts detected, and discusses what has been gained by this analysis compared to the $I(1)$ analysis. Section 8.10 concludes with a discussion of what the data were able to tell when allowed to speak freely.

8.2 The VAR model

The baseline VAR(2) model in its unrestricted form is given by:

$$\mathbf{x}_t = \mathbf{\Pi}_1 \mathbf{x}_{t-1} + \mathbf{\Pi}_2 \mathbf{x}_{t-2} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{8.1}$$

with:

$$\boldsymbol{\varepsilon}_t \sim N_p(\mathbf{0}, \mathbf{\Omega}),\ t = 1, \ldots, T,$$

where $\mathbf{x}_t' = [x_{1,t}, x_{2,t}, \ldots x_{p,t}]$ is a vector of p stochastic variables and $\mathbf{D}_t$ is a vector of deterministic variables, such as a constant, trend and various dummy variables. As the subsequent empirical VAR model has lag two, all results are given for the VAR(2) model. A generalization to higher lags should be straightforward.

In terms of likelihood, an equivalent formulation of (8.1) is the vector equilibrium correction form:

$$\Delta \mathbf{x}_t = \mathbf{\Gamma}_1 \Delta \mathbf{x}_{t-1} + \mathbf{\Pi} \mathbf{x}_{t-1} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{8.2}$$

where $\mathbf{\Gamma}_1 = -\mathbf{\Pi}_2$ and $\mathbf{\Pi} = -(\mathbf{I} - \mathbf{\Pi}_1 - \mathbf{\Pi}_2)$.

Alternatively, (8.1) can be formulated in acceleration rates, changes and levels:

$$\Delta^2 \mathbf{x}_t = \mathbf{\Gamma} \Delta \mathbf{x}_{t-1} + \mathbf{\Pi} \mathbf{x}_{t-1} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{8.3}$$

where $\mathbf{\Gamma} = -(\mathbf{I} - \mathbf{\Gamma}_1)$. As long as all parameters are unrestricted, the VAR model is no more than a convenient summary of the covariances of the data. As a result, most VAR models are heavily overparameterized and insignificant parameters need to be set to zero. The idea of general-to-specific modeling is to reduce the number of parameters by significance testing, with the final aim of finding a parsimonious parameterization with interpretable economic contents. Provided that the simplification search is statistically valid, the final restricted model will reflect the full information of the data. Thus, given the broad framework of a theory model, a correct CVAR analysis allows the data to speak freely about the underlying mechanisms that have generated the data.

All three models are equivalent from a likelihood point of view, but (8.1) would generally be chosen when $\mathbf{x}_t$ is $I(0)$, (8.2) when $\mathbf{x}_t$ is $I(1)$, and (8.3) when $\boldsymbol{x}_t$ is $I(2)$.

The hypothesis that $\mathbf{x}_t$ is $I(1)$ is formulated as a reduced rank restriction on the matrix $\mathbf{\Pi}$:

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}', \text{ where } \boldsymbol{\alpha}, \boldsymbol{\beta} \text{ are } p \times r, \tag{8.4}$$

and that $\mathbf{x}_t$ is $I(2)$ as an additional reduced rank restriction on the transformed matrix $\mathbf{\Gamma}$:

$$\boldsymbol{\alpha}_{\perp}'\mathbf{\Gamma}\boldsymbol{\beta}_{\perp} = \boldsymbol{\xi}\boldsymbol{\eta}', \text{ where } \boldsymbol{\xi}, \boldsymbol{\eta} \text{ are } (p-r) \times s_1, \tag{8.5}$$

where $\boldsymbol{\beta}_{\perp}, \boldsymbol{\alpha}_{\perp}$ are the orthogonal complements of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The first reduced rank condition is formulated on the variables in levels, the second on the variables in differences. Condition (8.4) tells us that the variables contain stochastic trends (unit roots) that can be canceled by linear combinations. Condition (8.5) tells us that the differenced process also contains unit roots when data are $I(2)$. However, in this case the linear combinations that cancel these roots are more complicated. Thus, when $\mathbf{x}_t \sim I(2)$, and hence $\Delta\mathbf{x}_t \sim I(1)$, it is not sufficient to impose the reduced rank restriction on the matrix $\mathbf{\Pi}$ to get rid of all (near) unit roots in the model. This is because $\Delta\mathbf{x}_t$ is also a unit root process and lowering the value of r does not remove the unit roots belonging to $\mathbf{\Gamma} = -\left(\mathbf{I} - \mathbf{\Gamma}_1\right)$. Therefore, even though the rank of $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ has been correctly determined, there will remain additional unit roots in the VAR model when the data are $I(2)$. As will be demonstrated below, this provides a good diagnostic tool for detecting $I(2)$ problems in the VAR analysis.

Inverting the VAR model gives us the moving average (MA) form. Under the reduced rank of (8.4) and the full rank of (8.5), the MA form is given by:

$$\mathbf{x}_t = \mathbf{C}\sum_{i=1}^{t}(\boldsymbol{\varepsilon}_i + \mathbf{\Phi}\boldsymbol{D}_i) + \mathbf{C}^*(L)(\boldsymbol{\varepsilon}_t + \mathbf{\Phi}\mathbf{D}_t) + \mathbf{A}, \tag{8.6}$$

where $\mathbf{C}^*(L)$ is a lag polynomial describing the impulse response functions of the empirical shocks to the system, $\mathbf{A}$ is a function of the initial values $\mathbf{x}_0, \mathbf{x}_{-1}, \mathbf{x}_{-2}$, and $\mathbf{C}$ is of reduced rank $p-r$:

$$\mathbf{C} = \boldsymbol{\beta}_{\perp}(\boldsymbol{\alpha}_{\perp}'\mathbf{\Gamma}\boldsymbol{\beta}_{\perp})^{-1}\boldsymbol{\alpha}_{\perp}' = \tilde{\boldsymbol{\beta}}_{\perp}\alpha_{\perp}', \tag{8.7}$$

with $\tilde{\boldsymbol{\beta}}_{\perp} = \boldsymbol{\beta}_{\perp}(\boldsymbol{\alpha}_{\perp}'\boldsymbol{\beta}_{\perp})^{-1}$.

Inverting the VAR under the reduced rank of both (8.4) and (8.5) will be discussed in section 8.6.

8.3 The persistent movements in real exchange rate data

Parity conditions are central to international finance and, more specifically, to many open economy macro-models, such as the Dornbusch (1976) sticky-price overshooting model with rational expectations (RE). One important implication of this model and its modifications is that PPP should hold as an equilibrium cointegrating relationship (see Frydman *et al.*, 2008, and references therein). The idea here is to formulate the PPP condition under a currency float into testable hypotheses

on the pushing and pulling forces of the cointegrated VAR model. Comparing assumed with actual behavior is then likely to pinpoint the empirical mechanisms underlying the puzzling behavior. Since the VAR model is just a reformulation of the covariance information in the data, the end results should be a set of empirical features which a theory model should be able to replicate in order to claim empirical relevance.

8.3.1 The long swings puzzle

PPP is defined as:

$$p_1 = p_2 + s_{12}, \tag{8.8}$$

where p_1 is the log of the domestic price level (here German), p_2 is the log of the foreign price level (here US), and s_{12} denotes the log of the spot exchange rate (here Dmk–$). Thus, the departure at time t from (8.8) is given by:

$$ppp_t = p_{1,t} - p_{2,t} - s_{12,t}.^{1} \tag{8.9}$$

An ocular inspection gives a first impression of the development over time of prices and the nominal exchange rate and illustrates what the puzzle is all about. Figure 8.1 (upper panel) shows that US prices have grown more than German prices, resulting in a downward sloping stochastic trend in relative prices. According to purchasing power parity, the nominal exchange rate should reflect this downward sloping trend. The figure shows that this is also approximately the case over the very long run. However, what is striking are the long swings around that downward sloping trend.

How can we use econometrics to learn about the mechanisms underlying these swings? The subsequent VAR analyis will demonstrate that the joint modeling of prices and exchange rates allows us to formulate much richer hypotheses about the empirical mechanisms behind the puzzle.

8.3.2 Pulling and pushing forces in the cointegrated VAR model

To provide the intuition for the VAR approach and to show how the results can be interpreted in terms of pulling and pushing forces, a hypothetical VAR analysis of the German–US PPP data will be used as an illustration. For simplicity, the discussion will be restricted to a bivariate $I(1)$ model for relative prices and the nominal exchange rate. Because the period of interest defines a currency float, a prior hypothesis is that the nominal exchange rate has been adjusting and prices pushing. Provided that the stochastic trend in nominal exchange rates reflects the stochastic trend in relative prices, it is easy to show that $ppp = p_1 - p_2 - s_{12} \sim I(0)$. Thus the stationarity of PPP and its adjustment dynamics can be formulated as a composite hypothesis: $(p_1 - p_2) = pp \sim I(1)$, $s_{12} \sim I(1)$, $ppp \sim I(0)$, s_{12} is adjusting, and p_1, p_2 are pushing.

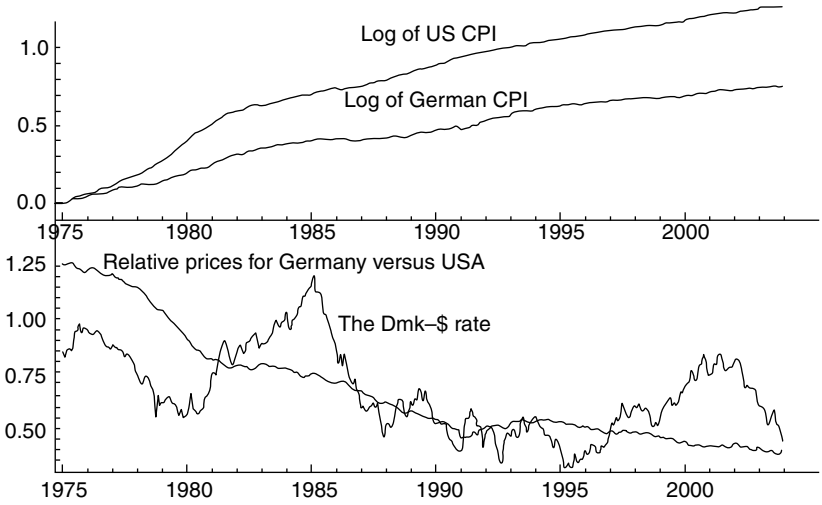


Figure 8.1 Time graphs of German and US prices (upper panel) and their relative prices and nominal exchange rate (lower panel)

The pulling forces are described by the vector equilibrium correction model:

$$\begin{bmatrix} \Delta pp_t \\ \Delta s_{12,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} (pp_{t-1} - s_{12,t-1} - \beta_0) + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix},$$

where $(pp_t - s_{12,t} - \beta_0) = \boldsymbol{\beta}' \mathbf{x}_t$ is the cointegration relation with $E(ppp_t) = \beta_0$. Thus an equilibrium position, defined as $pp_t - s_{12,t} = \beta_0$, can be given an interpretation as a resting point towards which the process is drawn after it has been pushed away. In this sense, an equilibrium position exists at all time points, t, contrary to the long-run value of the process, which is the value of the process in the limit as $t \to \infty$ and all shocks have been switched off.

The pushing forces are described by the corresponding common trends model:

$$\begin{bmatrix} pp_t \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} c \\ c \end{bmatrix} \boldsymbol{\alpha}'_{\perp} \sum_{i=1}^{t} \varepsilon_i + \mathbf{C}^*(L) \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix},$$

with $\boldsymbol{\alpha}'_{\perp} = \frac{1}{\alpha_1 - \alpha_2}[-\alpha_2, \alpha_1]$ and with $\boldsymbol{\alpha}'_{\perp} \sum_{i=1}^{t} \boldsymbol{\varepsilon}_i$ describing the common stochastic trend. Assume now that $\alpha' = [0, \alpha_2]$, i.e., only the nominal exchange rate is equilibrium correcting when $ppp_t - \beta_0 \neq 0$. In this case $\boldsymbol{\alpha}'_{\perp} = [1, 0]$ implies that the common stochastic trend originates from relative price shocks. This would conform to the theoretical prior for a period of floating exchange rates.

The question is now whether the empirical reality given by the observed variables in Figure 8.1 (lower panel) can be adequately represented by the above assumed pulling and pushing forces. Stationarity of ppp_t should imply that the nominal exchange rate would follow relative prices one-for-one apart from stationary noise. Figure 8.2 shows a cross-plot of the pp_t and $s_{12,t}$ variables. If the assumption that

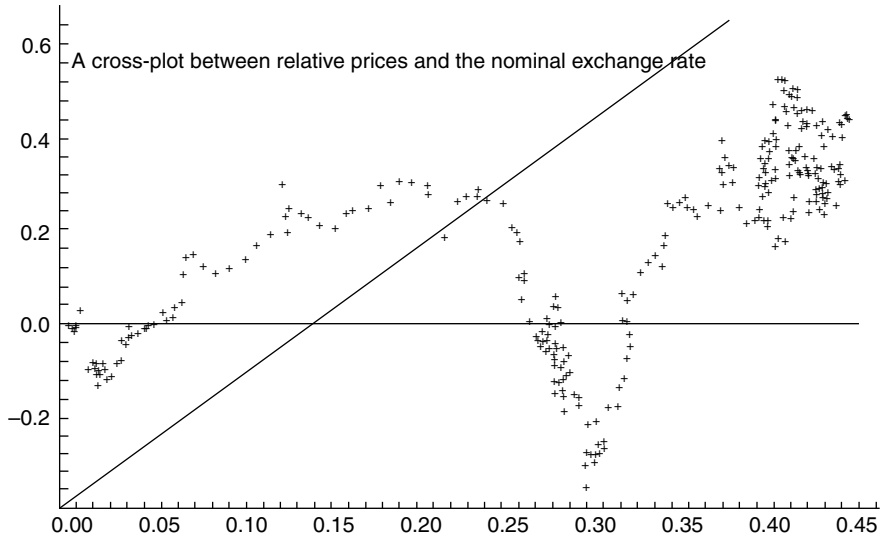


Figure 8.2 A cross-plot of US–German relative prices and the Dmk–$ rate for the period 1975:4–1998:12

$ppp_t \sim I(0)$ were correct, then the cross-plots should be randomly scattered around the 45° line defining the equilibrium position $pp_t = s_{12,t}$. Obviously, the cross-plots measuring the deviation from *ppp*, i.e., $\boldsymbol{\beta}'\mathbf{x}_t = pp_t - s_{12,t} - \beta_0$, are systematically scattered either above or below the 45° line. Thus the reality behind the observed real exchange rate looks very different from the assumed stationary PPP illustrating the puzzle. The non-stationarity of real exchange rates has been demonstrated in a number of studies (see Froot and Rogoff, 1995, and MacDonald, 1995, for surveys; Cheung and Lai, 1993; Juselius, 1995; Johansen and Juselius, 1992).

8.3.3 Approximating persistent behavior with $I(1)$ or $I(2)$

The above ocular analysis showed that the long swings puzzle is essentially a question of why nominal exchange rates have so persistently moved away from relative prices. The previous sub-section suggested that the cointegrated VAR model should be used to structure such data by the pulling and pushing forces. Section 8.3 defined the $I(1)$ and $I(2)$ models as reduced rank parameter restrictions on the $I(0)$ model, providing us with an empirically strong procedure for addressing behavioral macroeconomic problems. This is because the reduced rank parameterization of the CVAR allows us to group together components of similar persistence over the sample period. The characterization of the data into *empirically* $I(0)$, $I(1)$ and $I(2)$ components is a powerful organizing principle, allowing us to structure the data into long-run, medium-run and short-run behavior. An additional advantage is that inference is likely to become more robust than otherwise. For example, treating a near unit root as stationary tends to invalidate certain inferences based on the χ^2, F and t distributions unless we have a very long sample.[2]

This is a fairly pragmatic way of classifying data that allow a variable to be treated as $I(1)$ in one sample and $I(0)$ or even $I(2)$ in another. The idea is that, in a general equilibrium world, a persistent departure from a steady-state value of a variable or a relation should generate a similar persistent movement somewhere else in the economy. For example, if the Fisher parity holds as a stationary relation (stationary real interest rates) and we find that inflationary shocks have been very persistent, then we should expect interest rate shocks to have a similar persistence. Thus empirical persistence is a powerful property that can be used to investigate whether our prior hypothesis (the Fisher parity) is empirically relevant, and if not, which other variables have been co-moving in a similar manner, giving rise to new hypotheses.

From the outset, many economists would consider the idea that economic variables are $I(2)$ highly problematic. The argument is often that all inference on long-run values (the steady-state value a variable converges to when the errors are switched off) would lead to meaningless results. This is a valid argument provided one can argue that the order of integration is a structural parameter, which often seems doubtful. Nonetheless, there are cases when a structural interpretation is warranted. For example, Frydman *et al.* (2008) show that speculative behavior based on IKE is consistent with near $I(2)$ behavior; arbitrage theory suggests that a nominal market interest rate should be a martingale difference process, i.e., approximately a unit root process. Of course, in such cases a structural unit root should be invariant to the choice of sample period.

8.4 Modeling $I(2)$ data with the $I(1)$ model: does it work?

It often happens that $I(2)$ data are analyzed as if they were $I(1)$ because the $I(2)$ possibility was never checked, or one might have realized that the data exhibit $I(2)$ features but decided to ignore these signals in the data. For this reason, it is of some interest to ask whether the findings from such $I(1)$ analyses are totally useless, misleading, or can be trusted to some extent.

Before answering these questions, it is useful to examine the so-called **R**-model, in which short-run effects have been concentrated out. We consider first the simple VAR(2) model:

$$\begin{aligned}\Delta\mathbf{x}_t &= \boldsymbol{\Gamma}_1\Delta\mathbf{x}_{t-1} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &\sim N_p(\mathbf{0},\boldsymbol{\Omega}),\ t = 1,\ldots,T,\end{aligned} \tag{8.10}$$

and the corresponding **R**-model:

$$\mathbf{R}_{0t} = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{R}_{1t} + \boldsymbol{\varepsilon}_t, \tag{8.11}$$

where $\mathbf{R}_{0t}$ and $\mathbf{R}_{1t}$ are found by concentrating out the lagged short-run effects, $\Delta\mathbf{x}_{t-1}$:

$$\Delta\mathbf{x}_t = \hat{\boldsymbol{B}}_1\Delta\mathbf{x}_{t-1} + \hat{\boldsymbol{\mu}}_0 + \mathbf{R}_{0t}, \tag{8.12}$$

and:

$$\mathbf{x}_{t-1} = \hat{\boldsymbol{B}}_2\Delta\mathbf{x}_{t-1} + \hat{\boldsymbol{\mu}}_0 + \mathbf{R}_{1t}. \tag{8.13}$$

When $\mathbf{x}_t \sim I(2)$, both $\Delta\mathbf{x}_t$ and $\Delta\mathbf{x}_{t-1}$ contain a common $I(1)$ trend, which therefore cancels in the regression of one on the other, as in (8.12). Thus, $\mathbf{R}_{0t} \sim I(0)$ even if $\Delta\mathbf{x}_t \sim I(1)$. On the other hand, an $I(2)$ trend cannot be cancelled by regressing on an $I(1)$ trend and regressing $\mathbf{x}_{t-1}$ on $\Delta\mathbf{x}_{t-1}$ as in (8.13) does not cancel the $I(2)$ trend, so $\mathbf{R}_{1t} \sim I(2)$. Because $\mathbf{R}_{0t} \sim I(0)$ and $\boldsymbol{\varepsilon}_t \sim I(0)$, equation (8.11) can only hold if $\boldsymbol{\beta} = \mathbf{0}$ or, alternatively, if $\boldsymbol{\beta}'\mathbf{R}_{1t} \sim I(0)$. Thus, unless the rank is zero, the linear combination $\boldsymbol{\beta}'\mathbf{R}_{1t}$ transforms the process from $I(2)$ to $I(0)$.

The connection between $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ and $\boldsymbol{\beta}'\mathbf{R}_{1t}$ can be seen by inserting (8.13) into (8.11):

$$\begin{aligned}
\underbrace{\mathbf{R}_{0t}}_{I(0)} &= \boldsymbol{\alpha}\boldsymbol{\beta}'(\underbrace{\mathbf{x}_{t-1}}_{I(2)} - \mathbf{B}_2\underbrace{\Delta\mathbf{x}_{t-1}}_{I(1)} - \hat{\boldsymbol{\mu}}_0) + \boldsymbol{\varepsilon}_t \\
&= \boldsymbol{\alpha}(\underbrace{\boldsymbol{\beta}'\mathbf{x}_{t-1}}_{I(1)} - \boldsymbol{\beta}'\underbrace{\mathbf{B}_2\Delta\mathbf{x}_{t-1}}_{I(1)} - \boldsymbol{\beta}'\hat{\boldsymbol{\mu}}_0) + \boldsymbol{\varepsilon}_t \\
&= \boldsymbol{\alpha}(\underbrace{\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\omega}'\Delta\mathbf{x}_{t-1}}_{I(0)} - \boldsymbol{\beta}'\hat{\boldsymbol{\mu}}_0) + \boldsymbol{\varepsilon}_t, \qquad (8.14)
\end{aligned}$$

where $\boldsymbol{\omega}' = \boldsymbol{\beta}'\mathbf{B}_2$. It is now easy to see that the stationary relations $\boldsymbol{\beta}'\mathbf{R}_{1t}$ consist of two components, $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ and $\boldsymbol{\omega}'\Delta\mathbf{x}_{t-1}$. There are two possibilities:

1. $\boldsymbol{\beta}_i'\mathbf{x}_{t-1} \sim I(0)$ and $\boldsymbol{\omega}_i = \mathbf{0}$, where $\boldsymbol{\beta}_i$ and $\boldsymbol{\omega}_i$ denote the ith column of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, or
2. $\boldsymbol{\beta}_i'\mathbf{x}_{t-1} \sim I(1)$ cointegrates with $\boldsymbol{\omega}_i'\Delta\mathbf{x}_{t-1} \sim I(1)$ to produce the stationary relation $\boldsymbol{\beta}'\mathbf{R}_{1t} \sim I(0)$.

In the first case, we talk about directly stationary relations; in the second case, about polynomially cointegrated relations. Here we shall consider $\boldsymbol{\beta}'\mathbf{x}_t \sim I(1)$ without distinguishing between the two cases, albeit recognizing that some of the cointegration relations $\boldsymbol{\beta}'\mathbf{x}_t$ may be stationary by themselves.

We have demonstrated above that $\mathbf{R}_{0t} \sim I(0)$ and $\boldsymbol{\beta}'\mathbf{R}_{1t} \sim I(0)$ in (8.11), which is the model on which all $I(1)$ estimation and test procedures are derived. This means that the $I(1)$ procedures can be used even though data are $I(2)$, albeit with the following reservations:

1. the $I(1)$ rank test cannot say anything about the reduced rank of the $\boldsymbol{\Gamma}$ matrix, i.e., about the number of $I(2)$ trends. The determination of the reduced rank of the $\boldsymbol{\Pi}$ matrix, though asymptotically unbiased, might have poor small sample properties (Nielsen and Rahbek, 2007)
2. the $\boldsymbol{\beta}$ coefficients relating $I(2)$ variables are T^2 consistent and thus are precisely estimated. We say that the estimate of $\boldsymbol{\beta}$ is super-super consistent
3. the tests of hypotheses on $\boldsymbol{\beta}$ are not tests of cointegration from $I(1)$ to $I(0)$, but instead from $I(2)$ to $I(1)$, as is evident from (8.14), and a cointegration relation should in general be considered $I(1)$, albeit noting that a cointegration relation $\boldsymbol{\beta}_i'\mathbf{x}_t$ can be CI(2,2), i.e., be cointegrating from $I(2)$ to $I(0)$
4. the MA representation is essentially useless, as the once cumulated residuals cannot satisfactorily explain variables containing $I(2)$ trends, i.e., twice cumulated residuals.

Thus, one can test a number of hypotheses based on the $I(1)$ procedure even if $\mathbf{x}_t$ is $I(2)$, but the interpretation of the results has to be modified accordingly.

8.5 An $I(1)$ analysis of prices and exchange rates

8.5.1 Specification

The VAR model is based on the assumption of multivariate normality which, if correct, implies linearity in parameters as well as constancy of parameters. However, multivariate normality is seldom satisfied in a first tentatively estimated VAR model. There are many reasons for this, e.g., omission of relevant variables, inadequate measurements, interventions, reforms, etc. All this may have changed the data-generating mechanisms, thus producing structural breaks or resulting in extraordinary effects on some of the variables. In the present case, the reunification of East and West Germany in 1991:1 was a particularly important institutional event which is likely to have changed some of the properties of the VAR model. For example, Figure 8.1 shows that the nominal exchange rate may have experienced a change in its trending behavior at the reunification, as well as a shift in its level. Therefore, a consequence of merging the less productive East with West Germany is likely to have been a change in relative productivity, which needs to be accounted for by a change in the slopes of the linear trends in the VAR model.

Thus, in order to achieve a well-specified VAR model one usually has to control for major institutional events. Section 8.6.2 will provide a more detailed account of how to specify deterministic components in the $I(2)$ model. For the specification of such events in the $I(1)$ model the reader is referred to Juselius (2006, Ch. 6). Here they will be modeled by a trend with a changing slope at 1991:1 ($t_{91.1}$) and various dummy variables, as explained below:

$$\Delta\mathbf{x}_t=\boldsymbol{\Gamma}_1\Delta\mathbf{x}_{t-1}+\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1}+\boldsymbol{\mu}_0+\boldsymbol{\mu}_{01}D_{s,91.1,t}+\boldsymbol{\mu}_1 t+\boldsymbol{\mu}_{11}t_{91.1}+\boldsymbol{\Phi}_p\mathbf{D}_{p,t}+\boldsymbol{\varepsilon}_t, \quad (8.15)$$

where the sample period is 1975:09–1998:12 and $\mathbf{x}_t' = [p_{1,t}, p_{2,t}, s_{12,t}]$ with:

$p_{1,t}$ = log of German CPI,[3]
$p_{2,t}$ = log of US CPI, and
$s_{12,t}$ = log of the nominal Dmk–$ exchange rate.

The linear terms in (8.15) are defined as:

$\boldsymbol{\mu}_0$ is a vector of constant terms,
$\boldsymbol{\mu}_{01}$ is a vector measuring a change in the constant term at 1991:1,
$\boldsymbol{\mu}_1$ is a vector of linear trend slopes,
$\boldsymbol{\mu}_{11}$ is a vector measuring a change in the trend slope at 1991:1.

The dummy variables are defined as:

$D_p tax = 1$ in 1991:7, 1991:9, and 1993:1, zero otherwise
$Ds91.1_t$ is 1 for $t \geq$ 1991:1, 0 otherwise,

$\mathbf{D}'_{p,t} = [Dp80.7, Dp91.1, D_p tax, D_p 97.7]$ with
$DpXX.y_t = 1$ in 19XX:y, zero otherwise.

The tax dummy is needed to account for a series of commodity tax increases to pay for reunification, and the three dummies are needed to account for a big drop in the US inflation rate in 1980:7, a large change in the nominal exchange rate in 1991:1, and a large change in the Dmk–$ rate in 1997:7.

As discussed in more detail in section 8.6, the two trend components, the constant, and the shift dummy need to be appropriately restricted in the VAR model to avoid quadratic and cubic trends. The dummy variables have been specified exclusively to control for the extraordinary shock at the time of the intervention, but to leave the information of the observation intact through its lagged impact. Thus the dummies do not *remove* the outlying observation as is usually the case in a static regression model. Table 8.1 reports the estimated effects.

Conditional on the dummies, the VAR model becomes reasonably well-specified. The tests for multivariate residual autocorrelation at one lag, $\chi^2(9) = 11.0[0.28]$, and two lags, $\chi^2(9) = 14.2[0.12]$, were acceptable, as were the tests of multivariate ARCH of order one, $\chi^2(36) = 45.9[0.12]$, and order two, $\chi^2(72) = 87.2[0.11]$. However, multivariate normality was rejected based on $\chi^2(6) = 27.1[0.00]$. To get some additional information, Table 8.1 reports the univariate Jarque–Bera tests, as well as skewness (third moment around the mean) and kurtosis (fourth moment around the mean). It appears that the non-normality problems are mostly due to excess kurtosis in the US inflation rate. Since the VAR estimates have been shown to be reasonably robust to moderate deviations from normality due to excess kurtosis (Gonzalo, 1994), the baseline VAR model is considered to be a reasonably adequate characterization of the data.

8.5.2 Rank determination and general model properties

The determination of the cointegration rank is a crucial step in the analysis, as it structures the data into its pulling and pushing components. The so-called trace test (Johansen, 1996) is a likelihood ratio test for the cointegration rank. However, the trace test is derived under the null of $p - r$ unit roots, which does not always correspond to the null of the theory model, as illustrated in section 8.8 (see also

Table 8.1 Estimated outlier effects and misspecification tests

	Estimated outlier effects				*Misspecification tests*		
	$D_p tax$	$Dp80.7$	$D_s 91.1$	$D_p 97.7$	*Norm.*	*Skew.*	*Kurt.*
$\Delta p_{1,t}$	**0.01** [11.36]	−0.00 [−1.40]	0.00 [1.77]	**0.01** [4.15]	7.22[0.03]	0.35	3.62
$\Delta p_{2,t}$	−0.00 [−0.15]	**−0.01** [−4.90]	0.00 [0.16]	0.00 [0.37]	15.4[0.00]	−0.20	4.20
$\Delta s_{12,t}$	−0.02 [−1.04]	0.01 [0.39]	**0.01** [2.57]	**0.06** [1.98]	6.31[0.04]	0.10	3.66

Note: t-ratios in [ ].

Table 8.2 Determination of rank in the $I(1)$ model

r	$p-r$	τ_{p-r}	*4 largest characteristic roots*			
0	3	80.06 [57.9]	1.0	1.0	1.0	0.75
1	2	**32.65** [36.6]	1.0	1.0	0.99	0.53
2	1	**6.72** [18.5]	1.0	0.99	0.99	0.52
3	0		0.99	0.99	0.98	0.53

Note: 95% quantiles in [].

Tests of pushing and pulling variables

	r	p_1	p_2	s_{12}
No levels feedback	1	7.52 [0.01]	16.17 [0.00]	7.58 [0.01]
	2	23.83 [0.00]	32.74 [0.00]	8.66 [0.01]
Pure adjustment	1	21.40 [0.00]	11.26 [0.00]	34.27 [0.00]
	2	**2.74** [0.10]	**1.31** [0.25]	18.74 [0.00]

Note: p-values in [].

Juselius, 2006, Ch. 8). Therefore, the choice of rank suggested by the trace test needs to be checked for its consistency with other information in the model, such as the characteristic roots.

The trace tests reported in Table 8.2 suggest a borderline acceptance of $r = 1$ cointegration relation, and hence $p - r = 2$ common stochastic trends or, alternatively, a strong acceptance of $r = 2$, and hence, $p - r = 1$ common stochastic trend. Thus, from a statistical point of view, both choices can be defended. Section 8.8 will argue that $r = 2$ is the theory consistent choice. To find out which choice is econometrically preferable, we shall check the consistency of $r = 1, 2$ with the characteristic roots in the model and with the mean reversion of the cointegration relations.

An inspection of the characteristic roots of the model shows that there are three large roots of magnitude 0.99 in the unrestricted model. These are generally indistinguishable from unit roots, so the model seems to contain three unit roots. The choice of $r = 1$ leaves one near unit root and the choice of $r = 2$ two near unit roots in the model. Section 8.4 showed that, when one or several large roots remain in the model for any reasonable choice of r, it is a sign of $I(2)$ behavior in at least one of the variables.[4]

To check the consistency of the results with the $I(2)$ model, it is useful to divide the total number of stochastic trends into $I(1)$ and $I(2)$ trends, i.e., $p - r = s_1 + s_2$, where s_1 denotes the number of $I(1)$ trends (unit root processes), and s_2 the number of $I(2)$ trends (double unit root processes). Three (near) unit roots in the model would be consistent with either $\{r = 0, p - r = 3\}$ or $\{r = 1, s_1 = 1, s_2 = 1\}$, whereas $\{r = 2, s_1 = 0, s_2 = 1\}$ corresponds to two unit roots. Since the latter is less than the three near unit roots in the model, the choice $r = 2$ would not be consistent with the empirical information in the data.

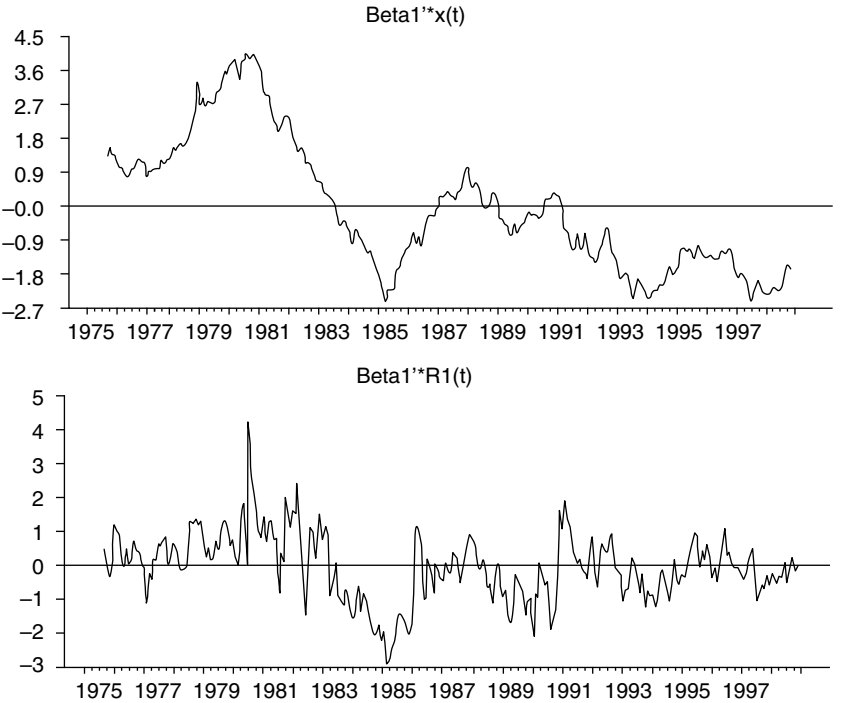


Figure 8.3 The graphs of the first cointegration relation ($\boldsymbol{\beta}'\mathbf{x}_t$ in the upper panel, $\boldsymbol{\beta}'\mathbf{R}_{1,t}$ in the lower panel)

Thus, by imposing $r = 1$, two of the big roots are restricted to unity, but the third would still be unrestricted in the $I(1)$ model, invalidating some of the interpretation of the empirical results discussed in section 8.4. The graphs of the first two cointegration relations, shown in Figures 8.3 and 8.4, illustrate the effect of a near unit root. Based on the graphs, it is difficult to argue that $\boldsymbol{\beta}_i'\mathbf{x}_t$, $i = 1, 2$, is mean-reverting as an equilibrium error should be. However, $\boldsymbol{\beta}_i'\mathbf{R}_{1,t}$ (in the lower panel) looks much more mean-reverting, at least for $r = 1$. This, of course, is exactly in accordance with (8.13). Thus, only $\{r = 1, s_1 = 1, s_2 = 1\}$ seems acceptable based on the characteristic roots of the model and the graphs of the cointegration relations.

It is also useful to investigate the general pulling and pushing properties of the model described by the test of a unit vector in $\boldsymbol{\alpha}$ and a zero row in $\boldsymbol{\alpha}$ (Juselius, 2006, Ch. 11) and how they would be affected by the choice of rank. In the lower part of Table 8.2 the tests of "no levels feedback" (a zero row in $\boldsymbol{\alpha}$) and "pure adjustment" (a unit vector in $\boldsymbol{\alpha}$) are reported for $r = 1$ and $r = 2$. For $r = 1$, none of the variables are found to purely pushing or pulling. For $r = 2$, there is some evidence that the two prices are exclusively adjusting (though the hypothesis that they are jointly adjusting is rejected). Altogether, the empirical evidence suggests that prices are

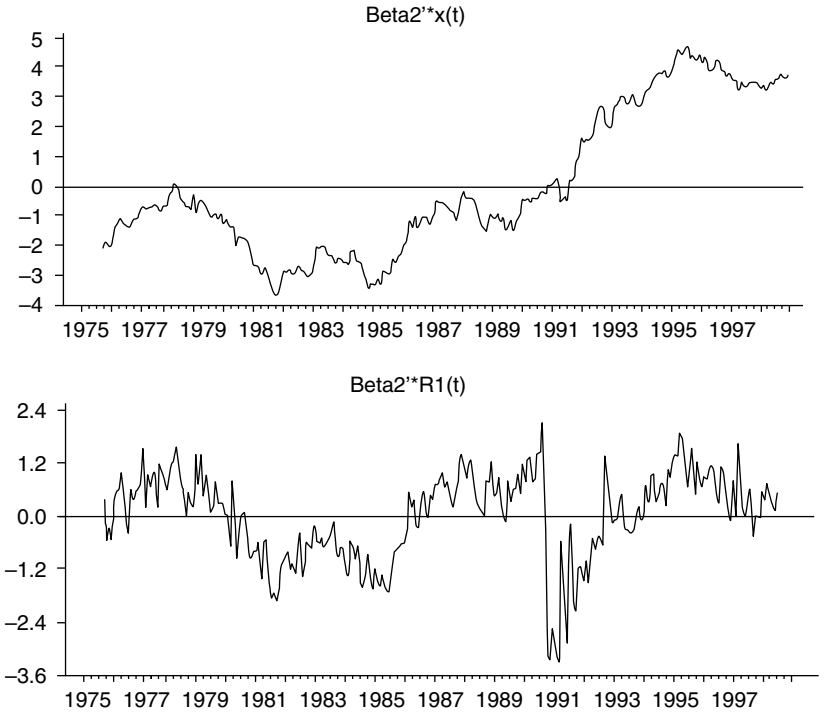


Figure 8.4 The graphs of the second cointegration relation ($\boldsymbol{\beta}'\mathbf{x}_t$ in the upper panel, $\boldsymbol{\beta}'\mathbf{R}_{1,t}$ in the lower panel)

"more" pulling than pushing, which is an interesting observation as one would expect the opposite during a currency float.

8.5.3 Estimating the long-run structure

Table 8.3 reports the estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}_1$ and $\boldsymbol{\Phi}$ for the choice of $r = 1$. The estimated $\boldsymbol{\beta}$ relation suggests that $p_{1,t}$ and $p_{2,t}$ are almost homogeneously related. Testing the hypothesis gives a test statistic $\chi^2(1) = 0.56[0.46]$ and, thus, price homogeneity of $\boldsymbol{\beta}'\mathbf{x}_t$ seems acceptable[5] when allowing for a broken trend. The presence of a broken linear trend might seem difficult to interpret but is probably a proxy for omitted variables effects, such as the effect of productivity differentials on relative prices, the so-called Balassa–Samuelson effect (Balassa, 1964; Samuelson, 1964). The change in the trend slope at reunification supports this interpretation. What is more surprising, however, is that the sign of the nominal exchange rate is opposite to the expected one. Based on Figure 8.1, it is easy to see why: over the sample period, relative prices and nominal exchange rates have frequently moved in opposite directions for extended periods of time. For this reason, the data do not support the *ppp* restrictions $(1, -1, -1)$ on $\boldsymbol{\beta}$.

Table 8.3 The estimated short-run dynamic adjustment structure in the $I(1)$ model

$$\underbrace{\begin{bmatrix} \Delta p_{1t} \\ \Delta p_{2,t} \\ \Delta s_{12,t} \end{bmatrix}}_{I(1)} = \underbrace{\begin{bmatrix} \underset{[4.50]}{\mathbf{0.21}} & \underset{[2.31]}{\mathbf{0.12}} & \underset{[2.06]}{\mathbf{0.01}} \\ \underset{[2.21]}{\mathbf{0.10}} & \underset{[10.23]}{\mathbf{0.52}} & \underset{[0.38]}{0.00} \\ \underset{[1.15]}{0.92} & \underset{[-1.59]}{-1.44} & \underset{[-0.18]}{-0.01} \end{bmatrix}}_{\Gamma_1} \underbrace{\begin{bmatrix} \Delta p_{1t-1} \\ \Delta p_{2,t-1} \\ \Delta s_{12,t-1} \end{bmatrix}}_{I(1)}$$

$$+ \underbrace{\begin{bmatrix} \underset{[-3.92]}{\mathbf{-0.01}} \\ \underset{[-5.92]}{\mathbf{-0.02}} \\ \underset{[-3.05]}{\mathbf{-0.17}} \end{bmatrix}}_{\alpha} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1' \mathbf{x}_{t-1} \end{bmatrix}}_{I(1)} + \underbrace{\begin{bmatrix} \underset{[1.77]}{\mathbf{0.00}} & \underset{[4.09]}{\mathbf{0.02}} \\ \underset{[0.16]}{0.00} & \underset{[6.21]}{\mathbf{0.03}} \\ \underset{[2.57]}{\mathbf{0.01}} & \underset{[2.96]}{\mathbf{0.22}} \end{bmatrix}}_{\Phi} \begin{bmatrix} D_s 91.1 \\ \mu_0 \end{bmatrix} + \underbrace{\begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}}_{I(0)}$$

where:

$$\boldsymbol{\beta}_1' x_t = \underset{}{\mathbf{1.0}} p_{1,t} \underset{[-7.76]}{-\mathbf{0.81}} p_{2,t} + \underset{[4.39]}{\mathbf{0.18}} s_{12,t} \underset{[-4.81]}{-\mathbf{0.0022}} t_{91.1} + \underset{[3.96]}{\mathbf{0.0022}} t,$$

and:

$$\Omega = \begin{bmatrix} 1.00 & & \\ 0.12 & 1.00 & \\ -0.58 & -0.07 & 1.00 \end{bmatrix}$$

The estimated $\boldsymbol{\alpha}$ coefficients show that German prices and nominal exchange rates have been equilibrium correcting to the estimated $\boldsymbol{\beta}$ relation whereas US prices have been increasing in the equilibrium errors. The overall behavior of the system is nevertheless stable as the other two variables compensate for the error-increasing behavior of US prices. The estimated coefficients of $\boldsymbol{\Gamma}_1$ show that lagged inflation rates are quite significant in the price equations, whereas the lagged depreciation/appreciation rate is only significant in the German price equation. As already demonstrated in section 8.4, the lagged changes of the $I(2)$ variables in $\boldsymbol{\Gamma}_1$ are needed to achieve stationarity of $\boldsymbol{\beta}_1' \mathbf{R}_{1,t}$.

The estimates of $\boldsymbol{\alpha}_{\perp 1}$, $\boldsymbol{\beta}_{\perp 1}$ and $\mathbf{C}$ in the MA representation of the $I(1)$ model are almost all insignificant and are not reported here. This is because the stochastic trends in the $I(1)$ model are measured by the once cumulated residuals, whereas the data are generated by second-order stochastic trends, measured by the twice cumulated residuals. Thus, when data are $I(2)$ the MA representation of the $I(1)$ model is completely uninformative.

Based on the above results, it would be hard to argue that the data are not empirically $I(2)$, and the next step is therefore to address the PPP puzzle in the correct framework of an $I(2)$ model.

8.6 Representing the $I(2)$ model

8.6.1 The basic structure

As discussed in section 8.2, formulation (8.3) is convenient when data are $I(2)$:

$$\Delta^2\mathbf{x}_t = \mathbf{\Gamma}\Delta\mathbf{x}_{t-1} + \mathbf{\Pi}\mathbf{x}_{t-1} + \boldsymbol{\mu}_0 + \boldsymbol{\mu}_{01}D_{s,91.1,t} + \boldsymbol{\mu}_1 t + \boldsymbol{\mu}_{11}t_{91.1} + \mathbf{\Phi}_p\mathbf{D}_{p,t} + \boldsymbol{\varepsilon}_t, \quad (8.16)$$

where the deterministic components are defined in section 8.5.1. Similar to the $I(1)$ model, we need to define the concentrated $I(2)$ model:[6]

$$\mathbf{R}_{0,t} = \mathbf{\Gamma}\mathbf{R}_{1,t} + \mathbf{\Pi}\mathbf{R}_{2,t} + \boldsymbol{\varepsilon}_t, \quad (8.17)$$

where $\mathbf{R}_{0,t}$, $\mathbf{R}_{1,t}$, and $\mathbf{R}_{2,t}$ are defined by:

$$\Delta^2\tilde{\mathbf{x}}_t = \hat{\mathbf{b}}_{10} + \hat{\mathbf{b}}_{11}t + \hat{\mathbf{B}}_{11}\mathbf{D}_{s,t} + \hat{\mathbf{B}}_{12}\mathbf{D}_{p,t} + \mathbf{R}_{0,t}, \quad (8.18)$$

$$\Delta\tilde{\mathbf{x}}_{t-1} = \hat{\mathbf{b}}_{20} + \hat{\mathbf{b}}_{21}t + \hat{\mathbf{B}}_{21}\mathbf{D}_{s,t} + \hat{\mathbf{B}}_{22}\mathbf{D}_{p,t} + \mathbf{R}_{1,t}, \quad (8.19)$$

$$\tilde{\mathbf{x}}_{t-1} = \hat{\mathbf{b}}_{30} + \hat{\mathbf{b}}_{31}t + \hat{\mathbf{B}}_{31}\mathbf{D}_{s,t} + \hat{\mathbf{B}}_{32}\mathbf{D}_{p,t} + \mathbf{R}_{2,t}, \quad (8.20)$$

and $\tilde{\mathbf{x}}_t$ indicates that $\mathbf{x}_t$ has been augmented with some deterministic components such as trend, constant, and shift dummy variables. The matrices $\mathbf{\Pi}$ and $\mathbf{\Gamma}$ are subject to the two reduced rank restrictions, $\mathbf{\Pi} = \boldsymbol{\alpha}'\boldsymbol{\beta}$, where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are $p \times r$, and $\boldsymbol{\alpha}_\perp'\mathbf{\Gamma}\boldsymbol{\beta}_\perp = \boldsymbol{\xi}\boldsymbol{\eta}'$, where $\boldsymbol{\xi}, \boldsymbol{\eta}$ are $(p-r)\times s_1$. The model in (8.16) contains an unrestricted constant with a shift, a broken trend and a few impulse dummies that will have to be adequately restricted to avoid undesirable effects, as discussed in section 8.6.2.

The moving average representation of the $I(2)$ model (Johansen, 1992, 1995, 1997) with unrestricted deterministic components is given by:

$$\begin{aligned}\mathbf{x}_t = {} & \mathbf{C}_2\sum_{j=1}^{t}\sum_{i=1}^{j}(\boldsymbol{\varepsilon}_i + \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 i + \boldsymbol{\mu}_{01}D_{s,91.1,i} + \mathbf{\Phi}_p\mathbf{D}_{p,i}) \\ & + \mathbf{C}_1\sum_{j=1}^{t}(\boldsymbol{\varepsilon}_j + \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 j + \boldsymbol{\mu}_{01}D_{s,91.1,j} + \mathbf{\Phi}_p\mathbf{D}_{p,j}) \\ & + \mathbf{C}^*(L)(\boldsymbol{\varepsilon}_t + \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t + \boldsymbol{\mu}_{01}D_{s,91.1,t} + \mathbf{\Phi}_p\mathbf{D}_{p,t}) + \mathbf{A} + \mathbf{B}t,\end{aligned} \quad (8.21)$$

where $\mathbf{A}$ and $\mathbf{B}$ are functions of the initial values $\mathbf{x}_0, \mathbf{x}_{-1}, \mathbf{x}_{-2}$, and the coefficient matrices satisfy:

$$\begin{aligned}\mathbf{C}_2 &= \boldsymbol{\beta}_{\perp 2}(\boldsymbol{\alpha}_{\perp 2}'\mathbf{\Psi}\boldsymbol{\beta}_{\perp 2})^{-1}\boldsymbol{\alpha}_{\perp 2}', \\ \boldsymbol{\beta}'\mathbf{C}_1 &= -\overline{\boldsymbol{\alpha}}'\mathbf{\Gamma}\mathbf{C}_2, \quad \boldsymbol{\beta}_{\perp 1}'\mathbf{C}_1 = -\overline{\boldsymbol{\alpha}}_{\perp 1}'(\mathbf{I} - \mathbf{\Psi}\boldsymbol{C}_2), \\ \mathbf{\Psi} &= \mathbf{\Gamma}\overline{\boldsymbol{\beta}}\overline{\boldsymbol{\alpha}}'\mathbf{\Gamma} + \mathbf{I} - \mathbf{\Gamma}_1\end{aligned} \quad (8.22)$$

where the notation $\overline{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\boldsymbol{\alpha}'\boldsymbol{\alpha})^{-1}$ is used throughout the chapter. To facilitate the interpretation of the $I(2)$ stochastic trends and how they load into the variables, it is useful to let $\tilde{\boldsymbol{\beta}}_{\perp 2} = \boldsymbol{\beta}_{\perp 2}(\boldsymbol{\alpha}_{\perp 2}'\mathbf{\Psi}\boldsymbol{\beta}_{\perp 2})^{-1}$, so that:

$$\mathbf{C}_2 = \tilde{\boldsymbol{\beta}}_{\perp 2}\boldsymbol{\alpha}_{\perp 2}'. \quad (8.23)$$

It is now easy to see that the $\mathbf{C}_2$ matrix has a similar reduced rank representation to $\mathbf{C}_1$ in the $I(1)$ model, so it is straightforward to interpret $\boldsymbol{\alpha}'_{\perp 2}\sum\sum\boldsymbol{\varepsilon}_i$ as a measure of the s_2 second-order stochastic trends which load into the variables $\mathbf{x}_t$ with the weights $\tilde{\boldsymbol{\beta}}_{\perp 2}$.

From (8.22) we note that the $\mathbf{C}_1$ matrix in the $I(2)$ model cannot be given a simple decomposition as it depends on both the $\mathbf{C}_2$ matrix and the other model parameters in a complex way. Johansen (2008) derives an analytical expression for $\mathbf{C}_1$, essentially showing that:

$$\mathbf{C}_1 = \boldsymbol{\omega}_0\boldsymbol{\alpha}' + \boldsymbol{\omega}_1\boldsymbol{\alpha}'_{\perp 1} + \boldsymbol{\omega}_2\boldsymbol{\alpha}'_{\perp 2}, \tag{8.24}$$

where $\boldsymbol{\omega}_i$ are complicated functions of the parameters of the model (not reproduced here).

To summarize the basic structures of the $I(2)$ model, Table 8.4 decomposes the vector $\mathbf{x}_t$ into the directions of $(\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1}, \boldsymbol{\beta}_{\perp 2})$ and the directions of $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\perp 1}, \boldsymbol{\alpha}_{\perp 2})$. The left-hand side of the table illustrates the $\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp}$ directions, where $\boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}'\Delta\mathbf{x}_t$ defines the stationary polynomially cointegrating relation, and $\boldsymbol{\beta}'_{\perp 1}\mathbf{x}_t$ the $CI(2,1)$ relation that can only become stationary by differencing. The $\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1}$ relations define the two stationary cointegration relations between the differenced variables, $\boldsymbol{\tau}'\Delta\mathbf{x}_t$. Finally, $\boldsymbol{\beta}'_{\perp 2}\mathbf{x}_t \sim I(2)$ is a non-cointegrating relation, which can only become stationary by differencing twice. The right-hand side of the table illustrates the corresponding decomposition into the $\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\perp}$ directions, where $\boldsymbol{\alpha}$ defines the dynamic adjustment coefficients to the polynomially cointegrating relation, whereas $\boldsymbol{\alpha}_{\perp 1}$ and $\boldsymbol{\alpha}_{\perp 2}$ define the first- and second-order stochastic trends as a linear function of the VAR residuals.

8.6.2 Deterministic components

A correct specification of the deterministic components, such as trends, constant and dummies, and how they enter the model, is mandatory for the $I(2)$ analysis. This is because the chosen specification is likely to strongly affect the reliability of the model estimates and to change the asymptotic distribution of the rank test. Because the typical smooth behavior of a stochastic $I(2)$ trend sometimes can be approximated with an $I(1)$ stochastic trend around a broken linear deterministic trend, one can in some cases avoid the $I(2)$ analysis altogether by allowing for sufficiently many breaks in the linear trend. Whether one specification is preferable

Table 8.4 Decomposing the data vector using the $I(2)$ model

	The $\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp}$ *decomposition of* $\mathbf{x}_t$	*The* $\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\perp}$ *decomposition*
$r = 1$	$[\underbrace{\boldsymbol{\beta}'_1\mathbf{x}_t}_{I(1)} + \underbrace{\boldsymbol{\delta}'_1\Delta\mathbf{x}_t}_{I(1)}] \sim I(0)$	$\boldsymbol{\alpha}_1$: short-run adjustment coefficients
$s_1 = 1$	$\boldsymbol{\beta}'_{\perp 1}\mathbf{x}_t \sim I(1)$	$\boldsymbol{\alpha}'_{\perp 1}\sum_{i=1}^{t}\boldsymbol{\varepsilon}_i$: $I(1)$ stochastic trend
$p - s_2 = 2$	$\boldsymbol{\tau}'\Delta\mathbf{x}_t = (\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1})'\Delta\mathbf{x}_t \sim I(0)$	
$s_2 = 1$	$\boldsymbol{\beta}'_{\perp 2}\mathbf{x}_t = \boldsymbol{\tau}'_{\perp}\mathbf{x}_t \sim I(2)$	$\boldsymbol{\alpha}'_{\perp 2}\sum_{s=1}^{t}\sum_{i=1}^{s}\boldsymbol{\varepsilon}_i$: $I(2)$ stochastic trend

to the other is difficult to know, but we need to pay sufficient attention to this question, as the choice is likely to influence the empirical results significantly.

In the present data, the reunification of Germany is likely to have affected German prices significantly, but not US prices. The raw data exhibit an extraordinary large shock in $\Delta^2 p_{1,t}$ due to the reunification in 1991:1. A big impulse in $\Delta^2 p_{1,t}$ cumulates to a level shift in $\Delta p_{1,t}$, and double cumulates to a broken linear trend in $p_{1,t}$. Thus, accounting for the extraordinary large shock at 1991:1 with a blip dummy in $\Delta^2 p_{1,t}$, a shift dummy in $\Delta p_{1,t}$ is econometrically consistent with broken linear trends in prices. Because such a broken linear trend may or may not cancel in $\boldsymbol{\beta}'\mathbf{x}_t$, the model should be specified to allow for a (testable) broken linear trend in $\boldsymbol{\beta}'\mathbf{x}_t$. Likewise, the level shift may (or may not) cancel in $\boldsymbol{\delta}'\Delta\mathbf{x}_t$ or $\boldsymbol{\tau}'\Delta\mathbf{x}_t$. Thus the model specification should allow for this possibility. Inspecting the graphs in Figure 8.1 shows an increasing trend in both prices and a downward sloping trend in relative prices, and the question is whether the latter is canceled by cointegration with the nominal exchange rate.

Whatever the case, quadratic or cubic trends will be excluded from the outset and the model specification should account for this.

To understand the role of the deterministic terms in the $I(2)$ model, it is useful to specify the mean of the stationary parts of (8.16) allowing for the above effects (so that they can be tested), while at the same time excluding cubic or quadratic trend effects.

The mean of $\Delta^2\mathbf{x}_t$ should be allowed to contain the impulse dummies as these do not double cumulate to quadratic trends, i.e.:

$$E\Delta^2\mathbf{x}_t = \boldsymbol{\Phi}_p\mathbf{D}_{p,t}.$$

The mean of the polynomially cointegrated relations should be allowed to have a trend and a broken linear trend in $\boldsymbol{\beta}'\mathbf{x}_t$ and a constant and a shift dummy in $\boldsymbol{\delta}'\Delta\mathbf{x}_t$, i.e.:

$$E(\boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}'\Delta\mathbf{x}_t) = \boldsymbol{\rho}_0 t + \boldsymbol{\rho}_{01} t_{91.1} + \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_{01} D_s 91.1_t. \tag{8.25}$$

The mean of the difference stationary relations $\boldsymbol{\tau}'\Delta\mathbf{x}_t$ should be allowed to contain a shift dummy and a constant, i.e.:

$$E(\boldsymbol{\tau}'\Delta\mathbf{x}_t) = \boldsymbol{\omega}_0 + \boldsymbol{\omega}_{01} D_s 91.1_t.$$

The question is now how to restrict $\boldsymbol{\mu}_0, \boldsymbol{\mu}_{01}, \boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_{11}$ in (8.16)[7] to allow for the deterministic components in the above mean values while suppressing any quadratic or cubic trend effects in the model. The general idea will only be demonstrated for the constant term $\boldsymbol{\mu}_0$ and the linear term $\boldsymbol{\mu}_1$, as the procedure is easily generalized to the step dummy and the broken trend. A more detailed discussion is given in Juselius (2006, Ch. 17).

First, the constant term $\boldsymbol{\mu}_0$ is decomposed into three components proportional to $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_{\perp 1}$ and $\boldsymbol{\alpha}_{\perp 2}$:

$$\boldsymbol{\mu}_0 = \boldsymbol{\alpha}\boldsymbol{\gamma}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\gamma}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2. \tag{8.26}$$

The shift dummy $\boldsymbol{\mu}_{01}$ is similarly decomposed:

$$\boldsymbol{\mu}_{01} = \boldsymbol{\alpha}\boldsymbol{\gamma}_{01} + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\gamma}_{11} + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_{21}.$$

To investigate the effect of an unrestricted constant on $\mathbf{x}_t$, (8.26) is then inserted in (8.21) using (8.23) and (8.24). The effect of cumulating the constant twice is given by:

$$\begin{aligned}\mathbf{C}_2 \sum_{j=1}^{t}\sum_{i=1}^{j} \boldsymbol{\mu}_0 &= \sum_{j=1}^{t}\sum_{i=1}^{j} \tilde{\boldsymbol{\beta}}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}(\boldsymbol{\alpha}\boldsymbol{\gamma}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\gamma}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2) \\ &= \tilde{\boldsymbol{\beta}}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2(t(t-1)/2), \end{aligned} \tag{8.27}$$

as $\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{1\perp} = \mathbf{0}$. Thus, an *unrestricted constant* term in the VAR model will allow for a quadratic trend in $\mathbf{x}_t$ so we need to restrict the $\boldsymbol{\alpha}_{\perp 2}$ component of $\boldsymbol{\mu}_0$ to avoid this. How to do this will be discussed below.

The effect of cumulating the constant term once is given by:

$$\begin{aligned}\mathbf{C}_1 \sum_{j=1}^{t} \boldsymbol{\mu}_0 &= (\boldsymbol{\omega}_0\boldsymbol{\alpha}' + \boldsymbol{\omega}_1\boldsymbol{\alpha}'_{\perp 1} + \boldsymbol{\omega}_2\boldsymbol{\alpha}'_{\perp 2}) \sum_{j=1}^{t} (\boldsymbol{\alpha}\boldsymbol{\gamma}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\gamma}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2) \\ &= \left[(\boldsymbol{\omega}_0 \underbrace{\boldsymbol{\alpha}'\boldsymbol{\alpha}\boldsymbol{\gamma}_0}_{\widetilde{\boldsymbol{\gamma}}_0} + \boldsymbol{\omega}_1 \underbrace{\boldsymbol{\alpha}'_{\perp 1}\boldsymbol{\alpha}_{\perp 1}\boldsymbol{\gamma}_1}_{\widetilde{\boldsymbol{\gamma}}_1} + \boldsymbol{\omega}_2 \underbrace{\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2}_{\widetilde{\boldsymbol{\gamma}}_2}) \right] t, \end{aligned} \tag{8.28}$$

as $\boldsymbol{\alpha}'\boldsymbol{\alpha}_{\perp 1} = \mathbf{0}$, $\boldsymbol{\alpha}'\boldsymbol{\alpha}_{\perp 2} = \mathbf{0}$ and $\boldsymbol{\alpha}'_{\perp 1}\boldsymbol{\alpha}_{\perp 2} = \mathbf{0}$. Thus, there are three different linear trends associated with the $\mathbf{C}_1$ components of the constant term.

Most applications of the $I(2)$ model are for nominal variables, implying that linear trends in the data are a natural starting hypothesis (as average nominal growth rates are generally non-zero). To achieve similarity in the rank test procedure (Nielsen and Rahbek, 2000), the model should allow for linear trends in all directions consistent with the specification of trend-stationarity as a starting hypothesis in (8.25). This means that $\boldsymbol{\mu}_1 t \neq \mathbf{0}$ and $\boldsymbol{\mu}_{11} t_{91.1} \neq \mathbf{0}$ in (8.16), so the vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_{11}$ need to be decomposed similarly to the constant term and the step dummy:

$$\boldsymbol{\mu}_1 = \boldsymbol{\alpha}\boldsymbol{\rho}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2,$$

and:

$$\boldsymbol{\mu}_{11} = \boldsymbol{\alpha}\boldsymbol{\rho}_{01} + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_{11} + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_{21}.$$

We now focus on the linear trend term. The effect of cumulating this term twice is given by:

$$\begin{aligned}\mathbf{C}_2 \sum_{j=1}^{t}\sum_{i=1}^{j} \boldsymbol{\mu}_1 i &= \sum_{j=1}^{t}\sum_{i=1}^{j} \boldsymbol{\beta}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}(\boldsymbol{\alpha}\boldsymbol{\rho}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2) i \\ &= \sum_{j=1}^{t}\sum_{i=1}^{j} \boldsymbol{\beta}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}\underbrace{\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2}_{=0} i. \end{aligned} \tag{8.29}$$

Thus, unless we restrict $\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2 = \mathbf{0}$ the model will allow for cubic trends in the data. The $I(2)$ procedure in CATS (Cointegration Analysis of Time Series) in RATS (Regression Analysis of Time Series) (Dennis, Johansen and Juselius, 2005) imposes this restriction. The effect of cumulating the linear trend term once is given by:

$$\begin{aligned}\mathbf{C}_1 \sum_{j=1}^{t} \boldsymbol{\mu}_1 j &= \sum_{j=1}^{t} (\boldsymbol{\omega}_0\boldsymbol{\alpha}' + \boldsymbol{\omega}_1\boldsymbol{\alpha}'_{\perp 1} + \boldsymbol{\omega}_2\boldsymbol{\alpha}'_{\perp 2})(\boldsymbol{\alpha}\boldsymbol{\rho}_0 + \boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_1 + \boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2) j \\ &= \sum_{j=1}^{t} (\boldsymbol{\omega}_0\boldsymbol{\alpha}' \underbrace{\boldsymbol{\alpha}\boldsymbol{\rho}_0}_{\neq 0} + \boldsymbol{\omega}_1\boldsymbol{\alpha}'_{\perp 1}\underbrace{\boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_1}_{=0} + \boldsymbol{\omega}_2\boldsymbol{\alpha}'_{\perp 2}\underbrace{\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2}_{=0}) j. \end{aligned} \tag{8.30}$$

It appears that all three $\mathbf{C}_1$ components of the linear trend will generate quadratic trends in the data. Based on (8.29), we already know that $\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\rho}_2 = \mathbf{0}$. Unless we are willing to accept linear trends in $\boldsymbol{\alpha}'_{\perp 1}\Delta\mathbf{x}_t$,[8] we should also restrict $\boldsymbol{\alpha}_{\perp 1}\boldsymbol{\rho}_1 = \mathbf{0}$. This leaves us with the $\boldsymbol{\alpha}$ component of $\mathbf{C}_1$, which cannot be set to zero, because $\boldsymbol{\alpha}\boldsymbol{\rho}_0 \neq \mathbf{0}$ is needed to allow for a linear trend in $\boldsymbol{\beta}'\mathbf{x}_t$. The problem is that a linear trend in a polynomially cointegrating relation, unless adequately restricted, generates a quadratic trend in $\mathbf{x}_t$. However, this can be solved by noticing that $\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2 \neq \mathbf{0}$ in (8.27) also generates a quadratic trend in $\mathbf{x}_t$, so that by restricting $\boldsymbol{\omega}_0\boldsymbol{\alpha}'\alpha\rho_0 = -\boldsymbol{\beta}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2$, the two trend components cancel and there will be no quadratic trends in the data. The trend-stationary polynomially cointegrated relation in Kongsted, Rahbek and Jørgensen (1999) was estimated subject to this constraint.

To summarize: to avoid quadratic and cubic trends in the $I(2)$ model we need to impose the following restrictions: $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_2 = \mathbf{0}$ and $\boldsymbol{\omega}_0\boldsymbol{\alpha}'\boldsymbol{\alpha}\boldsymbol{\rho}_0 = -\boldsymbol{\beta}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_2$, as well as $\boldsymbol{\rho}_{11} = \boldsymbol{\rho}_{21} = \mathbf{0}$ and $\boldsymbol{\omega}_0\boldsymbol{\alpha}'\boldsymbol{\alpha}\boldsymbol{\rho}_{01} = -\boldsymbol{\beta}_{\perp 2}\boldsymbol{\alpha}'_{\perp 2}\boldsymbol{\alpha}_{\perp 2}\boldsymbol{\gamma}_{21}$ to avoid broken quadratic and cubic trends.

8.7 Estimation in the $I(2)$ model

Johansen (1995) provided the solution to the two-step estimator and Johansen (1997) to the full maximum likelihood (ML) estimator. Even though the two-stage procedure gives asymptotically efficient ML estimates (Paruolo, 2000), the small sample properties of the ML estimates are generally superior (Nielsen and Rahbek, 2007), and all subsequent results are based on the ML procedure.

8.7.1 The ML procedure

Section 8.2 showed that there is an important difference between the first- and second-rank conditions. The former is formulated as a reduced rank condition directly on $\boldsymbol{\Pi}$, whereas the latter is on a transformed $\boldsymbol{\Gamma}$. This is the basic reason why the ML estimation procedure needs a different parameterization than the one in (8.3).

The full ML procedure exploits the fact that the $I(2)$ model contains $p-s_2$ cointegration relations, $\boldsymbol{\tau}'\mathbf{x}_t$, where $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1})$ define $r+s_1 = p-s_2$ directions in which the process is cointegrated from $I(2)$ to $I(1)$. This means that $\boldsymbol{\tau}$ can be determined by solving just one reduced rank regression, after which the vector $\mathbf{x}_t$ is decomposed into the $p-s_2$ directions $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1})$ in which the process is $I(1)$, and the s_2 directions $\boldsymbol{\tau}_\perp = \boldsymbol{\beta}_{\perp 2}$ in which it is $I(2)$.

Johansen (1997) does not make a distinction between stationary and nonstationary components in $\Delta\mathbf{x}_t$. For example, when $\mathbf{x}_t$ contains variables which are $I(2)$, e.g., prices, as well as $I(1)$, e.g., nominal exchange rates, then some of the differenced variables will be $I(0)$. As the latter do not contain any stochastic $I(1)$ trends, they are by definition redundant in the polynomially cointegrated relations. The idea behind the parameterization in Paruolo and Rahbek (1999) was to express the polynomially cointegrated relations exclusively in terms of the differences of the $I(2)$ variables. The model given below is based on the Paruolo and Rahbek parameterization. As discussed in section 8.6, the (broken) trend has been restricted to be proportional to $\boldsymbol{\alpha}$, and the constant and the shift dummy to be proportional to $\boldsymbol{\zeta}$.

$$\begin{aligned}\underbrace{\Delta^2\mathbf{x}_t}_{I(0)} = \boldsymbol{\alpha}\Bigg\{\underbrace{\Big[\boldsymbol{\beta}', \boldsymbol{\rho}_0, \boldsymbol{\rho}_{01}\Big]\begin{bmatrix}\mathbf{x}_{t-1}\\ t\\ t_{91.1}\end{bmatrix}}_{I(1)} + \underbrace{\Big[\boldsymbol{\delta}', \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_{01}\Big]\begin{bmatrix}\Delta\mathbf{x}_{t-1}\\ c\\ D_s 91.1_{t-1}\end{bmatrix}}_{I(1)}\Bigg\}_{I(0)} \\ + \underbrace{\boldsymbol{\zeta}\begin{bmatrix}\boldsymbol{\beta}', \boldsymbol{\rho}_0, \boldsymbol{\rho}_{01}\\ \underbrace{\boldsymbol{\beta}'_{\perp 1}, \tilde{\boldsymbol{\gamma}}_0, \tilde{\boldsymbol{\gamma}}_{01}}_{\boldsymbol{\tau}}\end{bmatrix}\begin{bmatrix}\Delta\mathbf{x}_{t-1}\\ c\\ D_s 91.1_{t-1}\end{bmatrix}}_{I(0)} + \boldsymbol{\Phi}_p\mathbf{D}_{p,t} + \boldsymbol{\varepsilon}_t,\end{aligned} \tag{8.31}$$

where $\boldsymbol{\varepsilon}_t \sim N_p(\mathbf{0}, \boldsymbol{\Omega})$, $t = 1, \ldots, T$, $\boldsymbol{\delta}' = \boldsymbol{\psi}'\boldsymbol{\tau}_\perp\boldsymbol{\tau}'_\perp$ with $\boldsymbol{\psi}' = -(\boldsymbol{\alpha}'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}$, $\boldsymbol{\zeta}' = \boldsymbol{\psi}'\boldsymbol{\tau} - \boldsymbol{\Omega}\boldsymbol{\alpha}_\perp(\boldsymbol{\alpha}'_\perp\boldsymbol{\Omega}\boldsymbol{\alpha}_\perp)^{-1}(\boldsymbol{\alpha}'_\perp\boldsymbol{\Gamma}\boldsymbol{\beta}, \boldsymbol{\xi})$ and $\boldsymbol{\xi}$ is defined in (8.5).

The relations $\boldsymbol{\beta}'\tilde{\mathbf{x}}_t + \tilde{\boldsymbol{\delta}}'\Delta\tilde{\mathbf{x}}_t$, with $\boldsymbol{\beta}' = [\boldsymbol{\beta}', \boldsymbol{\rho}_0, \boldsymbol{\rho}_{01}]$, $\tilde{\mathbf{x}}'_t = [\mathbf{x}'_t, t, t_{91.1}]$, $\tilde{\boldsymbol{\delta}}' = \Big[\boldsymbol{\delta}', \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_{01}\Big]$, and $\Delta\tilde{\mathbf{x}}'_t = [\Delta\mathbf{x}'_t, 1, D_s 91.1]$, define r stationary polynomially cointegrating relations, whereas the relations $\boldsymbol{\tau}'\Delta\tilde{\mathbf{x}}_t$ define $p-s_2$ stationary relations between the growth rates.

8.7.2 Linking $I(1)$ with $I(2)$

It is useful to see how the formulation (8.31) relates to the usual VAR formulation (8.3). Relying on results in Johansen (1997), the levels and difference components

of the unrestricted VAR model (8.3) can be decomposed as:

$$\begin{aligned}\boldsymbol{\Gamma}\Delta\mathbf{x}_{t-1}+\boldsymbol{\Pi}\mathbf{x}_{t-1} = (\boldsymbol{\Gamma}\overline{\boldsymbol{\beta}})\underbrace{\boldsymbol{\beta}'\Delta\mathbf{x}_{t-1}}_{I(0)} \\ +(\boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{\Gamma}\overline{\boldsymbol{\beta}}_{\perp 1}+\boldsymbol{\alpha}_{\perp 1})\underbrace{\boldsymbol{\beta}'_{\perp 1}\Delta\mathbf{x}_{t-1}}_{I(0)} \\ +(\boldsymbol{\alpha}\overline{\boldsymbol{\alpha}}'\boldsymbol{\Gamma}\overline{\boldsymbol{\beta}}_{\perp 2})\underbrace{\boldsymbol{\beta}'_{\perp 2}\Delta\mathbf{x}_{t-1}}_{I(1)} \\ +\boldsymbol{\alpha}\underbrace{\boldsymbol{\beta}'\mathbf{x}_{t-1}}_{I(1)},\end{aligned} \tag{8.32}$$

where $\overline{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}$ and $\overline{\boldsymbol{\alpha}}$ is similarly defined. The decomposition describes three types of linear relations between the growth rates, $\boldsymbol{\beta}'\Delta\mathbf{x}_{t-1}, \boldsymbol{\beta}'_{\perp 1}\boldsymbol{\Delta}\mathbf{x}_{t-1}$ and $\boldsymbol{\beta}'_{\perp 2}\Delta\mathbf{x}_{t-1}$, of which the first two define $I(0)$ relations, and the third an $I(1)$ relation. The coefficients in soft brackets define the corresponding adjustment coefficients.

Since $\boldsymbol{\beta}'_{\perp 2}\Delta\mathbf{x}_{t-1}$ is $I(1)$, it needs to be combined with another $I(1)$ variable to become stationary. An obvious candidate for this is $\boldsymbol{\beta}'\mathbf{x}_{t-1}$. It is now easy to see how the parameterization in (8.3) relates to the one in (8.31):

$$\boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{t-1}+(\overline{\boldsymbol{\alpha}}'\boldsymbol{\Gamma}\overline{\boldsymbol{\beta}}_{\perp 2})\,\boldsymbol{\beta}'_{\perp 2}\Delta\mathbf{x}_{t-1}) = \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{t-1}+\boldsymbol{\delta}'\Delta\mathbf{x}_{t-1}). \tag{8.33}$$

Finally, when $r > s_2$, the long-run matrix $\boldsymbol{\Pi}$ can be expressed as the sum of the two levels components:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}_0\boldsymbol{\beta}'_0 + \boldsymbol{\alpha}_1\boldsymbol{\beta}'_1,$$

where $\boldsymbol{\beta}'_0\mathbf{x}_{t-1}$ defines $r - s_2$ directly stationary $CI(2,2)$ relations, whereas $\boldsymbol{\beta}'_1\mathbf{x}_{t-1}$ defines s_2 non-stationary $CI(2,1)$ cointegrating relations, which needs to be combined with the differenced process to become stationary through polynomial cointegration.

Thus, the $I(2)$ model can distinguish between the $CI(2,1)$ relations between levels $\{\boldsymbol{\beta}'\mathbf{x}_t, \boldsymbol{\beta}'_{\perp 1}\mathbf{x}_t\}$, the $CI(1,1)$ relations between levels and differences $\{\boldsymbol{\beta}'\mathbf{x}_{t-1} + \boldsymbol{\delta}'\Delta\mathbf{x}_t\}$, and finally the $CI(1,1)$ relations between differences $\{\boldsymbol{\tau}'\Delta\mathbf{x}_t\}$. As a consequence, when discussing the economic interpretation of these components, the generic concept of a "long-run" equilibrium relation needs to be modified

accordingly. Juselius (2006, Ch. 17) proposed the following interpretation:

1. $\boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}'\Delta\mathbf{x}_t$ as r *dynamic long-run equilibrium relations*, or alternatively when $r > s_2$
   - $\boldsymbol{\beta}_0'\mathbf{x}_t$ as $r - s_2$ *static long-run equilibrium relations*, and
   - $\boldsymbol{\beta}_1'\mathbf{x}_t + \boldsymbol{\delta}_1\Delta\mathbf{x}_t$ as s_2 *dynamic long-run equilibrium relations,*
2. $\boldsymbol{\tau}'\Delta\mathbf{x}_t$ as *medium-run equilibrium relations*.

8.8 Two hypothetical scenarios

To be able to structure and interpret the empirical VAR results, it is useful to formulate a scenario for what we would expect to find in the VAR model, provided the reality is in accordance with the assumptions of the theoretical model. For example, the first scenario below is specified for the hypothesis: {$ppp_t \sim I(0)$, prices are pushing and the nominal exchange rate is pulling} under the assumption that $\mathbf{x}_t$ is empirically $I(2)$.

We shall discuss the following two cases: (1) $r = 2$, which corresponds to the theory consistent case, and (2) $r = 1$, which is what we find in the data. In both cases it will be assumed that long-run price homogeneity holds, i.e., $\boldsymbol{\beta}_{\perp 2}' = [c, c, 0]$.

Case 1 $\{r = 2, s_1 = 0, s_2 = 1\}$ is consistent with:

$$\begin{bmatrix} p_{1,t} \\ p_{2,t} \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} c \\ c \\ 0 \end{bmatrix} \sum_{j=1}^{t}\sum_{i=1}^{j} u_{1,i} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \sum_{i=1}^{j} u_{1,i} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}. \tag{8.34}$$

It is easy to see that $(p_{1,t} - p_{2,t}) \sim I(1)$ and $(p_{1,t} - p_{2,t} - s_{12,t}) \sim I(0)$ if $(b_1 - b_2) = b_3$. When the nominal exchange rate is adjusting (and price shocks are pushing), one would have it that $u_{1,t} = \boldsymbol{\alpha}_{\perp 1}'\boldsymbol{\varepsilon}_t$ with $\boldsymbol{\alpha}_{\perp 1}' = [a_1, a_2, 0]$. This scenario would imply two cointegrating relations, one of which is directly cointegrating, because $r - s_2 = 1$, and the other of which is polynomially cointegrating, because $s_2 = 1$. It is easy to show that the directly cointegrating relation is the *ppp* relation, i.e., $(p_{1,t} - p_{2,t} - s_{12,t}) \sim I(0)$. The polynomially cointegrated relation is more difficult to see and it is helpful to examine the system based on the nominal-to-real transformation (Kongsted, 2005):[9]

$$\begin{bmatrix} p_{1,t} - p_{2,t} \\ \Delta p_{1,t} \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} b_1 - b_2 \\ c \\ b_3 \end{bmatrix} \sum_{i=1}^{j} u_{1,i} + \begin{bmatrix} \tilde{\varepsilon}_{1,t} \\ \tilde{\varepsilon}_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

It is now straightforward to show that $\left\{p_{1,t} - p_{2,t} + \omega\Delta p_{1,t}\right\} \sim I(0)$, if $c = -(b_1 - b_2/\omega)$. Alternatively, if $c = -b_3/\omega$, then $\left\{s_{12,t} + \omega\Delta p_{1,t}\right\} \sim I(0)$. In both cases the polynomially cointegrating relation can be thought of as a dynamic equilibrium relation describing how the inflation rate adjusts when relative prices have been pushed apart, i.e., $\Delta p_{1,t} = -1/\omega\,(p_{1,t} - p_{2,t})$. It simply states the obvious, that the

inflation rates have to react in a non-homogeneous manner if relative prices move persistently apart.

Case 2 $\{r = 1, s_1 = 1, s_2 = 1\}$ is consistent with:

$$\begin{bmatrix} p_{1,t} \\ p_{2,t} \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} c \\ c \\ 0 \end{bmatrix} \sum_{j=1}^{t} \sum_{i=1}^{j} u_{1,i} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{j} u_{1,i} \\ \sum_{i=1}^{j} u_{2,i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

In this case one would not expect to find a directly cointegrating relation, as $r - s_2 = 0$. This result is easily seen from the nominal-to-real transformed system:

$$\begin{bmatrix} p_{1,t} - p_{2,t} \\ \Delta p_{1,t} \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} b_{11} - b_{21} & b_{12} - b_{22} \\ c & 0 \\ b_{31} & b_{32} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{j} u_{1,i} \\ \sum_{i=1}^{j} u_{2,i} \end{bmatrix} + \begin{bmatrix} \tilde{\varepsilon}_{1,t} \\ \tilde{\varepsilon}_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

It is now easy to see that stationarity of ppp_t can only be achieved in the special case when $b_{11} - b_{21} = b_{31}$ and $b_{12} - b_{22} = b_{32}$. But in general, empirical support for ppp_t can only be achieved by polynomial cointegration, i.e., in the form of a dynamic long-run adjustment relation. For example, if $b_{12} - b_{22} = b_{32}$ and $c = -(b_{11} - b_{21} - b_{31})/\omega$, then $\left\{p_{1,t} - p_{2,t} - s_{12,t} + \omega \Delta p_{1,t}\right\} \sim I(0)$. The latter can be interpreted as evidence of the following dynamic adjustment relation: $\Delta p_{1,t} = -1/\omega \left\{p_{1,t} - p_{2,t} - s_{12,t}\right\}$. In this case, either inflation rates or the currency depreciation/appreciation rate have to move in an offsetting direction when *ppp* has persistently deviated from its benchmark values.

Thus the outcome of testing rank indices in the $I(2)$ model has strong implications for whether support for a stationary relation can be found or not.

8.9 An $I(2)$ analysis of prices and exchange rates

8.9.1 Determining the two rank indices

The number of stationary multi-cointegrating relations, r, and the number of $I(1)$ trends, s_1, among the common stochastic trends, $p - r$, can be determined by the ML procedure in Nielsen and Rahbek (2007), where the trace test is calculated for all possible combinations of r and s_1 so that the joint hypothesis (r, s_1) can be tested, as explained below.

Table 8.5 reports the ML tests of the joint hypothesis of (r, s_1), which corresponds to the two reduced rank hypotheses in (8.4) and (8.5). The test procedure starts with the most restricted model $(r = 0, s_1 = 0, s_2 = 3)$ in the upper left-hand corner, continues to the end of the first row $(r = 0, s_1 = 3, s_2 = 0)$, and proceeds similarly row-wise from left to right until the first acceptance. Based on the tests, the first acceptance is at $(r = 1, s_1 = 1, s_2 = 1)$, which was also the preferred choice in section 8.4. The last column of the table correponds to the $I(1)$ trace test. When

Table 8.5 Determination of rank indices

r	p-r	$s_2 = 3$	$s_2 = 2$	$s_2 = 1$	$s_2 = 0$
0	3	527.6 [110.9]	293.9 [89.3]	118.10 [71.9]	80.06 [57.9]
1	2		96.88 [64.4]	**32.25** [48.5]	**32.65** [36.6]
2	1			**8.20** [28.7]	**6.72** [18.4]
The 4 largest characteristic roots, $r = 2$					
$s_1 = 0$	$s_2 = 1$		1.0	1.0	0.98 0.53
The 4 largest characteristic roots, $r = 1$					
$s_1 = 2$	$s_2 = 0$		1.0	1.0	0.99 0.53
$s_1 = 1$	$s_2 = 1$		1.0	1.0	1.0 0.53

Note: 95% quantiles in [].

the data are $I(2)$, determining the rank r exclusively by this test can often lead to incorrect results.

Our model has a broken linear trend restricted to the polynomially cointegrated relation and a shift dummy restricted to the differences. Because of this, the standard asymptotic trace test distributions (e.g., provided by CATS for RATS) are no longer correct. The critical values given in brackets below the test values have been kindly provided by Heino Nielsen using a simulation program described in Nielsen (2004) (see also Kurita, 2007). The inclusion of a broken linear trend in the cointegration relations shifts the distributions to the right, implying that the test will be undersized if one ignores the effect of the broken trend.

Table 8.5 also reports the characteristic roots in the VAR model for $r = 1$ and 2. For $\{r = 2, p - r = 1\}$ there is just one common stochastic trend, which has to be $I(2)$ if the data are $I(2)$. The choice of $\{r = 2, s_2 = 1\}$ will impose two unit root restrictions on the characteristic roots of the model. As already discussed in section 8.5.2 and confirmed in Table 8.5, this leaves one large unrestricted root, 0.98, in the model. Such a root is not statistically distinguishable from a unit root and would give problems if left unrestricted in the empirical model. When $r = 1$, the choice $\{r = 1, s_1 = 1, s_2 = 1\}$ accounts for all three near unit roots in the model with 0.53 as the largest unrestricted root, whereas the choice of $\{r = 1, s_1 = 0, s_2 = 2\}$ corresponds to four unit roots in the model and basically forces 0.53 to be a unit root. Altogether, the results strongly suggest that $\{r = 1, s_1 = 1, s_2 = 1\}$ is the correct choice.

That $r = 1$ is an important result, as the two scenarios in section 8.8 showed that a stationary ppp_t is inherently associated with *one* stochastic trend having generated prices and nominal exchange rates. Thus, the finding of $p - r = 2$ suggests that there exists another source of permanent shocks that have contributed to the persistent behavior in the data. A plausible explanation will be given in the concluding section.

8.9.2 The pulling forces

The scenarios above assume long-run price homogeneity. In section 8.6, this hypothesis was tested on $\boldsymbol{\beta}'\mathbf{x}_t$ and was accepted with high p-value. However, when

Table 8.6 The estimated short-run dynamic adjustment structure in the $I(2)$ model

$$\underbrace{\begin{bmatrix}\Delta^2 p_{1t}\\ \Delta^2 p_{2,t}\\ \Delta^2 s_{12,t}\end{bmatrix}}_{I(0)} = \underbrace{\begin{bmatrix}\mathbf{-0.01}\\ [-4.98]\\ \mathbf{-0.02}\\ [-8.66]\\ \mathbf{-0.13}\\ [-3.41]\end{bmatrix}}_{\boldsymbol{\alpha}} \underbrace{\begin{bmatrix}\boldsymbol{\beta}_1'\mathbf{x}_{t-1} + \boldsymbol{\delta}_1'\Delta\mathbf{x}_{t-1}\end{bmatrix}}_{I(0)}$$

$$+ \underbrace{\begin{bmatrix}\mathbf{-0.51} & \mathbf{-0.25}\\ [-11.97] & [-11.15]\\ \mathbf{0.29} & \mathbf{-0.14}\\ [7.25] & [-6.51]\\ 1.19 & 0.06\\ [1.66] & [0.15]\end{bmatrix}}_{\boldsymbol{\zeta}} \underbrace{\begin{bmatrix}\boldsymbol{\beta}_1'\Delta\mathbf{x}_{t-1}\\ \boldsymbol{\beta}_{\perp 1,1}'\Delta\mathbf{x}_{t-1}\end{bmatrix}}_{I(0)} + \underbrace{\begin{bmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{3,t}\end{bmatrix}}_{I(0)},$$

where:

$$\boldsymbol{\beta}_1'\mathbf{x}_t + \boldsymbol{\delta}_1'\Delta\mathbf{x}_t = \mathbf{1.0}p_{1,t} \underset{[-7.68]}{\mathbf{-0.85}}\, p_{2,t} + \underset{[15.08]}{\mathbf{0.19}}\, s_{12,t} - \underset{[-5.99]}{\mathbf{0.0025}}\, t_{91.1} + \underset{[8.34]}{\mathbf{0.0024}}t$$
$$+ \mathbf{2.61}\ \Delta p_{1,t} + \mathbf{5.21}\ \Delta p_{2,t} + \mathbf{9.31}\ \Delta s_{12,t} - \mathbf{0.10}\ D_s 91.1$$

$$\boldsymbol{\beta}_{\perp 1}'\Delta\mathbf{x}_{t-1} = \mathbf{1.01}\ \Delta p_{1,t} + \mathbf{1.0}\ \Delta p_{2,t} - \mathbf{0.84}\ \Delta s_{12,t} + \mathbf{0.01}\ \Delta t_{91.1} - \mathbf{0.01}\ \Delta t$$

Note: t-values in [].

$\mathbf{x}_t \sim I(2)$, long-run price homogeneity is defined on $\boldsymbol{\tau}'\mathbf{x}_t$, where $\boldsymbol{\tau}' = [\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1}]$. Hence (see Johansen, 2006b), long-run homogeneity on $\boldsymbol{\beta}$ is a necessary but not sufficient condition. When tested, long-run price homogeneity of $\boldsymbol{\tau}'\mathbf{x}_t$ was strongly rejected based on $\chi^2(2) = 22.95[0.00]$ and $\boldsymbol{\beta}_{\perp 1}'\mathbf{x}_t$ cannot be considered homogeneous in prices. As Table 8.6 demonstrates, the coefficients to prices on $\boldsymbol{\beta}_{\perp 1}$ are proportional to (1, 1) rather than (1, −1). This, of course, is just another piece of evidence associated with the PPP puzzle.

Table 8.6 also reports the estimates of short-run adjustment dynamics towards the estimated long-run equilibrium relations. The $I(2)$ model is parameterized according to (8.31). We note that the $I(2)$ model allows the VAR variables to adjust to a medium-run equilibrium error, $\boldsymbol{\beta}_{\perp 1}'\boldsymbol{\Delta}\tilde{\mathbf{x}}_{t-1}$, to a change in the long-run "static equilibrium" error, $\boldsymbol{\beta}'\Delta\tilde{\mathbf{x}}_{t-1}$, and to the long-run "dynamic equilibrium" error, $\boldsymbol{\beta}'\tilde{\mathbf{x}}_{t-1} + \boldsymbol{\delta}\boldsymbol{\Delta}\tilde{\mathbf{x}}_{t-1}$. In this sense, the $I(2)$ model offers a much richer dynamic adjustment structure than the $I(1)$ model.

When discussing the adjustment dynamics with respect to the polynomially cointegrating relations, it is useful to interpret the adjustment coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ as two levels of equilibrium correction. Consider, for example, the following model for the variable $x_{i,t}$:

$$\Delta^2 x_{i,t} = \cdots \sum_{j=1}^{r} \alpha_{ij}(\boldsymbol{\delta}_j'\Delta\mathbf{x}_{t-1} + \boldsymbol{\beta}_j'\mathbf{x}_{t-1}) + \cdots \tag{8.35}$$

If $\alpha_{ij}\delta_{ij} < 0$ for $j = 1, \ldots, r$, then the acceleration rates, $\Delta^2 x_{i,t}$ are equilibrium error correcting to the changes $\Delta x_{i,t}$, and if $\delta_{ij}\beta_{ij} > 0$ for $i = 1, \ldots, p$, then the changes $\Delta x_{i,t}$ are equilibrium error correcting to the levels $x_{i,t}$. In the interpretation below we shall pay special attention to whether a variable is equilibrium error correcting or increasing as defined above, as this is an important feature of the data.

Based on the estimates in Table 8.6, it appears that the acceleration rates of prices and nominal exchange rates are all equilibrium error correcting to their respective growth rates. When it comes to the relationship between growth rates and levels of variables, there is just one polynomial cointegration relation to check for equilibrium correction, but the check has to be done for all three growth rates. To make the equilibrium correction property more visible, the relation $\boldsymbol{\delta}'\Delta\mathbf{x}_{t-1} + \boldsymbol{\beta}'\mathbf{x}_{t-1}$ has been formulated in three alternative but equivalent ways:

$$\begin{aligned}
\Delta p_{1,t} &= -\mathbf{0.38}(p_{1,t} - \underset{[-7.68]}{\mathbf{0.85}}\, p_{2,t} + \underset{[15.08]}{\mathbf{0.19}}\, s_{12,t} - \underset{[-5.99]}{\mathbf{0.0025}}\, t_{91.1} + \underset{[8.34]}{\mathbf{0.0025}}t) \\
&\quad - \mathbf{2.0}\ \Delta p_{2,t} - \mathbf{3.5}\ \Delta s_{12,t} \\
\Delta p_{2,t} &= \mathbf{0.16}(p_{2,t} - \underset{[-7.68]}{\mathbf{1.15}}\, p_{1,t} - \underset{[15.08]}{\mathbf{0.25}}\, s_{12,t} + \underset{[-5.99]}{\mathbf{0.003}}\, t_{91.1} - \underset{[8.34]}{\mathbf{0.003}}t) \\
&\quad - \mathbf{0.50}\ \Delta p_{1,t} - \mathbf{1.8}\ \Delta s_{12,t} \\
\Delta s_{12,t} &= -\mathbf{0.02}(s_{12,t} - \mathbf{4.5}p_{2,t} + \underset{[7.68]}{\mathbf{5.5}}\, p_{1,t} + \underset{[-5.99]}{\mathbf{0.013}}\, t_{91.1} - \underset{[8.34]}{\mathbf{0.013}}t) \\
&\quad - \mathbf{0.28}\ \Delta p_{1,t} - \mathbf{0.56}\ \Delta p_{2,t}.
\end{aligned}$$

It appears that the polynomially cointegrated relation is consistent with equilibrium correction behavior in the German inflation rate and the Dmk–$ depreciation/appreciation rate, whereas the US inflation rate is error increasing. The lack of equilibrium error correction in US prices, already commented on in section 8.5.3, is an interesting empirical finding that is likely to be related to the PPP puzzle.

Ideally, one would like to interpret the above relations as dynamic adjustment of growth rates to a long-run static equilibrium relation, as described in the second scenario in section 8.8. In the present case, this is not straightforward because the nominal exchange rate has the wrong sign in $\boldsymbol{\beta}'\mathbf{x}_t$. Therefore, the latter cannot be given an approximate interpretation of a long-run *ppp* relation. Whatever the case, Figure 8.5 illustrates that the polynomially cointegrated relation is strongly mean-reverting.

Finally, the estimated adjustment coefficients, $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2]$, to the growth rate relations, $\boldsymbol{\beta}_1'\Delta\mathbf{x}_{t-1}$ and $\boldsymbol{\beta}_{\perp 1}'\Delta\mathbf{x}_{t-1}$, show that it is primarily the two prices that are adjusting. Both German and US prices are equilibrium adjusting to the first "growth rates" relation, $\boldsymbol{\beta}_1'\Delta\mathbf{x}_{t-1} = 1.0\Delta p_{1t} - 0.85\Delta p_{2t} + 0.20\Delta s_{12,t}$, but German prices more quickly so. The second "growth rates" relation, $\boldsymbol{\beta}_{\perp 1}'\Delta\mathbf{x}_{t-1} = 1.01\Delta p_{1,t} + 1.0\Delta p_{2,t} - 0.84\Delta s_{12,t}$, is more difficult to interpret. It essentially says that the change in the Dmk–$ rate has been proportional to the sum of German and US inflation rates, rather than to the inflation spread. As the coefficients of $\boldsymbol{\beta}_{\perp 1}$ are the

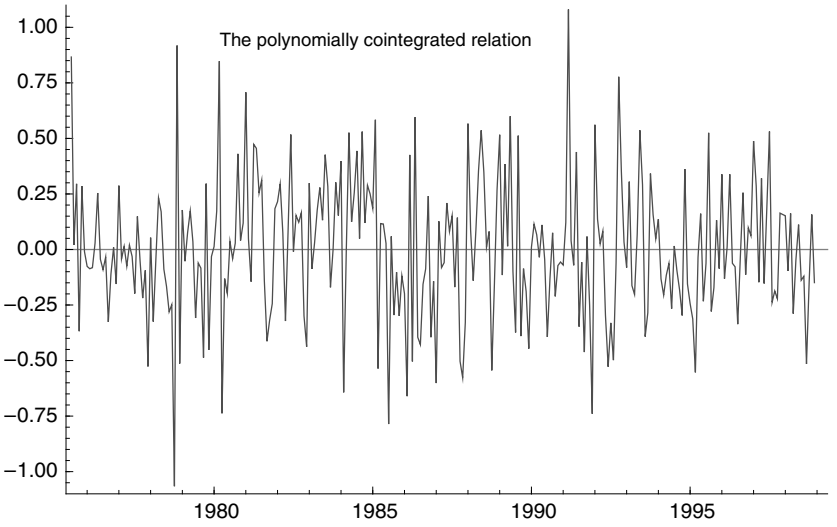


Figure 8.5 The graph of the polynomially cointegrated relation $\boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}'\Delta\mathbf{x}_t$

opposite of price homogeneity, the results explain why long-run price homogeneity in $\boldsymbol{\tau}$ was so strongly rejected.

That inflation rates are moving in opposite directions is a puzzling and even implausible result. Therefore, it is useful to check whether this result still holds for the combined estimates, $\boldsymbol{\zeta}\boldsymbol{\tau}'\Delta\mathbf{x}_t$, calculated below:

	$\Delta p_{1,t}$	$\Delta p_{2,t}$	$\Delta s_{12,t}$
$\Delta^2 p_{1t}$:	−0.75	0.18	0.10
$\Delta^2 p_{2,t}$:	0.44	−0.13	0.04
$\Delta^2 s_{12,t}$:	1.25	−1.00	0.17

Fortunately, the combined estimates are more plausible: German, as well as US, inflation rates are now equilibrium error-correcting to each other. The US inflation rate is equilibrium error-correcting to German price inflation with the correct sign, but to the Dmk–$ rate with an "incorrect" sign. However, the coefficient is very small and may not be significantly different from zero. Finally, the Dmk–$ rate is not equilibrium error correcting but even error increasing with the US–German inflation spread. Since the coefficients $\boldsymbol{\zeta}_{13}$ and $\boldsymbol{\zeta}_{23}$ were both insignificant, this is, however, not necessarily an empirically strong result.

To summarize, the VAR analysis has detected four puzzling results:

1. Nominal exchange rates tend to move in the opposite direction to relative prices for extended periods of time.
2. The US inflation rate is not equilibrium error correcting to $\boldsymbol{\beta}'\mathbf{x}_t$.

3. Changes in the nominal exchange rate either do not seem to have been significantly responding to movements in relative inflation rates or, if they have, in an equilibrium increasing manner.
4. The US inflation rate does not seem to have been responding to this "adverse" behavior of the change in the Dmk–$ rate.

8.9.3 The estimated driving forces

The scenario in section 8.8 can now be directly assessed based on the estimates of the MA representation in Table 8.7. The results clearly show that the empirical reality has deviated quite substantially from the assumed theoretical scenario. For example, the estimated loadings to the $I(2)$ trend, $\boldsymbol{\beta}_{\perp 2}$, show that the price coefficients are not even close to being equal, as assumed by the long-run homogeneity hypothesis. Given the previous rejection of long-run price homogeneity, this result should, of course, not come as a big surprise. However, what is more surprising is that the coefficient to the Dmk–$ rate is not even close to zero, suggesting that $s_{12,t}$ is empirically $I(2)$, rather than $I(1)$ as assumed in the scenario. Another surprising result is that, given the estimates of $\boldsymbol{\beta}_{\perp 2}$, the $I(2)$ trend does not seem to cancel in $ppp = p_1 - p_2 - s_{12}$. For this to be the case, the coefficients would need to be proportional to $\boldsymbol{\beta}'_{\perp 2} = [a, -a, 2a]$.

That the real exchange rate is empirically $I(2)$ would be hard to reconcile with standard theories. However, the theory of imperfect knowledge economics (Frydman and Goldberg, 2007) does in fact explain such a result. Frydman *et al.* (2008) demonstrate that, under highly plausible assumptions on agents' behavior, speculative transactions in the foreign exchange market are likely to generate pronounced persistence in nominal exchange rates that would be hard to distinguish from a near $I(2)$ process. Johansen *et al.* (2008) find strong evidence for this to be the case based on the same US–German (2008) data analyzed here, but extended with short- and long-term interest rates. They also find that the *ppp* transformed variable exhibits highly persistent behavior that can be considered either

Table 8.7 The common stochastic trends and their loadings

$$\begin{bmatrix} p_{1t} \\ p_{2,t} \\ s_{12,t} \end{bmatrix} = \begin{bmatrix} \mathbf{0.04} \\ \mathbf{0.09} \\ \mathbf{0.16} \end{bmatrix} \left[\boldsymbol{\alpha}'_{\perp 2,1} \sum \sum \hat{\boldsymbol{\varepsilon}} s \right]$$

$$+ \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}'_{\perp 2,1} \sum \hat{\boldsymbol{\varepsilon}}_i \\ \boldsymbol{\alpha}'_{\perp 1,1} \sum \hat{\boldsymbol{\varepsilon}}_i \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \begin{bmatrix} t_{91.1} \\ t \end{bmatrix}$$

where:

$$\boldsymbol{\alpha}'_{\perp 2,1} \hat{\boldsymbol{\varepsilon}}_t = \underset{[-4.03]}{\mathbf{-0.57}} \hat{\varepsilon}_{p_1,t} + \mathbf{1.0} \hat{\varepsilon}_{p_2,t} \underset{[-2.32]}{\mathbf{-0.09}} \hat{\varepsilon}_{s_{12},t}$$
$$\boldsymbol{\alpha}'_{\perp 1,1} \hat{\boldsymbol{\varepsilon}}_t = \underset{[6.79]}{\mathbf{0.25}} \hat{\varepsilon}_{p_1,t} + \underset{[2.52]}{\mathbf{0.14}} \hat{\varepsilon}_{p_2,t} \underset{[-1.82]}{\mathbf{-0.04}} \hat{\varepsilon}_{s_{12},t}$$

empirically near $I(2)$ or $I(1)$, depending on whether the emphasis is on size or power.

The estimated $\boldsymbol{\alpha}_{\perp 2}$ shows that it is shocks to relative prices (but with a larger weight on US prices) and to nominal exchange rates that seem to have generated the stochastic $I(2)$ trend. Contrary to the scenario, the coefficient to the nominal exchange rate is significant and the sign is opposite to the expected one. The estimated $\boldsymbol{\alpha}_{\perp 1}$, describing the stochastic $I(1)$ trend, shows that a weighted average of inflationary shocks in Germany and the US have generated the medium-run movements in prices and exchange rates.

These results seem to strengthen the previous conclusions: standard theories of price determination in the goods market cannot explain the PPP puzzle. The overriding impression is that it is the nominal exchange rate that is behaving oddly, suggesting that the long swings puzzle needs to be solved together with another international macro-puzzle, the "forward premium puzzle." This will be discussed in the concluding section.

8.9.4 What did we gain from the $I(2)$ analysis?

Section 8.5 reported estimates and tests using the $I(1)$ model even though data were empirically $I(2)$. The question is whether the $I(2)$ analysis has changed some of the previous conclusions, or provided new insight that could not have been obtained from the $I(1)$ analysis.

To facilitate a comparison of the $I(1)$ and $I(2)$ models, it is useful first to subtract $\Delta \mathbf{x}_{t-1}$ from both sides of equation (8.15) estimated in section 8.5. The vector process would then be formulated in second differences $\Delta^2 \mathbf{x}_t$, and $\boldsymbol{\Gamma}_1$ would become $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 - \mathbf{I}$. In terms of likelihood, the two models differ only with respect to $\boldsymbol{\Gamma}$, which is unrestricted in the $I(1)$ model but subject to one nonlinear parameter restriction in the $I(2)$ model.

The estimates of the $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ coefficients are very similar in the two models, but their standard errors are smaller in the $I(2)$ model, resulting in larger t-ratios. This is because in the $I(2)$ model the super-super consistency of $\boldsymbol{\beta}$ is adequately accounted for and because the $\boldsymbol{\beta}$ relation has been directly estimated as a polynomial cointegration relation. Also, the $\boldsymbol{\alpha}$ coefficients are not just measuring the adjustment to the levels relation, $\boldsymbol{\beta}' \mathbf{x}_{t-1}$, but to the levels and differences relation, $\boldsymbol{\beta}_1' \mathbf{x}_{t-1} + \boldsymbol{\delta}_1' \Delta \mathbf{x}_{t-1}$.

In the $I(1)$ model, the coefficient estimates of $\boldsymbol{\Gamma}_1$ are unrestricted, and there is not the same efficiency gain as in the $I(2)$ model, where the estimates are subject to the second reduced rank condition. In addition, the parameterization of the $I(1)$ model does not allow us to distinguish between $\boldsymbol{\beta}$ and $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp 1})$, and therefore not to decompose $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 - \mathbf{I}$ as in (8.32). So, even though we may have realized that the $\boldsymbol{\beta}$ relation is not mean-reverting by itself, and thus that it has to be combined with the differenced process $\boldsymbol{\delta}' \Delta \mathbf{x}_t$, we would not find the estimate of $\boldsymbol{\delta}$ without knowing the estimate of $\boldsymbol{\beta}_{\perp 1}$. Furthermore, the graphs of $\boldsymbol{\beta}_1' \mathbf{R}_{1,t}$ in Figure 8.3 and of $\boldsymbol{\beta}_1' \mathbf{x}_t + \boldsymbol{\delta}_1' \Delta \mathbf{x}_t$ in Figure 8.5 suggest that the latter relation is more precisely measured in terms of stationarity.

The hypothesis of long-run price homogeneity was adequately formulated as a test on $\boldsymbol{\tau}$ in the $I(2)$ model (and rejected), whereas in the $I(1)$ model it was formulated as a necessary but not sufficient test on $\boldsymbol{\beta}$ (and accepted). Thus, based on the $I(1)$ model, one might have been tempted to believe that long-run price homogeneity was acceptable even though it was strongly rejected. The rejection of homogeneity gave one of the clues as to why there are all these puzzles in international finance.

Finally, no useful results on the common driving trends could be obtained from the $I(1)$ model, whereas the MA analysis of the $I(2)$ model provided results on the $I(1)$ and $I(2)$ stochastic trends which suggested that we need to look closer at the determination of the nominal exchange rates.

To conclude, even though the $I(1)$ and $I(2)$ models are quite close in terms of likelihood, the $I(2)$ procedure is likely to insure against possible pitfalls in the statistical analysis when there is a double unit root in the data. Last, but not least, it also allows for a much richer structure and therefore more interesting interpretations of the information in the data.

8.10 Conclusions

The CVAR approach adopted in this chapter is based on general-to-specific modeling as a tool to uncover empirical regularities in the economy. Starting from a general unrestricted model representing the raw data and then testing down seems to be a useful way of extracting as much information as possible from the data without distorting them in a prespecified direction. In this vein, it is also important from the outset to untie any transformation of the variables, such as the real exchange rate transformation of prices and nominal exchange rates, assumed to hold rather than tested in the data. Such transformations, common in empirical economics, can often seriously distort signals in the data that otherwise might help to uncover important empirical regularities. This was also the case in this chapter, where the joint modeling of prices and exchange rates revealed empirical regularities in prices and the nominal exchange rate that were helpful in pinning down the underlying puzzling behavior in this period.

To effectively pull information from the data, this chapter has argued that the vector process should be classified into directions of similar persistence, dubbed empirically $I(0)$, $I(1)$ or $I(2)$. By following this route, one can achieve more precise inference and improve the interpretability of economic behavior in the short, medium and long run. However, the main advantage is the ability to associate persistent movements away from fundamental benchmark values in one variable/relation with similar persistent movements somewhere else in the economy. In a general equilibrium world, one would expect a persistent imbalance in one sector to generate a persistent departure in another. Thus, by characterizing the data according to the empirical order of integration, the CVAR approch offers a powerful tool with which to investigate the generating mechanisms underlying such puzzling behavior.

To distinguish between those empirical regularities which can be explained by the theory model and those which cannot, the chapter has demonstrated

the importance of first translating the basic assumptions of the theory model into testable assumptions on the CVAR model. As an illustration, the chapter showed how to translate the assumption of a stationary PPP and long-run price homogeneity, together with the assumption that prices are pushing and the exchange rate is pulling, into testable hypotheses in the CVAR model. This theory consistent scenario showed, among others, that a stationary real exchange rate is inherently associated with *one* stochastic trend having generated prices and nominal exchange rates. The finding of two (rather than one) stochastic trends was particularly important, as it suggested the existence of an additional source of permanent shocks that have contributed to the persistent behavior in the data. This additional shock seemed to be related to speculative behavior in the market for foreign exchange and pointed to the importance of addressing the long swings puzzle jointly with another puzzle in international finance, the "forward premium puzzle." Similar to the long swings puzzle, the forward premium puzzle also has to do with persistent movements in the data, now in the forward premium: $(R_{1,t} - R_{2,t} - E_t \Delta s_{12,t+m})$, where $R_{i,t}$ is an interest yield of maturity m.

Thus the two puzzles are connected, in that both stem from the determination of the nominal exchange rate in the foreign exchange market. Based on a cointegrated VAR analysis of German and US prices, the exchange rate, and interest rates, Juselius and MacDonald (2007) find that *ppp* and the real interest rate spread are cointegrating though individually $I(1)$, or even near $I(2)$. This is strong evidence against the implications of the Dornbusch (1976) sticky-price model with RE that PPP and real interest parity each should hold as equilibrium cointegrating relationships. A theoretical justification for this strong feature in the data is, however, provided by Frydman *et al*. (2008), who were able to show in a two-country monetary model with IKE that goods prices and exchange rates adjust to a long-run equilibrium relation, being a combination of the *ppp* and the real interest rate spreads.

Furthermore, Johansen *et al*. (2008) report results that point to the importance of inflationary expectations measured by the term spread which was found to be empirically $I(1)$. The latter finding again points to the importance of allowing for not just one, but at least two, stochastic trends in the term structure of interest rates (Giese, 2008), and thus to a reconsideration of the monetary policy interest rate channel.

This illustrates how the VAR approach can be used constructively. Starting with the basic information set, carefully structuring the information in the data, and adding more information if needed, might at an early stage suggest how to modify either the empirical or the economic model, or both.

The following passage from Hoover (2006) pinpoints the fundamental difference between an approach based on *a priori* theory and the general-to-specific approach to empirical economics:

> The Walrasian approach is totalizing. Theory comes first. Empirical reality must be theoretically articulated before it can be empirically observed. There is a sense that the Walrasian attitude is that to know anything, one must know everything.

> ... There is a fundamental problem: How do we come to our a priori knowledge? Most macroeconomists expect empirical evidence to be relevant to our understanding of the world. But if that evidence only can be viewed through totalizing *a priori* theory, then it cannot be used to revise the theory.
>
> ... The Marshallian approach is archaeological. We have some clues that a systematic structure lies behind the complexities of economic reality. The problem is how to lay this structure bare. To dig down to find the foundations, modifying and adapting our theoretical understanding as new facts accumulate, becoming ever more confident in our grasp of the super structure, but never quite sure that we have reached the lowest level of the structure.

For example, the significant finding of two shocks rather than one and the rejection of long-run price homogeneity are two examples of important information in the data, signaling the need to dig deeper in order to understand more. By taking this information in the data seriously instead of just ignoring it, we have been able to uncover more structure, and thus to improve our understanding, as demonstrated in Frydman *et al.* (2008) and Johansen *et al.* (2008). Needless to say, the need to dig deeper does not stop here.

Notes

1. Note that the *ppp* term is also the (logarithm) of the real exchange rate. We prefer to use the label *ppp* in this chapter because we are adopting a parity perspective and also because we do not model the real exchange rate in terms of so-called real fundamentals.
2. Johansen (2006a) demonstrated that valid inference on steady-state values requires more than 5,000 observations if the model contains a near unit root of 0.998.
3. German CPI has been additively mean corrected for the reunification in 1991:1 prior to the VAR analysis.
4. Note, however, that this diagnostic check is only reliable in a VAR model with a correct lag length. A VAR model with too many lags will often generate complex pairs of large (albeit insignificant) roots in the characteristic polynomial (Nielsen and Nielsen, 2006).
5. When the data are $I(2)$, price homogeneity of $\boldsymbol{\beta}'\mathbf{x}_t$ is a necessary but not sufficient condition, as will be discussed in section 8.8.
6. When the lag $k > 2$, there would also be lagged acceleration rates, $\Delta^2\mathbf{x}_{t-1}$, to concentrate out.
7. At this stage, $\boldsymbol{\Phi}_p$ will be left unrestricted in the model.
8. A linear trend in $\boldsymbol{\alpha}'_{\perp 1}\Delta\mathbf{x}_t$ would imply that the inflation rate, say, is allowed to grow with a linear trend, and thus prices with a quadratic trend. It would be hard to argue for such a specification, except possibly as a local approximation.
9. From a statistical point of view, an equivalent transformation would be achieved by replacing Δp_1 with Δp_2.

References

Balassa, B. (1964) The purchasing parity doctrine: a reappraisal. *Journal of Political Economy* **72**, 584–96.

Cheung, Y.W. and K.S. Lai (1993) Long-run purchasing power parity during the recent float. *Journal of International Economics* **34**, 181–92.

Dennis, J., S. Johansen and K. Juselius (2005) *CATS for RATS: Manual to Cointegration Analysis of Time Series*. Estima, Illinois.

Dornbusch, R. (1976) Expectations and exchange rate dynamics. *Journal of Political Economy*, December, 1161–74.

Froot, K. and K. Rogoff (1995) Perspectives on PPP and long-run real exchange rates. In E. Grossman and K. Rogoff (eds.), *Handbook of International Economics*, *Volume 3*. Amsterdam: North-Holland.

Frydman, R. and M. Goldberg (2007) *Imperfect Knowledge Economics: Exchange Rates and Risk*. Princeton: Princeton University Press.

Frydman, R., M. Goldberg, S. Johansen and K. Juselius (2008) Imperfect knowledge and a resolution of the purchasing power parity puzzle. Working Paper, Center an Capitalism and Society, Columbia University, and University of Copenhagen.

Giese, J. (2008) Level, slope and curvature: characterising the yield curve in a co-integrated VAR model. *Economics – The Open-Access, Open Assessment E-Journal* **2**, 28, http://www.economics-ejournal.org/economics/journalarticles/2008-28.

Gonzalo, J. (1994) Five alternative methods of estimating long-run equilibrium relationships. *Journal of Econometrics* **60**, 203–33.

Hoover, K. (2006) The past as the future: the Marshallian approach to post-Walrasian econometrics. In D. Colander (ed.), *Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model*, Ch. 12. Cambridge: Cambridge University Press.

Hoover, K., S. Johansen and K. Juselius, (2008) Allowing the data to speak freely: the macroeconometrics of the cointegrated vector autoregression. *Proceedings of the American Economic Association*.

Johansen, S. (1992) A representation of vector autoregressive processes integrated of order 2. *Econometric Theory* **8**, 188–202.

Johansen, S. (1995) A statistical analysis of cointegration for $I(2)$ variables. *Econometric Theory* **11**, 25–59.

Johansen, S. (1996) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Johansen, S. (1997) Likelihood analysis of the $I(2)$ model. *Scandinavian Journal of Statistics* **24**, 433–62.

Johansen, S. (2005) The interpretation of cointegrating coefficients in the cointegrated vector autoregressive model. *Oxford Bulletin of Economics and Statistics* **67**, 93–104.

Johansen, S. (2006a) Confronting the economic model with the data. In D. Colander (ed.), *Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic Equilibrium Model*, pp. 287–300. Cambridge: Cambridge University Press.

Johansen, S. (2006b) Statistical analysis of hypotheses on the cointegrating relations in the $I(2)$ model. *Journal of Econometrics* **132**, 81–115.

Johansen, S. (2008) Representation of cointegrated autoregressive processes with application to fractional processes. *Econometric Reviews*. Forthcoming.

Johansen, S. and K. Juselius (1992) Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics* **53**, 211–44.

Johansen, S., K. Juselius, R. Frydman and M.D. Goldberg (2008) Testing hypotheses in an $I(2)$ model with applications to the persistent long swings in the Dmk/$ Rate. *Journal of Econometrics*. Forthcoming.

Juselius, K. (1995) Do purchasing power parity and uncovered interest rate parity hold in the long run? An example of likelihood inference in a multivariate time-series model. *Journal of Econometrics* **69**, 211–40.

Juselius, K. (2006) *The Cointegrated VAR Model: Methodology and Applications*. Oxford: Oxford University Press.

Juselius, K. and M. Franchi (2007) Taking a DSGE model to the data meaningfully *Economics – The Open-Access, Open-Assessment E-Journal* **4**, http://www.economics-ejournal.org/economics/journalarticles/2007-4.

Juselius, K. and S. Johansen (2006) Extracting information from the data: a European view on empirical macro. In D. Colander (ed.), *Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic Equilibrium Model*, pp. 301–34. Cambridge: Cambridge University Press.

Juselius, K. and R. MacDonald (2007) International parity relationships between Germany and the United States: a joint modelling approach. In A. Morales-Zumaquero (ed.), *International Macroeconomics: Recent Developments*. Nova Science Publishers.

Kongsted, H.C. (2005) Testing the nominal-to-real transformation. *Journal of Econometrics* **124**, 205–25.

Kongsted, H.C., A. Rahbek and C. Jørgensen (1999) Trend stationarity in the $I(2)$ cointegration model. *Journal of Econometrics* **90**, 265–89.

Kurita, T. (2007) $I(2)$ cointegration analysis in the presence of deterministic shifts. Faculty of Economics, Fukuoka University.

MacDonald, R. (1995) Long-run exchange rate modelling: a survey of the recent evidence International Monetary Fund Staff Papers No. 42.

Nielsen, H.B. (2004) Cointegration analysis in the presence of outliers. *Econometrics Journal* **7**, 249–71.

Nielsen, B. and H.B. Nielsen (2006) The asymptotic distribution of the estimated characteristic roots in a second order autoregression. Preprint, Economics Department, University of Copenhagen.

Nielsen, B. and A. Rahbek (2000) Similarity issues in cointegration analysis. *Oxford Bulletin of Economics and Statistics* **62**, 5–22.

Nielsen, H.B. and A. Rahbek (2007) The likelihood ratio test for cointegration ranks in the $I(2)$ model. *Econometric Theory* **23**, 615–37.

Paruolo, P. (2000) Asymptotic efficiency of the two stage estimator in $I(2)$ systems. *Econometric Theory* **16**, 524–50.

Paruolo, P. (2002) Asymptotic inference on the moving average impact matrix in cointegrated $I(2)$ VAR systems. *Econometric Theory* **18**, 673–90.

Paruolo, P. and A. Rahbek (1999) Weak exogeneity in $I(2)$ VAR systems. *Journal of Econometrics* **93**, 281–308.

Rogoff, K. (1996) The purchasing power parity puzzle. *Journal of Economic Literature* **34**, 647–68.

Samuelson, P. (1964) Theoretical notes on trade problems. *Review of Economics and Statistics* **46**, 145–54.

9

Structural Time Series Models for Business Cycle Analysis

Tommaso Proietti

Abstract

The chapter deals with parametric models for the measurement of the business cycle in economic time series. It presents univariate methods based on parametric trend-cycle decompositions and multivariate models featuring a Phillips-type relationship between the output gap and inflation and the estimation of the gap using mixed frequency data. We finally address the issue of assessing the accuracy of the output gap estimates.

9.1 Introduction

The term "structural time series" refers to a class of parametric models that are specified directly in terms of unobserved components which capture essential features of the series, such as trends, cycles and seasonality. The approach is suitable for the analysis of macroeconomic time series, where latent variables, such as trends and cycles, and more specialized notions, such as the output gap, core inflation and the natural rate of unemployment, need to be measured.

One of the key issues economists have faced in characterizing the dynamic behavior of macroeconomic variables, such as output, unemployment and inflation, is separating trends from cycles. The decomposition of economic time series has a long tradition, dating back to the nineteenth century (see the first chapter of Mills, 2003, for an historical perspective). Along with providing a description of the salient features of a series, the distinction between what is permanent and what is transitory in economic dynamics has important implications for monetary and fiscal policy. The underlying idea is that trends and cycles can be ascribed to different economic mechanisms and an understanding of their determinants helps to define policy targets and instruments.

This chapter focuses on structural time series modeling for business cycle analysis and, in particular, for output gap measurement. The *output gap* is the deviation of the economy's realized output from its potential. Potential output is defined as the noninflationary level of output, i.e., as the level that can be attained using the available technology and productive factors at a stable inflation rate. The gap measures the presence and the extent of real disequilibria and constitutes an indicator of inflationary pressure in the short run: a positive output gap testifies to excess demand, while a negative output gap implies excess supply.

The output gap plays a central role in the transmission mechanism of monetary policy, since short-term interest rates influence aggregate demand and the latter affects inflation via a Phillips curve relationship. The Phillips curve establishes a trade-off between output and inflation over the short run, and provides the rationale for using the short run component in output as an indicator of demand-driven inflationary pressure. For instance, the Taylor rule (Taylor, 1999) explicitly links the central bank's policy to the output gap. On the other hand, the growth rate of potential output is a reference value for broad money growth. Other important uses of the output gap are in fiscal analysis, where it is employed to assess the impact of cyclical factors on budget deficits, and in the adjustment of exchange rates. The output gap is also related to cyclical unemployment, which is the deviation of unemployment from its trend, known as the non-accelerating inflation rate of unemployment (NAIRU).

The signal extraction problems relating to latent variables, such as the output gap, core inflation and the NAIRU, can be consistently formulated within a model-based framework and, in particular, within the class of unobserved components time series models, so matching the fundamental economic relationships with observable macroeconomic aggregates.

The chapter is divided into three main parts: the first (section 9.2) deals with univariate methods for cycle measurement. One approach is to formalize a model of economic fluctuations such that the different components are driven by specific shocks that are propagated via a dynamic transmission mechanism. We start by introducing the traditional trend-cycle structural decomposition, discussing the parametric representation of both components (sections 9.2.1–4) and the correlation between the trend and cycle disturbances (section 9.2.5). Another approach is to consider the cycle as the bandpass component of output, i.e., as those economic fluctuations which have a periodicity greater then a year and smaller than, say, eight years. We review the relationship between popular signal extraction filters, such as the Hodrick–Prescott and the Baxter and King filters, and the model-based Wiener–Kolmogorov filter. Particular attention is devoted to the implementation of bandpass filtering in a model-based framework (section 9.2.6). The advantages of this strategy are twofold: the components can be computed in real time using standard principles of optimal signal extraction, so that efficient algorithms, such as the Kalman filter and smoother, can be applied. Second, the reliability of the estimated components can be thoroughly assessed.

The second part, starting with section 9.3, deals with multivariate models for the measurement of the output gap. The above definition of the output gap as an indicator of inflationary pressures suggests that the minimal measurement framework is of a bivariate model for output and inflation. After reviewing the work done in this area (section 9.3.1), we illustrate the estimation of a bivariate model for the US economy, under both the classical and the Bayesian approaches, and incorporating the feature known as the "Great Moderation" of the volatility of economic fluctuations (section 9.3.2). In section 9.3.3 we review the multivariate extensions of the basic bivariate model and we conclude this part with an application which serves to illustrate the flexibility of the state-space methodology in accommodating data features such as missing data, nonlinearities and temporal aggregation. In particular, section 9.3.4 presents the results of fitting a four-variate monthly time series model for the US economy with mixed frequency data, as gross domestic product (GDP) is available only quarterly, whereas industrial production, the unemployment rate and inflation are monthly. The model incorporates the temporal aggregation constraints (which are nonlinear since the model is formulated in terms of the logarithm of the variables) and produces as a byproduct monthly estimates of GDP, along with their reliability, that are consistent with the quarterly observed values.

The third part, section 9.4, deals with the reliability of the output gap estimates. The assessment of the quality of the latter is crucial for the decision maker. We discuss the various sources of uncertainty (model selection, parameter estimation, data revision, estimation of unboserved components,

statistical revision), and discuss ways of dealing with them using the state-space methodology.

One of the objectives of this chapter is to provide an overview of the main state-space methods and to illustrate their application and scope. The description of the algorithms is relegated to an appendix and we refer to Harvey (1989), West and Harrison (1997), Kitagawa and Gersch (1996), Durbin and Koopman (2001), and the selection of readings in Harvey and Proietti (2005), for a thorough presentation of the main ideas and methodological aspects concerning state-space methods and unobserved components models. For the class of state-space models with Markov-switching, see Kim and Nelson (1999b), Frühwirth-Schnatter (2006) and Cappé, Moulines and Ryden (2005). An essential and up-to-date monograph on modeling trends and cycles in economics is Mills (2003).

9.2 Univariate methods

In univariate analysis, the output gap can be identified as the stationary or transitory component in a measure of aggregate economic activity, such as GDP. Estimating the output gap thus amounts to *detrending* the series and a large literature has been devoted to this very controversial issue (see, e.g., Canova, 1998; Mills, 2003).

We shall confine our attention to the additive decomposition (after a logarithmic transformation) of real output, y_t, into potential output, μ_t, and the output gap, ψ_t: $y_t = \mu_t + \psi_t$. This basic representation is readily extended to handle a seasonal component and other calendar components such as those associated with trading days and moving festivals, which for certain output series, e.g., industrial production, play a relevant role.

In the structural approach a parametric representation for the components is needed; furthermore, the specification of the model is completed by assumptions on the covariances among the various components. The first identifying restriction that will be adopted throughout is that μ_t is fully responsible for the non-stationary behavior of the series, whereas ψ_t is a transitory component.

9.2.1 The random walk plus noise model

The random walk plus noise (RWpN) model provides the most basic trend-cycle decomposition of output, such that the trend is a random walk process with normal and independently distributed (NID) increments, and the cycle is a pure white noise (WN) component. The structural specification is the following:

$$\begin{aligned} y_t &= \mu_t + \psi_t, \qquad t = 1, \ldots, n, \qquad & \psi_t &\sim \text{NID}(0, \sigma_\psi^2), \\ \mu_t &= \mu_{t-1} + \beta + \eta_t, & \eta_t &\sim \text{NID}(0, \sigma_\eta^2). \end{aligned} \tag{9.1}$$

When the drift is absent, i.e., when $\beta = 0$, the model is also known as the *local level model* (see Harvey, 1989). We assume throughout that $\text{E}(\eta_t \psi_{t-j}) = 0$ for all t and j, so that the two components are orthogonal.

If $\sigma_\eta^2 = 0$, μ_t is a deterministic linear trend. The one-sided Lagrange multiplier test of the null hypothesis $H_0 : \sigma_\eta^2 = 0$ against the alternative $H_1 : \sigma_\eta^2 > 0$, is

known as a stationarity test and is discussed in Nyblom and Mäkeläinen (1983). The nonparametric extension to the case when ψ_t is any indeterministic stationary process is provided by Kwiatkowski, Phillips, Schmidt and Shin (KPSS) (1992) (see also Harvey, 2001, for a review and extensions).

The reduced-form representation of (9.1) is an integrated moving average model of orders (1,1), or IMA(1,1): $\Delta y_t = \beta + \xi_t + \theta\xi_{t-1}, \xi_t \sim \text{NID}(0, \sigma^2)$, where $\Delta y_t = y_t - y_{t-1}$. The difference operator can be defined in terms of the lag operator L, such that $L^d y_t = y_{t-d}$, for an integer d, as $\Delta = (1 - L)$.

The moving average (MA) parameter is subject to the restriction $-1 \leq \theta \leq 0$. Equating the autocovariance generating functions of Δy_t implied by the IMA(1,1) and by the structural representation (9.1), it is possible to establish that $\sigma_\eta^2 = (1+\theta)^2\sigma^2$ and $\sigma_\psi^2 = -\theta\sigma^2$. Hence it is required that $\theta \leq 0$, so that persistence, $(1+\theta)$, cannot be greater than unity. The variance ratio $\lambda = \sigma_\psi^2/\sigma_\eta^2$ depends uniquely on θ, as $\lambda = -\theta/(1+\theta)^2$. The ratio provides a measure of relative smoothness of the trend: if λ is large, then the trend varies little with respect to the noise component, and thus it can be regarded as "smooth."

The RWpN model has a long tradition and a well-established role in the analysis of economic time series, since it provides the model-based interpretation for the popular forecasting technique known as *exponential smoothing*, which is widely used in applied economic forecasting and fares remarkably well in forecast competitions (see Muth, 1960, and the comprehensive reviews by Gardner, 1985, 2006).

Assuming a doubly infinite sample, the one-step-ahead predictions, $\tilde{\mu}_{t+1|t}$, and the filtered and smoothed estimates of the trend component, denoted $\tilde{\mu}_{t|\infty}$, are given, respectively, by:

$$\tilde{\mu}_{t+1|t} = \tilde{\mu}_{t|t} = (1+\theta)\sum_{j=0}^{\infty}(-\theta)^j y_{t-j}, \quad \tilde{\mu}_{t|\infty} = \frac{1+\theta}{1-\theta}\sum_{j=-\infty}^{\infty}(-\theta)^{|j|}y_{t-j}.$$

Here, $\tilde{\mu}_{t+1|t}$ denotes the expectation of μ_{t+1} based on the information available at time t, whereas $\tilde{\mu}_{t|\infty}$ is the expectation based on all of the information in the doubly infinite data set. The filter $\text{w}(L) = (1+\theta)(1+\theta L)^{-1} = (1+\theta)\sum_{j=0}^{\infty}(-\theta)^j L^j$ is known as a one-sided exponentially weighted moving average (EWMA). These expressions follow from applying the Wiener–Kolmogorov prediction and signal extraction formulae (see Appendix B). In terms of the structural form parameters, $\tilde{\mu}_{t|\infty} = \frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma_\psi^2|1-L|^2}y_t$, where $|1-L|^2 = (1-L)(1-L^{-1})$. The filter:

$$\text{w}_\mu(L) = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\psi^2|1-L|^2} = \frac{1+\theta}{1-\theta}\sum_{j=-\infty}^{\infty}(-\theta)^{|j|}L^j,$$

is known as a two-sided EWMA filter. In finite samples, the computations are performed by the Kalman filter and smoother (see Appendix C).

The parameter θ (or, equivalently, λ) is essential in determining the weights that are attached to the observations for signal extraction and prediction. When

$\theta = 0$, y_t is a pure random walk, and then the current observation provides the best estimate of the trend: $\tilde{\mu}_{t+1|t} = \tilde{\mu}_{t|t} = \tilde{\mu}_{t|\infty} = y_t$. When $\theta = -1$, the trend estimate, which is as smooth as possible, is a straight line passing through the observations.

The RWpN model provides a stripped to the bone separation of the transitory and permanent dynamics that depends on a single smoothness parameter, which determines the weights that are assigned to the available observations for forecasting and trend estimation. Its use as a misspecified model of economic fluctuations for out-of-sample forecasting, using multi-step (or adaptive) estimation, rather than maximum likelihood (ML) estimation, has been considered in the seminal paper by Cox (1961) and by Tiao and Xu (1993). Proietti (2005) discusses multi-step estimation of the RWpN model for the extraction of trends and cycles.

9.2.2 The local linear model and the Leser–HP filter

In the local linear trend model (LLTM) the trend μ_t is an integrated random walk:

$$\begin{array}{lllllll} y_t &=& \mu_t &+& \psi_t, & & \psi_t \sim \mathrm{NID}(0, \sigma_\psi^2), \quad t = 1, 2, \ldots, n, \\ \mu_t &=& \mu_{t-1} &+& \beta_{t-1} + \eta_t, & & \eta_t \sim \mathrm{NID}(0, \sigma_\eta^2), \\ \beta_t &=& & & \beta_{t-1} + \zeta_t, & & \zeta_t \sim \mathrm{NID}(0, \sigma_\zeta^2). \end{array} \tag{9.2}$$

It is assumed that ψ_t, η_t and ζ_t are mutually and serially uncorrelated. For $\sigma_\zeta^2 = 0$ the trend reduces to a random walk with constant drift, whereas for $\sigma_\eta^2 = 0$ the trend is an integrated random walk ($\Delta^2 \mu_t = \zeta_{t-1}$).

The above representation encompasses a deterministic linear trend, arising when both σ_η^2 and σ_ζ^2 are zero. Second, it is consistent with the notion that the real time estimate of the trend is coincident with the value of the eventual forecast function at the same time (see section 9.2.5 on the Beveridge–Nelson decomposition).

The LLTM is the model for which the Leser filter is optimal (see Leser, 1961). The latter is derived as the minimizer, with respect to $\mu_t, t = 1, \ldots, n$, of the penalized least squares (PLS) criterion:

$$PLS \;=\; \textstyle\sum_{t=1}^{n} (y_t - \mu_t)^2 + \lambda \sum_{t=3}^{n} (\Delta^2 \mu_t)^2.$$

The parameter λ governs the trade-off between fidelity and smoothness and it is referred to as the *smoothness* or *roughness penalty* parameter. The first addend of *PLS* measures the goodness-of-fit, whereas the second penalizes the departure from zero of the variance of the second differences (i.e., a measure of roughness). In matrix notation, if $\mathbf{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, and $\mathbf{D} = \{d_{ij}\}$ is the $n \times n$ matrix corresponding to a first difference filter, with $d_{ii} = 1$, $d_{i,i-1} = -1$ and zero otherwise, so that $\mathbf{D}\boldsymbol{\mu} = (\mu_2 - \mu_1, \ldots, \mu_n - \mu_{n-1})'$, we can write the criterion function as $PLS = (\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu}) + \lambda \boldsymbol{\mu}' \mathbf{D}^{2'} \mathbf{D}^2 \boldsymbol{\mu}$. Differentiating with respect to $\boldsymbol{\mu}$, the first-order conditions yield: $\tilde{\boldsymbol{\mu}} = (\mathbf{I}_n + \lambda \mathbf{D}^{2'} \mathbf{D}^2)^{-1} \mathbf{y}$. The rows of the matrix $(\mathbf{I}_n + \lambda \mathbf{D}^{2'} \mathbf{D}^2)^{-1}$ contain the filter weights for estimating the trend at a particular point in time. The solution arising for $\lambda = 1600$ is widely known in the analysis of quarterly macroeconomic time series as the Hodrick–Prescott filter (henceforth, HP; see Hodrick and Prescott, 1997); the choice of the smoothness parameter for

yearly and monthly time series is discussed in Ravn and Uhlig (2002) and Maravall and del Rio (2007).

We now show that the Leser filter is the optimal signal extraction filter for the LLTM (9.2) with $\sigma_\eta^2 = 0$ and $\lambda = \sigma_\psi^2/\sigma_\zeta^2$. In fact, apart from an additive term which does not depend on μ, *PLS* is proportional to $\ln f(\mathbf{y}, \boldsymbol{\mu}) = \ln f(\mathbf{y}|\boldsymbol{\mu}) + \ln f(\boldsymbol{\mu})$, where $f(\mathbf{y}, \boldsymbol{\mu})$, $f(\mathbf{y}|\boldsymbol{\mu})$ denote, respectively, the Gaussian joint density of the random vectors $\mathbf{y}$ and $\boldsymbol{\mu}$, and the conditional density of $\mathbf{y}$ given $\boldsymbol{\mu}$, whereas $f(\boldsymbol{\mu})$ is the joint density of $\boldsymbol{\mu}_t, t = 1, \ldots, n$. Now, $\ln f(\mathbf{y}|\boldsymbol{\mu})$ depends on $\boldsymbol{\mu}$ only via $(1/\sigma_\psi^2)\sum_{t=1}^n (y_t - \mu_t)^2$, whereas $\ln f(\boldsymbol{\mu}) = \ln f(\mu_3, \ldots, \mu_n|\mu_1, \mu_2) + \ln f(\mu_1, \mu_2)$. The first term depends on $\mu_t, t > 2$, only via $(1/\sigma_\zeta^2)\sum_{t=3}^n (\Delta^2 \mu_t)^2$. The contribution of the initial values vanishes under fixed initial conditions or diffuse initial conditions.[1] In conclusion, $\tilde{\boldsymbol{\mu}}$ maximizes, with respect to $\boldsymbol{\mu}$, the joint log-density $\ln f(\mathbf{y}, \boldsymbol{\mu})$ and thus the posterior log-density $\ln f(\boldsymbol{\mu}|\mathbf{y}) = \ln f(\mathbf{y}, \boldsymbol{\mu}) - \ln f(\mathbf{y})$. A consequence of this result is that the components can be efficiently computed using the Kalman filter and smoother (see Appendix C). The latter computes the mean of the conditional distribution $\boldsymbol{\mu}|\mathbf{y}$. As this distribution is Gaussian, the posterior mean is equal to the posterior mode. Hence, the smoother computes the mode of $f(\boldsymbol{\mu}|\mathbf{y})$, which is also the minimizer of the *PLS* criterion.

The equivalence $\lambda = \sigma_\psi^2/\sigma_\zeta^2$ makes clear that the roughness penalty measures the variability of the cyclical (noise) component relative to that of the trend disturbance, and regulates the smoothness of the long-term component. As σ_ζ^2 approaches zero, λ tends to infinity, and the limiting representation of the trend is a straight line. The Leser–HP detrended or cyclical component is the smoothed estimate of the component ψ_t in (9.2) and, although the maintained representation for the deviations from the trend is a WN component, the filter has been one of the most widely employed tools in macroeconomics to extract a measure of the business cycle. For the US GDP series (logarithms) this component is plotted in the top right-hand panel of Figure 9.4 (p. 404).

In terms of the reduced form of model (9.2), the IMA(2,2) model $\Delta^2 y_t = (1 + \theta_1 L + \theta_2 L^2)\xi_t$, $\xi_t \sim \text{NID}(0, \sigma^2)$, it can be shown that the restriction $\sigma_\eta^2 = 0$ implies $[(1 + \theta_2)\theta_2]/(1 - \theta_2)^2 = \lambda$ and $\theta_1 = -4\theta_2/(1 + \theta_2)$. Therefore, for $\lambda = 1600$, we have $\theta_1 = -1.778$ and $\theta_2 = 0.799$, so that $\theta(1) = 1 + \theta_1 + \theta_2 = 0.021$ and the MA polynomial is close to noninvertibility at the zero frequency.

The theoretical properties of the Leser–HP filter are better understood by assuming the availability of a doubly-infinite sample, $y_{t+j}, j = -\infty, \ldots, \infty$. In such a setting, the Wiener–Kolmogorov filter (see Whittle, 1983, and Appendix B) provides the minimum mean square linear estimator (MMSLE) of the trend, $\tilde{\mu}_{t|\infty} = \text{w}_\mu(L)y_t$, where:

$$\text{w}_\mu(L) = \frac{\sigma_\zeta^2}{\sigma_\zeta^2 + |1 - L|^4 \sigma_\psi^2} = \frac{1}{1 + \lambda|1 - L|^4}. \tag{9.3}$$

The frequency response function of the trend filter (see Appendix A) is:

$$\text{w}_\mu(e^{-\imath\omega}) = \frac{1}{1 + 4\lambda(1 - \cos\omega)^2}, \quad \omega \in [0, \pi];$$

notice that this is 1 at the zero frequency and decreases monotonically to zero as ω approaches π. This behavior enforces the interpretation of (9.3) as a lowpass filter, and the corresponding detrending filter, $1 - \mathrm{w}_\mu(L)$, is the highpass filter derived from it. We shall return to this issue in the next section.

9.2.3 Higher-order trends and lowpass filters

A *lowpass* filter is a filter that passes low-frequency fluctuations and reduces the amplitude of fluctuations with frequencies higher than a cut-off frequency ω_c (see, e.g., Percival and Walden, 1993). The frequency response function of an ideal lowpass filter takes the following form: for $\omega \in [0, \pi]$,

$$\mathrm{w}_{lp}(\omega) = \begin{cases} 1 & \text{if } \omega \leq \omega_c \\ 0 & \text{if } \omega_c < \omega \leq \pi. \end{cases}$$

The notion of a highpass filter is complementary, its frequency response function being $\mathrm{w}_{hp}(\omega) = 1 - \mathrm{w}_{lp}(\omega)$. The coefficients of the ideal lowpass filter are provided by the inverse Fourier transform of $\mathrm{w}_{lp}(\omega)$:

$$\mathrm{w}_{lp}(L) = \frac{\omega_c}{\pi} + \sum_{j=1}^{\infty} \frac{\sin(\omega_c j)}{\pi j}(L^j + L^{-j}).$$

A bandpass filter is a filter that passes fluctuations within a certain frequency range and attenuates those outside that range. Given lower and upper cut-off frequencies, $\omega_{1c} < \omega_{2c}$ in $(0, \pi)$, the ideal frequency response function is unity in the interval $[\omega_{1c}, \omega_{2c}]$ and zero outside. The notion of a bandpass filter is relevant to business cycle measurement: the traditional definition, ascribed to Burns and Mitchell (1946), considers all the fluctuations with a specified range of periodicities, namely those ranging from one and a half to eight years. Thus, if s is the number of observations in a year, fluctuations with periodicity between $1.5s$ and $8s$ are included. Baxter and King (1999; henceforth, BK) argue that the ideal filter for cycle measurement is a bandpass filter. Now, given the two business cycle frequencies, $\omega_{c1} = 2\pi/(8s)$ and $\omega_{c2} = 2\pi/(1.5s)$, the bandpass filter is:

$$\mathrm{w}_{bp}(L) = \frac{\omega_{c2} - \omega_{c1}}{\pi} + \sum_{j=1}^{\infty} \frac{\sin(\omega_{c2} j) - \sin(\omega_{c1} j)}{\pi j}(L^j + L^{-j}). \tag{9.4}$$

Notice that $w_{bp}(L)$ is the contrast between the two lowpass filters with cut-off frequencies ω_{c2} and ω_{c1}. The frequency response function of the ideal business cycle bandpass filter for quarterly observations ($s = 4$), which is equivalent to the gain function (see Appendix A), is plotted in Figure 9.3 (p. 403).

The ideal bandpass filter exists and is unique, but as it entails an infinite number of leads and lags, an approximation is required in practical applications. BK show that the K-terms approximation to the ideal filter (9.4), which is optimal in the sense of minimizing the integrated mean square approximation error, is obtained from (9.4) by truncating the lag distribution at a finite integer K. They propose

using a three-year window, i.e., $K = 3s$, as a valid rule of thumb for macroeconomic time series. They also constrain the weights to sum to zero, so that the resulting approximation is a detrending filter: denoting the truncated filter $\mathrm{w}_{bp,K}(L) = \mathrm{w}_0 + \sum_1^K \mathrm{w}_j(L^j + L^{-j})$, the weights of the adjusted filter will be $\mathrm{w}_j - \mathrm{w}_{bp,K}(1)/(2K+1)$. The gain of the resulting filter is displayed in Figure 9.3 (henceforth we shall refer to it as the BK filter). The ripples result from the truncation of the ideal filter and are referred to as the Gibbs phenomenon (see Percival and Walden, 1993, p. 177). BK do not entertain the problem of estimating the cycle at the extremes of the available sample; as a result the estimates for the first and last three years are unavailable. Christiano and Fitzgerald (2003) provide the optimal finite-sample approximations for the bandpass filter, including the real-time filter, using a model-based approach.

Within the class of parametric structural models, an important category of lowpass filter emerges from the application of Wiener–Kolmogorov optimal signal extraction theory to the following model:

$$\begin{array}{lcll} y_t & = & \mu_t + \psi_t, & t = 1, 2, \ldots, n, \\ \Delta^m \mu_t & = & (1+L)^r \zeta_t, \quad \zeta_t \sim \mathrm{NID}(0, \sigma_\zeta^2), & \\ \psi_t & \sim & \mathrm{NID}(0, \lambda\sigma_\zeta^2), \quad \mathrm{E}(\zeta_t, \psi_{t-j}) = 0, \forall j, & \end{array} \tag{9.5}$$

where μ_t is the signal or trend component, and ψ_t is the noise.

Assuming a doubly-infinite sample, the minimum mean square estimators of the components (see Appendix B) are, respectively, $\tilde{\mu}_t = \mathrm{w}_\mu(L) y_t$ and $\tilde{\psi}_t = y_t - \tilde{\mu}_t = [1 - \mathrm{w}_\mu(L)] y_t$, where:

$$\mathrm{w}_\mu(L) = \frac{|1+L|^{2r}}{|1+L|^{2r} + \lambda |1-L|^{2m}}. \tag{9.6}$$

The expression (9.6) defines a class of filters which depends on the order of integration of the trend (m, which regulates its flexibility), on the number of unit poles at the Nyquist frequency r, which *ceteris paribus* regulates the smoothness of $\Delta^m \mu_t$, and λ, which measures the relative variance of the noise component.

The Leser–HP filter arises for $m = 2, r = 0, \lambda = 1600$ (quarterly data). The two-sided EWMA filter arises for $m = 1, r = 0$. The filters arising for $m = r$ are Butterworth filters of the tangent version (see, e.g., Gómez, 2001). The analytical expression of the gain is:

$$\mathrm{w}_\mu(\omega) = \left\{ 1 + \left[\frac{\tan(\omega/2)}{\tan(\omega_c/2)} \right]^{2m} \right\}^{-1},$$

and depends solely on m and ω_c. As $m \to \infty$ the gain converges to the frequency response function of the ideal lowpass filter.

The previous discussion enforces the interpretation of the trend filter $\mathrm{w}_\mu(L)$ as a lowpass filter. Its cut-off frequency depends on the triple (m, r, λ). Frequency domain arguments can be advocated for designing these parameters so as to select the fluctuations that lie in a predetermined periodicity range. In particular, let us consider the Fourier transform of the trend filter (9.6), $\mathrm{w}_\mu(\omega) = \mathrm{w}_\mu(e^{-\imath\omega})$, $\omega \in [0, \pi]$,

which also expresses the gain of the filter. The latter is monotonically decreasing with λ; it takes the value 1 at the zero frequency and, if $r > 0$, it is zero at the Nyquist frequency. The trend filter will preserve to a great extent those fluctuations at frequencies for which the gain is greater than 1/2 and reduce to a given extent those for which the gain is below 1/2. This simple argument justifies the definition of a lowpass filter with cut-off frequency ω_c if the gain halves at that frequency; see Gómez (2001, sec. 1). Usually the investigator sets the cut-off frequency to a particular value, e.g., $\omega_c = 2\pi/(8s)$ and chooses the values of m and r (e.g., $m = 2, r = 0$ for the Leser–HP filter). Solving the equation $\mathrm{w}_\mu(\omega_c) = 1/2$, the parameter λ can be obtained in terms of the cut-off frequency and the orders m and r:

$$\lambda = 2^{r-m}\left[\frac{(1+\cos\,\omega_c)^r}{(1-\cos\,\omega_c)^m}\right]. \tag{9.7}$$

9.2.4 The cyclical component

In the previous section we considered some of the most popular decompositions of a time series into a trend and pure white noise component. Hence, the previous models are misspecified. In the analysis of economic time series it is more interesting to entertain a trend-cycle decomposition such that the trend is due to the accumulation of supply shocks that are permanent, whereas the cycle is ascribed to nominal or demand shocks that are propagated by a stable transmission mechanism. Clark (1987) and Harvey and Jaeger (1993), e.g., replace the irregular component by a stationary stochastic cycle, which is parameterized as an AR(2) or an ARMA(2,1) process, such that the roots of the AR polynomial are a pair of complex conjugates. The model for the cycle is a stationary process capable of reproducing widely acknowledged stylized facts, such as the presence of strong autocorrelation, determining the recurrence and alternation of phases, and the dampening of fluctuations, or zero long-run persistence.

In particular, the model adopted by Clark (1987) is:

$$\psi_t = \phi_1\psi_{t-1} + \phi_2\psi_{t-2} + \kappa_t, \quad \kappa_t \sim \mathrm{NID}(0, \sigma_\kappa^2),$$

where κ_t is independent of the trend disturbances. Harvey (1989) and Harvey and Jaeger (1993) use a different representation:

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho\begin{bmatrix} \cos\,\varpi & \sin\,\varpi \\ -\sin\,\varpi & \cos\,\varpi \end{bmatrix}\begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}, \tag{9.8}$$

where $\kappa_t \sim \mathrm{NID}(0, \sigma_\kappa^2)$ and $\kappa_t^* \sim \mathrm{NID}(0, \sigma_\kappa^2)$ are mutually independent and independent of the trend disturbance, $\varpi \in [0, \pi]$ is the frequency of the cycle and $\rho \in [0, 1)$ is the damping factor. The reduced form of (9.8) is the ARMA(2,1) process:

$$(1 - 2\rho\cos\,\varpi L + \rho^2L^2)\psi_t = (1 - \rho\cos\,\varpi L)\kappa_t + \rho\sin\,\varpi\kappa_{t-1}^*.$$

When ρ is strictly less than one the cycle is stationary with $\mathrm{E}(\psi_t) = 0$ and $\sigma_\psi^2 = \mathrm{Var}(\psi_t) = \sigma_\kappa^2/(1-\rho^2)$; the autocorrelation at lag j is $\rho^j\cos\,\varpi j$. For $\varpi \in (0, \pi)$ the

roots of the AR polynomial are a pair of complex conjugates with modulus ρ^{-1} and phase ϖ; correspondingly, the spectral density displays a peak around ϖ.

Harvey and Trimbur (2003) further extend the model specification by proposing a general class of model-based filters for extracting trend and cycles in macroeconomic time series, showing that the design of lowpass and bandpass filters can be considered as a signal extraction problem in an unobserved components framework. In particular, they consider the decomposition $y_t = \mu_{mt} + \psi_{kt} + \epsilon_t$, where $\epsilon_t \sim \text{NID}(0, \sigma_\epsilon^2)$. The trend is specified as an mth-order stochastic trend:

$$\begin{array}{rcl} \mu_{1t} &=& \mu_{1,t-1} + \zeta_t \\ \mu_{it} &=& \mu_{i,t-1} + \mu_{i-1,t}, \quad i = 2, \ldots, m. \end{array} \tag{9.9}$$

This is the recursive definition of an $m-1$-fold integrated random walk, such that $\Delta^m \mu_{mt} = \zeta_t$. The component ψ_{kt} is a kth-order stochastic cycle, defined as:

$$\begin{bmatrix} \psi_{1t} \\ \psi_{1t}^* \end{bmatrix} = \rho \begin{bmatrix} \cos\varpi & \sin\varpi \\ -\sin\varpi & \cos\varpi \end{bmatrix} \begin{bmatrix} \psi_{1,t-1} \\ \psi_{1,t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \psi_{it} \\ \psi_{it}^* \end{bmatrix} = \rho \begin{bmatrix} \cos\varpi & \sin\varpi \\ -\sin\varpi & \cos\varpi \end{bmatrix} \begin{bmatrix} \psi_{i,t-1} \\ \psi_{i,t-1}^* \end{bmatrix} + \begin{bmatrix} \psi_{i-1,t} \\ 0 \end{bmatrix}. \tag{9.10}$$

The reduced form representation for the cycle is:

$$(1 - 2\rho\cos\varpi L + \rho^2 L^2)^k \psi_{kt} = (1 - \rho\cos\varpi L)^k \kappa_t.$$

Harvey and Trimbur show that, as m and k increase, the optimal estimators of the trend and the cycle approach the ideal lowpass and bandpass filter, respectively.

9.2.5 Models with correlated components

Morley, Nelson and Zivot (2003; henceforth, MNZ) consider the following unobserved components model for US quarterly GDP:

$$\begin{array}{rcll} y_t &=& \mu_t + \psi_t & t = 1, 2, \ldots, n, \\ \mu_t &=& \mu_{t-1} + \beta + \eta_t, & \\ \psi_t &=& \phi_1\psi_{t-1} + \phi_2\psi_{t-2} + \kappa_t, & \\ \begin{pmatrix} \eta_t \\ \kappa_t \end{pmatrix} &\sim& \text{NID}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\eta^2 & \sigma_{\eta\kappa} \\ \sigma_{\eta\kappa} & \sigma_\kappa^2 \end{pmatrix}\right], & \sigma_{\eta\kappa} = r\sigma_\eta\sigma_\kappa. \end{array} \tag{9.11}$$

It should be noticed that the trend and cycle disturbances are allowed to be contemporaneously correlated, with r being the correlation coefficient. The reduced form of model (9.11) is the ARIMA(2,1,2) model: $\Delta y_t = \beta + \frac{\theta(L)}{\phi(L)}\xi_t, \quad \xi_t \sim \text{NID}(0, \sigma^2)$, where $\theta(L) = 1 + \theta_1 L + \theta_2 L^2$ and $\phi(L) = 1 - \phi_1 L - \phi_2 L^2$. The structural form is exactly identified, both it and the reduced form have six parameters. The orthogonal trend-cycle decomposition considered by Clark (1987) imposes the overidentifying restriction $r = 0$.

We estimate this model for the US GDP series using the sample period 1947:1–2006:4. For comparison we also fit an unrestricted ARIMA(2,1,2) model and the restricted version imposing $r = 0$, which will be referred to henceforth as the Clark model. Estimation of the unknown parameters is carried out by frequency domain ML estimation (see Nerlove, Grether and Carvalho, 1995; Harvey, 1989, sec. 4.3, for the derivation of the likelihood function and a discussion on the nature of the approximation involved). Given the availability of the differenced observations $\Delta y_t, t = 1, 2, \ldots, n$, and denoting by $\omega_j = 2\pi j/n$, $j = 0, 1, \ldots, (n-1)$, the Fourier frequencies, the Whittle likelihood is defined as follows:

$$\text{loglik} = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{j=0}^{n-1}\left[\log f(\omega_j) + \frac{I(\omega_j)}{f(\omega_j)}\right], \tag{9.12}$$

where $I(\omega_j)$ is the sample spectrum:

$$I(\omega_j) = \frac{1}{2\pi}\left[c_0 + 2\sum_{k=1}^{n-1} c_k \cos(\omega_j k)\right],$$

c_k is the sample autocovariance of Δy_t at lag k, and $f(\omega_j)$ is the parametric spectrum of the implied stationary representation of the MNZ model, $\Delta y_t = \beta + \eta_t + \Delta\psi_t, t = 1, \ldots, n$, evaluated at the Fourier frequency ω_j. In particular:

$$f(\omega) = f_{\Delta\mu}(\omega) + f_{\Delta\psi}(\omega) + f_{\Delta\mu,\Delta\psi}(\omega),$$

with:

$$f_{\Delta\mu}(\omega) = \frac{\sigma_\eta^2}{2\pi}, \quad f_{\Delta\psi}(\omega) = \frac{1}{2\pi}\frac{2(1-\cos\omega)\sigma_\kappa^2}{\phi(e^{-\imath\omega})\phi(e^{\imath\omega})},$$

$$f_{\Delta\mu,\Delta\psi}(\omega) = \frac{(1-e^{-\imath\omega})\phi(e^{\imath\omega}) + (1-e^{\imath\omega})\phi(e^{-\imath\omega})}{2\pi\phi(e^{-\imath\omega})\phi(e^{\imath\omega})} r\sigma_\eta\sigma_\kappa,$$

$e^{-\imath\omega} = \cos\omega - \imath\sin\omega$, where $\imath$ is the imaginary unit, is the complex exponential, and $\phi(e^{-\imath\omega}) = 1 - \phi_1 e^{-\imath\omega} - \phi_2 e^{-2\imath\omega}$. The last term is the cross-spectrum of $(\Delta\psi_t, \Delta\mu_t)$ and, of course, it vanishes if $r = 0$. For the Clark model the parametric spectrum is given by the above expression with $f_{\Delta\mu,\Delta\psi}(\omega) = 0$, whereas for the unrestricted ARIMA(2,1,2) it is given by $f(\omega) = \sigma^2\theta(e^{-\imath\omega})\theta(e^{\imath\omega})[\phi(e^{-\imath\omega})\phi(e^{\imath\omega})]^{-1}$.

Figure 9.1 displays the quarterly growth rates, Δy_t, of US GDP in the first panel. The next panel plots the profile likelihood for the correlation parameter against the value of r in $[-1,1]$ and shows the presence of two modes, the first around $-.9$ and the second around zero. The parameter estimates, along with their estimated standard errors, and the value of the maximized likelihood, are reported in Table 9.1.[2] It should be noticed that the unrestricted ARIMA(2,1,2) is exactly coincident with the reduced form of the MNZ model, as the two models yield the same likelihood and the AR and MA parameters are the mapping of the structural parameters. Second,

Table 9.1 Frequency domain ML estimation results for quarterly US real GDP, 1947:1–2006:4

	ARIMA	*MNZ*	*Clark*
ϕ_1	1.34 (0.07)	1.34 (0.07)	1.49 (0.05)
ϕ_2	−0.76 (0.16)	−0.76 (0.16)	−0.56 (0.11)
θ_1	−1.08 (0.11)		
θ_2	0.59 (0.20)		
σ^2	0.8224 (0.08)		
r		−0.93 (0.28)	0(r)
σ^2_η		1.2626 (0.08)	0.3478 (0.15)
σ^2_κ		0.3556 (0.33)	0.4120 (0.16)
loglik	−315.76	−315.76	−317.14

the estimated correlation coefficient is high and negative (−0.93) and the likelihood ratio test of the hypothesis $r = 0$ has a p-value equal to 0.097. MNZ interpret the negative disturbance correlation as strengthening the case for the importance of real shocks in the macroeconomy: real shocks tend to shift the long run path of output, so short-term fluctuations will largely reflect adjustments toward a shifting trend if real shocks play a dominant role.

The bottom left panel of Figure 9.1 displays the sample spectrum $I(\omega_j)$ of Δy_t along with the estimated parametric spectral densities for the MNZ model (which is, of course, coincident with that of the ARIMA(2,1,2) model) and the Clark restricted model ($r = 0$). For the ARIMA(2,1,2) and the MNZ models the roots of the AR polynomial are a pair of complex conjugates that imply a spectral peak for Δy_t at the frequency 0.68, corresponding to a period of nine quarters. As a matter of fact, a dominant feature of Δy_t is the presence of a cyclical component with a period of roughly two years. On the other hand, the spectral density implied by the Clark model peaks at the frequency 0.09, corresponding to a period of 68 quarters (i.e., a medium-run cycle).

A closer inspection of the sample spectrum reveals the presence of two consecutive periodogram ordinates, corresponding to a cycle of roughly two years, that are highly influential on the estimation results (they are circled in Figure 9.1). It is indeed remarkable that when these are not used in the estimation, the correlation coefficient turns positive ($\hat{r} = 0.35$). The last panel of the figure presents the leave-two-out cross-validation estimates of the correlation coefficient, which are obtained by maximizing Whittle's likelihood after deleting two consecutive periodogram ordinates at the frequencies ω_j and ω_{j+1}. This is a special case of weighted likelihood estimation, where each summand in (9.12) receives a weight equal to 1 if the frequency ω_j is retained and 0 if it is deleted.

The real-time and the smoothed estimates of the cyclical component arising from the MNZ model, $\tilde{\psi}_{t|t} = \mathrm{E}(\psi_t|Y_t)$ and $\tilde{\psi}_{t|n} = \mathrm{E}(\psi_t|Y_n)$, respectively, are reported in Figure 9.2, along with the 95% interval estimates; here Y_t denotes the information available up to and including time t. The bottom panels display the weights $\mathrm{w}_{\psi,j}$

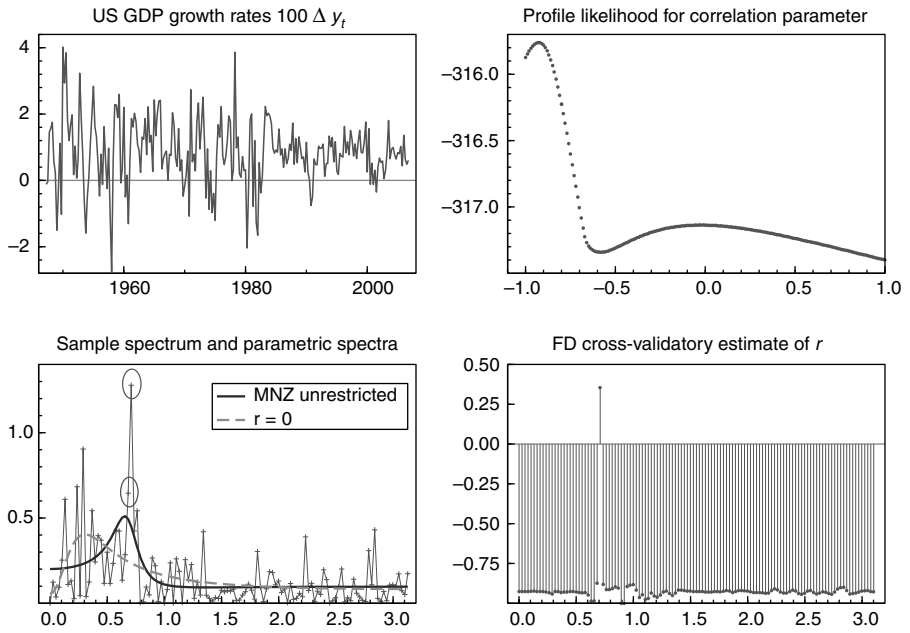


Figure 9.1 Quarterly US real growth, 1947:2–2006:4. Sample spectrum and parametric spectral fit of trend-cycle model with correlated components

of the signal extraction filters $\sum_j w_{\psi,j} L^j y_t$ that yield the cycle estimates in the two cases.

The real-time estimates support the view that most of the variation in GDP is permanent, i.e., it should be ascribed to changes in the trend component, whereas little variance is attributed to the transitory component. In fact, the amplitude of $\tilde{\psi}_{t|t}$ is small and the interval estimates of ψ_t in real time are never significantly different from zero. When we analyze the smoothed estimates the picture changes quite radically: the cycle estimates are much more variable and there is a dramatic reduction in the estimation error variance, so that the contribution of the transitory component to the variation in GDP is no longer negligible. The real-time estimates provide a gross underestimation of the cyclical component and are heavily revised as the future missing observations become available. As a matter of fact, the final estimates depend heavily on future observations, as can be seen from the pattern of the weights in the last panel of Figure 9.2. That this behavior is typical of the MNZ model when $\hat{r}$ is high and negative is documented in Proietti (2006a).

The real-time estimates of the trend and cyclical components are coincident with the Beveridge and Nelson (1981; henceforth, BN) components defined for the ARIMA(2,1,2) reduced form. The BN decomposition defines the trend component at time t as the value of the eventual forecast function at that time, or, equivalently, the value that the series would take if it were on its long-run path (see also Brewer, 1979). For an ARIMA$(p, 1, q)$ process, this argument defines the

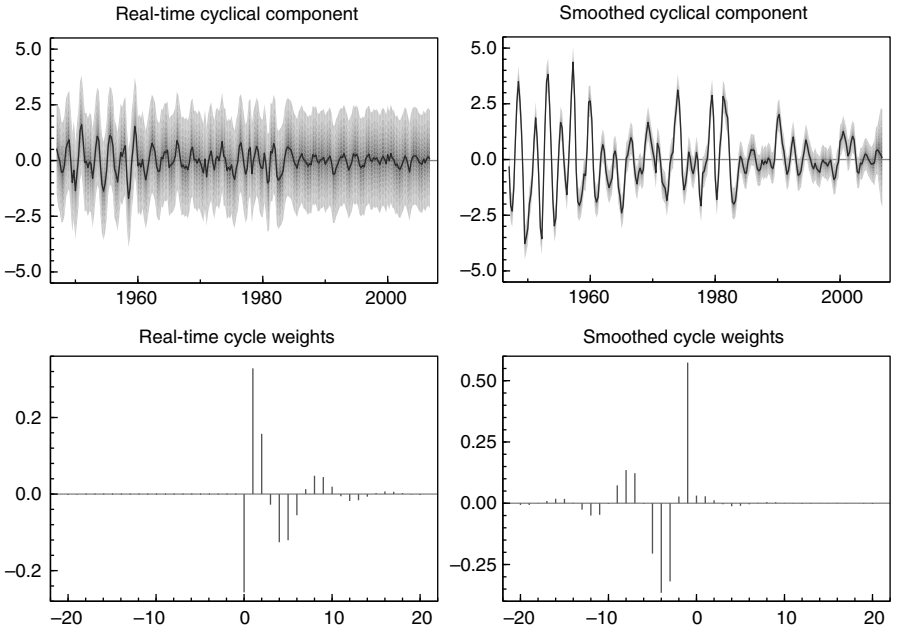


Figure 9.2 Trend-cycle decomposition with correlated disturbances. Real-time and smoothed estimates of the cyclical components

trend as a random walk driven by the innovations $\xi_t = y_t - \mathrm{E}(y_t|Y_{t-1})$. Writing the ARIMA representation for y_t as $\Delta y_t = \beta + \psi(L)\xi_t$, $\psi(L) = \theta(L)/\phi(L)$, where $\phi(L)$ is a stationary AR polynomial of order p and $\theta(L)$ an invertible MA polynomial of order q, the BN decomposition can be written as: $y_t = m_t + c_t, t = 1, \ldots, n$, where m_t is the BN trend, and c_t is the cyclical component.

The trend is defined as $\lim_{l\to\infty}[\tilde{y}_{t+l|t} - l\beta]$, with $\tilde{y}_{t+l|t} = \mathrm{E}(y_{t+l}|Y_t)$. Writing $y_{t+l} = y_{t+l-1} + \beta + \psi(L)\xi_t$, taking the conditional expectation and rearranging, it is easily shown to give $m_t = m_{t-1} + \beta + \psi(1)\xi_t$, where $\psi(1) = \theta(1)/\phi(1)$ is the persistence parameter, as it measures the fraction of the innovation at time t that is retained in the trend. In terms of the observations, $m_t = \mathrm{w}_m(L)y_t$, where $\mathrm{w}_m(L)$ is the one-sided filter:

$$w_m(L) = \frac{\psi(1)}{\psi(L)} = \frac{\theta(1)}{\phi(1)}\frac{\phi(L)}{\theta(L)}.$$

The sum of the weights is one, i.e., $w_m(1) = 1$.

The *transitory component* is defined residually as $c_t = y_t - m_t = \psi^*(L)\xi_t$, where $\Delta\psi^*(L) = \psi(L) - \psi(1)$. Alternative representations in terms of the observations y_t and of the innovations ξ_t are, respectively:

$$c_t = \frac{\phi(1)\theta(L) - \theta(1)\phi(L)}{\phi(1)\theta(L)}y_t, \qquad c_t = \frac{\phi(1)\theta(L) - \theta(1)\phi(L)}{\phi(1)\phi(L)\Delta}\xi_t. \tag{9.13}$$

The first expression shows that the weights for the extraction of the cycle sum to zero. Since $\phi(1)\theta(L) - \theta(1)\phi(L)$ must have a unit root, we can write $\phi(1)\theta(L) - \theta(1)\phi(L) = \Delta\vartheta(L)$, and substituting this into (9.13), the ARMA representation for this component can be established as $\phi(L)c_t = \vartheta(L)[\phi(1)]^{-1}\xi_t$. As the order of $\vartheta(L)$ is $\max(p,q) - 1$, the cyclical component has a stationary ARMA$(p, \max(p,q) - 1)$ representation. For the ARIMA(2,1,2) model fitted to US GDP, the cycle has the ARMA(2,1) representation:

$$\phi(L)c_t = (1 + \vartheta L)\left[1 - \frac{\theta(1)}{\phi(1)}\right]\xi_t, \quad \vartheta = -\frac{\phi_2\theta(1) + \theta_2\phi(1)}{\phi(1) - \theta(1)}. \tag{9.14}$$

It is apparent that the two components are driven by the innovations, ξ_t; the fraction $\theta(1)/\phi(1)$, known as *persistence*, is integrated in the trend, and its complement to 1 drives the cycle. The sign of the correlation between the trend and the cycle disturbances is provided by the sign of $\phi(1) - \theta(1)$; when persistence is less (greater) than one then the trend and cycle disturbances are positively (negatively) and perfectly correlated.

9.2.6 Model-based bandpass filters

As we said before, macroeconomic time series such as GDP do not usually admit the decomposition $y_t = \mu_t + \psi_t$, with ψ_t being a purely irregular component; nevertheless, applications of the class of filters (9.6) is widespread, as the popularity of the Hodrick–Prescott filter testifies. However, when the available series y_t cannot be modeled as (9.2), it is not immediately clear how the components should be defined and how inferences about them should be made. In particular, the Kalman filter and the associated smoothing algorithms no longer provide the minimum mean square estimators of the components nor their mean square error. The discussion of model-based bandpass filtering in a more general setting will be the theme of this section.

The trend-cycle decompositions dealt with in the two previous sections are models of economic fluctuations, such that the components are driven by random disturbances which are propagated according to a transmission mechanism. In this section we start from a reduced form model (as in the case of the BN decomposition) and define parametric trend-cycle decompositions that are less loaded with structural interpretation, since they just represent the lowpass and highpass components in the series. The aim is to motivate and extend the use of signal extraction filters of the class (9.6) to a more general and realistic setting than (9.5). For this approach to the definition of bandpass filters see Gómez (2001) and Kaiser and Maravall (2005). The following treatment is based on Proietti (2004).

Let y_t denote a univariate time series with ARIMA(p,d,q) representation, which we write as:

$$\phi(L)(\Delta^d y_t - \beta) = \theta(L)\xi_t, \quad \xi_t \sim \text{NID}(0, \sigma^2),$$

where β is a constant, $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ is the AR polynomial with stationary roots, and $\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ is invertible. We are going to exploit the fundamental idea that we can uniquely decompose the WN disturbance ξ_t into

two orthogonal stationary processes as follows:

$$\xi_t = \frac{(1+L)^r \zeta_t + (1-L)^m \kappa_t}{\varphi(L)}, \tag{9.15}$$

where ζ_t and κ_t are two mutually and serially independent Gaussian disturbances, $\zeta_t \sim \text{NID}(0, \sigma^2)$, $\kappa_t \sim \text{NID}(0, \lambda\sigma^2)$, and:

$$|\varphi(L)|^2 = \varphi(L)\varphi(L^{-1}) = |1+L|^{2r} + \lambda|1-L|^{2m}. \tag{9.16}$$

We assume that λ is known. Equation (9.16) is the spectral factorization of the lag polynomial on the right-hand side of (9.15); the existence of the polynomial $\varphi(L) = \varphi_0 + \varphi_1 L + \cdots + \varphi_{q^*} L^{q^*}$, of degree $q^* = \max(m, r)$, is guaranteed by the fact that the Fourier transform of the right-hand side is never zero over the entire frequency range (see Sayed and Kailath, 2001, for details).

According to (9.15), for given values of λ, m and r, the innovation ξ_t is decomposed into two ARMA(2,2) processes, characterized by the same AR polynomial, but by different MA components. The first component will drive the lowpass component of y_t and its spectral density is proportional to $\sigma^2 \text{w}_\mu(\omega)$, where $\text{w}_\mu(\omega)$ is the gain of the filter (9.6). If $r > 0$ the MA representation is noninvertible at the π frequency. Notice that, as m and r increase, the transition from the pass band to the stop band is sharper.

Substituting (9.15)–(9.16) into the ARIMA representation, the series can be decomposed into two orthogonal components:

$$\begin{aligned} y_t &= \mu_t + \psi_t, \\ \phi(L)\varphi(L)(\Delta^d \mu_t - \beta) &= (1+L)^r \theta(L)\zeta_t, \quad \zeta_t \sim \text{NID}(0, \sigma^2) \\ \phi(L)\varphi(L)\psi_t &= \Delta^{m-d}\theta(L)\kappa_t, \quad \kappa_t \sim \text{NID}(0, \lambda\sigma^2). \end{aligned} \tag{9.17}$$

The trend or lowpass component has the same order of integration as the series (regardless of m), whereas the cycle or highpass component is stationary provided that $m \geq d$, which will be assumed throughout.

Given the availability of a doubly infinite sample, the Wiener–Kolmogorov estimators of the components are $\tilde{\mu}_t = \text{w}_\mu(L) y_t$ and $\tilde{\psi}_t = [1 - \text{w}_\mu(L)] y_t$, where the impulse response function of the optimal filters is given by (9.6). Hence, the signal extraction filter for the central data points will continue to be represented by (9.6), regardless of the properties of y_t, but this is the only feature that is invariant to the nature of the time series and its ARIMA representation. The mean square error of the smoothed components, as a matter of fact, depends on the ARIMA model for y_t. In finite samples, the estimators and their mean square errors will be provided by the Kalman filter and smoother associated with the model (9.17), and thus will depend on the ARIMA model for y_t.

Bandpass filters can also be constructed from the principle of decomposing the lowpass component in (9.17). Let us consider fixed values of m and r and two cut-off frequencies, ω_{c1} and $\omega_{c2} > \omega_{c1}$, with corresponding values of the smoothness

parameter λ_1 and λ_2, determined according to (9.7). Obviously $\lambda_1 > \lambda_2$. The trend-cycle decomposition corresponding to the triple m, r, λ_2 (or, equivalently, m, r, ω_{c2}), is as in (9.17):

$$\begin{aligned} y_t &= \mu_{2t} + \epsilon_t, \\ \Delta^d \mu_{2t} &= \beta + \frac{(1+L)^r}{\varphi_2(L)} \frac{\theta(L)}{\phi(L)} \zeta_{2t}, \quad \zeta_{2t} \sim \text{NID}(0, \sigma^2) \\ \epsilon_t &= \frac{(1-L)^m}{\varphi_2(L)} \frac{\theta(L)}{\Delta^d \phi(L)} \kappa_{2t}, \quad \kappa_{2t} \sim \text{NID}(0, \lambda_2 \sigma^2), \end{aligned} \tag{9.18}$$

with $|\varphi_2(L)|^2 = |1+L|^{2r} + \lambda_2 |1-L|^{2m}$.

We can similarly define the trend-cycle decomposition corresponding to the triple m, r, λ_1 (or, equivalently, m, r, ω_{c1}), $y_t = \mu_{1t} + \psi_t$. As $\lambda_1 > \lambda_2$ this decomposition features a lower cut-off frequency, ω_{c1}, thereby yielding a smoother trend. The components μ_{1t} and ψ_t are defined as in (9.18), with $\varphi_1(L)$, $\zeta_{1t} \sim \text{NID}(0, \sigma^2)$ and $\kappa_{1t} \sim \text{NID}(0, \lambda_1 \sigma^2)$ replacing, respectively, $\varphi_2(L)$, ζ_{2t} and κ_{2t}. The polynomial $\varphi_1(L)$ is such that $|\varphi_1(L)|^2 = |1+L|^{2r} + \lambda_1 |1-L|^{2m}$.

The lowpass component, μ_{2t}, can, in turn, be decomposed using the orthogonal decomposition of the disturbance ζ_{2t}:

$$\zeta_{2t} = \frac{\varphi_2(L)}{\varphi_1(L)} \zeta_{1t} + \frac{(1-L)^m}{\varphi_1(L)} \kappa_{1t}, \tag{9.19}$$

with:

$$\zeta_{1t} \sim \text{NID}(0, \sigma^2), \quad \kappa_{1t} \sim \text{NID}\left(0, (\lambda_1 - \lambda_2)\sigma^2\right), \quad \text{E}(\zeta_{1j} \kappa_{1t}) = 0, \quad \forall j, t.$$

Under this setting, the spectrum of both sides of (9.19) is constant and equal to $\sigma^2/2\pi$.

Substituting (9.19) into (9.18), and writing $\mu_{2t} = \mu_{1t} + \psi_t$, enables y_t to be decomposed into three components, representing the lowpass (μ_{1t}), bandpass (ψ_t) and highpass (ϵ_t) components, respectively.

$$\begin{aligned} y_t &= \mu_{1t} + \psi_t + \epsilon_t, \\ \Delta^d \mu_{1t} &= c + \tfrac{(1+L)^r}{\varphi_1(L)} \tfrac{\theta(L)}{\phi(L)} \zeta_{1t}, \qquad \zeta_{1t} \sim \text{NID}(0, \sigma^2) \\ \psi_t &= \tfrac{(1+L)^n (1-L)^m}{\varphi_1(L)\varphi_2(L)} \tfrac{\theta(L)}{\Delta^d \phi(L)} \kappa_{1t}, \quad \kappa_{1t} \sim \text{NID}\left(0, (\lambda_1 - \lambda_2)\sigma^2\right), \end{aligned} \tag{9.20}$$

and ϵ_t, given in (9.18), is the highpass component of the decomposition (9.20). The model can be cast in state-space form and the Kalman filter and smoother (see Appendix C) will provide the optimal estimates of the components and their standard errors.

Figure 9.3 shows the gain of an ideal bandpass filter and the BK filter. The dashed line is the gain of the model-based bandpass filter which is optimal for ψ_t in (9.20)

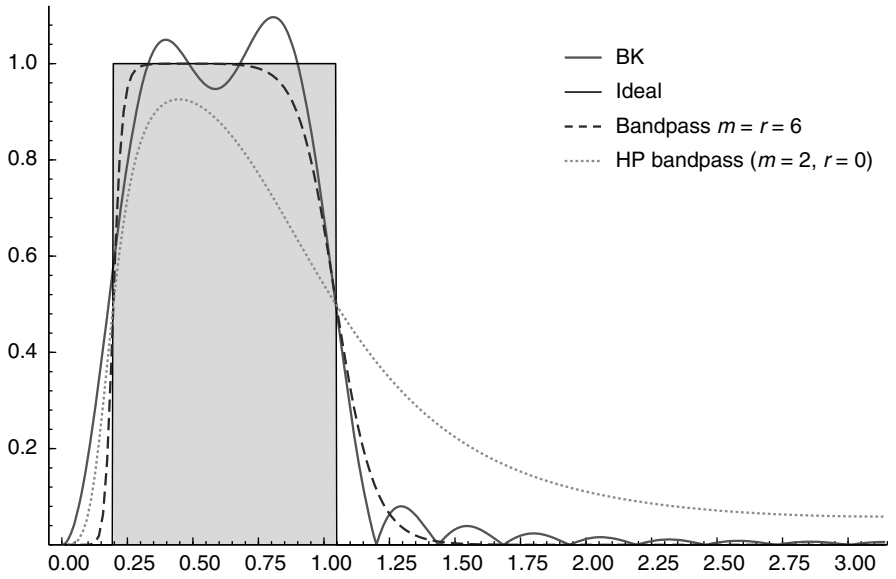


Figure 9.3 Gain function of the ideal business cycle bandpass filter, the BK filter and two model-based filters

using $m = r = 6$ and the two cut-off frequencies $\omega_{c1} = 2\pi/32$ (corresponding to a period of eight years for quarterly data) and $\omega_{c2} = 2\pi/6$ (one and a half years); such large values of the parameters yield a gain which is close to the ideal boxcar function. The HP bandpass curve is the gain of the Wiener–Kolmogorov filter for extracting the component ψ_t in (9.20) with $m = 2, r = 0$, and ω_{c1}, ω_{c2} given above. In this case the leakage is larger but, as shown in Proietti (2004), taking large values of m and r is detrimental to the reliability of the end of sample estimates.

9.2.7 Applications of model-based filtering: bandpass cycles and the estimation of recession probabilities

We present two applications of the model-based filtering approach outlined in the previous section. Our first illustration deals with the estimation and the assessment of the reliability of the deviation cycle in US GDP. The cycle is defined as the highpass component extracting the fluctuations in the level of log GDP that have a periodicity smaller than ten years (40 quarters). To evaluate model uncertainty, we fit three models to the logarithm of GDP, namely a simple random walk, or ARIMA(0,1,0) model ($\hat{\sigma}^2 = 0.8570$), an ARIMA(1,1,0) model (the estimated first-order autoregressive coefficient is $\hat{\phi} = 0.33$ and $\hat{\sigma}^2 = 0.8652$), and finally, we consider the ARIMA(2,1,2) model fitted in section 9.2.5, whose parameter estimates were reported in Table 9.1.

The estimates of the lowpass component corresponding to the three models obtained by setting $m = 2, r = 0$ (and thus $\lambda = 1600$ and $\omega_c = 0.158279$) are displayed in the top right-hand panel of Figure 9.4, along with the Leser–HP cycle.

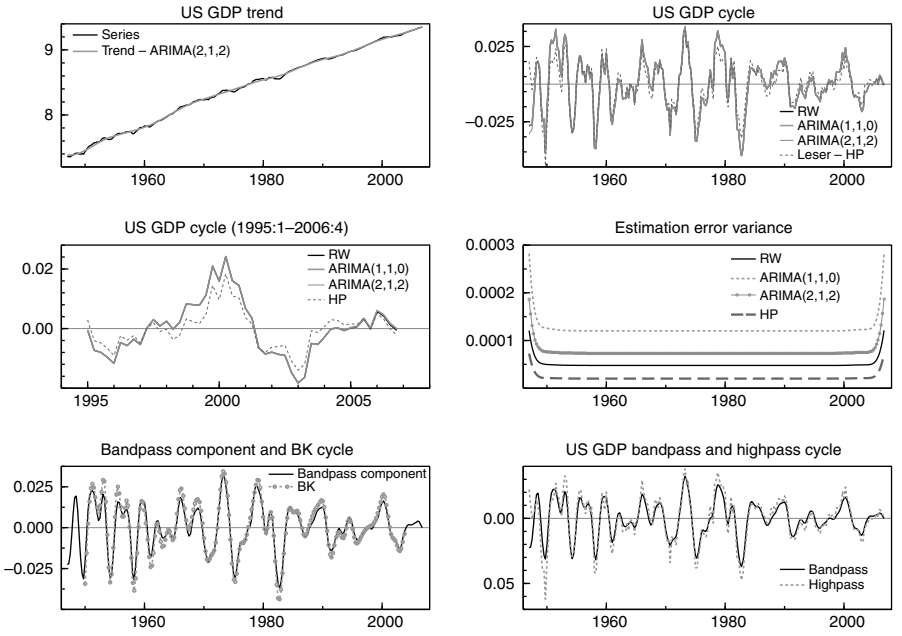


Figure 9.4 Model-based filtering. Estimates of the lowpass component (using the ARIMA(2,1,2) model) and of the highpass and bandpass components in US GDP, and their comparison with the Leser–HP and BK cycles

The estimates for the three models are obtained as the conditional mean of ψ_t given the observations by applying the Kalman filter and smoother to the representation (9.17); the algorithm also provides their estimation error variance. It must be stressed that the Leser–HP filter is optimal for a restricted IMA(2,2) process and thus it does not yield the minimum mean square estimator of the cycle, nor its standard error. In general, also looking at the middle panel, which displays the estimated cycles for the last 12 years, the model-based estimates are almost indistinguishable, and are quite close to the Leser–HP cycle estimates in the middle of the sample. Large differences with the latter arise at the beginning, where the lowpass component had greater amplitude, and at the end of the sample period.

The particular model that is chosen matters little as far as the point estimates of ψ_t are concerned. Nevertheless, it is relevant for the assessment of the accuracy of the estimates, as can be argued from the right middle panel of the figure, which shows the estimation error variance, $\mathrm{Var}(\psi_t|Y_n)$, for the three models of US GDP. It is also evident that the standard errors obtained for the Leser–HP filter would underestimate the uncertainty of the estimates.

We conclude this first illustration by estimating the deviation cycle as a bandpass component, assuming that the true model is the ARIMA(2,1,2) and using the cut-off frequencies $\omega_{c1} = 2\pi/32$, $\omega_{c2} = 2\pi/6$, and the values $m = 2, r = 0$; as a consequence, the component ψ_t in (9.20) selects all the fluctuations in a range

of periodicity that goes from one and a half years (6 quarters) to eight years (32 quarters). The gain of the filter is displayed in Figure 9.3. The estimates of ψ_t are compared to the BK cycle in the bottom left panel of Figure 9.4 and to the corresponding highpass estimates ($\psi_t + \epsilon_t$). With respect to the BK cycle, the estimates are available also in real time.

The conclusion is that model-based filtering improves the quality of the estimated lowpass component, providing estimates at the boundary of the sample period that are automatically adapted to the series under investigation, and enables the investigator to assess the reliability of the estimates (conditional on a particular reduced form).

The second application deals with assessing the uncertainty in estimating the business cycle chronology. According to the classical definition, the business cycle is a recurrent sequence of expansions and contractions in the aggregate level of economic activity (see Burns and Mitchell, 1946, p. 3). Dating the business cycle amounts to establishing a set of reference dates that mark the phases or states of the economy. Usually two phases, recessions and expansions, are considered, that are delimited by peaks and troughs in economic activity. Dating is carried out by an algorithm, such as that due to Bry and Boschan (1971), or that proposed by Artis, Marcellino and Proietti (2004), which aims at estimating the location of turning points, enforcing the alternation of peaks and troughs and minimum duration ties for the phases and the full cycle. Downturns and upturns have to be persistent to qualify as cycle phases; thus, they need to fulfill minimum duration constraints, such as at least two quarters for each phase; moreover, to separate it from seasonality, a complete sequence, recession-expansion or expansion-recession, i.e., a full cycle, has to last longer than one year. Depth restrictions, motivated by the fact that only major fluctuations qualify for the phases, should also be enforced.

An integral part of the dating algorithm is prefiltering, which is necessary in order to isolate fluctuations in the series with period greater than the minimum cycle duration. For instance, in the quarterly case we need to abstract from all the fluctuations with periodicity less than five quarters, i.e., from high-frequency fluctuations that do not satisfy the minimum cycle duration. In lieu of the *ad hoc* and old-fashioned moving averages adopted by Bry and Boschan, one can use model-based lowpass signal extraction filters.

The advantages are twofold: first, it is possible to select the cut-off frequency so as to match the minimum cycle duration; e.g., in our case $\omega_c = 2\pi/5$. Second, the uncertainty in dating arising from prefiltering can be assessed by Monte Carlo simulation, by means of an algorithm known as the simulation smoother (see de Jong and Shephard, 1995; Durbin and Koopman, 2002; and Appendix C, section 9.7.4). This repeatedly draws simulated samples from the posterior distribution of the lowpass component with a cut-off frequency corresponding to five quarters, $\tilde{\mu}_t^{(i)} \sim \mu_t | Y_n$, so that by repeating the draws a sufficient number of times we can get Monte Carlo estimates of different aspects of the marginal and joint distribution of the lowpass component, intended here as the level of output devoid of all fluctuations with a periodicity smaller that five quarters.

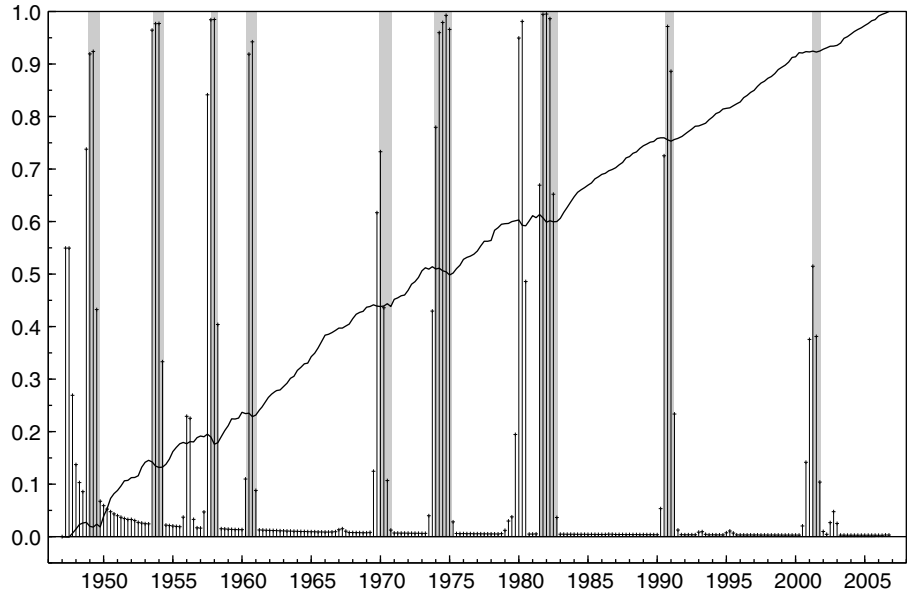


Figure 9.5 Relative number of times each quarter is classified as a recessionary period, using 5,000 simulated samples. The shaded areas represent NBER-classified recessions

Figure 9.5 plots the recession frequencies, i.e., the relative number of times each quarter was classified as a recessionary period. For this purpose 5,000 draws from the conditional distribution of $\boldsymbol{\mu}|\mathbf{y}$ were extracted; each quarter was classified as recession or expansion according to the Artis, Marcellino and Proietti (2004) Markov chain dating algorithm. There is a close agreement with the NBER chronology, which is not based on GDP alone, and the last recession, which started in March 2001 and ended in October 2001, was really mild in terms of GDP; in fact, the recession frequency is only in one quarter greater than 0.5.

9.2.8 *Ad hoc* filtering and the Slutsky–Yule effect

A filter is *ad hoc* when it is invariant to the properties of the time series under investigation. An instance is provided by the Leser–HP filter with a fixed smoothing parameter, and another example is the BK filter. The potential danger associated with an *ad hoc* cycle extraction filter is that the filtered series displays cyclical features that are absent from the original series. The risk of extracting spurious cycles is known in the time series literature as the Slutsky–Yule effect.

The distortionary effects of the Leser–HP filter have been discussed by King and Rebelo (1993), Harvey and Jaeger (1993), and Cogley and Nason (1995). These authors document that, when the series to which the filter is applied is difference stationary (e.g., a random walk, or an integrated random walk), the detrended series can display spurious cyclical behavior. As a matter of fact, the transfer function will display a distinctive peak at business cycle frequencies, which is only due to the

leakage from the non-stationary component. Moreover, the filter seriously distorts the evidence for the comovements among detrended series.

The issue of spuriousness is problematic, at least, if not tautological. The main difficulty stems from the fact that it ties in with a more fundamental question concerning what is indeed the cycle in economic time series. If we adhere to the bandpass paradigm of viewing the cycle as consisting of those fluctuations within a give range of periodicity, than the case for spuriousness is much less compelling.

Another source of concern among practitioners, especially for the conduct of monetary policy, relates to the end-of-sample behavior of the Leser–HP filter: the real-time estimates would be subject to "end-of-sample bias," since they result from the application of a one-sided filter and will suffer from both phase shifts and amplitude distortions. One has to separate two issues: as we hinted before, the IMA(2,2), for which the Leser–HP filter is optimal, is usually misspecified for macroeconomic time series. As a result, the cycle estimates have no optimality properties. Model-based bandpass filtering is aimed at overcoming this limitation. Having said that, it is a fact of life that, for a correctly specified model, the optimal real time signal extraction filter will be one-sided and thus will produce phase shifts and amplitude distortions.

9.3 Multivariate models

Information on the output gap is contained in macroeconomic variables other than aggregate output, either because those variables provide alternative measures of production, or because they are functionally related to the output gap. In this section we start from the consideration of a bivariate model that, along with an output decomposition, includes an inflation equation. We then extend the model to include other variables, such as the unemployment rate and industrial production, and consider the estimation of a monthly model using quarterly observations on real GDP.

9.3.1 Bivariate models of real output and inflation

Price inflation carries relevant information for the output gap. The definition of the latter as an indicator of inflationary pressure and, correspondingly, of potential output as the level of output consistent with stable inflation, makes clear that a rigorous measurement can be made within a bivariate model of output and inflation, embodying a Phillips curve relationship. The Phillips curve establishes a relation between the nominal price, or wage, inflation rate, Δp_t, where, e.g., p_t is the logarithm of the consumer price index (CPI), and an indicator of excess demand, typically the output gap (ψ_t).

A general specification is the following:

$$\delta(L)\Delta p_t = c + \theta_\psi(L)\psi_t + \boldsymbol{\gamma}(L)'\mathbf{x}_t + \xi_{pt}, \tag{9.21}$$

where c is a constant, $\mathbf{x}_t$ denotes a set of exogenous supply shocks, such as changes in energy prices and terms of trade, and ξ_{pt} is WN. Often the restriction is imposed

that the sum of the AR coefficients on lagged inflation is unity, $\delta(L) = \Delta\delta^*(L)$, where $\delta^*(L)$ is a stationary AR polynomial; the gap enters the equation with more than one lag to capture the change in demand, since we can rewrite $\theta_\psi(L) = \theta_\psi(1) + \Delta\theta^*_\psi(L)$. This is known as Gordon's "triangle" model of inflation (see Gordon, 1997), since it features the three main driving forces: inertia (or inflation persistence, via $\delta(L)$), endogenous demand shocks (via ψ_t), and exogenous supply shocks (via $\mathbf{x}_t$). If $\delta(L)$ has a unit root and $\theta_\psi(1) \neq 0$ the output gap has permanent effects on the inflation rate. If, instead, $\theta_\psi(1) = 0$, then the output gap is neutral in the long run and the inflation rate shares a common cycle in the levels with output. Harvey, Trimbur and Van Dijk (2007) consider the Bayesian estimation of a bivariate model of output and inflation, where the cycle in inflation is driven by the output gap plus an idiosyncratic cycle.

Kuttner (1994) estimated potential output and the output gap for the US using a bivariate model of real GDP and CPI inflation. The output equation was specified as in the Clark (1987) model, i.e., $y_t = \mu_t + \psi_t$, such that potential output is a random walk with drift and the output gap is an AR(2) process driven by orthogonal disturbances. The equation for the inflation rate is a variant of Gordon's triangle model:

$$\Delta p_t = c + \gamma\Delta y_{t-1} + \theta_\psi\psi_{t-1} + \nu(L)\xi_{pt},$$

according to which the inflation rate is linearly related to the lagged output gap and to lagged GDP growth; inflation persistence is captured by the MA feature, $\nu(L)\xi_{pt}$, where the disturbance ξ_{pt} is allowed to be correlated with the output gap disturbance, κ_t. The inclusion of lagged real growth is not formally justified by Kuttner, and the correlation between ξ_{pt} and κ_t makes the dynamic relationship between the output gap and inflation more elaborate than it appears at first sight (e.g., inflation depends on the contemporaneous value of the gap). Moreover, permanent shocks are allowed to drive inflation via the term $\Delta y_{t-1} = \beta + \eta_{t-1} + \Delta\psi_{t-1}$, so that it cannot be maintained that μ_t is the noninflationary level of output. Planas, Rossi and Fiorentini (2007) consider the Bayesian estimation of Kuttner's bivariate model, with the only variant being that the MA feature is replaced by an autoregressive feature: $\delta(L)\Delta p_t = c + \gamma\Delta y_{t-1} + \theta_\psi\psi_{t-1} + \xi_{pt}$.

Gerlach and Smets (1999) again use a bivariate model of output and inflation, but the output gap generating equation takes the form of an aggregate demand equation featuring the lagged real interest rate as an explanatory variable. The inflation equation is specified as in (9.21) with $\delta(L) = \Delta$.

The Gordon triangle model may be interpreted as a reduced form of a structural model of inflation that embodies expectations; the presence of lagged inflation in the specification reflects backward looking inflation expectations. In the New Keynesian approach the Phillips curve is forward-looking, as inflation depends on expected future inflation. Doménech and Gómez (2006) estimate a multivariate model of output fluctuations including a forward-looking Phillips curve specified as follows:

$$\Delta p_t = c + \delta\mathrm{E}(\Delta p_{t+1}|\mathcal{F}_t) + \theta_\psi(L)\psi_t + \xi_{pt},$$

where $\mathcal{F}_t$ is the information set at time t. Basistha and Nelson (2007) estimate a bivariate model of output and inflation where the output equation features the MNZ decomposition with correlated components, and in the inflation equation, survey-based expectations replace $\mathrm{E}(\Delta p_{t+1}|\mathcal{F}_t)$.

9.3.2 A bivariate quarterly model of output and inflation for the US

This section is devoted to the estimation of a bivariate model for US quarterly real GDP and the quarterly rate of inflation Δp_t, where p_t is the logarithm of quarterly CPI for the US, using data from the first quarter of 1950 to the fourth quarter of 2006. The KPSS test conducted on the inflation series leads to the rejection of the null of stationarity against a random walk for all values of the lag truncation parameter up to 5; if a linear trend is considered and stationarity is tested against a random walk with drift, then the null is also rejected for much higher values of the lag truncation parameter. In the sequel, inflation will be taken to be integrated of order one. The model has the following specification:

$$\begin{array}{lcll} y_t & = & \mu_t + \psi_t, & t = 1, \ldots, n, \\ \mu_t & = & \mu_{t-1} + \beta_t + \eta_t, & \eta_t \sim \mathrm{NID}(0, \sigma^2_\eta) \\ \psi_t & = & \phi_1 \psi_{t-1} + \phi_2 \psi_{t-2} + \kappa_t, & \kappa_t \sim \mathrm{NID}(0, \sigma^2_\kappa) \\ & & & \\ \Delta p_t & = & \tau_t + \varepsilon_{pt} & \varepsilon_{pt} \sim \mathrm{NID}(0, \sigma^2_{p\varepsilon}) \\ \tau_t & = & \tau_{t-1} + \theta_\psi(L)\psi_t + \eta_{\tau t} & \eta_{\tau t} \sim \mathrm{NID}(0, \sigma^2_{\tau\eta}), \end{array} \tag{9.22}$$

where η_t, κ_t, ε_{pt}, and η_{pt} are mutually independent.

The output equation is the usual decomposition into orthogonal components; the inflation equation is a decomposition into a core component, τ_t, and a transitory one. The changes in the core component are driven by the output gap and by the idiosyncratic disturbances $\eta_{\tau t}$. The lag polynomial $\theta_\psi(L) = \theta_{\psi 0} + \theta_{\psi 1} L$ can be rewritten as $\theta_\psi(1) - \theta_{\psi 1}\Delta$, which enables us to isolate the level effect of the gap from the change effect, which we expect to be positive, that is we expect $\theta_{\psi 1} < 0$. If $\theta_\psi(1) = 0$, the inflation equation can be rewritten as $\Delta p_t = \tau^*_t - \theta_{\psi 1}\psi_t + \varepsilon_t$, with $\Delta\tau^*_t = \eta_{\tau t}$, so that output and inflation would share a common cycle.

We also extend the specification of model (9.22) to take into account an important stylized fact, known as the "Great Moderation" of the business cycle, and which consists of a substantive reduction in the volatility of GDP growth. This feature is visible from the plot of Δy_t in Figure 9.1. The date when the structural break in volatility occurred is identified as the first quarter of 1984 (see Kim and Nelson, 1999a; McConnell and Perez Quiros, 2000; Stock and Watson, 2003).

Let S_t denote an indicator variable which takes the value 1 in the high volatility state (which we label regime *a*) and 0 in the low volatility state (regime *b*). The trend and cycle disturbance variances are time varying and the model will be specified as in (9.22) with:

$$\eta_t \sim \mathrm{N}\left(0, S_t\sigma^2_{\eta a} + (1 - S_t)\sigma^2_{\eta b}\right), \quad \kappa_t \sim \mathrm{N}\left(0, S_t\sigma^2_{\kappa a} + (1 - S_t)\sigma^2_{\kappa b}\right).$$

This will be referred to as the GM specification. We shall consider two cases: (i) the sequence S_t is deterministic, taking the value 1 before 1984:1, and 0 thereafter; (ii) S_t is a random process, which we model as a first-order Markov chain with initial probability $p(S_0 = 1) = 1$, i.e., we know for certain that the process started in a high variance state, and transition probabilities $P(S_t = j|S_{t-1} = i) = \mathcal{T}_{ij}, i = 0, 1$, with $\mathcal{T}_{ij} = 1 - \mathcal{T}_{ii}$ for $j \neq i$.

9.3.2.1 ML estimation

The bivariate model and its GM extension under assumption (i) were estimated by ML in the time domain. The likelihood is evaluated by the Kalman filter (see Appendix C for details). The parameter estimates and the associated standard errors are reported in Table 9.2. The estimated trend and cycle disturbance variances are smaller after 1984:1 (regime b), as expected, and the likelihood ratio test of the homogeneity hypothesis, $H_0 : \sigma^2_{\eta a} = \sigma^2_{\eta b}, \sigma^2_{\kappa a} = \sigma^2_{\kappa b}$, clearly leads to a rejection. The roots of the AR polynomial for the output gap are complex and the loadings of core inflation on the output gap are significantly different from zero at the 5% level. The table also reports the Wald test for the null of long-run neutrality of the output gap, $H_0 : \theta_\psi(1) = 0$, which is accepted under both specifications, with p-values equal to 0.16 and 0.19. The evidence is thus that the output gap has only transitory effects on the level of inflation.

Figure 9.6 displays the point and 95% interval estimates of the output gap and the core component of inflation for both specifications. It is interesting that the explicit consideration of the Great Moderation of volatility makes the estimates of the output gap after the 1984 break more precise. In interpreting this result, we must stress that the interval estimates make no allowance for parameter uncertainty and for the uncertainty in dating the transition from the high volatility state to the low volatility one.

9.3.2.2 Bayesian estimation

Let us focus on the standard bivariate model (9.22) first and denote by $\mathbf{y}$ the stack of the observations $(y_t, \Delta p_t)$ for $t = 1, \ldots, n$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \ldots, \boldsymbol{\alpha}_n')'$, where the state vector at time t is $\boldsymbol{\alpha}_t = (\mu_t, \beta_t, \psi_t, \psi_{t-1}, \tau_t)$. Also, let $\boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\kappa}$ denote, respectively, the stack of potential output, the output gap, the disturbances η_t, and the cycle disturbances, where, e.g., $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$, and let $\Xi = [\phi_1, \phi_2, \sigma^2_\eta, \sigma^2_\kappa, \sigma^2_{p\varepsilon}, \sigma^2_{\tau\eta}, \theta_{\psi 0}, \theta_{\psi 1}]$ denote the vector of hyperparameters.[3] Notice that knowledge of $\boldsymbol{\alpha}$ implies knowledge of both the individual state components and the disturbances. Our main interest lies in aspects of the posterior marginal densities of the states given the observations, e.g., $f(\boldsymbol{\psi}|\mathbf{y})$ and $f(\Xi|\mathbf{y})$: e.g., $\mathrm{E}[h(\boldsymbol{\psi})] = \int h(\boldsymbol{\psi}) f(\boldsymbol{\psi}|\mathbf{y}) d\boldsymbol{\psi}$, for some function $h(\cdot)$ such as $h(\boldsymbol{\psi}) = \psi_t$. The computation of the integral is carried out by stochastic simulation: given a sample $\psi_t^{(i)}, i = 1, \ldots, M$, drawn from the posterior $f(\boldsymbol{\psi}|\mathbf{y})$, $\mathrm{E}[h(\boldsymbol{\psi})]$ is approximated by $M^{-1} \sum_i h\left(\psi_t^{(i)}\right)$. The required sample is obtained by Monte Carlo Markov chain methods and, in particular, by a Gibbs sampling (GS) scheme that we now discuss in detail. This scheme produces correlated random draws from the joint posterior density $f(\boldsymbol{\alpha}, \Xi|\mathbf{y})$, and thus from $f(\boldsymbol{\psi}|\mathbf{y})$, by repeatedly sampling an

Table 9.2 ML estimation results for bivariate models of quarterly US log GDP (y_t) and the consumer price inflation rate (Δp_t), 1950:1–2006:4

	Bivariate		*Great Moderation*	
	Parameter	*Std. error*	*Parameter*	*Std. error*
		y_t equation		
σ^2_η	0.33	0.14		
$\sigma^2_{\eta a}$			0.58	0.27
$\sigma^2_{\eta b}$			0.13	0.05
σ^2_κ	0.38	0.15		
$\sigma^2_{\kappa a}$			0.47	0.24
$\sigma^2_{\kappa b}$			0.06	0.04
ϕ_1	1.47	0.06	1.55	0.06
ϕ_2	−0.54	0.10	−0.60	0.09
		Δp_t equation		
$\sigma^2_{p\varepsilon}$	0.11	0.03	0.12	0.03
$\sigma^2_{\tau\eta}$	0.05	0.02	0.05	0.02
$\theta_{\psi 0}$	0.12	0.05	0.12	0.06
$\theta_{\psi 1}$	−0.10	0.05	−0.10	0.06
	Wald tests of restriction $\theta_\psi(1) = 0$			
	2.00		1.68	
loglik	−447.79		−415.53	

ergodic Markov chain whose invariant distribution is the target density (see Chib, 2001, and the references therein).

This is achieved by the following iterative scheme. Specify an initial value $\boldsymbol{\alpha}^{(1)}, \Xi^{(1)}$. For $i = 1, 2, \ldots, M$:

1. Generate $\boldsymbol{\alpha}^{(i)} \sim f(\boldsymbol{\alpha}|\Xi^{(i-1)}, \mathbf{y})$ using the simulation smoother (see Appendix C, section 9.7.4)
2. Generate $\Xi^{(i)} \sim f(\Xi^{(i)}|\boldsymbol{\alpha}^{(i)}, \mathbf{y})$. This block is divided into smaller components, whose full conditional distribution is available for sampling. In particular:
   (a) Generate $(\phi_1^{(i)}, \phi_2^{(i)})'$ from the full conditional $(\phi_1, \phi_2)'|\boldsymbol{\psi}, \sigma_\kappa^{2(i-1)}$ (this distribution is conditionally independent of $\mathbf{y}$, given $\boldsymbol{\psi}$). Assuming a Gaussian prior distribution, $\mathrm{N}(\mathbf{m}_{\phi 0}, \Sigma_{\phi 0})$, $(\phi_1, \phi_2)'|\boldsymbol{\psi}, \sigma_\kappa^{2(i-1)} \sim \mathrm{N}(\mathbf{m}_{\phi 1}, \Sigma_{\phi 1})$ where, denoting $\boldsymbol{\chi}_{t-1} = (\psi_{t-1}^{(i-1)}, \psi_{t-2}^{(i-1)})'$,

$$\Sigma_{\phi 1} = \left(\Sigma_{\phi 0}^{-1} + \frac{1}{\sigma_\kappa^{2(i-1)}} \sum_t \boldsymbol{\chi}_{t-1} \boldsymbol{\chi}_{t-1}' \right)^{-1},$$

$$\mathbf{m}_{\phi 1} = \Sigma_{\phi 1} \left(\Sigma_{\phi 0}^{-1} \mathbf{m}_{\phi 0} + \frac{1}{\sigma_\kappa^{2(i-1)}} \sum_t \boldsymbol{\chi}_{t-1} \psi_t \right).$$

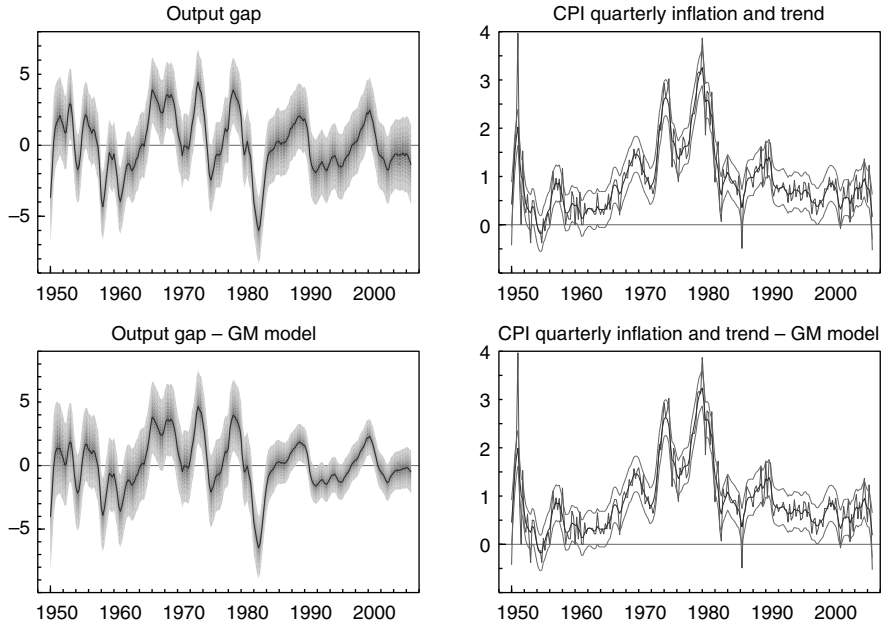


Figure 9.6 Estimates of the output gap and core inflation using the ML estimates of the parameters of the bivariate models of output and inflation under two specifications

The generations are repeated until a draw falls inside the stationarity region.

(b) Generate $\sigma_\eta^{2(i)}$ from the full conditional inverse gamma (IG) distribution:

$$\sigma_\eta^2 | \boldsymbol{\eta}^{(i-1)} \sim IG\left(\frac{\nu_\eta + n}{2}, \frac{\delta_\eta + \sum_t \eta_t^{(i-1)^2}}{2}\right).$$

This assumes that the prior distribution is $\sigma_\eta^2 \sim IG(\nu_\eta/2, \delta_\eta/2)$.

(c) Generate $\sigma_\kappa^{2(i)}$ from the full conditional IG distribution:

$$\sigma_\kappa^2 | \boldsymbol{\kappa}^{(i-1)} \sim IG\left(\frac{\nu_\kappa + n}{2}, \frac{\delta_\kappa + \sum_t \kappa_t^{(i-1)^2}}{2}\right).$$

This assumes that the prior distribution is $\sigma_\kappa^2 \sim IG(\nu_\kappa/2, \delta_k/2)$.

(d) Generate $(\theta_{\psi 0}^{(i)}, \theta_{\psi 1}^{(i)})'$. Assuming the Gaussian prior $(\theta_{\psi 0}, \theta_{\psi 1})' \sim \mathrm{N}(\mathbf{m}_{\theta 0}, \Sigma_{\theta 0})$, the full posterior is $(\theta_{\psi 0}, \theta_{\psi 1})' | \boldsymbol{\tau}, \sigma_{\tau\eta}^{2(i-1)} \sim \mathrm{N}(\mathbf{m}_{\theta 1}, \Sigma_{\theta 1})$, where

$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, and:

$$\Sigma_{\theta 1} = \left(\Sigma_{\theta 0}^{-1} + \frac{1}{\sigma_{\tau\eta}^{2(i-1)}} \sum_t \boldsymbol{\chi}_t \boldsymbol{\chi}_t' \right)^{-1},$$

$$\mathbf{m}_{\phi 1} = \Sigma_{\phi 1} \left(\Sigma_{\theta 0}^{-1} \mathbf{m}_{\phi 0} + \frac{1}{\sigma_{\tau\eta}^{2(i-1)}} \sum_t \boldsymbol{\chi}_t \Delta\tau_t \right).$$

(e) Generate $\sigma_{p\varepsilon}^{2(i)}$ from the full conditional IG distribution:

$$\sigma_{p\varepsilon}^2 | \boldsymbol{\varepsilon}_p^{(i-1)} \sim IG\left(\frac{\nu_\varepsilon + n}{2}, \frac{\delta_\varepsilon + \sum_t (\varepsilon_t^{(i-1)})^2}{2} \right).$$

Here $\boldsymbol{\varepsilon}_p$ is the stack of the inflation equation measurement disturbances, and we assume the prior $\sigma_{p\varepsilon}^2 \sim IG(\nu_\varepsilon/2, \delta_\varepsilon/2)$.

(f) Generate $\sigma_{\tau\eta}^{2(i)}$ from the full conditional IG distribution:

$$\sigma_{\tau\eta}^2 | \boldsymbol{\eta}_\tau^{(i-1)} \sim IG\left(\frac{\nu_\tau + n}{2}, \frac{\delta_\tau + \sum_t \eta_{\tau t}^{(i-1)^2}}{2} \right),$$

where $\boldsymbol{\eta}_\tau$ is the stack of the inflation equation core level disturbances, and we assume the prior $\sigma_{\eta\tau}^2 \sim IG(\nu_\tau/2, \delta_\tau/2)$.

The above GS scheme defines a homogeneous Markov chain such that the transition kernel is formed by the full conditional distributions and the invariant distribution is the unavailable target density.

The IG prior for the variance parameter is centered around the ML estimate and is not very informative ($\nu_\eta = \nu_\kappa = \nu_\varepsilon = \nu_\tau = 4$, and $n = 426$); for the AR parameters and the loadings impose a standard normal prior. The number of samples is $M = 2{,}000$ after a burn-in sample of size 1,000. Figure 9.7 displays the posterior means and the 95% interval estimates of the output gap (first panel), along with a nonparametric estimate of the posterior density of the variance parameters σ_η^2 and σ_κ^2 (top right panel); the modes are not far from the ML estimates. The bottom left panel shows the M draws $(\phi_1^{(i)}, \phi_2^{(i)})$ from the posterior of the AR parameter distribution. The triangle delimits the stationary region of the parameter space; the posterior means are 1.48 for ϕ_1 and -0.57 for ϕ_2. Finally, the last panel shows the posterior distribution of the change effect, $-\theta_{\psi 1}$, and the level effect $\theta_\psi(1)$. The 95% confidence interval for the latter is $(-0.01, 0.05)$, which confirms that the output gap has only transitory effects on inflation. The posterior mean of ψ_t does not differ from the point estimates arising from the classical analysis. However, the classical confidence intervals in Figure 9.6 are constructed by replacing Ξ with the ML estimates and thus do not take into account parameter uncertainty (see also section 9.4.2). It cannot be maintained that the classical estimates are more reliable.

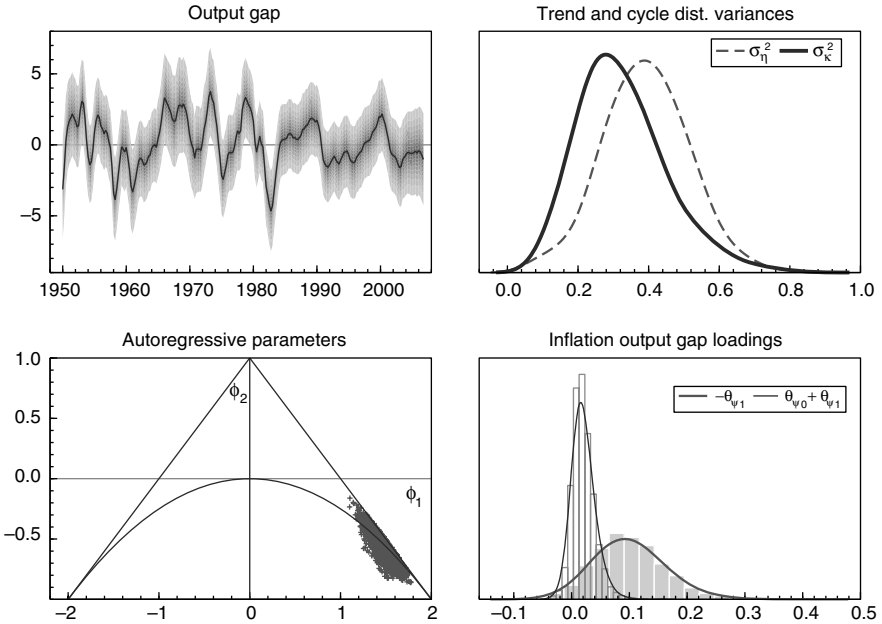


Figure 9.7 Bayesian estimation of the standard bivariate output gap model. Point and 95% interval estimates of the output gap; posterior densities of variance and loadings parameters; draws from the posterior of the AR parameters

For the GM model, the parameter set Ξ is such that the trend and cycle disturbance variances are replaced by the variances in the two regimes, $\sigma^2_{\eta a}, \sigma^2_{\eta b}, \sigma^2_{\kappa a}, \sigma^2_{\kappa b}$, and under the Markov switching specification (ii), according to which S_t is a first-order Markov chain, includes the transition probabilities $\mathcal{T}_{11}, \mathcal{T}_{00}$.

The steps of the GS algorithm need to be amended. An additional step is necessary to draw a sample from the distribution of $\mathbf{S} = (S_0, \ldots, S_n)$ conditional on $\boldsymbol{\alpha}$ and Ξ. Notice that this distribution depends on these random vectors only via $\boldsymbol{\eta}, \boldsymbol{\kappa}$, and the elements of Ξ, $\sigma^2_{\eta a}, \sigma^2_{\eta b}, \sigma^2_{\kappa a}, \sigma^2_{\kappa b}, \mathcal{T}_{11}, \mathcal{T}_{00}$. Sampling from the full posterior of the indicator variable $\mathbf{S}$ is achieved by the following algorithm (Carter and Kohn, 1994):

1. Sample $S_n^{(i)}$ from the filtered state probability distribution $P(S_n|\boldsymbol{\alpha}, \Xi, \mathbf{y}) = P(S_n|\boldsymbol{\eta}, \boldsymbol{\kappa}, \Xi)$.
2. For $t = n-1, \ldots, 1, 0$, sample $S_t^{(i)}$ from the conditional probability distribution:

$$P(S_t|S_{t+1}^{(i)}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \Xi) = \frac{P(S_{t+1}^{(i)}|S_t, \Xi)P(S_t|\boldsymbol{\eta}^t, \boldsymbol{\kappa}^t, \Xi)}{\sum_{S_t} P(S_{t+1}^{(i)}|S_t, \Xi)P(S_t|\boldsymbol{\eta}^t, \boldsymbol{\kappa}^t, \Xi)},$$

where $\boldsymbol{\eta}^t = (\eta_0, \ldots, \eta_t)$ and $\boldsymbol{\kappa}^t = (\kappa_0, \ldots, \kappa_t)$.

The filtered probabilities, $P(S_t|\boldsymbol{\eta}^t, \boldsymbol{\kappa}^t, \Xi)$, are obtained by the following discrete filter:

(i) The filter is started with the initial distribution $P(S_0 = 1|\boldsymbol{\eta}^0, \boldsymbol{\kappa}^0, \Xi) = 1, P(S_0 = 0|\boldsymbol{\eta}^0, \boldsymbol{\kappa}^0, \Xi) = 0$: that is, we impose that S_t started in the high volatility regime.
(ii) For $t = 1, 2, \ldots, n$, compute the one-step-ahead probability distribution $P(S_t|\boldsymbol{\eta}^{t-1}, \boldsymbol{\kappa}^{t-1}, \Xi) = \sum_{S_{t-1}} P(S_t|S_{t-1}, \Xi)P(S_{t-1}|\boldsymbol{\eta}^{t-1}, \boldsymbol{\kappa}^{t-1}, \Xi)$.
(iii) Compute the filtered probabilities:

$$P(S_t|\boldsymbol{\eta}^t, \boldsymbol{\kappa}^t, \Xi) = \frac{f(\eta_t, \kappa_t|S_t, \Xi)P(S_t|\boldsymbol{\eta}^{t-1}, \boldsymbol{\kappa}^{t-1}, \Xi)}{\sum_{S_t} f(\eta_t, \kappa_t|S_t, \Xi)P(S_t|\boldsymbol{\eta}^{t-1}, \boldsymbol{\kappa}^{t-1}, \Xi)},$$

where $f(\eta_t, \kappa_t|S_t, \Xi)$ is the product of two independent Gaussian densities with time-varying scale parameters.

Gerlach, Carter and Kohn (2000) have proposed an alternative sampling scheme for the indicator variable S_t which generates samples from $P(S_t|S_{j\neq t}, \mathbf{y}, \Xi)$ without conditioning on the states or the disturbances. This is more efficient than the above sampler if S_t is highly correlated with the states or the disturbances, which is not the case in our particular application.

Steps 1 and 2 of the GS algorithm are similar but the full posteriors are understood to be conditional on $\mathbf{S}^{(i-1)}$ as well. Furthermore, an additional step, 2(g), is added for sampling from the full conditionals of the transition probabilities, $\mathcal{T}_{11}, \mathcal{T}_{22}$, and the steps 2(b) and 2(c) are replaced as follows:

(b) Generate $\sigma^{2(i)}_{\eta a}$ and $\sigma^{2(i)}_{\eta b}$ from:

$$\sigma^2_{\eta a}|\boldsymbol{\eta}^{(i-1)}, \mathbf{S}^{(i-1)} \sim IG\left(\frac{\nu_\eta + \sum_t S_t^{(i-1)}}{2}, \frac{\delta_\eta + \sum_t S_t^{(i-1)}\eta_t^{(i-1)^2}}{2}\right),$$

$$\sigma^2_{\eta b}|\boldsymbol{\eta}^{(i-1)}, \mathbf{S}^{(i-1)} \sim IG\left(\frac{\nu_\eta + \sum_t (1 - S_t^{(i-1)})}{2}, \frac{\delta_\eta + \sum_t (1 - S_t^{(i-1)})\eta_t^{(i-1)^2}}{2}\right).$$

This assumes that the prior distribution is $\sigma^2_{\eta a}$ and $\sigma^2_{\eta b} \sim IG(\nu_\eta/2, \delta_\eta/2)$.

(c) Generate $\sigma^{2(i)}_{\kappa a}$ and $\sigma^{2(i)}_{\kappa b}$ from:

$$\sigma^2_{\kappa a}|\boldsymbol{\kappa}^{(i-1)}, \mathbf{S}^{(i-1)} \sim IG\left(\frac{\nu_\kappa + \sum_t S_t^{(i-1)}}{2}, \frac{\delta_\kappa + \sum_t S_t^{(i-1)}\kappa_t^{(i-1)^2}}{2}\right),$$

$$\sigma^2_{\kappa b}|\boldsymbol{\kappa}^{(i-1)}, \mathbf{S}^{(i-1)} \sim IG\left(\frac{\nu_\kappa + \sum_t (1 - S_t^{(i-1)})}{2}, \frac{\delta_\kappa + \sum_t (1 - S_t^{(i-1)})\kappa_t^{(i-1)^2}}{2}\right).$$

This assumes that the prior distribution is $\sigma^2_{\kappa a}$ and $\sigma^2_{\kappa b} \sim IG(\nu_\kappa/2, \delta_\kappa/2)$.

(g) Generate $\mathcal{T}_{11}^{(i)}, \mathcal{T}_{10}^{(i)} = 1 - \mathcal{T}_{11}^{(i)}$ and $\mathcal{T}_{00}^{(i)}, \mathcal{T}_{01}^{(i)} = 1 - \mathcal{T}_{00}^{(i)}$ from the posterior:

$$\mathcal{T}_{11}^{(i)}|\mathbf{S}^{(i-1)} \sim B\left(a_1 + N_{11}^{(i-1)}, b_1 + N_{10}^{(i-1)}\right),$$
$$\mathcal{T}_{00}^{(i)}|\mathbf{S}^{(i-1)} \sim B\left(a_0 + N_{00}^{(i-1)}, b_0 + N_{01}^{(i-1)}\right),$$

where $B(a, b)$ is the Beta distribution, $N_{ij}^{(i-1)}$ is the number of transitions from $S_t^{(i-1)} = i$ to $S_{t+1}^{(i-1)} = j$, and $a_i, b_i, i = 0, 1$, are the parameters of the Beta prior distributions (set equal to $a_1 = b_1 = b_0 = 1, a_0 = 5$). Notice that the transition probabilities are conditionally independent of $\boldsymbol{\alpha}$ and the other elements of Ξ, given $\mathbf{S}$.

Figure 9.8 summarizes aspects of the posterior distribution of the cycle, the indicator S_t, and some important parameters using a sample of $M = 2{,}000$ draws from the GS scheme outlined above with a burn-in of 2,000 iterations. Interestingly, the output gap interval estimates are more widely dispersed than in the original specification with no Markov-switching in the disturbance variances. This is so since the GM specification has a further source of variation and uncertainty, related to the state of Markov chain S_t, which in turn drives the changes in the volatility regime. As a result, the Bayesian interval estimates cannot be compared with the classical ones reported in the bottom left panel of Figure 9.6, since those were derived under the assumption that S_t was deterministic and known, and they make no allowance for parameter uncertainty. The estimated posterior probabilities of being in a high volatility regime confirm the general finding that the main stylized fact is a relatively sharp change point taking place in the first quarter of 1984, although there remains some uncertainty around that date. The nonparametric estimates of the posterior distribution of the transition probabilities $\mathcal{T}_{11}$ and $\mathcal{T}_{00}$ are displayed in the last panel of the figure. The posterior distributions of the variance parameters for the trend and cycle disturbances strongly confirm the Great Moderation hypothesis, and quantify it further, as both the permanent and transitory disturbances underwent a significant volatility reduction. The posterior means do not differ from the ML estimates reported in Table 9.2: $\mathrm{E}(\sigma_{\eta a}^2|\mathbf{y}) = 0.60, \mathrm{E}(\sigma_{\eta b}^2|\mathbf{y}) = 0.14$ and $\mathrm{E}(\sigma_{\kappa a}^2|\mathbf{y}) = 0.51, \mathrm{E}(\sigma_{\kappa b}^2|\mathbf{y}) = 0.09$. As far as the inflation equation is concerned, the overall conclusion is unchanged: the output gap is a significant source of variation (the value $-\theta_{\psi 1} = 0$ is estimated to be the 2.6 percentile of the posterior distribution of $-\theta_{\psi 1}$, which measures the change effect, but it drives inflation only in the short run, as the null of long-run neutrality is accepted (the 95% credible set for $\theta_\psi(1)$ is the interval $(-0.01, 0.05)$).

9.3.3 Multivariate extensions

The output gap is related to the deviations of the unemployment rate, u_t, from its "natural rate" or NAIRU via Okun's law. Okun (1962) defined natural unemployment as that level of unemployment occurring when output is equal to its

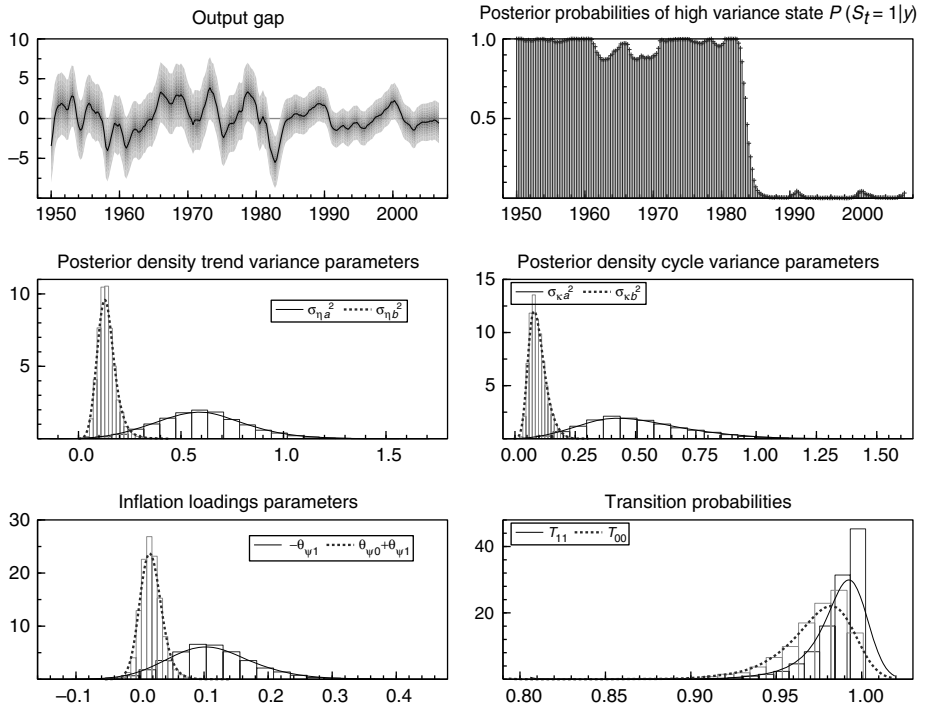


Figure 9.8 Bayesian estimation of the bivariate output gap model with Markov-switching in the variances of the trend and cycle disturbances (GM specification). Point and 95% interval estimates of the output gap; posterior probabilities of the high volatility state, $P(S_t = 1|\mathbf{y})$, and posterior densities of variance and loadings parameters

potential, and established an empirical law of strict proportionality between cyclical unemployment and the output gap. Hence, Okun's law is meant to imply that output and the unemployment rate share a common cycle.

Against this background, Clark (1989) estimated a bivariate model of US real output and unemployment such that output and unemployment are decomposed into two unrelated permanent components and the comovements between the two series result from the presence of a common cycle, represented as an AR(2) stationary component. Apel and Jansson (1999) obtained estimates of the NAIRU and potential output for the UK, the US and Canada, based on an unobserved components model of output, inflation and unemployment rates.

Another important multivariate extension of the basic bivariate model is the *production function approach* (PFA) to the estimation of potential output and the output gap, according to which potential output is obtained from the trend, or "noninflationary," levels of its structural determinants, such as productivity and factor inputs. This approach is currently one of the most popular methods of measuring potential output among statistical agencies being employed by the OECD (2001),

the International Monetary Fund (DeMasi, 1997), the Congressional Budget Office (2001), and the European Commission (see McMorrow and Roeger, 2001).

The PFA assumes that technology can be represented by a Cobb–Douglas production function with constant returns to scale on labor, measured by hours worked or by the number of employed persons, and capital:

$$y_t = f_t + \alpha h_t + (1-\alpha)k_t, \tag{9.23}$$

where f_t is the Solow residual, h_t is hours worked, k_t is the capital stock (all variables expressed in logarithms), and α is the elasticity of output with respect to labor ($0 < \alpha < 1$).

To achieve the decomposition $y_t = \mu_t + \psi_t$, the variables on the right-hand side of equation (9.23) are broken down additively into their permanent (denoted by the superscript P) and transitory (denoted by the superscript T) components, giving:

$$f_t = f_t^{(P)} + f_t^{(T)}, \quad h_t = h_t^{(P)} + h_t^{(T)}, \quad k_t = k_t^{(P)}. \tag{9.24}$$

It should be noticed that potential capital is always assumed to be equal to its actual value; this is so since capacity utilization is absorbed in the cyclical component of the Solow residual. Only survey-based measures of capacity utilization for the manufacturing sector are available for the euro-area.

Hence potential output is the value corresponding to the permanent values of factor inputs and the Solow residual, while the output gap is a linear combination of the transitory components:

$$\begin{aligned} \mu_t &= f_t^{(P)} + \alpha h_t^{(P)} + (1-\alpha)k_t, \\ \psi_t &= f_t^{(T)} + \alpha h_t^{(T)}. \end{aligned} \tag{9.25}$$

Hours worked can be separated into four components that are affected differently by the business cycle, as can be seen from the identity $h_t = n_t + pr_t + er_t + hl_t$, where n_t is the logarithm of working-age population (i.e., population of age 15–64), pr_t is the logarithm of the labor force participation rate (defined as the ratio of the labor force to the working-age population), er_t is the logarithm of the employment rate (defined here as the ratio of employment to the labor force), and hl_t is the logarithm of labor intensity (i.e., average hours worked). Each of these determinants is in turn decomposed into its permanent and transitory component in order to obtain the decomposition:

$$h_t^{(P)} = n_t + pr_t^{(P)} + er_t^{(P)} + hl_t^{(P)}, \quad h_t^{(T)} = pr_t^{(T)} + er_t^{(T)} + hl_t^{(T)}. \tag{9.26}$$

The idea is that population dynamics are fully permanent, whereas labor force participation, employment and average hours are also cyclical. Moreover, since the employment rate can be restated in terms of the unemployment rate, we can relate the output gap to cyclical unemployment and potential output to structural unemployment. As a matter of fact, since the unemployment rate is one minus the employment rate, $u_t = \log(1 - \exp(er_t))$, the variable $cur_t = -er_t$ (the contribution

of the unemployment rate, using a terminology due to Rünstler, 2002), is the first-order Taylor approximation to the unemployment rate. Thus, $cur_t^{(P)}$ can be assimilated to the NAIRU and $cur_t^{(T)}$ to the unemployment gap.

The PFA has the appealing feature that it uses a lot of economic information on the determinants of potential output; however, apart from the stringent data requirements (in particular, it requires the capital stock and hours worked), it requires the decomposition of the series involved into their permanent and transitory components. Proietti, Musso and Westermann (2007) propose a structural time series model-based implementation of the PFA approach, and Proietti and Musso (2007) extend it to carry out a growth accounting analysis for the euro-area.

9.3.4 A multivariate model with mixed frequency data

This section presents the results of fitting a multivariate monthly time series model for the US economy, using quarterly observations for GDP and monthly observations for industrial production, ip_t, the unemployment rate, u_t, and CPI inflation, Δp_t. The equation for the logarithm of GDP is the usual decomposition $y_t = \mu_t + \psi_t$ as in (9.22), with the important difference that the model is now formulated at the monthly frequency. The CPI equation is also specified as in (9.22).

Industrial production is included since it is an important timely coincident indicator: the time series model for ip_t is the trend-cycle decomposition $ip_t = \mu_{ip,t} + \theta_{ip}\psi_t + \psi_{ip,t}$, where $\mu_{ip,t} = \mu_{ip,t-1} + \beta_{ip} + \eta_{ip,t}$, and we assume that the trend disturbance is contemporaneously correlated with the GDP trend disturbance, η_t, $\eta_{ip,t} \sim \mathrm{N}(0, \sigma^2_{\eta,ip})$, $\mathrm{E}(\eta_t \eta_{ip,t}) = \sigma_{y,ip}$. The cyclical component is the combination of a common cycle and the idiosyncratic cycle $\psi_{ip,t} = \phi_{ip,1}\psi_{ip,t-1} + \phi_{ip,1}\psi_{ip,t-2} + \kappa_{ip,t}$.

The unemployment rate, u_t, is decomposed into the NAIRU, $\mu_{u,t}$, and cyclical unemployment, which is a distributed lag combination of the output gap plus an idiosyncratic component, ψ_{ut}, $u_t = \mu_{u,t} + \theta_{u0}\psi_t + \theta_{u1}\psi_{t-1} + \psi_{ut}$, where the NAIRU is a random walk without drift, $\mu_{u,t} = \mu_{u,t-1} + \eta_{u,t}$, and we assume that $\eta_{u,t} \sim \mathrm{NID}(0, \sigma^2_{\eta,u})$ is independent of any other disturbance in the model, whereas $\psi_{ut} = \phi_{u1}\psi_{u,t-1} + \phi_{u1}\psi_{u,t-2} + \kappa_{ut}$, with $\kappa_{ut} \sim \mathrm{NID}(0, \sigma^2_{\kappa u})$, independently of any other disturbance.

The link between the individual time series equations is provided by the output gap, ψ_t, which acts as the common cycle driving the short-run fluctuations; furthermore, the trend disturbances of y_t and ip_t are correlated. As GDP is quarterly, y_t is unobserved, whereas the available observations consist of the aggregated quarterly levels $Y_\tau = \exp(y_{3\tau}) + \exp(y_{3\tau-1}) + \exp(y_{3\tau-2})$, $\tau = 1, 2, \ldots, [n/3]$, where $[\cdot]$ is the integer part of the argument. For the statistical treatment it is useful to convert temporal aggregation into a systematic sampling problem, which is achieved by constructing a cumulator variable, generated by the following time-varying recursion (see Harvey, 1989): $Y_t^C = \varrho_t Y_{t-1}^C + \exp(y_t)$, where ϱ_t is the cumulator coefficient, defined as follows:

$$\varrho_t = \begin{cases} 0 & t = 3(\tau - 1) + 1, \quad \tau = 1, \ldots, [n/3] \\ 1 & \text{otherwise.} \end{cases}$$

Only a systematic sample of the cumulator variable Y_τ^c is available; in particular, the end of quarter value is observed, for $t = 3, 6, 9, \ldots, [n/3]$.

The model is represented in state-space form (see Appendix C) with the cumulator variable included in the state vector in the following way. The transition equation $Y_t^c = \varrho_t Y_{t-1}^c + \exp(y_t)$ is nonlinear, but it can be linearized around a trial estimate $\tilde{y}_t^*$ by a first-order Taylor series expansion:

$$Y_t^c = \varrho_t Y_{t-1}^c + \exp(\tilde{y}_t^*)[1 - \tilde{y}_t^*] + \exp(\tilde{y}_t^*) y_t;$$

replacing $y_t = \mu_t + \psi_t = \mu_{t-1} + \beta + \phi_1 \psi_{t-1} + \phi_2 \psi_{t-2}$ in the previous expression, Y_t^c can be given a first-order inhomogeneous Markovian representation, and thus the model can be cast in state-space form, so that conditionally on $\tilde{y}_t^*$ the model is linear and Gaussian.

The fixed interval smoother (see Appendix C, section 9.7.3) can be applied to the linearized model to yield estimates of the components μ_t and ψ_t of the unobserved monthly GDP (on a logarithmic scale), denoted μ_t^* and ψ_t^*. The latter provides a new $y_t^* = \mu_t^* + \psi_t^*$ value, which is used to build a new linearized Gaussian model, by a first-order Taylor series expansion of Y_t^c around y_t^*. Iterating this process until convergence yields an estimate of the component and of monthly GDP that satisfies the aggregation constraints (see Proietti, 2006b, for the theory and applications).

The model was estimated by ML using data from January 1950 to December 2006. The estimated parameters for the output gap (standard errors in parentheses) are $\hat{\phi}_1 = 1.73$ (0.021), $\hat{\phi}_2 = -0.744$ (0.037), and $\hat{\sigma}_\kappa^2 = 43 \times 10^{-7}$. For potential output $\hat{\beta} = .003$, $\hat{\sigma}_\eta^2 = 204 \times 10^{-7}$. The specific cycles for ip_t and u_t are estimated with zero variance, so that the cyclical components of industrial production and unemployment are related to the output gap. The estimated loading of ip_t on ψ_t is $\hat{\theta}_{ip} = 2.454$ (0.186); furthermore, the ip trend disturbances have variance $\hat{\sigma}_{\eta,ip}^2 = 3.74 \times 10^{-7}$, and are positively correlated (with coefficient 0.38) with the GDP trend disturbances. As far as unemployment is concerned, the estimated loadings on ψ_t are $\hat{\theta}_0 = -4.771$ (0.204) and $\hat{\theta}_1 = -2.904$ (0.267); moreover, $\hat{\sigma}_{\eta,u}^2 = 9304 \times 10^{-7}$, whereas the irregular disturbance variance was set to zero.

For the inflation equation the output gap loadings are estimated as $\hat{\theta}_{\tau 0} = 0.051$ (0.012) and $\hat{\theta}_{\tau 1} = -0.048$ (0.012); the Wald test for long-run neutrality, $H_0 : \theta_{\tau 0} + \theta_{\tau 1} = 0$ takes the value 1.401 with a p-value of 0.24, providing again evidence that the output gap has only transitory effects on inflation. The change effect, $-\theta_{\tau 1}$, is significant and has the expected sign. Finally, the trend disturbance variance for inflation was $\hat{\sigma}_\upsilon^2 = 2 \times 10^{-7}$.

Figure 9.9 presents the smoothed estimates of potential output, the output gap, the NAIRU and core inflation. As a by product, our model produces estimates of monthly GDP that are consistent with the quarterly observed values (the temporal aggregation constraints are satisfied exactly at convergence) and incorporate the information from related series.

Comparing the output gap estimates with those arising from the bivariate quarterly model, it can be argued that the use of an unemployment series makes a

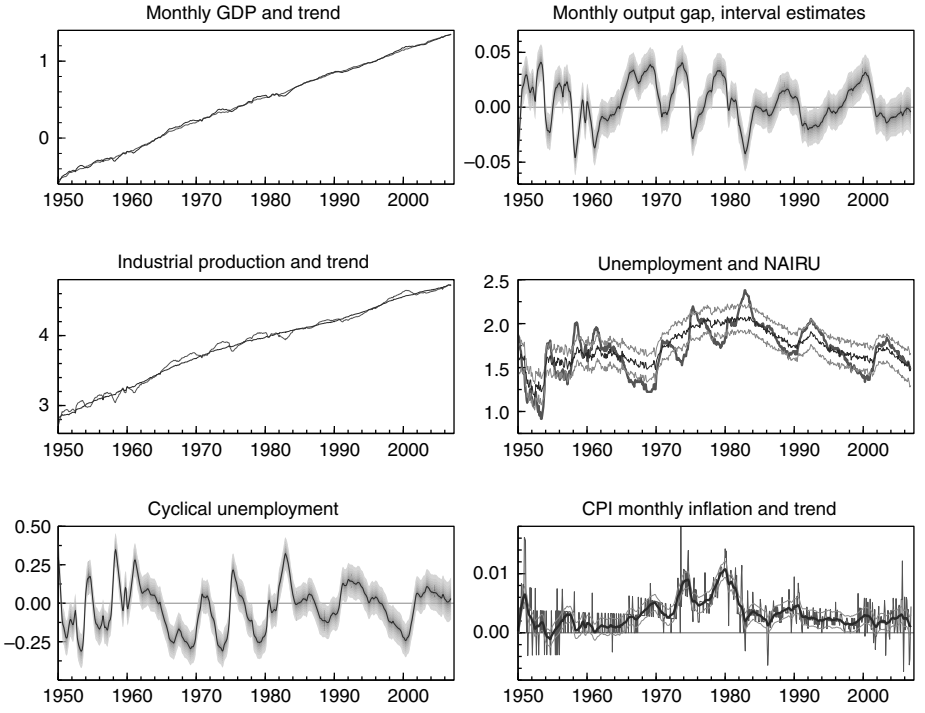


Figure 9.9 Monthly multivariate output gap model with temporal aggregation constraints. Smoothed estimates of monthly GDP, potential output, output gap, NAIRU, cyclical unemployment and core inflation

significant difference at the end of the sample. Also, enlarging the information set is beneficial to the reliability of the output gap estimates.

If the model is extended to allow for correlation between the output gap disturbance κ_t and the trend disturbance η_t, as in section 9.2.5, but in a multivariate set-up, the estimated correlation is $\hat{r} = 0.10$ and does not significantly differ from zero. In fact, the model with correlated disturbances has a likelihood of 7263.51, whereas the maximized likelihood of the restricted model ($r = 0$) is 7263.28. Thus, the LR test of $H_0 : r = 0$ takes the value 0.459, with p-value 0.50.

9.4 The reliability of the output gap measurement

The reliability of the output gap measurement is the subject of rich debate, and also has strong implications for optimal monetary policy. Orphanides and van Norden (2002) and Camba-Méndez and Rodriguez-Palenzuela (2003) discuss the different sources of uncertainty and their empirical assessment. The former conclude that the real-time estimates are unreliable. This conclusion echoes that by Staiger, Stock and Watson (1997) and Laubach (2001) concerning the NAIRU, obtained from a variety of methods. Somewhat different conclusions are reached by Planas

and Rossi (2004) and Proietti, Musso and Westermann (2007). The implications of the uncertainty surrounding the output gap estimates for monetary policy are considered in Orphanides *et al.* (2000) and Ehrmann and Smets (2003), among others.

A full assessment of the output gap reliability is complicated by the very nature of the measurement which, like the NAIRU, core inflation, and so forth, refers to a latent variable, for which there is no underlying "true value" to be elicited by other data collection techniques.

The previous sections have presented different parametric methods that can be used to measure the underlying signals. Assume that there is a true output gap ψ_t and that there is an approximating model, denoted by $\mathcal{M}$, providing a representation for it. The model specifies how the observations are related to the object of the measurement. Let us denote by $\psi_{t,\mathcal{M}}$ this (parametric) representation. Now, let $\tilde{\psi}_{t,\mathcal{M}}$ denote the estimator of ψ_t based on model $\mathcal{M}$, i.e., using the representation $\psi_{t,\mathcal{M}}$. We assume that $\tilde{\psi}_{t,\mathcal{M}}$ is the optimal signal extraction method for $\psi_{t,\mathcal{M}}$. How do we judge the reliability of $\tilde{\psi}_{t,\mathcal{M}}$? Reliability is a statement concerning the closeness of $\tilde{\psi}_{t,\mathcal{M}}$ and ψ_t. Following Boumans (2007), two key features are accuracy and precision, as the discrepancy $\tilde{\psi}_{t,\mathcal{M}} - \psi_t$ can be broken down into two components: $(\tilde{\psi}_{t,\mathcal{M}} - \psi_{t,\mathcal{M}}) + (\psi_{t,\mathcal{M}} - \psi_t)$, which are associated respectively to the precision of the method, and to the accuracy or validity of the representation chosen. Given the information set $\mathcal{F}$, precision is measured by (the inverse of) $\mathrm{Var}(\psi_{t,\mathcal{M}}|\mathcal{F}) = \mathrm{E}[(\tilde{\psi}_{t,\mathcal{M}} - \tilde{\psi}_{t,\mathcal{M}})^2|\mathcal{F}]$.

9.4.1 Validity

Validity is usually difficult to ascertain, as it is related to the appropriateness of $\psi_{t,\mathcal{M}}$ as a model for the signal ψ_t. This is a complex assessment, involving many subjective elements, such as any prior available information and the original motivation for signal extraction. The issue is indissolubly entwined with the nature of ψ_t: the previous paragraphs have considered two main perspectives. The first regards ψ_t as the component of the series that results from the transmission mechanism of demand or nominal shocks. The second view considers ψ_t as the bandpass component of output.

Recently, there has been a surge of interest in model uncertainty and in model averaging. The individual estimates $\tilde{\psi}_{t,\mathcal{M}_i}, i = 1, 2, \ldots, K$, may be combined linearly, giving $\tilde{\psi}_t = \sum_i c_i \tilde{\psi}_{t,\mathcal{M}_i}$, where the coefficients c_i are proportional to the precision of the methods, or the posterior probability in a Bayesian setting.

It is more viable to assess two other aspects of validity, namely concurrent and predictive validity. The first is concerned with the contemporaneous relationship between the measure $\tilde{\psi}_{t,\mathcal{M}}$ and a related alternative measure of the same phenomenon. Such measures are rarely available. Although business surveys are implemented with the objective of collecting informed opinions on latent variables, such as the state of the business cycle, they can hardly be considered as providing a measure of the "true" underlying state of the economy.

Predictive validity relates to the ability to forecast future realizations of y_t or related variables; evaluating the mean forecast error yields useful insight on its

predictive validity, as possible bias would emerge. This criterion is adopted by a number of authors; e.g., Camba-Mendez and Rodriguez-Palenzuela (2003) and Proietti, Musso and Westermann (2007) assess the reliability of alternative output gap estimates through their capability to predict future inflation.

9.4.2 Precision

A measurement method is precise if repeated measurements of the same quantity are in close agreement. Loosely speaking, precision is inversely related to the uncertainty of an estimate. In the measurement of immaterial constructs the sources of uncertainty would include: (i) *parameter uncertainty*, due to the fact that the core parameters Ξ characterizing model $\mathcal{M}$, such as the variance of the disturbances driving the components and the impulse response function, are unknown and have to be estimated; (ii) *estimation error*, the latent components are estimated with a positive variance even if a doubly infinite sample on y_t is available; (iii) *statistical revision*, as new observations become available, the estimate of a signal is updated so as to incorporate the new information; (iv) *data revision*.

The first source can be assessed by various methods in the classical approach; it is automatically incorporated in the interval estimates of the output gap if a Bayesian approach is adopted, as in section 9.3.2.2. The methods rely on the fundamental result that, under regularity conditions, the ML estimator of Ξ has the asymptotic distribution $\tilde{\Xi} \sim \mathrm{N}(\Xi, \mathbf{V})$, where $\mathbf{V}$ is the inverse of the information matrix. Hamilton (1986) proposed a Bayesian marginalization approach, which uses $\tilde{\Xi} \sim \mathrm{N}(\Xi, \mathbf{V})$ as a normal approximation to the posterior distribution of Ξ, given the available data. Then, a measure of the uncertainty of the smoothed estimates of the output gap, which embodies parameter uncertainty, is:

$$\widehat{\mathrm{Var}}(\psi_t|\mathcal{F}) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Var}(\psi_t|\mathcal{F}, \tilde{\Xi}^{(k)})] + \frac{1}{K}\sum_{k=1}^{K}\left[\mathrm{E}(\psi_t|\mathcal{F}, \tilde{\Xi}^{(k)}) - \hat{\mathrm{E}}(\psi_t|\mathcal{F})\right]^2, \quad (9.27)$$

where $\hat{\mathrm{E}}(\psi_t|\mathcal{F}) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}(\psi_t|\mathcal{F}, \tilde{\Xi}^{(k)})$, and the $\tilde{\Xi}^{(k)}$s are independent draws from the multivariate normal density $\mathrm{N}(\tilde{\Xi}, \tilde{\mathbf{V}}), k = 1, \ldots, K$, where $\tilde{\mathbf{V}}$ is evaluated at $\tilde{\Xi}$. According to the *delta method* proposed by Ansley and Kohn (1986), expressing the output gap estimates as a linear function of the parameter estimation error $\Xi - \tilde{\Xi}$ gives:

$$\widehat{\mathrm{Var}}(\psi_t|\mathcal{F}) = \mathrm{Var}(\psi_t|\mathcal{F}, \tilde{\Xi}) + \mathbf{d}(\tilde{\Xi})'\tilde{\mathbf{V}}\mathbf{d}(\tilde{\Xi}), \qquad \mathbf{d}(\tilde{\Xi}) = \frac{\partial}{\partial \boldsymbol{\Xi}}\mathrm{E}(\psi_t|\mathcal{F}, \tilde{\Xi})\,\big|_{\Xi=\tilde{\Xi}}, \quad (9.28)$$

where the derivatives in $\mathbf{d}(\tilde{\Xi})$ are evaluated numerically using the support of the Kalman filter and smoother. Similar methods apply for the real-time estimates, with the ML estimator being based on the information set $\mathcal{F}_t$. Quenneville and Singh (2000) evaluate and compare the two methods, and propose enhancements in a Bayesian perspective.

In an unobserved component framework the Kalman filter and smoother provide all the relevant information for assessing (ii) and (iii). For the latter, we can keep track of revisions by using a *fixed-point smoothing* algorithm (see Anderson and Moore, 1979; de Jong, 1989).

The sources (ii) and (iii) typically arise because the individual components are unobserved and are dependent through time. The availability of additional time series observations helps to improve the estimation of an unobserved component. Multivariate methods are more reliable as they use repeated measures of the same underlying latent variable and this increases the precision of the estimates. It is important to measure the uncertainty that surrounds the real time, or concurrent, estimates, $\mathsf{Var}(\psi_t|\mathcal{F}_t, \tilde{\Xi})$, which are conditional on the information set available to economic agents and policy makers at the time of making the assessment of the state of the economy, as opposed to the historical, or final, estimates. Comparing $\mathsf{Var}(\psi_t|\mathcal{F}_t, \tilde{\boldsymbol{\theta}})$ with the final estimation error variance, $\mathsf{Var}(\psi_t|\mathcal{F}_n, \tilde{\Xi}), n \to \infty$, gives a clue about the magnitude of the revision of the estimates as new observations become available.

In the absence of structural breaks, statistical revisions are sound and a fact of life (i.e., a natural consequence of optimal signal extraction). There is, however, great concern about revisions, especially for policy purposes; Orphanides and van Norden (2002) propose temporal consistency as a yardstick for assessing the reliability of output gap estimates; temporal consistency occurs when real-time (filtered) estimates do not differ significantly from the final (smoothed) estimates.

Finally, an additional source of uncertainty is data revision, which concerns y_t. Timely economic data are only provisional and are revised subsequently with the accrual of more complete information. Data revision is particularly relevant for national accounts aggregates, which require integrating statistical information from different sources and balancing it so as to produce internally consistent estimates.

9.5 Appendix A: Linear filters

A linear filter applied to a univariate series y_t is a weighted linear combination of its consecutive values. A time invariant filter can be represented as:

$$\mathrm{w}(L) = \sum_j \mathrm{w}_j L^j, \tag{9.29}$$

with w_j representing the filter weights. The above filter is symmetric if $\mathrm{w}_j = \mathrm{w}_{-j}$, in which case we can write $\mathrm{w}(L) = w_0 + \sum_j \mathrm{w}_j(L^j + L^{-j})$.

Applying $\mathrm{w}(L)$ to y_t yields $\mathrm{w}(L)y_t$ and has two consequences: the amplitude of the original fluctuations will be compressed or enhanced and the displacement over time of the original fluctuations will be altered. These effects can be fully understood in the frequency domain by considering the frequency response function (FRF) associated with the filter, which is defined as the Fourier transform of (9.29): $\mathrm{w}(e^{-\imath\omega}) = \sum_j \mathrm{w}_j e^{-\imath\omega j} = \mathrm{w}_{\mathcal{R}}(\omega) + \imath w_{\mathcal{I}}(\omega)$, where $\mathrm{w}_{\mathcal{R}}(\omega) = \sum_j \mathrm{w}_j \cos \omega j$ and $\mathrm{w}_{\mathcal{I}}(\omega) = \sum_j \mathrm{w}_j \sin \omega j$. The last equality stresses that, in general, the FRF is a complex quantity, with $w_{\mathcal{R}}(\omega)$ and $\mathrm{w}_{\mathcal{I}}(\omega)$ representing its real and complex part, respectively. The polar representation of the FRF, $\mathrm{w}(e^{-\imath\omega}) = G(\omega)e^{-\imath Ph(\omega)}$, is written in terms of two crucial quantities, the *gain*, $G(\omega) = |\mathrm{w}(e^{-\imath\omega})| = \sqrt{\mathrm{w}_{\mathcal{R}}(\omega)^2 + \mathrm{w}_{\mathcal{I}}(\omega)^2}$,

and the *phase* $Ph(\omega) = \arctan(-\mathrm{w}_{\mathcal{I}}(\omega)/\mathrm{w}_{\mathcal{R}}(\omega))$. The former measures the amplitude effect of the filter, so that if at some frequencies the gain is less than one, then those frequency components will be attenuated in the filtered series; the latter measures the displacement, or the phase shift, of the signal.

If $f_y(\omega)$ denotes the spectrum of y_t, the spectrum of $\mathrm{w}(L)y_t$ is equal to $|\mathrm{w}(e^{-i\omega})|^2 f_y(\omega)$, and therefore the square of the gain function (also known as the *power transfer function*) provides the factor by which the spectrum of the input series is multiplied to obtain that of the filtered series. In the important special case when $\mathrm{w}(L)$ is symmetric, the phase displacement is zero, and the gain is simply $G(\omega) = |\mathrm{w}_0 + 2\sum_{j=1}^{m} \mathrm{w}_j \cos \omega j|$.

9.6 Appendix B: The Wiener–Kolmogorov filter

The classical Wiener–Kolmogorov prediction theory, which is restricted to stationary processes, deals with optimal signal extraction of an unobserved component. Letting μ_t denote some stationary signal and y_t an indeterministic linear process with Wold representation $y_t = \nu(L)\xi_t, \nu(L) = 1 + \nu_1 L + \nu_2 L^2 + \cdots, \sum |\nu_j| < \infty$, $\xi_t \sim \mathrm{WN}(0, \sigma^2)$, the minimum mean square linear estimator of μ_{t+l} based on a semi-infinite sample $y_{t-j}, j = 0, 1, \ldots, \infty$, is:

$$\tilde{\mu}_{t+l|t} = \frac{1}{\sigma^2 \nu(L)} \left[\frac{g_{\mu y}(L)}{\nu(L^{-1})} L^{-l} \right]_+ y_t. \tag{9.30}$$

Here, $g_{\mu y}(L)$ denotes the cross-covariance generating function of μ_t and y_t, $g_{\mu y}(L) = \sum_j \gamma_{\mu y, j} L^j$, where $\gamma_{\mu y, j}$ is the cross-covariance at lag j, $\mathrm{E}[(\mu_t - \mathrm{E}(\mu_t))(y_{t-j} - \mathrm{E}(y_t))]$, and for $h(L) = \sum_{j=-\infty}^{\infty} h_j L^j$, $[h(L)]_+ = \sum_{j=0}^{\infty} h_j L^j$, i.e., a polynomial containing only non-negative powers of L; see Whittle (1983, p. 42). The formula for $l \leq 0$ provides the weights for signal extraction (contemporaneous filtering for $l = 0$).

If an infinite realization of future y_t was also available, the minimum mean square linear estimator is:

$$\tilde{\mu}_{t|\infty} = \frac{g_{\mu y}(L)}{g_y(L)} y_t,$$

where $g_y(L)$ is the autocovariance generating function of y_t, $g_y(L) = |\nu(L)|^2 \sigma^2$, and $|\nu(L)|^2 = \nu(L)\nu(L^{-1})$. If the series is decomposed into two orthogonal components, $y_t = \mu_t + \psi_t$, $g_{\mu y}(L) = g_\mu(L)$ (see Whittle, 1983, Ch. 5).

These formulae also hold when y_t and μ_t are non-stationary (see Pierce, 1979). As an example of their application, the expressions for the final and concurrent estimators of the trend component for model (9.2), with $\sigma_\eta^2 = 0$ and $\sigma_\psi^2/\sigma_\zeta^2 = \lambda$ (Leser–HP filter), are:

$$\tilde{\mu}_{t|\infty} = \frac{1}{1 + \lambda|1 - L|^4} y_t, \quad \tilde{\mu}_{t|t} = \frac{\theta(1)}{\theta(L)} y_t,$$

and the corresponding detrending filters are:

$$\tilde{\psi}_{t|\infty} = \frac{\lambda|1-L|^4}{1+\lambda|1-L|^4}y_t, \qquad \tilde{\psi}_{t|t} = \frac{\theta(L)-\theta(1)}{\theta(L)}y_t.$$

Here, $\theta(L) = 1+\theta_1 L+\theta_2 L^2$ is the reduced form MA polynomial of the local linear trend model (9.2). The numerator of the filtered detrended series can be rewritten: $\theta(L)-\theta(1) = \Delta\theta^*(1)L + \Delta^2\theta_0^*$, with $\theta^*(L) = \theta_0^* + \theta_1^* L = -(\theta_1+\theta_2) - \theta_2 L$.

The expression for $\tilde{\psi}_{t|\infty}$ is sometimes mistakenly taken to imply that the Leser–HP cycle filter makes stationary series that are integrated up to the fourth order, due to the presence of $|1-L|^4 = (1-L)^2(1-L^{-1})^2$ in the numerator of the filter. It should be recalled that the above formula holds true only for a doubly-infinite sample, and the real-time filter for extracting $\tilde{\psi}_{t|t}$ contains only the factor Δ^2.

9.7 Appendix C: State-space models and methods

The models considered in this chapter admit the state-space representation:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{G}_t\boldsymbol{\epsilon}_t, \quad t = 1,2,\ldots,n, \\ \boldsymbol{\alpha}_t &= \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{H}_t\boldsymbol{\eta}_t, \end{aligned} \tag{9.31}$$

where $\boldsymbol{\epsilon}_t \sim \text{NID}(\mathbf{0},\mathbf{I})$, $\boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0},\mathbf{I})$, and $\text{E}(\boldsymbol{\epsilon}_t\boldsymbol{\eta}_t') = \mathbf{0}$. The initial conditions are specified as follows: $\boldsymbol{\alpha}_0 = \tilde{\boldsymbol{\alpha}}_{0|0}^* + \mathbf{W}_0\boldsymbol{\delta} + \mathbf{H}_0\boldsymbol{\eta}_0$, so that $\boldsymbol{\alpha}_1|\boldsymbol{\delta} \sim \text{N}(\tilde{\boldsymbol{\alpha}}_{1|0}^* + \mathbf{W}_1\boldsymbol{\delta}, \mathbf{P}_{1|0}^*)$, where $\tilde{\boldsymbol{\alpha}}_{1|0}^* = \mathbf{T}_1\tilde{\boldsymbol{\alpha}}_{0|0}^*, \mathbf{W}_1 = \mathbf{T}_1\mathbf{W}_0$, and $\mathbf{P}_{1|0}^* = \mathbf{H}_1\mathbf{H}_1' + \mathbf{T}_1\mathbf{H}_0\mathbf{H}_0'\mathbf{T}_1'$. The random vector $\boldsymbol{\delta}$ captures the initial conditions for non-stationary state components and is assumed to have a diffuse distribution, $\boldsymbol{\delta} \sim \text{N}(\mathbf{0}, \Sigma_\delta)$, with $\Sigma_\delta^{-1} \to 0$. The matrices $\mathbf{Z}_t, \mathbf{G}_t, \mathbf{T}_t, \mathbf{H}_t, \mathbf{W}_0$ are deterministically related to a set of hyperparameters, Ξ.

For instance, for the bivariate model of output and inflation considered in section 9.3.1, $\mathbf{y}_t$ is a bivariate time series, $\boldsymbol{\alpha}_t = (\mu_t, \beta_t, \psi_t, \psi_{t-1}, \tau_t)'$, $\mathbf{Z}_t = \mathbf{Z} = (\mathbf{z}_y, \mathbf{z}_p)'$, $\mathbf{z}_y' = (1,0,1,0,0)$, $\mathbf{z}_p' = (0,0,0,0,1)$, $\boldsymbol{\epsilon}_t = \varepsilon_t/\sigma_\varepsilon$, $\mathbf{G}_t = \mathbf{G} = (0,\sigma_\varepsilon)'$, $\boldsymbol{\eta}_t = (\eta_t/\sigma_\eta, \kappa_t/\sigma_\kappa, \upsilon_t/\sigma_\upsilon)'$, $\boldsymbol{\delta} = (\mu_0, \beta_0, \tau_0)'$, $\tilde{\boldsymbol{\alpha}}_{0|0}^* = \mathbf{0}$,

$$\mathbf{T}_t = \mathbf{T} = \begin{pmatrix} \mathbf{T}_\mu & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\psi & \mathbf{0} \\ \mathbf{0}' & \mathbf{t}_p' & 1 \end{pmatrix}, \mathbf{T}_\mu = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \mathbf{T}_\psi = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}, \mathbf{t}_p = \begin{pmatrix} \theta_{\tau 0}\phi_1 + \theta_{\tau 1} \\ \theta_{\tau 0}\phi_2 \end{pmatrix},$$

$$\mathbf{H}_t = \mathbf{H} = \begin{pmatrix} \sigma_\eta & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \sigma_\kappa & 0 \\ 0 & 0 & 0 \\ 0 & \theta_{\tau 0}\sigma_\kappa & 0 \end{pmatrix}, \mathbf{W}_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \boldsymbol{\eta}_0 \sim \text{N}(\mathbf{0}, \mathbf{I}_2), \mathbf{H}_0 = \begin{pmatrix} \mathbf{0} \\ \mathbf{C}_\psi \\ 0 \end{pmatrix},$$

where $\mathbf{C}_\psi$ is such that $\text{E}(\boldsymbol{\psi}_0\boldsymbol{\psi}_0') = \mathbf{C}_\psi\mathbf{C}_\psi'$, $\boldsymbol{\psi}_0 = (\psi_0, \psi_{-1})'$.

9.7.1 The augmented Kalman filter

The Kalman filter (KF) is a fundamental algorithm for the statistical treatment of a state-space model. Under the Gaussian assumption it produces the minimum mean square estimator of the state vector along with its mean square error matrix, conditional on past information; this is used to build the one-step-ahead predictor of $\mathbf{y}_t$ and its mean square error matrix. Due to the independence of the one-step-ahead prediction errors, the likelihood can be evaluated via the prediction error decomposition.

The case when $\boldsymbol{\delta}$ is a fixed vector (fixed initial conditions) has been considered by Rosenberg (1973). He showed that $\boldsymbol{\delta}$ can be concentrated out of the likelihood function and that its generalized least square estimate is obtained from the output of an augmented KF. The diffuse case has been dealt with by de Jong (1989).

Defining $\mathbf{A}_{1|0} = -\mathbf{W}_1$, $q_0 = 0, \mathbf{s}_0 = \mathbf{0}, \mathbf{S}_0 = \mathbf{0}$, the augmented KF is given by the following recursive formulae and definitions for $t = 1, \ldots, n$:

$$
\begin{array}{ll}
\mathbf{v}_t^* = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}^*, & \mathbf{V}_t = -\mathbf{Z}_t\mathbf{A}_{t|t-1}, \\
\mathbf{F}_t^* = \mathbf{Z}_t\mathbf{P}_{t|t-1}^*\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t', & \mathbf{K}_t = \mathbf{T}_{t+1}\mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}, \\
q_t = q_{t-1} + \mathbf{v}_t^{*'}\mathbf{F}_t^{*-1}\mathbf{v}_t^*, \;\; \mathbf{s}_t = \mathbf{s}_{t-1} + \mathbf{V}_t'\mathbf{F}_t^{*-1}\mathbf{v}_t^*, & \mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{V}_t'\mathbf{F}_t^{*-1}\mathbf{V}_t, \\
\tilde{\boldsymbol{\alpha}}_{t+1|t}^* = \mathbf{T}_{t+1}\tilde{\boldsymbol{\alpha}}_{t|t-1}^* + \mathbf{K}_t\mathbf{v}_t^*, & \mathbf{A}_{t+1|t} = \mathbf{T}_{t+1}\mathbf{A}_{t|t-1} + \mathbf{K}_t\mathbf{V}_t \\
\mathbf{P}_{t+1|t}^* = \mathbf{T}_{t+1}\mathbf{P}_{t|t-1}^*\mathbf{T}_{t+1}' + \mathbf{H}_{t+1}\mathbf{H}_{t+1}' - \mathbf{K}_t\mathbf{F}_t^*\mathbf{K}_t'. &
\end{array}
\tag{9.32}
$$

The diffuse likelihood is defined as follows (de Jong, 1991):

$$
\ell(\mathbf{y}_1, \ldots, \mathbf{y}_n; \Xi) = -\frac{1}{2}\left(\sum_t \ln|\mathbf{F}_t^*| + \ln|\mathbf{S}_n| + q_n - \mathbf{s}_n'\mathbf{S}_n^{-1}\mathbf{s}_n\right). \tag{9.33}
$$

Denoting $\mathbf{Y}_t = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$, the innovations, $\mathbf{v}_t = \mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\mathbf{Y}_{t-1})$, the conditional covariance matrix $\mathbf{F}_t = \mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})$, the one-step-ahead prediction of the state vector $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1})$, and the corresponding covariance matrices, $\mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) = \mathbf{P}_{t|t-1}$, are given by:

$$
\begin{array}{lcllcl}
\mathbf{v}_t & = & \mathbf{v}_t^* - \mathbf{V}_t\mathbf{S}_{t-1}^{-1}\mathbf{s}_{t-1}, & \mathbf{F}_t & = & \mathbf{F}_t^* + \mathbf{V}_t\mathbf{S}_{t-1}^{-1}\mathbf{V}_t', \\
\tilde{\boldsymbol{\alpha}}_{t|t-1} & = & \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\mathbf{S}_{t-1}^{-1}\mathbf{s}_{t-1}, & \mathbf{P}_{t|t-1} & = & \mathbf{P}_{t|t-1}^* + \mathbf{A}_{t|t-1}\mathbf{S}_{t-1}^{-1}\mathbf{A}_{t|t-1}'.
\end{array}
\tag{9.34}
$$

9.7.2 Real-time (updated) estimates

The *updated* (or *real-time*, *filtered*) estimates of the state vector, $\tilde{\boldsymbol{\alpha}}_{t|t} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_t)$, and the covariance matrix of the real-time estimation error are, respectively:

$$
\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{t|t} &= \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\mathbf{S}_t^{-1}\mathbf{s}_t + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{-1}(\mathbf{v}_t^* - \mathbf{V}_t\mathbf{S}_t^{-1}\mathbf{s}_t), \\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1}^* - \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}^* + (\mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{V}_t)\mathbf{S}_t^{-1}(\mathbf{A}_{t|t-1} \\
&\quad + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{V}_t)'.
\end{aligned}
$$

9.7.3 Smoothing

Smoothing deals with the estimation of the components and the disturbances based on the full sample of observations. In the Gaussian case the *fixed interval smoother* provides the minimum mean square estimator of $\boldsymbol{\alpha}_t$ using $\mathbf{Y}_n$, $\tilde{\boldsymbol{\alpha}}_{t|n} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}_n)$, along with its covariance matrix $\mathbf{P}_{t|n} = \mathrm{E}[(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|n})(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|n})'|\mathbf{Y}_n]$. The computations can be carried out efficiently using the following backwards recursive formulae, given by Bryson and Ho (1969) and de Jong (1989), starting at $t = n$, with initial values $\mathbf{r}_n = 0, \mathbf{R}_n = \mathbf{0}$ and $\mathbf{N}_n = 0$:

$$\mathbf{r}_{t-1} = \mathbf{L}_t'\mathbf{r}_t + \mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{v}_t, \quad \mathbf{R}_{t-1} = \mathbf{L}_t'\mathbf{R}_t + \mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{V}_t, t = n-1, \ldots, 1.$$

$$\mathbf{N}_{t-1} = \mathbf{L}_t'\mathbf{N}_t\mathbf{L}_t + \mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{Z}_t,$$

$$\tilde{\boldsymbol{\alpha}}_{t|n} = \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\mathbf{S}_n^{-1}\mathbf{s}_n + \mathbf{P}_{t|t-1}^*(\mathbf{r}_{t-1} - \mathbf{R}_{t-1}\mathbf{S}_n^{-1}\mathbf{s}_n),$$

$$\mathbf{P}_{t|n} = \mathbf{P}_{t|t-1}^* - \mathbf{P}_{t|t-1}^*\mathbf{N}_{t-1}\mathbf{P}_{t|t-1}^* + (\mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{R}_{t-1})\mathbf{S}_n^{-1}(\mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{R}_{t-1})', \tag{9.35}$$

where $\mathbf{L}_t = \mathbf{T}_{t+1} - \mathbf{K}_t\mathbf{Z}_t'$. A preliminary forward KF pass is required to store the quantities $\tilde{\boldsymbol{\alpha}}_{t|t-1}^*$, $\mathbf{A}_{t|t-1}$, $\mathbf{P}_{t|t-1}^*$, $\mathbf{v}_t^*$, $\mathbf{V}_t$, $\mathbf{F}_t^*$ and $\mathbf{K}_t$.

The smoothed estimates of the disturbances are given by $\mathbf{H}_t\tilde{\boldsymbol{\eta}}_t = \mathrm{E}(\mathbf{H}_t\boldsymbol{\eta}_t|\mathbf{Y}_n) = \mathbf{H}_t\mathbf{H}_t'(\mathbf{r}_{t-1} - \mathbf{R}_{t-1}\mathbf{S}_n^{-1}\mathbf{s}_n)$, and $\mathbf{G}_t\tilde{\boldsymbol{\epsilon}}_t = \mathrm{E}(\mathbf{G}_t\boldsymbol{\epsilon}_t|\mathbf{Y}_n) = \mathbf{G}_t\mathbf{G}_t'\left[\mathbf{F}_t^{*-1}(\mathbf{v}_t - \mathbf{V}_t\mathbf{S}_n^{-1}\mathbf{s}_n) + \mathbf{K}_t'(\mathbf{r}_t - \mathbf{R}_t\mathbf{S}_n^{-1}\mathbf{s}_n)\right]$.

9.7.4 The simulation smoother

The simulation smoother is an algorithm which draws samples from the conditional distribution of the states and the disturbances given the observations and the hyperparameters. Carlin, Polson and Stoffer (1992) proposed a single-move state sampler, by which the states are sampled one at a time. This proves to be inefficient in the presence of highly autocorrelated state components. Gamerman (1998) proposed a single move disturbance sampler, which is more efficient since the disturbances driving the components are much less persistent and autocorrelated over time. Along with reparameterization, an effective strategy is blocking, through the adoption of a multi-move sampler as in Carter and Kohn (1994) and Früwirth-Schnatter (1994), who focus on sampling the states. Again, a more efficient multi-move sampler can be constructed by focusing on the disturbances, rather than the states. This is the idea underlying the simulation smoother proposed by de Jong and Shephard (1996).

Let $\boldsymbol{\varsigma}_t = \mathbf{C}[\boldsymbol{\epsilon}_t', \boldsymbol{\eta}_t']'$ denote a sub-set of the disturbances of the series, with $\mathbf{C}$ being a selection matrix. The structure of the state-space model model is such that the states are a (possibly singular) linear transformation of the disturbances and that $\mathbf{G}_t\boldsymbol{\epsilon}_t$ can be recovered from $\mathbf{H}_t\boldsymbol{\eta}_t$ via the measurement equation, which implies that the distribution of $(\boldsymbol{\epsilon}', \boldsymbol{\eta}')'|\mathbf{Y}_n$ is singular. Hence, to achieve efficiency and to avoid degeneracies, we need to focus on a suitably selected sub-set of the disturbances. The simulation smoother hinges on the following factorization of

the joint posterior density:

$$f(\varsigma_0, \ldots, \varsigma_n|\mathbf{Y}_n) = f(\varsigma_n|\mathbf{y}) \prod_{t=0}^{n-1} f(\varsigma_t|\varsigma_{t+1}, \ldots, \varsigma_n; \mathbf{Y}_n).$$

Conditional random vectors are generated recursively: in the forward step the Kalman filter is run and the innovations, their covariance matrix and the Kalman gain are stored. In the backwards sampling step conditional random vectors are generated recursively from $\varsigma_t|\varsigma_{t+1}, \ldots, \varsigma_n; \mathbf{y}$; the algorithm keeps track of all the changes in the mean and the covariance matrix of these conditional densities. The simulated disturbances are then inserted into the transition equation to obtain a sample from $\boldsymbol{\alpha}|\mathbf{Y}_n$.

A more efficient simulation smoother has been developed by Durbin and Koopman (2002). The gain in efficiency arises from the fact that only the first conditional moments of the states or the disturbances need to be evaluated. Let us redefine $\varsigma_t = (\boldsymbol{\epsilon}_t', \boldsymbol{\eta}_t')'$ and let $\tilde{\varsigma} = \mathrm{E}(\varsigma|\mathbf{Y}_n)$, where ς is the stack of the vectors ς_t; $\tilde{\varsigma}$ is computed by the disturbance smoother (see Koopman, 1993, and Appendix C, section 9.7.3). We can write $\varsigma = \tilde{\varsigma} + \varsigma^*$, where $\varsigma^* = \varsigma - \tilde{\varsigma}$ is the disturbance smoothing error, with conditional distribution $\varsigma^*|\mathbf{Y}_n \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$, such that the covariance matrix $\mathbf{V}$ does not depend on the observations, and thus does not vary across the simulations (the diagonal blocks are computed by the smoothing algorithm in Appendix C, section 9.7.3). A sample from $\varsigma^*|\mathbf{Y}_n$ is constructed as follows: we first draw the disturbances from their unconditional Gaussian distribution $\varsigma^+ \sim \mathrm{NID}(0, \mathbf{I})$ and construct the pseudo observations $\mathbf{y}^+$ recursively from $\boldsymbol{\alpha}_t^+ = \mathbf{T}_t\boldsymbol{\alpha}_{t-1}^+ + \mathbf{H}_t\boldsymbol{\eta}_t^+, \mathbf{y}_t^+ = \mathbf{Z}_t\boldsymbol{\alpha}_t^+ + \mathbf{G}_t\boldsymbol{\epsilon}_t^+, t = 1, 2, \ldots, n$, where the initial draw is $\boldsymbol{\alpha}_0^+ \sim \mathrm{N}(\mathbf{0}, \mathbf{H}_0\mathbf{H}_0')$. The Kalman filter and the smoothing algorithm computed on the simulated observations $\mathbf{y}_t^+$ will produce $\tilde{\varsigma}_t^+$ and $\tilde{\boldsymbol{\alpha}}_t^+$, and $\varsigma_t^+ - \tilde{\varsigma}_t^+$ will be the desired draw from $\varsigma^*|\mathbf{Y}_n$. Hence, $\tilde{\varsigma} + \varsigma_t^+ - \tilde{\varsigma}_t^+$ is a sample from $\varsigma|\mathbf{Y}_n \sim \mathrm{N}(\tilde{\varsigma}, \mathbf{V})$.

Acknowledgments

The author wishes to thank Andrew Harvey, Terence Mills and Alberto Musso for their useful suggestions.

Notes

1. Assuming $\boldsymbol{\mu}_* = (\mu_1, \mu_2)' \sim \mathrm{N}(\mathbf{0}, \Sigma_\mu)$, and that the process μ_t has started in the indefinite past, $\Sigma_\mu^{-1} \to \mathbf{0}$, and thus the quadratic form $\boldsymbol{\mu}_*'\Sigma_\mu^{-1}\boldsymbol{\mu}_*$ converges to zero.
2. All the computations in this chapter have been performed using Ox version 4 (see Doornik, 2006).
3. The slope parameter is included in the state vector; the transition equation is $\beta_t = \beta_{t-1}$, with β_0 being a diffuse parameter (see Appendix C, section 9.7).

References

Anderson, B.D.O. and J.B. Moore (1979) *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall.
Ansley, C. and R. Kohn (1986) Prediction mean square error for state space models with estimated paramaters. *Biometrika* **73**, 467–73.

Apel, M. and P. Jansson (1999) A theory-consistent system approach for estimating potential output and the NAIRU. *Economics Letters* **64**, 271–5.

Artis, M., M. Marcellino and T. Proietti (2004) Dating business cycles: a methodological contribution with an application to the euro area. *Oxford Bulletin of Economics and Statistics* **66**, 537–74.

Basistha, A. and C.R. Nelson (2007) New measures of the output gap based on the forward-looking New Keynesian Phillips curve. *Journal of Monetary Economics* **54**, 498–511.

Baxter, M. and R.G. King (1999) Measuring business cycles: approximate band-pass filters for economic time series. *Review of Economics and Statistics* **81**, 575–93.

Beveridge, S. and C.R. Nelson (1981) A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "Business cycle." *Journal of Monetary Economics* **7**, 151–74.

Brewer, K.R.W. (1979) Seasonal adjustment of ARIMA series. *Économie Appliquée* **1**, 7–22.

Boumans, M. (2007) *Measurement in Economics: A Handbook*. Amsterdam: Elsevier.

Bry, G. and C. Boschan (1971) *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. New York: NBER.

Bryson, A.E. and Y.C. Ho (1969) *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, Mass.: Blaisdell Publishing.

Burns, A.F. and W.C. Mitchell (1946) *Measuring Business Cycles*. New York: NBER.

Camba-Mendez, G. and D. Rodriguez-Palenzuela (2003) Assessment criteria for output gap estimates. *Economic Modelling* **20**, 529–62.

Canova, F. (1998) Detrending and business cycle facts. *Journal of Monetary Economics* **41**, 475–512.

Cappé, O., E. Moulines and T. Ryden (2005) *Inference in Hidden Markov Models*. Springer Series in Statistics. New York: Springer.

Carlin, B.P., N.G. Polson and D.S. Stoffer (1992) A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association* **87**, 493–500.

Carter, C.K. and R. Kohn (1994) On Gibbs sampling for state space models. *Biometrika* **81**, 541–53.

Chib, S. (2001) Markov chain Monte Carlo methods: computation and inference. In J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Volume 5*, pp. 3569–649. Amsterdam: North-Holland.

Christiano, L.J. and T.J. Fitzgerald (2003) The band pass filter. *International Economic Review* **44** 435–65.

Clark, P.K. (1987) The cyclical component of U.S. economic activity. *Quarterly Journal of Economics* **102**, 797–814.

Clark, P.K. (1989) Trend reversion in real output and unemployment. *Journal of Econometrics* **40**, 15–32.

Cox, D.R. (1961) Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society*, *Series B* **23**, 414–22.

Cogley, T. and J.M. Nason (1995) Effects of the Hodrick–Prescott filter on trend and difference stationary time series. Implications for business cycle research. *Journal of Economic Dynamics and Control* **19**, 253–78.

Congressional Budget Office (2001) *CBO's Method for Estimating Potential Output: An Update*. CBO Memorandum, Washington, DC.

de Jong, P. (1989) Smoothing and interpolation with the state space model. *Journal of the American Statistical Association* **84**, 1085–8.

de Jong, P. (1991) The diffuse Kalman filter. *Annals of Statistics* **19**, 1073–83.

de Jong, P. and N. Shephard (1995) The simulation smoother. *Biometrika* **2**, 339–50.

DeMasi, P. (1997) IMF estimates of potential output: theory and practice. IMF Working Paper No. 97/177. Washington, DC: IMF.

Doménech, R. and V. Gómez (2006) Estimating potential output, core inflation and the NAIRU as latent variables. *Journal of Business and Economic Statistics* **24**, 354–65.

Doornik, J.A. (2006) *Ox: An Object-Oriented Matrix Programming Language*. London: Timberlake Consultants Press.
Durbin, J. and S.J. Koopman (2001) *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
Durbin, J. and S.J. Koopman (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**, 603–15.
Ehrmann, M. and F. Smets (2003) Uncertain potential output: implications for monetary policy. *Journal of Economic Dynamics and Control* **27**, 1611–38.
Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.
Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York: Springer.
Gamerman, D. (1998) Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* **85**, 215–27.
Gardner, E.S. (1985) Exponential smoothing: the state of the art. *Journal of Forecasting* **4**, 1–28.
Gardner, E.S. (2006) Exponential smoothing: the state of the art. Part II. *International Journal of Forecasting* **22**, 637–66.
Gerlach, R., C. Carter and R. Kohn (2000) Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association* **95**, 819–28.
Gerlach, S. and F. Smets (1999) Output gaps and monetary policy in the EMU area. *European Economic Review* **43**, 801–12.
Gómez, V. (2001) The use of Butterworth filters for trend and cycle estimation in economic time series. *Journal of Business and Economic Statistics* **19**(3), 365–73.
Gordon, R.J. (1997) The time-varying NAIRU and its implications for economic policy. *Journal of Economic Perspectives* **11**, 11–32.
Hamilton, J.D. (1986) A standard error for the estimated state vector of a state space model. *Journal of Econometrics* **33**, 387–97.
Harvey, A.C. (1989) *Forecasting, Structural Time Series and the Kalman Filter*. Cambridge: Cambridge University Press.
Harvey, A.C. (2001) Testing in unobserved components models. *Journal of Forecasting* **20**, 1–19.
Harvey, A.C. and A. Jaeger (1993) Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics* **8**, 231–247.
Harvey, A.C. and T. Proietti (2005) *Readings in Unobserved Components Models*. Advanced Texts in Econometrics. Oxford: Oxford University Press.
Harvey, A.C. and T.M. Trimbur (2003) General model-based filters for extracting trends and cycles in economic time series. *Review of Economics and Statistics* **85**, 244–55.
Harvey, A.C., T.M. Trimbur and H.K. Van Dijk (2007) Trends and cycles in economic time series: a Bayesian approach. *Journal of Econometrics* **140**, 618–49.
Hodrick R.J. and E.C. Prescott (1997) Postwar U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking* **29**, 1–16.
Kaiser, R. and A. Maravall (2005) Combining filter design with model-based filtering: an application to business-cycle estimation. *International Journal of Forecasting* **21**, 691–710.
King, R.G. and Rebelo, S.T. (1993) Low frequency filtering and real business cycles. *Journal of Economic Dynamics and Control* **17**, 207–31.
Kim, C.J. and C.R. Nelson (1999a). Has the U.S. economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. *Review of Economics and Statistics* **81**, 608–16.
Kim, C.J., and C. Nelson (1999b) *State Space Models with Regime Switching*. Cambridge, Mass.: MIT Press.
Kitagawa, G. and W. Gersch (1996) *Smoothness Priors Analysis of Time Series*. Berlin: Springer-Verlag.
Koopman, S.J. (1993) Disturbance smoother for state space models. *Biometrika* **80**, 117–26.

Kuttner, K.N. (1994) Estimating potential output as a latent variable. *Journal of Business and Economic Statistics* **12**, 361–8.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* **54**, 159–78.

Laubach, T. (2001) Measuring the NAIRU: evidence from seven economies. *Review of Economics and Statistics* **83**, 218–31.

Leser, C.E.V. (1961) A simple method of trend construction. *Journal of the Royal Statistical Society, Series B* **23**, 91–107.

Maravall, A. and A. del Rio (2007) Temporal aggregation, systematic sampling, and the Hodrick-Prescott filter. *Computational Statistics & Data Analysis* **52**, 975–98.

McConnell, M.M. and G.P. Perez Quiros (2000) Output fluctuations in the United States: what has changed since the early 1980s? *American Economic Review* **90**, 1464–76.

McMorrow, K. and W. Roeger (2001) Potential output: measurement methods, "new" economy influences and scenarios for 2001–2010. A comparison of the EU15 and the US. *European Economy – Economic Papers* No. 150. Commission of the European Communities. Directorate-General for Economic and Financial Affairs, Brussels.

Mills, T.C. (2003) *Modelling Trends and Cycles in Economic Time Series*. Basingstoke: Palgrave Macmillan.

Morley, J.C., C.R. Nelson and E. Zivot (2003) Why are Beveridge-Nelson and unobserved-component decompositions of GDP so different? *Review of Economics and Statistics* **85**, 235–43.

Muth, J.F. (1960) Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* **55**, 299–306.

Nerlove, M.L., D.M. Grether and J.L. Carvalho (1995) *Analysis of Economic Time Series: A Synthesis* (revised edition). New York: Academic Press Inc.

Nyblom, J. and T. Mäkeläinen (1983) Comparison of tests for the presence of random walk coefficients in a simple linear model. *Journal of the American Statistical Association* **78**, 856–64.

OECD (2001) *OECD Economic Outlook*, No. 69, June 2001. Paris: OECD.

Okun, A. (1962) Potential GNP: its measurement and significance. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*. Reprinted in A. Okun, *The Political Economy of Prosperity*. pp. 132–5. New York: Norton, 1970.

Orphanides, A., R. Porter, D. Reifschneider, R. Tetlow and F. Finan (2000) Errors in the measurement of the output gap and the design of monetary policy. *Journal of Economics and Business* **52**, 117–43.

Orphanides, A. and S. van Norden (2002) The unreliability of output gap estimates in real time. *Review of Economics and Statistics* **84**, 569–83.

Percival, D.B. and A.T. Walden (1993) *Spectral Analysis for Physical Applications. Multitaper and Conventional Univariate Techniques*. Cambridge: Cambridge University Press.

Pierce, D.A. (1979) Signal extraction error in nonstationary time series. *Annals of Statistics* **7**, 1303–20.

Planas, C. and A. Rossi (2004) Can inflation data improve the real-time reliability of output gap estimates? *Journal of Applied Econometrics* **19**, 121–33.

Planas, C., A. Rossi and G. Fiorentini (2007) Bayesian analysis of output gap. *Journal of Business and Economic Statistics* **26**(1), 18–32.

Proietti, T. (2004) On the model based interpretation of filters and the reliability of trend-cycle estimates. Forthcoming in 2009 in *Econometric Reviews*.

Proietti, T. (2005) Forecasting and signal extraction with misspecified models. *Journal of Forecasting* **24**, 539–56.

Proietti T. (2006a) Trend-cycle decompositions with correlated components. *Econometric Reviews* **25**, 61–84.

Proietti T. (2006b) On the estimation of nonlinearly aggregated mixed models. *Journal of Computational and Graphical Statistics* **15**, 18–38.

Proietti, T. and A. Musso (2007) Growth accounting for the euro area: a structural approach. ECB Working Paper Series No. 804, Frankfurt: ECB.

Proietti T., A. Musso and T. Westermann (2007) Estimating potential output and the output gap for the euro area: a model-based production function approach. *Empirical Economics* **33**, 85–113.

Quenneville, B. and A.C. Singh (2000) Bayesian prediction mean squared error for state space models with estimated parameters. *Journal of Time Series Analysis* **21**, 219–36.

Ravn, M.O. and H. Uhlig (2002) On adjusting the Hodrick Prescott filter for the frequency of observations. *Review of Economics and Statistics* **84**, 371–6.

Rosenberg, B. (1973) Random coefficient models: the analysis of a cross-section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement* **2**, 399–428.

Rudebusch, G. and L.E.O. Svensson (1998) Policy rules for inflation targeting. In J.B. Taylor (ed.), *Monetary Policy Rules*, NBER Business Cycle Series, Volume 31. Chicago: Chicago University Press.

Rünstler, G. (2002) The information content of real-time output gap estimates an application to the euro area. ECB Working Paper No. 182. Frankfurt, ECB.

Sayed, A.H. and T. Kailath (2001) A survey of spectral factorization methods. *Numerical Linear Algebra with Applications* **8**, 467–96.

Staiger, D., J.H. Stock and M.W. Watson (1997) The NAIRU, unemployment and monetary policy. *Journal of Economic Perspectives* **11**, 33–50.

Stock, J. and M. Watson (2003) Has the business cycle changed and why? *NBER Macroeconomics Annual 2002* **17**, 159–218.

Taylor, J.B. (1999) A historical analysis of monetary policy rules. In J.B. Taylor (ed.), *Monetary Policy Rules*, pp. 319–41. Chicago: University of Chicago Press.

Tiao, G.C. and D. Xu (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika* **80**, 623–41.

West, M. and J. Harrison (1997), *Bayesian Forecasting and Dynamic Models* (second edition). New York: Springer-Verlag.

Whittle, P. (1983). *Prediction and Regulation by Linear Least Squares Methods* (second edition). Oxford: Blackwell.

10

Fractional Integration and Cointegration: An Overview and an Empirical Application

Luis A. Gil-Alana and Javier Hualde

Abstract

In this chapter we first review the theoretical and empirical work on fractional integration and cointegration, placing special emphasis on the estimation procedures for fractionally cointegrated systems. An empirical application is then carried out using some of the more recently developed techniques in this area. In particular, we investigate the purchasing power parity hypothesis for four bivariate price series, the US ("domestic") versus the "foreign" countries Australia, Canada, Italy and the UK. Fractional cointegration is found in the US–UK relationship.

10.1 Introduction

One characteristic of many economic and financial time series is its non-stationary nature. There exists a great variety of models to describe such non-stationarity.

Until the 1980s a standard approach was to impose a deterministic (linear or quadratic) function of time, thus assuming that the residuals from the regression model were stationary. Later on, and especially after the seminal work of Nelson and Plosser (1982), there was a general agreement that the non-stationary component of most series was stochastic, and unit roots (or first differences) were commonly adopted. However, the unit root is merely one particular model to describe such behavior. In fact, the number of differences required to get to stationarity may not necessarily be an integer value but any point in the real line.[1] In such a case, the process is said to be fractionally integrated or $I(d)$. The $I(d)$ models belong to a wider class of processes called long memory, which we can define in either the time or the frequency domains.

Let us consider a zero mean process $\{x_t,\ t = 0, \pm 1, \ldots\}$ with $\gamma_u = E(x_t x_{t+u})$. The time domain definition of long memory states that:

$$\sum_{u=-\infty}^{\infty} |\gamma_u| = \infty.$$

Now, assuming that x_t has an absolutely continuous spectral distribution, so that it has spectral density function:

$$f(\lambda) = \frac{1}{2\pi}\left(\gamma_0 + 2\sum_{u=1}^{\infty}\gamma_u \cos(\lambda u)\right),$$

the frequency domain definition of long memory states that the spectral density function is unbounded at some frequency in the interval $[0, \pi)$. Most of the empirical literature has concentrated on the case where the singularity, or pole, in the spectrum takes place at the zero frequency. This is the standard case of $I(d)$ models of the form:

$$(1 - L)^d x_t = u_t, \quad t = 0, \pm 1, \ldots, \tag{10.1}$$

where L is the lag operator ($Lx_t = x_{t-1}$) and u_t is $I(0)$. However, fractional integration may also occur at some other frequencies away from zero, as in the case of seasonal/cyclical models.

In the multivariate case, the natural extension of fractional integration is the concept of fractional cointegration. Though the original idea of cointegration, as espoused by Engle and Granger (1987), allows for fractional orders of integration, all the empirical work carried out during the 1990s was restricted to the case of integer degrees of differencing. Only in recent years have fractional values also been taken into account.

In this chapter we review fractional integration and cointegration, placing special emphasis on the latter concept, which has recently emerged in the time series literature. We also present an empirical application using some of the most novel techniques in this area. The outline of the chapter is as follows. Section 10.2 concentrates on fractional integration and some of its most recent developments. Section 10.3 deals with fractional cointegration, while section 10.4 is devoted to an empirical example.

10.2 Fractional integration

10.2.1 Concept and modelization

The idea of fractional integration was introduced by Granger and Joyeux (1980), Granger (1980, 1981) and Hosking (1981), though Adenstedt (1974) and Taqqu (1975) earlier showed an awareness of its representation. Assuming that x_t is given by equation (10.1), the first point we have to deal with is the treatment of pre-sample observations. In short memory contexts (that is, $d = 0$), different initial value conventions lead to parameter estimates which typically share the same first-order asymptotic properties but have different finite-sample properties. In empirical work zeros or the sample mean often initiate the series, with early observations being thrown away. Different conventions have also been followed in non-stationary series with an autoregressive unit root. In stationary fractional processes, infinitely many pre-sample values have to be chosen, so the potential divergence between rival methods and parameter estimates is greater, though first-order asymptotic properties are again robust. In fractional contexts, two definitions of fractional integration have been employed. From equation (10.1), for $d < 1/2$:

$$x_t = \Delta^{-d} u_t, t = 0, \pm 1, \ldots,$$

where $\Delta = 1 - L$. For integer $d_a \geq 0$:

$$z_{at} = \Delta^{-d_a} x_t^{\#}, \quad t = 0, \pm 1, \ldots,$$

is called a Type I $(d + d_a)$ process, where the # superscript attached to a scalar or vector sequence has the meaning $w_t^{\#} = w_t 1(t > 0)$, $1(\cdot)$ being the indicator function. Similarly:

$$z_{bt} = \Delta^{-d_a - d} u_t^{\#}, \quad t = 0, \pm 1, \ldots,$$

is called a Type II $(d + d_a)$ process. When $d = 0$, $x_t^{\#} = u_t^{\#}$ and hence $z_{at} = z_{bt}$, so both definitions are equivalent in non-fractional contexts.[2] Note that the polynomial on the left-hand side of (10.1) can be expanded as:

$$(1 - L)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j L^j = 1 - dL + \frac{d(d-1)}{2} L^2 - \ldots.$$

Thus, if d is an integer value, x_t will be a function of a finite number of past observations, while if d is real, x_t depends upon values of the time series far in the past. The higher the value of d, the higher the level of association between the observations.

There exist several sources that might produce $I(d)$ processes, aggregation being the usual argument. Robinson (1978) and Granger (1980) showed that fractionally integrated data could arise as a result of aggregation when: (i) data are aggregated across heterogeneous autoregressive (AR) processes, and (ii) data involving heterogeneous dynamic relationships at the individual level are then aggregated to form the time series. Cioczek-Georges and Mandelbrot (1995), Taqqu, Willinger and

Sherman (1997) and Chambers (1998) also use aggregation to motivate long memory processes, while Parke (1999) uses a closely related discrete time error duration model. More recently, Diebold and Inoue (2001) proposed another source of long memory based on regime-switching models.[3]

Note that u_t in (10.1) may also include some type of weak dependence structure: for example, a stationary and invertible autoregressive moving average (ARMA) process of the form:

$$\phi(L)u_t = \theta(L)\varepsilon_t, \quad t = 0, \pm 1, \ldots,$$

where ε_t is an independent and identically distributed (i.i.d.) sequence. Thus, when $d < 1/2$, x_t in (10.1) can be written as:

$$\phi(L)(1-L)^d x_t = \theta(L)\varepsilon_t, \quad t = 0, \pm 1, \ldots, \tag{10.2}$$

which is usually called an autoregressive fractionally integrated moving average (ARFIMA) process. Sowell (1992a) analyzed the exact maximum likelihood (ML) estimator of the parameters of the ARFIMA model (10.2) in the time domain, using a recursive procedure that allows quick evaluation of the likelihood function, which is given by:

$$(2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} X_n' \Sigma^{-1} X_n\right),$$

where $X_n = (x_1, x_2, \ldots, x_n)'$ and $X_n \sim N(0, \Sigma)$. Other parametric methods of estimating d in the frequency domain were proposed, among others, by Fox and Taqqu (1986) and Dahlhaus (1989). Small sample properties of these and other estimators were examined in Smith, Taylor and Yadav (1997) and Hauser (1999). In the former, several semiparametric procedures were compared with Sowell's (1992a) ML estimation method, finding that Sowell's procedure outperforms the semiparametric ones in terms of bias and mean square error. Hauser also compares Sowell's procedure with others based on the exact and the Whittle likelihood function in the time and the frequency domain, and shows that Sowell's procedure dominates the others in the case of fractionally integrated models. A semiparametric frequency domain estimator is the log-periodogram estimator proposed by Geweke and Porter-Hudak (1983).[4] Other parametric and semiparametric methods have been proposed: see, for example, Robinson (1994a, 1995a, 1995b), Tanaka (1999), Velasco (1999a, 1999b), and Phillips and Shimotsu (2004, 2005).[5]

So far we have focused on the case where the singularity occurs at the zero frequency. Let us consider now the following process:

$$(1 - 2\cos w_r L + L^2)^d x_t = u_t, \quad t = 0, \pm 1, \ldots, \tag{10.3}$$

where w_r is a real value equal to $2\pi r/n$, with $r = n/c$. In this context, if $d > 0$, the process is also fractionally integrated, although the pole (unboundedness) in the spectrum now occurs at a (cyclical) frequency $\lambda \neq 0$, and c will be an indicator of the number of periods per cycle. These processes were introduced by Gray, Yhang and Woodward (1989, 1994), who showed that the polynomial in (10.3) can be

expressed in terms of the Gegenbauer polynomial such that, for all $d \neq 0$, and setting $\mu = \cos w_r$:

$$(1 - 2\mu L + L^2)^{-d} = \sum_{j=0}^{\infty} C_{j,d}(\mu) L^j,$$

where:

$$C_{j,d}(\mu) = \sum_{k=0}^{j} \frac{(-1)^k (d)_{j-k} (2\mu)^{j-2k}}{k!(j-2k)!}; \quad (d)_j = \frac{\Gamma(d+j)}{\Gamma(d)},$$

and $\Gamma(x)$ is the Gamma function. Alternatively, we can use the recursive formula $C_{0,d}(\mu) = 1$, $C_{1,d}(\mu) = 2\mu d$, and:

$$C_{j,d}(\mu) = 2\mu \left(\frac{d-1}{j} + 1 \right) C_{j-1,d}(\mu) - \left(2\frac{d-1}{j} + 1 \right) C_{j-2,d}(\mu), \quad j = 2, 3, \ldots.$$

(See, for instance, Magnus, Oberhettinger and Soni, 1966, or Rainville, 1960, for further details on Gegenbauer polynomials.) Gray, Yhang and Woodward (1989) showed that x_t in (10.3) is stationary if $d < 0.5$ for $|\mu = \cos w_r| < 1$ and if $d < 0.25$ for $|\mu| = 1$. Lobato and Robinson (1998) proposed a semiparametric approach for testing this type of model, and Dalla and Hidalgo (2005) suggested a parametric test where the unbounded frequency in the spectrum is assumed to be unknown.[6]

As mentioned above, these processes are characterized by an unbounded spectral density function at a single frequency. There may be cases, however, where the spectrum is unbounded at several frequencies simultaneously, the most typical corresponding to a seasonal process. We can consider a model of the form:

$$(1 - L^s)^d x_t = u_t, \quad t = 0, \pm 1, \ldots, \tag{10.4}$$

where s refers to the number of time periods per year (that is, $s = 4$ with quarterly data, $s = 12$ with monthly). Similarly to (10.1), the (seasonal) fractional polynomial above can be expressed as:

$$(1 - L^s)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j L^{js} = 1 - dL^s + \frac{d(d-1)}{2} L^{2s} - \ldots,$$

for all real d, so that d becomes crucial for describing the degree of seasonal persistence. The notion of fractional Gaussian noise with seasonality was initially suggested by Abrahams and Dempster (1979) and Jonas (1981), and extended to a Bayesian framework by Carlin, Dempster and Jonas (1985) and Carlin and Dempster (1989). Note that if, for example, $s = 4$, the polynomial $(1 - L^4)$ can be decomposed into $(1 - L)(1 + L)(1 + L^2)$, which is thus a case of multiple poles in the spectrum (at zero, π, and $\pi/2$ ($3\pi/2$) of a 2π cycle), all of them, according to (10.4), with the same order of integration d.

10.2.2 Empirical evidence of fractional integration

The empirical literature on fractional integration is large. In the 1960s, Granger (1966) and Adelman (1965) had pointed out that most aggregate economic time

series have a typical shape where by the spectral density increases dramatically as the frequency approaches zero. However, differencing the data frequently leads to overdifferencing at the zero frequency. Some 15 years later, Robinson (1978) and Granger (1980) showed that aggregation could be a source of fractional integration. Since then, fractional processes have been widely employed to describe the dynamics of many time series. Given the vast amount of empirical work, this section is divided into various sub-sections according to the different nature of the series under examination.

10.2.2.1 Applications to macroeconomics

Diebold and Rudebusch (1989) and Sowell (1992b) analyzed US quarterly post-war real output data and obtained estimates of d below unity. Nevertheless, their results were in line with Rudebusch's (1993) and Christiano and Eichembaum's (1990) conclusions in the sense that their confidence intervals for d included the unit root and, in Sowell's case, also the trend-stationary $I(0)$ representation. Haubrich and Lo (1991) examined US real output using R/S techniques and found little evidence of long-range dependence in the business cycle. Using Bayesian techniques, Koop *et al.* (1997) also examined real US GNP and found some evidence of long memory, although their results also reflected model uncertainty. Michelacci and Zaffaroni (2000) showed that GDP per capita in 16 OECD countries exhibited long memory. Mayoral (2006) also finds evidence of fractional integration in real GNP and GNP per capita in the US, showing that their results are robust to the presence of structural breaks in the deterministic components.[7]

Long memory in inflation rates is another topic that has been widely examined in the empirical literature. Much of the evidence supports the view that inflation is fractionally integrated with a differencing parameter that is significantly different from zero or unity. For US monthly data, Backus and Zin (1993) found a fractional degree of integration. They argue that aggregation across agents with heterogeneous beliefs results in long memory in inflation. Hassler (1993) and Delgado and Robinson (1994) provide strong evidence of long memory in the Swiss and Spanish inflation rates respectively. Baillie, Chung and Tieslau (1996) examined monthly post-World War II CPI inflation in ten countries, and found evidence of long memory with mean-reverting (with smaller memory than one) behavior in all countries except Japan. Similar evidence was found in Hassler and Wolters (1995) and Baum, Barkoulas and Caglayan (1999). In the context of structural breaks, Bos, Franses and Ooms (1999, 2001) examined inflation in the G7 countries, finding that long memory is quite resistant to level shifts, although, for a few inflation rates, they found that the evidence for long memory disappeared. Evidence of long memory behavior in the conditional mean of inflation is found in Baillie, Chung and Tieslau (1996) and Baillie, Han and Kwon (2002). Other recent papers relating long memory and structural breaks in inflation rates are Gadea, Sabate and Serrano (2004), Franses, Hyung and Penn (2006) and Gil-Alana (2008a), and forecasting issues are examined in Franses and Ooms (1997) and Barkoulas and Baum (2006).

Other variables, such as consumption and income, have been analyzed from a fractional viewpoint in Diebold and Rudebusch (1991), Haubrich (1993) and Dolado and Marmol (2004). Finally, Crato and Rothman (1994a) and Gil-Alana and Robinson (1997) analyzed updated versions of Nelson and Plosser's (1982) dataset, which has 14 US macroeconomic series, and found evidence of fractional integration in practically all series.

10.2.2.2 Applications to exchange rates

The theory of purchasing power parity (PPP) occupies a central place in international economics, being a key building block in monetary models of exchange rate determination. In a flexible-price monetary model, PPP is assumed to hold continuously. In a sticky-price model, PPP does not hold, but is a maintained assumption for the long run. The question of interest is to determine if deviations from PPP are transitory or permanent. Applying R/S techniques to daily rates for the British pound, French franc and Deutsche mark, Booth, Kaen and Koveos (1982) found positive memory during the flexible exchange rate period (1973–79) but negative memory (that is, anti-persistence) during the fixed exchange rate period (1965–71). Later, Cheung (1993) also found evidence of long memory behavior in foreign exchange markets during the managed floating regime. On the other hand, Baum, Barkoulas and Caglayan (1999) estimated ARFIMA models for real exchange rates in the post-Bretton Woods era and found almost no evidence to support long-run PPP. Additional papers on exchange rate dynamics using fractional integration are Fang, Lai and Lai (1994), Crato and Ray (2000) and Wang (2004). The volatility dynamics in foreign exchange rates with fractional integration has been examined with the FIGARCH-model, introduced by Baillie, Bollerslev and Mikkelsen (1996), and subsequent papers using this approach are Andersen and Bollerslev (1997, 1998), Tse (1998), Baillie, Cecen and Han (2000), Kihc (2004) and Morana and Beltratti (2004).

10.2.2.3 Applications to interest rates

Shea (1991) investigated the consequences of long memory in interest rates for tests of the expectations hypothesis of the term structure. He found that allowing for the possibility of long memory significantly improves the performance of the model, even though the expectations hypothesis cannot be fully resurrected. In related work, Backus and Zin (1993) observed that the volatility of bond yields does not decline exponentially when the maturity of the bond increases; in fact, they noticed that the decline was hyperbolic, consistent with the fractionally integrated specification. Lai (1997) and Phillips (1998) provided evidence, based on semiparametric methods, that *ex ante* and *ex post* US real interest rates are fractionally integrated. Tsay (2000) employs an ARFIMA model to provide evidence that the US real interest rate can be described as an $I(d)$ process. Further evidence can be found in Barkoulas and Baum (1997), Meade and Maier (2003) and Gil-Alana (2004a, 2004b). Couchman, Gounder and Su (2006) estimated ARFIMA models for *ex post* and *ex ante* interest rates from 16 countries. Their results suggest that, for

the majority of countries, the fractional differencing parameter lies between 0 and 1, and it seems to be considerably smaller for *ex post* real rates than for *ex ante* rates.

10.2.2.4 Applications to stock markets

Long memory analysis was first conducted in stock return series in Greene and Fielitz (1977). They report evidence of persistence in daily US stock return series using R/S methods. However, Aydogan and Booth (1988) concluded that there was no significant evidence of long memory in common stock returns. Lo (1991) used the modified R/S method, along with spectral regression methods, and found no evidence of long memory in stock returns. Many other authors have found little or no evidence of long memory in stock markets (see, for example, Hiemstra and Jones, 1997). On the other hand, Crato (1994), Cheung and Lai (1995), Barkoulas and Baum (1996), Barkoulas, Baum, and Travlos (2000), Sadique and Silvapulle (2001), Henry (2002), Tolvi (2003) and Gil-Alana (2006) are among those who find evidence of long memory in monthly, weekly, and daily stock market returns.

Several papers use the Standard & Poor (S&P) 500 index over a long span of daily observations. Granger and Ding (1995a, 1995b) focus on power transformations of the absolute value of the returns (which they use as a proxy for volatility). They estimate a long memory process to study persistence in volatility, and establish some stylized facts concerning the temporal and distributional properties of absolute returns. However, in a related study, Granger and Ding (1996) find that the parameters of the long memory model vary considerably from one sub-series to the next. The issue of fractional integration with structural breaks in stock markets has been examined in Mikosch and Starica (2000) and Granger and Hyung (2004). Stochastic volatility models using fractional integration have been implemented in Crato and de Lima (1994), Bollerslev and Mikkelsen (1996), Ding and Granger (1996), Breidt, Crato, and de Lima (1997, 1998), Arteche (2004) and Baillie *et al.* (2007).

10.2.2.5 Applications to geophysics and other sciences

Fractional integration has also been applied in many other areas. Examples include meteorology (Haslett and Raftery, 1989; Bloomfield, 1992; Hussain and Elbergali, 1999; Gil-Alana, 2005a, 2008b); etternet (and internet) traffic traces (Abry and Veitch, 1998; Karagiannis, Molle and Faloutsos, 2004); hydrology (Montanari, Rosso and Taqqu, 1997, 2000; Rao and Bhattacharya, 1999; Wang *et al.*, 2005; Wang *et al.*, 2007); and political sciences (Box-Steffensmeier and Smith, 1996, 1998; Byers, Davidson and Peel, 1997, 2000; Dolado, Gonzalo and Mayoral, 2003).

10.2.2.6 Applications using seasonal and cyclical FI models

This review has so far focused exclusively on models with the pole or singularity in the spectrum occurring at the zero frequency. In this sub-section we briefly review the empirical literature on seasonal and cyclical fractional models. Starting with the seasonal model in (10.4), Porter-Hudak (1990) applied a seasonally fractionally integrated model to quarterly US monetary aggregates, concluding

that a fractional model could be more appropriate than standard ARIMAs. Advantages of seasonally fractionally integrated models for forecasting are illustrated in Ray (1993) and Sutcliffe (1994), and other empirical applications using quarterly seasonal models can be found in Gil-Alana and Robinson (2001) and Gil-Alana (2005b). Monthly data in the context of seasonal fractional integration have been examined in Gil-Alana (1999) and Ooms and Franses (2001).

Applications using the cyclical model based on the Gegenbauer process described by equation (10.3) can be found, for example, in Arteche and Robinson (2000), Bierens (2001) and Gil-Alana (2001), and empirical work based on multiple cyclical structures (*k*-factor Gegenbauer processes) can be found in Ferrara and Guegan (2001), Sadek and Khotanzad (2004) and Gil-Alana (2007).

10.3 Fractional cointegration

The concept of fractional integration leads naturally to an extension of the standard notion of cointegration (which involves series with integer orders of integration) to the fractional case, where equilibrium relations among fractional processes could be captured. In the present section, we introduce this concept, give an overview of the different estimation methods proposed so far in the literature, and give evidence of the empirical relevance of this idea.

10.3.1 The concept and modelization of fractional cointegration

Engle and Granger (1987) suggested that, if two processes x_t and y_t are both $I(d)$, then it is generally true that, for a certain scalar $a \neq 0$, a linear combination $w_t = y_t - ax_t$ will also be $I(d)$, although it is possible that w_t be $I(d-b)$ with $b > 0$. This idea characterizes the concept of cointegration, which they adapted from Granger (1981) and Granger and Weiss (1983). They provided the following definition for multivariate series. Given two real numbers d, b, the components of the vector z_t are said to be cointegrated of order d, b, denoted $z_t \sim CI(d, b)$, if:

(i) all the components of z_t are $I(d)$,
(ii) there exists a vector $\alpha \neq 0$ such that $w_t = \alpha' z_t \sim I(d-b)$, $b > 0$.

Here, α and w_t are called the cointegrating vector and error respectively. Engle and Granger (1987) offered some intuition behind this crucial concept in modern time series econometrics, suggesting the existence of forces which tend to keep series not too far apart. Given a vector of economic variables z_t, and a certain vector $\alpha \neq 0$, economic theory would say that the variables are in equilibrium if $\alpha' z_t = 0$. This is a very tight notion of equilibrium, and it is a very narrow view that this equality could hold for every time period t. Alternatively, we might think of an equilibrium error as $w_t = \alpha' z_t$, which accommodates deviations from equilibrium. If, for example, in Engle and Granger's (1987) definition $d = b = 1$, what characterizes cointegration as a "long-run equilibrium" relationship is that a linear combination of $I(1)$ processes is $I(0)$, so that the series in z_t cannot drift too far apart.

To be fair, the idea of equilibrium between $I(1)$ processes was hinted at long before in the statistical literature. In the AR model:

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad t > 0; \quad y_t = 0, \quad t \leq 0,$$

ε_t being a sequence of independent normally distributed random variables with mean 0 and finite variance, Dickey and Fuller (1979) studied the properties of the regression estimate of ρ, $\hat{\rho}$, under the assumption that $\rho = 1$. In fact, this represents a situation of cointegration between the $I(1)$ processes y_t and y_{t-1}, as the linear combination $y_t - y_{t-1}$ is $I(0)$. This is a particular case of what Park (1992) called "singular cointegration," which is characterized by cointegrating errors being linear combinations of innovations which generate the regressors.

Engle and Granger (1987) introduced another important concept. If the multivariate $I(d)$ process z_t has $p > 2$ components, there may be several linearly independent cointegrating vectors, representing the case where several equilibrium relations drive the joint movement of the variables in z_t. It is easy to see that the maximum number of linearly independent cointegrating vectors is $r \leq p - 1$, the value r defining the "cointegrating rank" of z_t.

Even considering only integer orders of integration, a more general definition of cointegration than the one given by Engle and Granger (1987) is possible, one that allows for a multivariate process with components having different orders of integration. Denoting d_1 and d_p to be the largest and smallest of these orders, respectively, Johansen (1996) proposed that any vector z_t such that $\alpha' z_t \sim I(d_w)$, with $d_w < d_1$, is a cointegrating vector. Flôres and Szafarz (1996) narrowed Johansen's definition, proposing instead that the vector series is cointegrated if there is a non-trivial linear combination of its components (with at least a non-zero scalar multiplying d_1) which is integrated of order $d_w < d_1$. Alternatively, Robinson and Marinucci (2003) define z_t to be cointegrated if there exists a vector $\alpha \neq 0$ such that $\alpha' z_t \sim I(d_w)$, with $d_w < d_p$, which is a much stronger requirement. Robinson and Yajima (2002) offered an alternative (although rather more involved) definition and a comparison of the different definitions that have appeared in the literature.

Once fractional integration is defined, the concept of fractional cointegration appears as a natural extension of traditional cointegration. In fact, the standard definition of cointegration by Engle and Granger (1987) does not necessarily refer to integer orders of integration. In the simple bivariate case, two series y_t, x_t, sharing the same order of integration, say δ, are cointegrated if there exists a vector $\alpha \neq 0$ such that $\alpha' z_t \sim I(\gamma)$, with $\gamma < \delta$, with $z_t = (y_t, x_t)'$. This prompts consideration of an extension of Phillips' (1991a) triangular system, which, for a very simple bivariate case, is:

$$y_t = \nu x_t + u_{1t}(-\gamma), \tag{10.5}$$

$$x_t = u_{2t}(-\delta), \tag{10.6}$$

for $t = 0, \pm 1, \ldots$, where, for any vector or scalar sequence w_t and any c, we introduce the notation $w_t\,(c) = \Delta^{c} w_t^{\#}$. $u_t = (u_{1t}, u_{2t})'$ is a bivariate zero mean covariance stationary $I(0)$ unobservable process and $\nu \neq 0$, $\gamma < \delta$. The truncation in (10.6)

ensures that x_t has finite variance, and implies that $x_t = 0$, $t \leq 0$. This restriction is unnecessary if $\gamma < 1/2$ because, in that case, $y_t - \nu x_t$ is covariance stationary without it and "asymptotically covariance stationary" with it, but it is imposed for the sake of a uniform treatment, implying that $y_t = 0$, $t \leq 0$. Under (10.5)–(10.6), x_t is $I(\delta)$, as is y_t by construction, while the cointegrating error $y_t - \nu x_t$ is $I(\gamma)$. Model (10.5)–(10.6) reduces to the bivariate version of Phillips' triangular form when $\gamma = 0$ and $\delta = 1$, which is one of the most popular models displaying $CI(1,1)$ cointegration considered in both the empirical and theoretical literatures. This model allows greater flexibility in representing equilibrium relationships between economic variables than the traditional $CI(1,1)$ prescription. On the one hand, it is plausible that there exists long-run co-movements between non-stationary series which are not precisely $I(1)$. On the other hand, there is usually no *a priori* reason for restricting analysis to just $I(0)$ cointegrating errors, as the convergence to equilibrium of any cointegrating relation could be much slower than the adjustment implied by, for example, a finite ARMA cointegrating error. Furthermore, we could also consider cointegration among (asymptotically) stationary variables, with some linear combinations producing cointegrating errors characterized by having weaker memory than that of the observed series. Also, it could be that the cointegrating error is purely non-stationary but mean-reverting, so that a certain long-run equilibrium among non-mean-reverting observables holds.[8]

There are various directions in which model (10.5)–(10.6) has been generalized. Robinson and Iacone (2005), still within a bivariate framework, allow for deterministic components, extending (10.5)–(10.6) to:

$$y_t = \nu x_t + \sum_{j=1}^{p_1} \mu_{1j} t^{\phi_{1j} - 1/2} + u_{1t}(-\gamma), \tag{10.7}$$

$$x_t = \sum_{j=1}^{p_2} \mu_{2j} t^{\phi_{2j} - 1/2} + u_{2t}(-\delta), \tag{10.8}$$

where $\delta > \max(\gamma, 0.5)$, and the ϕ_{ij} are real numbers satisfying $\phi_{11} > \cdots > \phi_{1p_1} > 0$; $\phi_{21} > \cdots > \phi_{2p_2} > 0$, noting that an intercept appears in (10.7)–(10.8) when $\phi_{1j} = \phi_{2j} = 1/2$, while integer powers are also possible.

Kim and Phillips (2002) proposed a multivariate version of (10.5)–(10.6), employing the Type I definition of fractionally integrated processes instead, so that:

$$\tilde{y}_t = \nu \tilde{x}_t + v_{1t}^{(\gamma)}, \quad t \geq 1, \tag{10.9}$$

$$\tilde{x}_t = v_{21}^{(\delta)} + \cdots + v_{2t}^{(\delta)}, \quad t \geq 1, \tag{10.10}$$

where $v_{1t}^{(\gamma)}$ and $v_{2t}^{(\delta)}$ are jointly stationary Type I fractionally integrated processes of orders γ and $\delta - 1$, respectively, with $|\gamma| < 0.5$, $0.5 < \delta < 1.5$, $\tilde{y}_t$ and $\tilde{x}_t$ are $p \times 1$ and $q \times 1$ vectors, respectively, and ν is a $p \times q$ matrix of cointegrating parameters. Note that, when $p = q = 1$ and $\gamma = 0$, $\delta = 1$, $\left(v_{1t}^{(\gamma)}, v_{2t}^{(\delta)}\right)' \equiv \left(u_{1t}, u_{2t}\right)'$ implies that $\left(\tilde{x}_t, \tilde{y}_t\right)' = \left(x_t, y_t\right)'$, but more generally this is not the case. Model (10.9)–(10.10) is

also a particular case of the fractional setting proposed by Jeganathan (1999, 2001). This multivariate modelization is a straightforward generalization of the bivariate setting, given that the time series involved still depend on two integration orders, γ and δ. A richer structure is proposed by Hualde and Robinson (2006), who consider the model:

$$\Psi z_t = \Delta^{-1}(\delta)\, u_t^{\#}, \tag{10.11}$$

where $\delta = \left(\delta_1, \ldots, \delta_p\right)'$, z_t is a $p \times 1$ vector observable process, $\Delta(\delta) = diag\left(\Delta^{\delta_1}, \ldots, \Delta^{\delta_p}\right)$ and u_t is a $p \times 1$ zero mean covariance stationary unobservable process. The presence of the various integration orders in δ poses additional difficulties and, unless δ and Ψ are restricted, (10.11) does not ensure cointegration and identification. Hualde and Robinson (2006) impose restrictions (simplifying matters substantially in an already complicated setting) which ensure identifiability and imply that z_t is $I\left(\delta_p\right)$ (δ_p is assumed to be the largest fractional order among those in δ). The restrictions also imply that there are r cointegrating relations among the elements of z_t, and allow the integration orders of the r cointegrating errors to vary, unlike in previous models of multivariate fractional cointegration. In a different setting, Chen and Hurvich (2003a) model a $q \times 1$ time series, z_t, such that its $(p-1)$th difference (where p is an integer), y_t, is a covariance stationary process with common memory parameter $d \in (-p + 0.5, 0.5)$. They assume that y_t has the common-components representation:

$$y_t = Ax_t + Bu_t, \tag{10.12}$$

where, for $1 \leq r \leq q-1$, A, B are $q \times (q-r)$ and $q \times r$ unknown deterministic matrices of ranks $q-r$ and r, respectively, and x_t, u_t, are $(q-r) \times 1$ and $r \times 1$, unknown unobserved processes with memories d and d_u, respectively, where $d_u \in (-p + 0.5,\ 0.5) < d$. Basically, these conditions imply that y_t is cointegrated, with cointegrating space identified by the null space of A' ($Ker(A')$), noting that, if $\alpha \in Ker(A')$, then $\alpha' y_t = \alpha' B u_t$, which has at most memory d_u. Chen and Hurvich (2006) further enrich this setting by means of a common components model in which the components have different memory parameters, while still allowing the $q \times 1$ vector of observed series to have just one common memory parameter. Thus, using notation similar to Chen and Hurvich (2003a), they set:

$$y_t = A_0 u_t^{(0)} + A_1 u_t^{(1)} + \cdots + A_s u_t^{(s)}, \tag{10.13}$$

where, for $k = 1, \ldots, s$, the A_k are $q \times a_k$ unknown deterministic matrices with $a_0 = q - r$, $\sum_{j=1}^{s} a_j = r$, the $u_t^{(k)}$ are $a_k \times 1$ unobservable processes with memory d_k, such that $-p + 0.5 < d_s < \cdots < d_0 < 0.5$, and all rows of A_0 are non-zero. This setting ensures that all the components in y_t have common memory d_0, the cointegrating rank is r, such that $1 \leq r < q$, and there are s cointegrating subspaces, with $1 \leq s \leq r$. In this framework, Chen and Hurvich (2006) define s different cointegrating sub-spaces B_k, $k = 1, \ldots, s$, with the main characteristic being that, if $\beta \in B_k$, $\beta' A_l = 0, l = 0, \ldots, k-1$, and $\beta' A_k \neq 0$.

Finally, Johansen (2008) proposes a new vector autoregressive model:

$$\Delta^d z_t = \alpha\beta' \Delta^{d-b} L_b z_t + \sum_{i=1}^{k} \Gamma_i \Delta^d L_b^i z_t + \varepsilon_t, \tag{10.14}$$

where $L_b = 1 - \Delta^b$ is a particularly useful lag operator with $0 < b \leq d$. Here, if there is a unit root in an associated characteristic polynomial, (10.14) generates a fractional process z_t of order d, for which the $r \times 1$ vector $\beta' z_t$ is fractional of order $d - b$.

10.3.2 Estimation methods for fractional cointegration

Depending on the cointegrating model considered, different estimation techniques have been proposed. For simplicity, we deal initially with the estimation of ν in the bivariate system (10.5)–(10.6), although the estimation methods below can be straightforwardly generalized to cover multivariate situations where the observables and cointegrating errors still depend on two integration orders, δ and γ, respectively.

The most obvious proposal is to estimate ν in (10.5)–(10.6) by the ordinary least squares (OLS) estimator:

$$\hat{\nu}_{\text{OLS}} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}. \tag{10.15}$$

Here, in the standard cointegrating setting, with $\gamma = 0$ and $\delta = 1$, it has been shown (see, for example, Phillips and Durlauf, 1986) that $\hat{\nu}_{\text{OLS}}$ is n-consistent with non-standard asymptotic distribution, in general. In fractional settings, the properties of OLS could be very different from those in this standard framework. For example, Robinson (1994b) showed the inconsistency of $\hat{\nu}_{\text{OLS}}$ when $\delta < 0.5$, which has been termed stationary cointegration (with special importance in finance, see section 10.3.3).[9] When the observables are purely non-stationary (so that $\delta \geq 0.5$), consistency of $\hat{\nu}_{\text{OLS}}$ is retained, but its rate of convergence and asymptotic distribution depends crucially on γ and δ. In particular, Robinson and Marinucci (2001) showed, for a model slightly more general than (10.5)–(10.6), that if $\delta \geq 0.5$, $\gamma \geq 0$, the rate of convergence of OLS is $n^{\min(2\delta-1,\delta-\gamma)}$, except when $\gamma > 0$ and $\gamma + \delta = 1$, where OLS is $n^{\delta-\gamma}/\log n$-consistent. In all these cases OLS has non-standard limiting distributions in general. An alternative method of estimating ν was developed from the following observation. Equation (10.15) is obviously a time-domain representation of the estimate, but it can easily be shown that:

$$\hat{\nu}_{\text{OLS}} = \frac{\sum_{j=0}^{n-1} I_{xy}\left(\lambda_j\right)}{\sum_{j=0}^{n-1} I_x\left(\lambda_j\right)}, \tag{10.16}$$

where $\lambda_j = 2\pi j/n$, $j = 1, \ldots, n$, are the Fourier frequencies, and for arbitrary sequences ξ_t, ζ_t, (possibly the same as ξ_t), we define the discrete Fourier transform

and (cross-) periodogram:

$$w_\xi(\lambda) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} \xi_t e^{it\lambda}, \quad I_{\xi\zeta}(\lambda) = w_\xi(\lambda) w'_\zeta(-\lambda), \quad I_\xi(\lambda) = I_{\xi\xi}(\lambda).$$

Here, the discrete Fourier transform at a given frequency captures the components of the series related to this particular frequency. Thus, noting that cointegration is a long-run phenomenon, when estimating ν one could concentrate just on low frequencies, which are precisely those representing the long-run components of the series, hence neglecting information from high frequencies, associated with the short run, which could have a distorting effect on estimation. Robinson (1994b) proposed the narrow band least squares (NBLS) estimator, which is related to the band estimator proposed by Hannan (1963), and is given by:

$$\hat{\nu}_{\text{NBLS}} = \frac{\sum_{j=0}^{m} s_j \text{Re} I_{xy}\left(\lambda_j\right)}{\sum_{j=0}^{m} s_j I_x\left(\lambda_j\right)}, \tag{10.17}$$

where $1 \leq m \leq n/2$, $s_j = 1$ for $j = 0, n/2$, $s_j = 2$ otherwise, and $(1/m) + (m/n) \to 0$ as $n \to \infty$. Robinson (1994b) showed the consistency of this estimator under stationary cointegration, using the fact that focusing on a degenerating band of low frequencies reduces the bias due to the contemporaneous correlation between u_{1t} and u_{2t}, which was precisely the reason why OLS was inconsistent. For the case $\delta < 0.5$, $\gamma \geq 0$, Robinson and Marinucci (2003) conjectured the rate of convergence to be $(n/m)^{\delta-\gamma}$ and, later, in a similar framework, Christensen and Nielsen (2006) showed that the better rate of convergence $m^{1/2}(n/m)^{\delta-\gamma}$ (and, in fact, asymptotic normality) was achievable if the coherency at frequency zero between u_{1t} and u_{2t} was zero, a restriction that is not satisfied in general by standard weak dependent processes (like, for example, ARMA processes).

In the non-stationary setting, Robinson and Marinucci (2001) showed that, if $\delta + \gamma < 1$ or $\delta + \gamma = 1$ with $\gamma > 0$, the rates of convergence previously given for OLS can be improved, being now $m^{\delta+\gamma-1} n^{\delta-\gamma}$ if $\delta + \gamma < 1$ and $n^{\delta-\gamma}/\log m$ if $\delta + \gamma = 1$ with $\gamma > 0$. As with OLS, NBLS has a non-standard limiting distribution in general.

With the aim of obtaining estimates of ν having improved asymptotic properties (optimal rate of convergence, median unbiasedness, asymptotic mixed-normality leading to standard inference procedures), more developed techniques to estimate ν have been proposed in the fractional setting. These are related to the work of Johansen (1988, 1991), Phillips and Hansen (1990), Phillips (1991a, 1991b), Phillips and Loretan (1991), Saikkonen (1991), Park (1992) and Stock and Watson (1993), who all proposed estimators with optimal asymptotic properties (under Gaussianity) in the standard cointegrating setting with $\gamma = 0$, $\delta = 1$. However, in all these estimators knowledge of γ, δ, was assumed (usually after pretesting), and in fractional circumstances this is hard to justify. Dolado and Marmol (1996) proposed an extension to the fractional setting of the fully modified (FM)-OLS

estimator of Phillips and Hansen (1990), assuming knowledge of γ and δ. Kim and Phillips (2002) considered an alternative extension to FM-OLS, and analyzed its relationship to Gaussian ML estimation in (10.9)–(10.10), assuming parametric autocorrelation in $\left(v_{1t}^{(\gamma)}, v_{2t}^{(\delta)}\right)'$. Jeganathan (1999, 2001) considered ML estimation of (10.9)–(10.10) assuming knowledge of the distribution of the innovations and also of γ and δ, although he did include some discussion of their estimation (as also did Kim and Phillips, 2002).

In model (10.5)–(10.6), and assuming the bivariate process u_t has a parametric spectral density $f(\lambda) = f(\lambda;\theta)$, where θ is an unknown vector of short memory parameters, Robinson and Hualde (2003), based on generalized least squares (GLS) type corrections, propose time and frequency domain methods to estimate optimally (under Gaussianity) ν when $\delta - \gamma > 0.5$ (denoted strong cointegration). For simplicity, we just present the frequency domain approach, which is asymptotically equivalent (to first-order properties) to that of the time domain, and which will be applied to empirical data in section 10.4. Denoting:

$$z_t(c,d) = (y_t(c), x_t(d))', \quad \zeta = (1,0)', \quad p(\lambda;h) = \zeta' f(\lambda;h)^{-1},$$

$$a(c,d,h) = \sum_{j=1}^{n} p(\lambda_j;h) w_{x(c)}(-\lambda_j) w_{z(c,d)}(\lambda_j), \quad q(\lambda;h) = \zeta' f(\lambda;h)^{-1} \zeta,$$

$$b(c,d) = \sum_{j=1}^{n} q(\lambda_j;h) I_{x(c)}(\lambda_j),$$

and defining:

$$\hat{\nu}(c,d,h) = \frac{a(c,d,h)}{b(c,h)},$$

they considered five different estimators, given by:

$$\hat{\nu}(\gamma,\delta,\theta), \hat{\nu}(\gamma,\delta,\hat{\theta}), \hat{\nu}(\gamma,\hat{\delta},\hat{\theta}), \hat{\nu}(\hat{\gamma},\delta,\hat{\theta}), \hat{\nu}(\hat{\gamma},\hat{\delta},\hat{\theta}), \tag{10.18}$$

where $\hat{\gamma}, \hat{\delta}, \hat{\theta}$ are corresponding estimators of the nuisance parameters γ, δ, θ. The estimators in (10.18) reflect different knowledge about the structure of the model, the first being in general infeasible, the second assuming just knowledge of the integration orders (as was done previously in the standard cointegrating literature), whereas the last estimator represents the most realistic situation. Under regularity conditions,[10] Robinson and Hualde (2003) showed that any of the estimators in (10.18) is $n^{\delta-\gamma}$-consistent with identical mixed-Gaussian asymptotic distributions, leading to Wald tests on the parameter ν:

$$W(\gamma,\delta,\theta), W(\gamma,\delta,\hat{\theta}), W(\gamma,\hat{\delta},\hat{\theta}), W(\hat{\gamma},\delta,\hat{\theta}), W(\hat{\gamma},\hat{\delta},\hat{\theta}), \tag{10.19}$$

where $W(c,d,h) = b(c,h)\{\hat{\nu}(c,d,h) - 1\}^2$, with a chi-squared limit.

Hualde and Robinson (2007) propose an estimator of ν in (10.5)–(10.6) under the more adverse situation $\delta - \gamma < 0.5$ (denoted weak cointegration). Assuming u_t

is generated by a vector autoregressive (VAR) process:

$$u_t = \sum_{j=1}^{p} B_j u_{t-j} + \varepsilon_t,$$

and defining:

$$Z_t(c,d) = \left(x_t(c), x_t(d), w'_{t-1}(c,d), \ldots, w'_{t-p}(c,d)\right)',$$

where $w_t(c,d) = \left(x_t(c), x_t(d), y_t(c)\right)'$, they analyzed the behavior of the estimators $\hat{\nu}(\gamma,\delta), \hat{\nu}(\hat{\gamma},\hat{\delta})$, where $\hat{\gamma},\hat{\delta}$ are corresponding estimators of γ,δ, and $\hat{\nu}(c,d) = i_1' G(c,d)^{-1} g(c,d)$. Here, $i_1 = (1,0,\ldots,0)'$:

$$G(c,d) = Q\frac{1}{n}\sum_{t=p+1}^{n} Z_t(c,d) Z_t'(c,d) Q', \quad g(c,d) = Q\frac{1}{n}\sum_{t=p+1}^{n} Z_t(c,d) y_t(c),$$

where Q caters for the possibility of prior zero restrictions on the B_j's. As in Robinson and Hualde (2003), this method is based on a GLS-type of correction. Hualde and Robinson (2007) showed that both estimators are $\sqrt{n}$-consistent and asymptotically normal, but with different asymptotic variance, which in the case of $\hat{\nu}(\hat{\gamma},\hat{\delta})$ depends on how γ, δ, are estimated. Here, it is important to note that, in order to get a $\sqrt{n}$-consistent $\hat{\nu}(\hat{\gamma},\hat{\delta})$, it is essential that γ, δ, are $\sqrt{n}$-consistently estimated, and Hualde and Robinson (2007) proposed a feasible method of estimation for the case where B_j is upper-triangular for all $j = 1,\ldots,p$.

Robinson and Iacone (2005) consider the data-generating process (10.7)–(10.8), and deal with estimation of ν and of those elements in μ_{1j}, $j = 1,\ldots,p_1$, μ_{2j}, $j = 1,\ldots,p_2$, whose associated deterministic trends are not dominated (in the precise way defined in Robinson and Iacone, 2005) by stochastic components. They consider three different estimators of the parameters of interest. First, they analyze the properties of (10.16), concluding that when the deterministic term in (10.7) dominates the stochastic and deterministic components in x_t, $\hat{\nu}_{OLS}$ is not even consistent. Otherwise, consistency is retained, and the authors offer a very detailed analysis of the different possibilities involved. The second scenario refers to estimating by OLS ν and μ_1 in (10.7), where μ_1 collects the μ_{1j}'s associated with the deterministic terms not dominated by $u_{1t}(-\gamma)$ (which are the only ones which could be consistently estimated). Here, due to the accounting of the deterministic components, the estimate of ν is always consistent, with asymptotic properties very dependent on $(\gamma,\delta,\phi_{2\forall})$, where $\phi_{2\forall}$ is the maximum value of those ϕ_{2j}'s for which $\mu_{2j} \neq 0$. Finally, they also consider GLS estimation, taking into account the deterministic terms in (10.7)–(10.8), thus extending the time and frequency domain estimators of Robinson and Hualde (2003), also for $\delta - \gamma > 0.5$. Under identical regularity conditions, when $\delta > \phi_{2\forall}$, the estimate of ν has identical asymptotic properties to that of Robinson and Hualde (2003); when $\delta = \phi_{2\forall}$, the rate of convergence remains the same, but the limiting distribution changes, whereas if $\delta < \phi_{2\forall}$,

a higher rate of convergence is achieved and the limiting distribution is normal, which is natural given that now the deterministic component in x_t is dominant.[11]

In the multivariate setting (10.11), by considering the spectral density of u_t to be a nonparametric function, Hualde and Robinson (2006) propose an extension of the estimators of Robinson and Hualde (2003), allowing for the simultaneous presence of strong and weak cointegrating relations. Because of the generality of the framework, the representation of the estimators and asymptotic results are rather involved but, essentially, the same properties as in Robinson and Hualde (2003) are achieved by the estimators of the cointegrating parameters in strong cointegrating relations, whereas the estimators of parameters in weak cointegrating relations are asymptotically normal, but with a slower rate of convergence than the parametric one ($\sqrt{n}$) given by $m^{1/2}(n/m)^{\delta_p-\gamma}$, where δ_p, γ denote the common integration order of the observables and that of the cointegrating error of the particular weak cointegrating relation, respectively. This result leads to Wald statistics for testing linear restrictions among the elements of Ψ in (10.11) having a standard null chi-squared limit distribution, irrespective of the type of cointegrating relations present in the model.

In the different setting of Chen and Hurvich (2003a), whose focus is on estimating the space of cointegration and not cointegrating regressions, then, on assuming the cointegrating rank is r, this space is estimated by the eigenvectors corresponding to the r smallest eigenvalues of the averaged tapered periodogram matrix of y_t. The intuition behind this result is the following. Noting (10.12):

$$\sum_{j=1}^{m} \mathrm{Re}\, I_y\left(\lambda_j\right) = A \sum_{j=1}^{m} \mathrm{Re}\, I_x\left(\lambda_j\right) A' + A \sum_{j=1}^{m} \mathrm{Re}\, I_{xu}\left(\lambda_j\right) B' + B \sum_{j=1}^{m} \mathrm{Re}\, I_{ux}\left(\lambda_j\right) A' + B \sum_{j=1}^{m} \mathrm{Re}\, I_u\left(\lambda_j\right) B'.$$

Given that $d > d_u$, the right-hand side of the above equation is dominated by the first term, and setting conditions ensuring that $\sum_{j=1}^{m} \mathrm{Re} I_x(\lambda_j)$ is positive definite with probability approaching one (basically meaning that there is no cointegration among the elements of x_t), then, with probability approaching one, the cointegrating space ($Ker(A')$) is the space of eigenvectors of $A \sum_{j=1}^{m} \mathrm{Re} I_x(\lambda_j) A'$, with corresponding eigenvalues equal to zero. Finally, Chen and Hurvich (2006), in the more general setting (10.13), estimate separately each cointegrating sub-space using appropriate sets of eigenvectors of an averaged periodogram matrix of tapered observations.

10.3.3 Evidence of fractional cointegration

Since the early 1990s fractional cointegration has attracted the attention of many empirical researchers working in different fields. We detail below some of the most relevant applications.

10.3.3.1 Applications to exchange rates

The initial applications of the concept of fractional cointegration were devoted to analyzing PPP. This refers to the tendency for nominal exchange rates and prices to adjust in such a way that the real exchange rate reverts (perhaps slowly) to its parity value. Thus the (log) real exchange rate could be viewed as the cointegrating error in a linear combination of (log) nominal exchange rates and (log) prices with cointegrating vector $(1, -1)'$. Here the literature mainly provided evidence of weak fractional cointegration ($\delta - \gamma < 0.5$), with (approximately) unit root observables with non-stationary but mean-reverting cointegrating errors, as presented by Diebold, Husted and Rush (1991). Although the authors approximated the log of the real exchange rate in a particular way, and not as the difference of the logs of the nominal exchange rate and prices, their empirical analysis, taking into account that the (log) nominal exchange rate is $I(1)$ (see, for example, Baillie and Bollerslev, 1994a, 1994b), reported some cases where the estimated memory of the real exchange rates (for example, France–Germany, Germany–UK) were non-stationary mean-reverting, while for others there was evidence of stationary long memory. In a similar framework, Cheung and Lai (1993) proposed checking the PPP hypothesis via a regression of a foreign price index, converted to domestic (US) currency units, on a domestic price index, with the errors of this relation capturing deviations from PPP. While they provided evidence of the unit root character of the observables, they stated that PPP will be characterized by stationarity, or at least mean reversion, in the cointegrating error. They estimated the degree of memory of the cointegrating error for different countries and bandwidths and, in 11 out of 15 cases, these estimates were suggestive of $\delta - \gamma < 0.5$ instead of $\delta - \gamma > 0.5$.

In a similar setting to Diebold, Husted and Rush (1991), Crato and Rothman (1994b) provided estimates of the (log) real bilateral sterling exchange rates for different countries. Of the nine countries analyzed, for only two of them (Netherlands and Italy) was there very clear evidence that $\delta - \gamma < 0.5$, whereas for another two (Canada and Sweden) there were conflicting results.

Chou and Shih (1997) investigated long-run PPP in the relationship of four Asian countries (Hong Kong, Singapore, South Korea and Taiwan) with the US dollar in an unrestricted trivariate model (involving the logs of the nominal exchange rate and domestic and foreign price levels), allowing for a linear trend and using quarterly data from 1965 to 1992. Using the Geweke and Porter-Hudak (1983) (GPH) test for fractional cointegration (that is, estimating the memory of the cointegrating error by the GPH estimate applied to the cointegrating regression residuals and testing whether this memory is significantly smaller than the integration order of the observables, assumed to be one in this case), they found some evidence of fractional cointegration, with $\delta - \gamma > 0.5$ for the South Korea–US relationship.

Choudhry (1999a) investigated PPP between the US and four high-inflation Eastern European countries (Poland, Romania, Russia and Slovenia), using monthly data from 1991 to 1997. The author estimated a regression model of the log of the nominal exchange rate on the log ratio of domestic to foreign prices, the absolute

version of PPP being characterized by a slope equal to 1, while evidence of fractional cointegration implies a weaker (relative) version of PPP. Using the GPH test, Choudhry provided evidence of relative PPP for Russia and Slovenia (with unit root observables and estimates of the integration order of the cointegrating error close to 0 and 0.5, respectively), but failed to find evidence for absolute PPP.

In a different setting, Baillie and Bollerslev (1994a) argued whether seven spot exchange rates appear to be tied together in the long run or not, taking into account that there does not seem to be much discussion in the literature about the unit root character of these series, which makes much more fragile the idea that they are cointegrated (see, for example, Sephton and Larsen, 1991; Diebold, Gardeazabal and Yilmaz, 1994). Baillie and Bollerslev's (1994a) explanation was that unit root tests, which serve traditionally to detect the presence of unit roots, have very low power against fractional alternatives, so that a situation of fractional cointegration with long memory cointegrating error could be hidden. In fact, their estimate of the memory of the cointegrating error was 0.89, over five standard errors away from 1, thus providing evidence of fractional cointegration with $\delta - \gamma < 0.5$. Similarly, Pan and Liu (1999), using the same nominal exchange rate data as Baillie and Bollerslev (1994a), analyzed the presence of cointegration in different sub-samples by means of the GPH test. Interestingly, they only found evidence of fractional cointegration (with $\delta - \gamma < 0.5$) for the 1980–84 sample, whereas standard cointegration was found for the most recent period 1985–92, supporting the conjecture that the fractional cointegration feature could vary across different time spans.[12]

Another important topic in the literature of exchange rates is the analysis of the forward premium, $f_t - s_t$, where s_t and f_t are logs of the spot exchange rate and of the forward rate respectively. Here, noting the "overwhelming" evidence of unit roots in spot exchange rates, the difference $f_t - s_t$ could be considered as a cointegrating error with cointegrating vector $(1, -1)'$. Baillie and Bollerslev (1994b) claimed that unit root tests generally reject that the forward premium is $I(0)$, which is paradoxical as, given that the forward premium is associated with risk, it seems hard to see any theoretical reason for an $I(1)$ risk premium. The purpose of their paper was to show that the forward premium is indeed mean-reverting, the estimates of the memory of the forward premium for Canada, Germany and UK (with respect to US) being 0.45, 0.77 and 0.55 respectively, suggesting stronger evidence in favor of weak cointegration relations.

Choudhry (1999b) analyzed, by means of the GPH test, nine forward premiums (with respect to US), showing evidence of (weak) cointegration for three of them (Canada, Hong Kong and Italy). He also tested the unbiased forward rate hypothesis (in short, that the forward exchange rate is an unbiased predictor of the corresponding future spot rate), which was examined by analyzing the existence of fractional cointegration in a regression of s_{t+k} on f_t (although two other alternative specifications were also considered), and testing for a unit slope, which ensures that the forward rate is an unbiased predictor of the future spot rate. Evidence of cointegration was found but not of the unbiasedness hypothesis (with the exception of South Africa).

Still concerning the forward premium, much effort has also been devoted to explanations of the so-called "forward premium anomaly." This refers to surprising negative estimates from regressions of the change in the log of the spot exchange rate on the forward premium, where theory predicts a value of 1 for that slope (see, for example, Bekaert, 1996; Bekaert, Hodrick and Marshall, 1997). Baillie and Bollerslev (2000) consider this issue to be a statistical problem caused by the different integration orders of the dependent variable and regressor in the equation mentioned above. They indicate that the spot exchange rate is approximately a unit root, whereas there seems to be evidence in favor of a mean-reverting (but non-stationary) forward premium. Under these circumstances, Maynard and Phillips (2001) showed that the slope coefficient of such a regression (with a short memory dependent variable and a non-stationary but mean-reverting regressor) converges in probability to zero, the long left tail of the asymptotic distribution of this estimator giving further support to the puzzling negative values obtained in the literature. Baillie, Han and Koul (2002) provided further evidence regarding the imbalance of the forward premium regression equation with high-frequency data, so that the forward premium anomaly seems to be an intrinsic property of exchange rates, and they conjectured that the phenomenon is not due to regime shifts or structural breaks.

Finally, there is a more recent literature trying to explain, by means of fractional cointegration, the connection between exchange rates and fundamentals. Caporale and Gil-Alana (2004a) examined the issue of whether real exchange rates were cointegrated with real interest rates and labor productivity differentials in the DM–US$, Yen–US$ relations using quarterly data (1975–98). They provided evidence of unit roots in the observables and also, by estimating parametrically the memory of the cointegrating error from cointegrating residuals, conjectured the existence of fractional cointegration, with the estimated integration order of the residuals fluctuating between 0 and 0.5 in the case of Germany and between 0.1 and 0.6 in the case of Japan.

Dufrénot *et al.* (2006) explored the real exchange rate misalignments of five European countries during the period 1979–99. They posited an equilibrium relation between real exchange rates and macroeconomic fundamentals (terms of trade, prices, foreign assets, fiscal wedge, interest rate differential), with each variable representing the value in a particular country related to a weighted average of the same variable for other countries. They estimated the memory of the cointegrating residuals by means of the modified R/S statistic of Lo (1991), the GPH estimator and the exact ML estimator of Sowell (1992a). In view of their results, there seems to be strong evidence of fractional cointegration for the Netherlands, with mixed evidence for France and the UK.

10.3.3.2 Applications to financial series

Within this area of research, the main focus has been the study of the volatility of financial series, providing evidence of long memory covariance stationary observables with weakly dependent cointegrating errors. The first explicit reference

to this type of cointegration, denoted stationary cointegration, appears in Robinson (1994b). Robinson and Marinucci (2003) investigated further this situation, indicating that the phenomenon of cointegration between stationary variables had recently emerged in finance, and emphasizing the difficulty of distinguishing between a unit root process and a stationary long memory process with an autoregressive part having a root near the unit circle.

Andersen *et al.* (2001) examined "realized" daily equity return volatilities and correlations obtained from high-frequency transaction prices on individual stocks in the Dow Jones Industrial Average. They provided evidence of long memory for certain time series of logarithmic standard deviations and correlations, and stressed the evidence of comovements in volatility across assets. Christensen and Nielsen (2006) make a similar point and argue for the existence of stationary cointegration between the volatility implied in option prices and the subsequent realized return volatility of the underlying asset, since, in their view, the observables (log-volatilities) were integrated of order between 0.35 and 0.40, whereas the cointegrating error seemed weakly dependent. By using an NBLS estimator, they obtained a much higher value for the estimate of the slope of their cointegration relation than that obtained in similar work by Christensen and Prabhala (1998), who used OLS, which, as shown by Robinson (1994b), is inconsistent in the case of stationary cointegration.

Brunetti and Gilbert (2000) proposed a bivariate cointegrated fractional volatility (FIGARCH) model, and applied it to the volatility (measured in terms of squared and absolute returns) of the New York NYMEX and London IPE crude oil markets. They concluded that both processes were highly persistent, with a common degree of fractional integration (around 0.4), and were fractionally cointegrated. Using similar series, Robinson and Yajima (2002) analyzed stationary cointegration in the context of testing for cointegration rank, finding support for this type of behavior in spot closing prices of crude oil.

Beltratti and Morana (2006) also provided evidence of fractional cointegration in stock market volatility. They analyzed the relationship between S&P 500 returns volatility and that of some macroeconomic variables over 1970–2001 using monthly data and allowing for both long memory and structural breaks. They found evidence of long memory and structural change in the volatility, the break possibly being related to break processes in the volatility of the macroeconomic factors, and carried out a fractional cointegration analysis on break-free processes. They found a common memory parameter of 0.25 for the series and concluded in favor of the existence of three cointegrating relations among the variables using the cointegrating test of Robinson and Yajima (2002).

Finally, Caporale and Gil-Alana (2004b) tested the present value model by checking for cointegration between stock prices and dividends using annual data for the period 1871–1995 (updating the series employed by Campbell and Shiller, 1987). They provided evidence of the unit root nature of these series and, by applying Robinson's (1994a) test to OLS residuals, concluded that the series were fractionally cointegrated with long memory cointegrating error and mixed evidence about the type of cointegrating relation (weak or strong) which characterizes the data.

10.3.3.3 Applications to interest rates

This has not been a very popular application of fractional cointegration, but we can mention at least two relevant works. First, Dueker and Startz (1998) proposed an ARFIMA model and discussed its estimation, the main feature being to provide joint estimates of the common integration order of the observables and cointegrating error. They illustrated their method by analyzing the relation between US and Canadian bond rates (using monthly data for 1987–97). The authors suggested that it is not desirable to rely on an assumed value for the order of integration of the observables (usually one), as traditionally was done in previous empirical analyses related to fractional cointegration. They provided evidence of fractional cointegration, their estimates of the memory of the observables and the cointegrating error being 0.674 and 0.200, respectively, which could be evidence in favor of a weak cointegration relation.

In a different setting, Barkoulas, Baum and Oguz (1996) analyzed cointegration among long-term interest rates from five countries (the US, Canada, Germany, the UK and Japan) using monthly data for the period 1967–90. They justified the unit root condition of the series and examined the possibility of cointegration by means of the GPH test applied to different sub-systems of observables, concluding in favor of the existence of strong co-movements between Canadian and US interest rates.

10.3.3.4 Applications to electricity prices

This is probably the most recent (and one of the most promising) application of fractional integration and cointegration techniques. Haldrup and Nielsen (2006) mention the possibility of cointegration in a regime-switching model which allows for fractional cointegration in each of the regime states. They analyzed hourly spot electricity prices (January 2000–October 2003) for the Nord Pool area (Mid Norway, South Norway, West Denmark, East Denmark, Sweden and Finland). Two different regimes are allowed, congestion (where prices differ across areas) and non-congestion (where prices are identical for every area). Two main conclusions can be drawn from their results. First, the memory properties of the individual series seem to differ substantially across regimes (although, in all cases, series appear to be stationary). Second, the use of a non-switching model could lead to wrong conclusions regarding the cointegration of the series (which are analyzed in pairs), which could be driven by the extreme type of cointegration which characterizes the data when the series are in a non-congestion state.

10.3.3.5 Applications to political studies

There are different political issues which have been analyzed in recent times by fractional integration and cointegration techniques. As Robinson (1978) and Granger (1980) demonstrated, fractional integration could originate from aggregation of data which exhibits heterogeneous dynamic behavior at the individual level. This has an important appeal for political data, where series are obtained by aggregating the opinions of possibly very heterogeneous individuals. The first topic

where fractional cointegration has attracted attention is the analysis of the linkage between political opinions and economic indicators. Box-Steffensmeier and Tomlinson (2000) analyzed the relation between congressional approval and economic expectations in the US, using quarterly data from 1974 to 1993. They estimated parametrically the integration orders of both series (0.72 and 0.86 respectively), computed by OLS the relationship between them, and estimated parametrically the memory of the cointegrating error (0.40). They considered their results to be inconclusive about the existence of cointegration because of large standard errors.

Similarly, Davidson (2003), using quarterly data from 1955 to 1996, examined the relation between political opinion (measured as the end-of-quarter difference between support for the governing party and that for the opposition) and economic indicators in the UK by means of bootstrap methods. The author provides support for the non-stationary, but mean-reverting, nature of the political opinion variable, concluding that there is little or no evidence of linkages between political and economic cycles.

Another interesting issue is the relationship between governing party and prime ministerial support. In principle, it seems plausible that both variables are cointegrated, and different studies have provided support for this conjecture. Clarke and Lebo (2003), using monthly data from 1979 to 1996, linked governing party support, prime ministerial approval and four subjective economic evaluations. Using different methods, the authors concluded that the series are non-stationary but mean-reverting, suggesting the possibility of cointegration between governing party support and prime ministerial approval. Additionally, they found that, whereas personal economic evaluations were influential, national ones are not significant. Similarly, Davidson, Peel and Byers (2006) proposed two variants of a fractionally integrated vector error correction model and applied them to the relationship between the respective performances of prime minister and government in the UK. Evidence of cointegration was provided.

Finally, a different issue was addressed by Lebo and Moore (2003), who analyzed an action–reaction model of foreign policy behavior for different pairs of countries, including Egypt–Israel and US–Russia. They provided strong evidence that those foreign policy series are fractionally integrated (mainly stationary and in all cases mean-reverting), and suggested the possibility of cointegration in the Egypt–Israel relation during the period 1948–76.

10.4 The empirical investigation: the PPP hypothesis

Numerous empirical studies have cast significant doubt on the PPP hypothesis with respect to the short run, but have yielded mixed evidence with respect to the long run (see, for example, Corbae and Ouliaris, 1988; Enders, 1988; Kim, 1990; Taylor, 1988). As mentioned earlier, Cheung and Lai (1993) proposed a fractional version of the PPP specification, essentially (10.5)–(10.6) with x_t representing the domestic price index and y_t the foreign price index, converted to domestic currency units. The coefficient ν in (10.5) is unity according to the absolute or homogeneous

version of PPP, so this is testable by the Wald statistic of (10.19). Using unit root tests, Cheung and Lai (1993) failed to reject the hypothesis of $\delta = 1$ and then, using differenced OLS residuals, computed semiparametric log-periodogram estimates of $\gamma - 1$ and tested the non-cointegration null hypothesis of $\delta - \gamma = 0$ against the alternative $\delta - \gamma > 0$, using critical values computed by simulation. They found evidence of cointegration in a number of bivariate series, but did not test $\nu = 1$. We employ a step-by-step approach, first testing whether the integration orders, δ_x and δ_y, of x_t and y_t are the same, then testing for the presence of cointegration, then testing for $\delta - \gamma > 0.5$ and, finally, given that all these hurdles have been crossed, testing $\nu = 1$. In the first three steps we use semiparametric procedures (as did Cheung and Lai, 1993; Marinucci and Robinson, 2001), while in the final step we identify parametric models for the autocorrelation in u_t and hence compute estimates of ν and Wald statistics.

The semiparametric estimates of integration orders were all Robinson's (1995b) versions of log-periodogram estimates, but without trimming, using first differences and then adding back 1. We estimated δ_x and δ_y separately, and then tested $\delta_x = \delta_y (= \delta)$ by an adaptation of Robinson and Yajima's (2002) statistic $\hat{T}_{ab}$ to log-periodogram estimation, with their trimming sequence $h(n)$ chosen as $b^{-1/(5+2i)}$ for $i = 1, \ldots, 4$, with b the bandwidth used in the estimation. Given $\delta_x = \delta_y$ is not rejected, we performed the Hausman test for no-cointegration of Marinucci and Robinson (2001), comparing the estimate $\tilde{\delta}_x$ of δ_x with the more efficient bivariate one of Robinson (1995b), which uses the information that $\delta_x = \delta_y$. Given cointegration is not rejected, the null $\delta - \gamma = 0.5$ was rejected in favor of $\delta - \gamma > 0.5$ if and only if a studentized $\tilde{\delta}_x - \tilde{\gamma} - 0.5$ was significantly large relative to the standard normal distribution, where $\tilde{\gamma}$ is the estimate of γ using OLS residuals.

Using annual data (as is relevant to the long-run version of PPP) of Obstfeld and Taylor (2002) for the period 1870–1992 (with $n = 123$), we applied the above methodology to four bivariate series, the US ("domestic") versus the "foreign" countries, Australia, Canada, Italy and the UK. Strong evidence against equality of integration orders was found in the case of Australia and Italy, and against cointegration in the case of Canada. However, the UK "passed" all three initial tests. Across the range $b = 10, \ldots, 29$, $(\tilde{\delta}_x, \tilde{\delta}_y)$ varied between the extremes (1.341, 1.095) and (1.572, 1.376), and across $b = 16, \ldots, 25$ and the four $h(n)$ choices, $\delta_x = \delta_y$ was rejected in only 9 out of 40 cases, and these were all at the 10% level. For the same b, no-cointegration was rejected at the 10% level in all cases, at 5% in 4 cases, and at 1% in 3 cases, while $\delta - \gamma = 0.5$ was rejected against $\delta - \gamma > 0.5$ at the 1% level in all cases.

For the US–UK data, we identified parametric models for $f(\lambda)$ as follows. We consider:

$$u_t = A(L)\varepsilon_t, \tag{10.20}$$

where ε_t is considered to be an i.i.d. process. Throughout, $A(L)$ in (10.20) was diagonal, and u_{1t}, u_{2t} treated separately. They were proxied by $\Delta^{\tilde{\gamma}}(y_t - \hat{v}_{OLS} x_t)$, $\Delta^{\tilde{\delta}_x} x_t$, for each of the extreme $\tilde{\gamma}, \tilde{\delta}_x$, namely $\tilde{\gamma} = 0.374$, 0.698 and $\tilde{\delta}_x = 1.572$,

1.341, and then Box-Jenkins-type procedures identified models within the ARMA class. This resulted in AR(1) and ARMA(1,1) models for u_{1t} and white noise and ARMA(1,1) models for u_{2t}, and we fitted all four combinations. We also fitted bivariate versions of Bloomfield's (1973) model, where:

$$A(s) = diag\left\{\exp\left(\sum_{j=1}^{p}\theta_{1j}s^{j}\right), \exp\left(\sum_{j=1}^{p}\theta_{2j}s^{j}\right)\right\},$$

for $p = 1, 2, 3$. For each model we applied the univariate Whittle procedure of Velasco and Robinson (2000), using untapered, differenced data and adding back 1. We summarize the seven models and the resulting $(\hat{\delta}, \hat{\gamma})$ as follows:

Model 1: u_{1t} is AR(1) and u_{2t} is white noise	$(\hat{\delta}, \hat{\gamma}) = (1.612, 0.669)$
Model 2: u_{1t} is AR(1) and u_{2t} is ARMA(1,1)	$(\hat{\delta}, \hat{\gamma}) = (1.408, 0.669)$
Model 3: u_{1t} is ARMA(1,1) and u_{2t} is white noise	$(\hat{\delta}, \hat{\gamma}) = (1.612, 0.660)$
Model 4: u_{1t} is ARMA(1,1) and u_{2t} is ARMA(1,1)	$(\hat{\delta}, \hat{\gamma}) = (1.408, 0.660)$
Model 5: u_t is bivariate Bloomfield with $p = 1$	$(\hat{\delta}, \hat{\gamma}) = (1.214, 0.710)$
Model 6: u_t is bivariate Bloomfield with $p = 2$	$(\hat{\delta}, \hat{\gamma}) = (1.434, 0.701)$
Model 7: u_t is bivariate Bloomfield with $p = 3$	$(\hat{\delta}, \hat{\gamma}) = (1.323, 0.547)$

The $\hat{\gamma}$ seem very robust to the short memory specification, the $\hat{\delta}$ rather less so. We also took the opportunity to examine a further question which, in one form or another, always arises with applications of fractional models, and perhaps most acutely when non-stationary data are involved. This is the matter of truncation. When estimated innovations from a stationary fractional model are computed, the (infinite) AR representation has to be truncated because the data begins at time "1," not at time "$-\infty$." Now, in our model (10.5)–(10.6) for non-stationary data, the truncation is actually inherent in the model, so strictly speaking there is no "error" associated with it. However, the model reflects the time when the data begin, and if we were to drop the first observation, say, and start the model off at the second, the degree of filtering applied to all subsequent observations would change, and it is possible that this could have a marked effect, especially with non-stationary data. Thus, in Table 10.1 we report computations of our estimates $\hat{v}(\hat{\gamma}, \hat{\delta}, \hat{\theta}) = \bar{v}_i$ and Wald statistics:

$$b(\hat{\gamma}, \hat{\theta})\{\hat{v}(\hat{\gamma}, \hat{\delta}, \hat{\theta}) - 1\}^2 = W_i,$$

for models $i = 1, \ldots, 7$, based on the last $n' = n - j$ observations, for $j = 0, 1, \ldots, 10$, in order to explore sensitivity to starting value. Substantial variation is evident across the larger n', with all $\bar{v}_i$ exceeding 1 and the homogeneity hypothesis being strongly rejected when $n' = 123$ across all seven models, but as n' decreases, things stabilize. For $n' \leq 119$ some sensitivity to the u_{2t} specification was found, the white-noise cases (Models 1 and 3) providing estimates of v less than 0.9, whereas for the other models they all exceed 0.9, with the largest values for Model 7. For $n' \leq 122$ the homogeneity hypothesis $v = 1$ is never rejected even at the 10% level.

From certain perspectives, practitioners could consider our empirical analysis simplistic, as we do not take into account possible alternative features of our

Table 10.1 PPP empirical example estimates of ν and Wald tests of $\nu = 1$ for Models 1–7 computed from the last $n' = 113, \ldots, 123$ observations of US/UK

n'	123	122	121	120	119	118	117	116	115	114	113
$\bar{\nu}_1$	1.139	1.050	1.014	.952	.889	.875	.871	.867	.864	.875	.875
W_1	26.23	.352	.017	.163	.759	.940	.986	1.035	1.082	.903	.890
$\bar{\nu}_2$	1.294	.959	1.030	.995	.949	.941	.941	.938	.936	.944	.943
W_2	117.3	.231	.078	.002	.159	.208	.206	.226	.243	.181	.182
$\bar{\nu}_3$	1.113	1.084	1.017	.955	.889	.871	.866	.863	.859	.871	.868
W_3	18.64	1.070	.027	.161	.823	1.079	1.138	1.196	1.251	1.051	1.059
$\bar{\nu}_4$	1.290	.966	1.028	.997	.950	.939	.939	.936	.934	.942	.939
W_4	122.6	.178	.078	.001	.170	.241	.240	.263	.281	.212	.227
$\bar{\nu}_5$	1.274	1.042	1.025	.986	.940	.933	.932	.931	.929	.939	.936
W_5	112.2	.225	.055	.014	.230	.283	.283	.296	.306	.223	.239
$\bar{\nu}_6$	1.278	.960	1.015	.983	.939	.932	.931	.930	.927	.937	.935
W_6	114.9	.211	.019	.020	.241	.292	.292	.306	.325	.246	.255
$\bar{\nu}_7$	1.298	.999	1.048	1.024	.975	.961	.962	.956	.956	.963	.958
W_7	116.9	.000	.279	.052	.047	.109	.105	.138	.136	.096	.122

data. In particular, we did not check for the possibility of structural breaks or nonlinearities in our long time series. Admittedly, these are relevant issues, whose linkages with fractional processes are mainly undiscovered, but which have already attracted the attention of some researchers. For example, Granger (1999) showed that structural break processes could produce "long memory" properties of the data, while he suggested that, among nonlinear time series, there could be other plausible alternatives to $I(d)$ processes. Undoubtedly, a very rigorous and exhaustive analysis of the PPP hypothesis should contemplate these issues but, at this stage, our intention was simply to propose a sensible methodology incorporating the techniques developed in the literature and which, at the same time, motivated our testing problem appropriately.

Acknowledgments

The two authors gratefully acknowledge financial support from the Ministerio de Educación y Ciencia through the SEJ2005-07657/ECON project, and the second author also through a Ramón y Cajal contract. Thanks to Peter M. Robinson for useful comments on section 10.4 and to Alan M. Taylor for providing the data we employ in that section.

Notes

1. For the purpose of the present work we define an $I(0)$ process as a covariance stationary process with spectral density function that is positive and bounded at any frequency. Alternatively, a time domain definition corresponds to a process where the infinite sum of the autocovariances is finite.
2. The Type I definition of fractional integration has been used by Sowell (1990), Hurvich and Ray (1995), Chan and Terrin (1995), Jeganathan (1999), Velasco (1999a, 1999b), Marinucci (2000), Velasco and Robinson (2000) and others, whilst the Type II definition has been used by Robinson and Marinucci (2001), Kim and Phillips (2002), Robinson and Hualde (2003) and others.

3. Excellent surveys can be found in Beran (1994), Baillie (1996), Doukhan, Oppenheim, and Taqqu (2003) and Robinson (2003).
4. Much earlier, Hurst (1951) proposed the adjusted rescaled range, or R/S statistic. This specific estimator of d can be found in Mandelbrot and Wallis (1968) and its properties were analyzed in Mandelbrot and Wallis (1969), Mandelbrot (1972, 1975) and Mandelbrot and Taqqu (1979): see also Lo (1991) and Giraitis *et al.* (2003).
5. Multivariate methods of fractional integration (not involving cointegrating relationships) have been examined by Gil-Alana (2003a, 2003b) and Nielsen (2004, 2005).
6. See also Hidalgo and Soulier (2004) and Hidalgo (2005) for recent developments in this area.
7. The issue of fractional integration in the context of structural breaks has received increasing attention in recent years. For a review, see Banerjee and Urga (2005).
8. Note that a normalization has been carried out in (10.5), the cointegration vector corresponding to Engle and Granger's (1987) definition being now $(1, -\nu)'$. As demonstrated by Phillips and Loretan (1991), (10.5)–(10.6) with $\gamma = 0$, $\delta = 1$, represents "a typical cointegrated system" in structural form. (10.5) could be regarded as a stochastic version of the partial equilibrium $y_t - \nu x_t$, with $u_{1t}\,(-\gamma)$ representing deviations from this equilibrium. (10.6) is a reduced form equation.
9. He did not analyze exactly the model (10.5)–(10.6), but a similar one where y_t and x_t were covariance stationary long memory processes.
10. These conditions refer to the smoothness of f and convergence rates of the estimates of the nuisance parameters.
11. Chen and Hurvich (2003b) also propose estimators of the cointegrating parameter in a system with deterministic trends, but their framework is much more specific than that of Robinson and Iacone (2005), and their objective is different, because by means of tapering and differentiating the data appropriately, they present a tapered NBLS which is invariant with respect to deterministic polynomial trends in the series.
12. Kim and Phillips (2002) also provided a similar analysis to the one by Baillie and Bollerslev (1994a), assuming also the memory of certain series of exchange rates to be one.

References

Abrahams, M.D. and A. Dempster (1979) Research on seasonal analysis. Progress report of the ASA/Census project on seasonal adjustment. Technical report, Department of Statistics, Harvard University, Cambridge, Mass.

Abry, P. and D. Veitch (1998) Wavelet analysis of long range dependence traffic. *IEEE Transactions on Information Theory* **44**, 2–15.

Adelman, I. (1965) Long cycles: fact or artifacts. *American Economic Review* **55**, 444–63.

Adenstedt, R.K. (1974) On large sample estimation for the mean of a stationary random sequence. *Annals of Statistics* **2**, 259–72.

Andersen, T.G. and T. Bollerslev (1997) Heterogeneous information arrivals and return volatility dynamics: uncovering the long run in high frequency returns. *Journal of Finance* **52**, 975–1005.

Andersen, T.G. and T. Bollerslev (1998) Deutsche Mark–Dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* **53**, 219–65.

Andersen, T.G., T. Bollerslev, F.X. Diebold and H. Ebens (2001) The distribution of stock return volatility. *Journal of Financial Economics* **61**, 43–76.

Arteche, J. (2004) Gaussian semiparametric estimation in long memory in stochastic volatility and signal plus noise models. *Journal of Econometrics* **119**, 131–54.

Arteche, J. and P.M. Robinson (2000) Semiparametric inference in seasonal and cyclical long memory processes. *Journal of Time Series Analysis* **21**, 1–25.

Aydogan, K. and G.G. Booth (1988) Are there long cycles in common stock returns? *Southern Economic Journal* **55**, 141–9.
Backus, D. and S. Zin (1993) Long memory inflation uncertainty. Evidence from the term structure of interest rates. *Journal of Money, Credit and Banking* **25**, 681–700.
Baillie, R.T. (1996) Long memory processes and fractional integration in econometrics. *Journal of Econometrics* **73**, 5–59.
Baillie, R.T. and T. Bollerslev (1994a) Cointegration, fractional cointegration and exchange rate dynamics. *Journal of Finance* **49**, 737–45.
Baillie, R.T. and T. Bollerslev (1994b) The long memory of the forward premium. *Journal of International Money and Finance* **13**, 565–71.
Baillie, R.T. and T. Bollerslev (2000) The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance* **19**, 471–88.
Baillie, R.T., T. Bollerslev and H.O. Mikkelsen (1996) Fractionally integrated generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **74**, 3–30.
Baillie, R.T., A.A. Cecen and Y.W. Han (2000) High frequency Deutsche mark–US dollar return. FIGARCH representations and non-linearities. *Multinational Finance Journal* **4**, 247–67.
Baillie, R.T., C.F. Chung and M.A. Tieslau (1996) Analyzing inflation by the fractionally integrated ARFIMA-GARCH Model. *Journal of Applied Econometrics* **11**, 23–40.
Baillie, R.T., Y.W. Han and Koul (2002) A high frequency perspective on the forward premium anomaly. Preprint.
Baillie, R.T., Y.W. Han and T.G. Kwon (2002) Further long memory properties of inflationary shocks. *Southern Economic Journal* **68**, 496–510.
Baillie, R.T., Y.W. Han, R.J. Myers and J. Song (2007) Long memory models for daily and high frequency commodity future returns. *Journal of Future Markets* **27**, 643–68.
Banerjee, A. and G. Urga (2005) Modelling structural break, long memory and stock market volatility. *Journal of Econometrics* **129**, 1–34.
Barkoulas, J.T. and C.F. Baum (1996) Long term dependence in stock returns. *Economics Letters* **53**, 253–9.
Barkoulas, J.T. and C.F. Baum (1997) Fractional differencing modeling and forecasting of eurocurrency deposit rates. *Journal of Financial Research* **20**, 355–72.
Barkoulas, J.T. and C.F. Baum (2006) Long memory forecasting of US monetary indices. *Journal of Forecasting* **25**, 291–302.
Barkoulas, J., C.F. Baum and S. Oguz (1996) Fractional cointegration analysis of long term international interest rates. Manuscript, Boston College.
Barkoulas, J.T., C.F. Baum and N. Travlos (2000) Long memory in the Greek stock market. *Applied Financial Economics* **10**, 177–84.
Baum, C.F., J. Barkoulas and M. Caglayan (1999) Persistence in the international inflation rates. *Southern Economic Journal* **65**, 900–13.
Bekaert, G. (1996) The time variation of risk and return in foreign exchange markets. A general equilibrium perspective. *Review of Financial Studies* **9**, 427–70.
Bekaert, G., R.J. Hodrick and D. Marshall (1997) The implications of first order risk aversion for asset market risk premiums. *Journal of Monetary Economics* **40**, 3–39.
Beltratti, A. and C. Morana (2006) Breaks and persistency. Macroeconomic causes of stock market volatility. *Journal of Econometrics* **131**, 151–77.
Beran, J. (1994) *Statistics for Long Memory Processes*. New York: Chapman and Hall.
Bierens, H.J. (2001) Complex unit roots and business cycles: are they real? *Econometric Theory* **17**, 962–83.
Bloomfield, P. (1973) An exponential model in the spectrum of a scalar time series. *Biometrika* **60**, 217–26.
Bloomfield, P. (1992) Trends in global temperatures. *Climatic Changes* **21**, 1–26.
Bollerslev, T. and H.O. Mikkelsen (1996) Modeling and pricing long memory in stock market volatility. *Journal of Econometrics* **73**, 151–84.
Booth, G.G., F.R. Kaen and P.E. Koveos (1982) R/S analysis of foreign exchange markets under two international monetary regimes. *Journal of Monetary Economics* **10**, 407–15.

Bos, C., P.H. Franses and M. Ooms (1999) Long memory and level shifts: reanalyzing inflation rates. *Empirical Economics* **24**, 427–50.
Bos, C., P.H. Franses and M. Ooms (2001) Inflation, forecast intervals and long memory regression models. *International Journal of Forecasting* **18**, 243–64.
Box-Steffensmeier, J.M. and R.M. Smith (1996) The dynamics of aggregate partisanship. *American Political Science Review* **90**, 567–80.
Box-Steffensmeier, J.M. and R.M. Smith (1998) Investigating political dynamics using fractional integration methods. *American Journal of Political Sciences* **42**, 661–89.
Box-Steffensmeier, J.M. and A.R. Tomlinson (2000) Fractional integration methods in political sciences. *Electoral Studies* **19**, 63–76.
Breidt, F., N. Crato and P. de Lima (1997) Modeling persistent volatility of asset returns. *Computational Inteligence for Financial Engineering* **23**, 266–72.
Breidt, F., N. Crato and P. de Lima (1998) The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* **83**, 325–48.
Brunetti, C. and C.L. Gilbert (2000) Bivariate FIGARCH and fractional cointegration. *Journal of Empirical Finance* **7**, 509–30.
Byers, D., J. Davidson and D. Peel (1997) Modelling political popularity: an analysis of long range dependence in opinion poll series. *Journal of the Royal Statistical Society, Series A* **160**, 471–80.
Byers, D., J. Davidson and D. Peel (2000) The dynamics of aggregate political popularity: evidence from eight countries. *Electoral Studies* **19**, 49–62.
Campbell, J. and R.J. Shiller (1987) Cointegration and tests of present value models. *Journal of Political Economy* **95**, 1062–88.
Caporale, G.M. and L.A. Gil-Alana (2004a) Fractional cointegration and real exchange rates. *Review of Financial Economics* **13**, 327–40.
Caporale, G.M. and L.A. Gil-Alana (2004b) Fractional cointegration and tests of present value models. *Review of Financial Economics* **13**, 245–58.
Carlin, J.B. and A.P. Dempster (1989) Sensitivity analysis of seasonal adjustments: empirical cases studies. *Journal of the American Statistical Association* **84**, 6–20.
Carlin, J.B., A.P. Dempster and A.B. Jonas (1985) On methods and moments for Bayesian time series analysis. *Journal of Econometrics* **30**, 67–90.
Chambers, M. (1998) Long memory and aggregation in macroeconomic time series. *International Economic Review* **39**, 1053–72.
Chan, N.H. and N. Terrin (1995) Inference for long memory processes with application to fractionally unit root autoregression. *Annals of Statistics* **23**, 1662–83.
Chen, W. and C. Hurvich (2003a) Semiparametric estimation of multivariate fractional cointegration. *Journal of the American Statistical Association* **98**, 629–42.
Chen, W. and C. Hurvich (2003b) Estimating fractional cointegration in the presence of a polynomial trend. *Journal of Econometrics* **117**, 95–121.
Chen, W. and C. Hurvich (2006) Semiparametric estimation of fractional cointegrating subspaces. *Annals of Statistics* **27**, 2939–79.
Cheung, Y.-W. (1993) Long memory in foreign exchange rates. *Journal of Business and Economic Statistics* **11**, 93–101.
Cheung, Y.-W. and K. Lai (1993) A fractional cointegration analysis of purchasing power parity. *Journal of Business and Economics Statistics* **11**, 103–12.
Cheung, Y.-W. and K.S. Lai (1995) A search for long memory in international stock market returns. *Journal of International Money and Finance* **14**, 597–615.
Chou, W. and Y. Shih (1997) Long run Purchasing Power Parity and long term memory. Evidence from Asian newly industrialized countries. *Applied Economics Letters* **4**, 575–8.
Choudhry, T. (1999a) Purchasing Power Parity in high inflation Eastern European countries. Evidence from fractional and Harris-Inder cointegration tests. *Journal of Macroeconomics* **21**, 293–308.
Choudhry, T. (1999b) Re-examining forward market efficiency. Evidence from fractional and Harris-Inder cointegration tests. *International Review of Economics and Finance* **8**, 433–53.

Christensen, B.J. and N.O. Nielsen (2006) Asymptotic normality of narrowed band least squares in the stationary fractional cointegration model and volatility forecasting. *Journal of Econometrics* **133**, 343–71.
Christensen, B.J. and N.R. Prabhala (1998) The relation between implied and realized volatility. *Journal of Financial Economics* **50**, 125–50.
Christiano, L.J. and M. Eichembaum (1990) Unit roots in real GNP. Do we know and do we care? *Carnegie-Rochester Conference Series on Public Policy* **32**, 7–61.
Cioczek-George, R. and B.B. Mandelbrot (1995) A class of micropulses and anti persistent fractional Brownian motion. *Stochastic Processes and Their Applications* **60**, 1–18.
Clarke, M.D. and M. Lebo (2003) Fractional (co)-integration and governing party support in Britain. *British Journal of Political Science* **33**, 283–301.
Corbae, D. and S. Ouliaris (1988) Cointegration and tests of Purchasing Power Parity. *Review of Economics and Statistics* **70**, 508–11.
Couchman, J., R. Gounder and J.J. Su (2006) Long memory properties of real interest rates for 16 countries. *Applied Financial Economics Letters* **2**, 25–30.
Crato, N. (1994) Some international evidence regarding the stochastic behaviour of stock returns. *Applied Financial Economics* **4**, 33–9.
Crato, N. and P.J.F. de Lima (1994) Long range dependence in the conditional variance of stock returns. *Economics Letters* **45**, 281–5.
Crato, N. and B.K. Ray (2000) Memory in returns and volatilities of future's contracts. *Journal of Futures Markets* **20**, 525–43.
Crato, N. and P. Rothman (1994a) Fractional integration analysis of long run behaviour for US macroeconomic time series. *Economics Letters* **45**, 287–91.
Crato, N. and P. Rothman (1994b) A reappraisal of parity reversion for UK real exchange rates. *Applied Economics Letters* **1**, 134–41.
Dahlhaus, R. (1989) Efficient parameter estimation for self-similar process. *Annals of Statistics* **17**, 1749–66.
Dalla, V. and J. Hidalgo (2005) A parametric bootstrap test for cycles. *Journal of Econometrics* **129**, 219–61.
Davidson, J. (2003) Testing for fractional cointegration. The relationship between government popularity and economic performance in the UK. In C. Diebolt and C. Kyrtsov (eds.), *New Trends in Macroeconomics*, pp. 147–71. Berlin and London: Springer.
Davidson, J., D. Peel and D. Byers (2006) Support for governments and leaders: fractional cointegration analysis of poll evidence from the UK, 1960–2004. *Studies in Non-linear Dynamics and Econometrics* **10**(1), Article 3.
Delgado, M. and P.M. Robinson (1994) New methods for the analysis of long memory time series. Application to Spanish inflation. *Journal of Forecasting* **13**, 97–107.
Dickey, D. and W.A. Fuller (1979) Distributions of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**, 427–31.
Diebold, F.X., J. Gardeazabal and K. Yilmaz (1994) On cointegration and exchange rate dynamics. *Journal of Finance* **49**, 727–45.
Diebold, F.X., S. Husted and M. Rush (1991) Real exchange rates under the gold standard. *Journal of Political Economy* **99**, 1252–71.
Diebold, F.X. and A. Inoue (2001) Long memory and regime switching. *Journal of Econometrics* **105**, 131–59.
Diebold, F.X. and G.D. Rudebusch (1989) Long memory and persistence in the aggregate output. *Journal of Monetary Economics* **24**, 189–209.
Diebold, F.X. and G.D. Rudebusch (1991) Is consumption too smooth? Long memory and the Deaton paradox. *Review of Economics and Statistics* **73**, 1–9.
Ding, Z. and C.W.J. Granger (1996) Modeling volatility persistence of speculative returns: a new approach. *Journal of Econometrics* **73**, 185–215.
Dolado, J.J., J. Gonzalo and L. Mayoral (2003) Long range dependence in Spanish political opinion pool series. *Journal of Applied Econometrics* **18**, 137–55.

Dolado, J.J. and F. Marmol (1996) Efficient estimation of cointegrating relationships among higher orders and fractionally integrated processes. Banco de España, Servicio de Estudios, Madrid.

Dolado, J.J. and F. Marmol (2004) Asymptotic inference results for multivariate long memory processes. *Econometrics Journal* **7**, 168–90.

Doukhan, P., G. Oppenheim and M.S. Taqqu (2003) *Theory and Applications of Long Range Dependence*. Basel: Birkhäuser.

Dueker, M. and R. Startz (1998) Maximum likelihood estimation of fractional cointegration with an application to US and Canadian bond rates. *Review of Economics and Statistics* **80**, 420–6.

Dufrénot, G., L. Mathieu, V. Mignon and A. Peguin-Feissolle (2006) Persistent misalignments of the European exchange rates: some evidence from non-linear cointegration. *Applied Economics* **38**, 203–29.

Enders, W. (1988) Arima and cointegration tests of Purchasing Power Parity under fixed and flexible exchange rate regimes. *Review of Economics and Statistics* **70**, 504–8.

Engle, R.F. and C.W.J. Granger (1987) Cointegration and error correction model. Representation, estimation and testing. *Econometrica* **55**, 251–76.

Fang, H., K.S. Lai and M. Lai (1994) Fractal structure in currency futures price dynamics. *Journal of Futures Markets* **14**, 169–81.

Ferrara, L. and D. Guegan (2001) Forecasting with k-factor Gegenbauer processes. Theory and applications. *Journal of Forecasting* **20**, 581–601.

Flôres, R. and A. Szafarz (1996) An enlarged definition of cointegration. *Economics Letters* **50**, 193–5.

Fox, R. and Taqqu, M. (1986) Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Annals of Statistics* **14**, 517–32.

Franses, N. Hyung and J. Penn (2006) Structural breaks and long memory in US inflation rates. Do they matter for forecasting? *Research in International Business and Finance* **20**, 95–110.

Franses, P.H. and M. Ooms (1997) A periodic long memory model for quarterly UK inflation. *International Journal of Forecasting* **13**, 117–26.

Gadea, M.D., M. Sabate and J.M. Serrano (2004) Structural breaks and their trace in the memory. Inflation rate series in the long run. *Journal of International Financial Markets, Institutions and Money* **14**, 117–34.

Geweke, J. and S. Porter-Hudak (1983) The estimation and application of long memory time series models. *Journal of Time Series Analysis* **4**, 221–38.

Gil-Alana, L.A. (1999) Fractional integration with monthly data. *Economic Modelling* **16**, 613–29.

Gil-Alana, L.A. (2001) Testing stochastic cycles in macroeconomic time series. *Journal of Time Series Analysis* **22**, 411–30.

Gil-Alana, L.A. (2003a) Multivariate tests of nonstationary hypotheses. *South African Statistical Journal* **37**, 1–28.

Gil-Alana, L.A. (2003b) A fractional multivariate long memory model for the US and the Canadian real output. *Economics Letters* **81**, 355–9.

Gil-Alana, L.A. (2004a) Long memory in the interest rates in some Asian countries. *International Advances in Economic Research* **9**, 257–67.

Gil-Alana, L.A. (2004b) Long memory in the US interest rate. *International Review of Financial Analysis* **13**, 265–76.

Gil-Alana, L.A. (2005a) Statistical model of the temperatures in the northern hemisphere using fractionally integrated techniques. *Journal of Climate* **18**, 5357–69.

Gil-Alana, L.A. (2005b) Deterministic seasonality versus seasonal fractional integration. *Journal of Statistical Planning and Inference* **134**, 445–61.

Gil-Alana, L.A. (2006) Fractional integration in daily stock market returns. *Review of Financial Economics* **15**, 28–48.

Gil-Alana, L.A. (2007) Testing the existence of multiple cycles in financial and economic time series. *Annals of Economics and Finance* **1**, 1–20.
Gil-Alana, L.A. (2008a) Fractional integration and structural breaks at unknown periods of time. *Journal of Time Series Analysis* **29**(1), 163–85.
Gil-Alana, L.A. (2008b) Warming break trends and fractional integration in the northern, southern and global temperature anomaly series. *Journal of the Atmospheric Oceanic Technology* **25**(4), 570–8.
Gil-Alana, L.A. and P.M. Robinson (1997) Testing of unit roots and other nonstationary hypotheses in macroeconomic time series. *Journal of Econometrics* **80**, 241–68.
Gil-Alana, L.A. and P.M. Robinson (2001) Testing of seasonal fractional integration in the UK and Japanese consumption and income. *Journal of Applied Econometrics* **16**, 95–114.
Giraitis, L., P. Kokosza, P. Leypus and R. Teyssière (2003) Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics* **112**, 265–94.
Granger, C.W.J. (1966) The typical spectral shape of an economic variable. *Econometrica* **37**, 150–61.
Granger, C.W.J. (1980) Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* **14**, 227–38.
Granger, C.W.J. (1981) Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* **16**, 121–30.
Granger, C.W.J. (1999) Aspects of research strategy for time series analysis, unpublished presentation to New Developments in Time Series Economics conference, New Haven, Conn.
Granger, C.W.J. and Z. Ding (1995a) Some properties of absolute returns. An alternative measure of risk. *Annales d'Economie et de Statistique* **40**, 67–91.
Granger, C.W.J. and Z. Ding (1995b) Stylized facts on the temporal and distributional properties of daily data from speculative markets. UCSD Working Paper.
Granger, C.W.J. and Z. Ding (1996) Varieties of long memory models. *Journal of Econometrics* **73**, 61–78.
Granger, C.W.J. and N. Hyung (2004) Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* **11**, 399–421.
Granger, C.W.J. and R. Joyeux (1980) An introduction to long memory time series and fractionally differencing. *Journal of Time Series Analysis* **1**, 15–29.
Granger, C.W.J. and A. Weiss (1983) Time series analysis of error-correcting models. In S. Karlin (ed.), *Studies in Econometrics, Time Series and Multivariate Statistics*, pp. 255–78. New York: Academic Press.
Gray, H.L., Yhang, N. and W.A. Woodward (1989) On generalized fractional processes. *Journal of Time Series Analysis* **10**, 233–57.
Gray, H.L., Yhang, N. and Woodward, W.A. (1994) On generalized fractional processes. A correction. *Journal of Time Series Analysis* **15**, 561–2.
Greene, M.T. and B.D. Fielitz (1977) Long term dependence in common stock returns. *Journal of Financial Economics* **5**, 339–49.
Haldrup, N. and N.O. Nielsen (2006) A regime switching long memory model for electricity prices. *Journal of Econometrics* **135**, 349–76.
Hannan, E.J. (1963) Regression for time series. In M. Rosenblatt (ed.), *Time Series Analysis*, pp. 17–37. New York: John Wiley.
Hassler, U. (1993) Regression of spectral estimators with fractionally integrated time series. *Journal of Time Series Analysis* **14**, 369–80.
Hassler, U. and J. Wolters (1995) Long memory in inflation rates. International evidence. *Journal of Business and Economic Statistics* **13**, 37–45.
Hasslet, J. and A.E. Raftery (1989) Space-time modeling with long memory dependence: assessing Ireland's wind power resource. *Applied Statistics* **38**, 1–50.

Haubrich, J.G. (1993) Consumption and fractional differencing. Old and new anomalies. *Review of Economics and Statistics* **75**, 767–72.
Haubrich, J.G. and A. Lo (1991) The sources and nature of long-term memory in the business cycle. Federal Reserve Bank of Cleveland, Working Paper No. 9116.
Hauser, M.A. (1999) Maximum likelihood estimators for ARFIMA models: a Monte Carlo study. *Journal of Statistical Planning and Inference* **8**, 223–55.
Henry, O.T. (2002) Long memory in stock returns. Some international evidence. *Applied Financial Economics* **12**, 725–9.
Hidalgo, J. (2005) Semiparametric estimation for stationary processes whose spectra have an unknown pole. *Annals of Statistics* **35**, 1843–89.
Hidalgo, J. and P. Soulier (2004) Estimation of the location and exponent of the spectral singularity of a long memory process. *Journal of Time Series Analysis* **25**, 55–81.
Hiemstra, C. and J.D. Jones (1997) Another look at long memory in common stock returns. *Journal of Empirical Finance* **4**, 373–401.
Hosking, J.R.M. (1981) Fractional differencing. *Biometrika* **68**, 165–76.
Hualde, J. and P.M. Robinson (2006) Semiparametric inference in multivariate fractionally integrated systems. Preprint.
Hualde, J. and P.M. Robinson (2007) Root-n-consistent estimation of weak fractional cointegration. *Journal of Econometrics* **140**, 450–84.
Hurvich, C.M. and B.K. Ray (1995) Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis* **16**, 17–41.
Hurst, H.E. (1951) Long-term storage capacity of reservoirs. *Transactions of the American Society Civil Engineers* **116**, 770–9.
Hussain, S. and A. Elbergali (1999) Fractional order estimation and testing. Application to Swedish temperature data. *Environmetrics* **10**, 339–49.
Jeganathan, P. (1999) On asymptotic inference on cointegrated time series with fractionally integrated errors. *Econometric Theory* **15**, 583–621.
Jeganathan, P. (2001) Correction to: "On asymptotic inference on cointegrated time series with fractionally integrated errors." Preprint. Department of Statistics, University of Michigan.
Johansen, S. (1988) Statistical analysis of cointegrating vectors. *Journal of Economics Dynamics and Control* **12**, 231–54.
Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551–80.
Johansen, S. (1996) *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
Johansen, S. (2008) A representation theory for a class of vector autoregressive models for fractional processes. *Econometric Theory* **24**, 651–76.
Jonas, A.B. (1981) Long memory self similar time series models. Unpublished manuscript, Harvard University, Department of Economics.
Karagiannis, T., M. Molle and M. Faloutsos (2004) Long range dependence: ten years of internet traffic modeling. *IEEE Internet Computing* **8**, 57–64.
Kihc, R. (2004) On the long memory properties of emerging capital markets. Evidence from Istanbul exchange. *Applied Financial Economics* **14**, 915–22.
Kim, S.C. and P.C.B. Phillips (2002) Fully modified estimation of fractional cointegration models. Preprint.
Kim, Y. (1990) Purchasing Power Parity in the long run. A cointegration approach. *Journal of Money, Credit and Banking* **22**, 491–503.
Koop, G., E. Ley, J. Osiewalski and M.F.J. Steel (1997) Bayesian analysis of long memory and persistence using ARFIMA models. *Journal of Econometrics* **76**, 149–69.
Lai, K.S. (1997) Long term persistence in the real interest rate. Some evidence of a fractional unit root. *International Journal of Finance and Economics* **2**, 225–35.

Lebo, M. and W.M. Moore (2003) Dynamic foreign policy. *Journal of Conflict Resolution* **74**, 13–32.
Lo, A.W. (1991) Long-term memory in stock prices. *Econometrica* **59**, 1279–313.
Lobato, I. and P.M. Robinson (1998) A non-parametric test for I(0). *Review of Economics and Statistics* **65**, 475–95.
Magnus, W., F. Oberhettinger and R.P. Soni (1966) *Formulas and Theorems for the Special Functions of Mathematical Physics*. Berlin: Springer.
Mandelbrot, B.B. (1972) Statistical methodology for non-periodic cycles: from the covariance to R/S analysis. *Annals of Economics and Social Measurement* **1**, 259–90.
Mandelbrot, B.B. (1975) Limit theorems on the self-normalized range for weakly and strongly dependent processes. *Z. Wahrscheinlinchkeitstheorie verw.* **31**, 271–85.
Mandelbrot, B.B. and M.S. Taqqu (1979) Robust R/S analysis of long run serial correlation. *Proceedings of the 42nd Session of the International Statistical Institute*, Manila.
Mandelbrot, B.B. and J.R. Wallis (1968) Noah, Joseph and operational hydrology. *Water Resources Research* **4**, 909–18.
Mandelbrot, B.B. and J.R. Wallis (1969) Some long run properties of geophysical records. *Water Resources Research* **5**, 321–40.
Marinucci, D. (2000) Spectral regression for cointegrated time series with long memory innovations. *Journal of Time Series Analysis* **21**, 685–705.
Marinucci, D. and P.M. Robinson (2001) Semiparametric fractional cointegration analysis. *Journal of Econometrics* **105**, 225–47.
Maynard, A. and P.C.B. Phillips (2001) Rethinking an old empirical puzzle: econometric evidence on the forward premium anomaly. *Journal of Applied Econometrics* **16**, 671–708.
Mayoral, L. (2006) Further evidence on the statistical properties of real GNP. *Oxford Bulletin of Economics and Statistics* **68**, 901–20.
Meade, N. and M.R. Maier (2003) Evidence of long memory is short term interest rates. *Journal of Forecasting* **22**, 553–68.
Michelacci, C. and P. Zaffaroni (2000) (Fractional) Beta convergence, *Journal of Monetary Economics* **45**, 129–53.
Mikosch, T. and C. Starica (2000) Change of structure in financial time series, long range dependence and the GARCH model. Centre for Analytical Finance, University of Aarhus, Working Paper Series No. 58.
Montanari, A., R. Rosso and M.S. Taqqu (1997) Fractionally differenced ARIMA models applied to hydrological time series. Identification, estimation and simulation. *Water Resources Research* **33**, 1035–44.
Montanari, A., R. Rosso and M.S. Taqqu (2000) A seasonal fractional ARIMA model applied to the Nile river monthly at Aswan. *Water Resources Research* **36**, 1249–59.
Morana, C. and A. Beltratti (2004) Structural change and long range dependence in volatility of exchange rates: either, neither or both? *Journal of Empirical Finance* **11**, 629–58.
Nelson, C.R. and C.I. Plosser (1982) Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* **10**, 139–62.
Nielsen, M.O. (2004) Efficient inference in multivariate fractionally integrated time series models. *Econometrics Journal* **7**, 63–97.
Nielsen, M.O. (2005) Multivariate Lagrange Multiplier tests for fractional integration. *Journal of Financial Econometrics* **3**, 372–98.
Obstfeld, M. and A.M. Taylor (2002) Global capital markets: integration, crisis and growth. Japan-US Center Sanwa *Monographs on International Financial Markets*. Cambridge: Cambridge University Press.
Ooms, M. and P.H. Franses (2001) A seasonal periodic long memory model for monthly river flows. *Environmental Modelling and Software* **16**, 559–69.
Pan, M.P. and Y.A. Liu (1999) Fractional cointegration, long memory and exchange rate dynamics. *International Review of Economics and Finance* **8**, 305–16.
Park, J.Y. (1992) Canonical cointegrating regression. *Econometrica* **60**, 119–44.
Parke, W.R. (1999) What is fractional integration? *Review of Economics and Statistics* **81**, 632–8.

Phillips, P.C.B. (1991a) Optimal inference in cointegrating systems. *Econometrica* **59**, 283–306.
Phillips, P.C.B. (1991b) Spectral regresssions for cointegrated time series. In W.A. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.
Phillips, P.C.B. (1998) Econometric analysis of Fisher's equation. Yale University, Cowles Foundation Discussion Paper 1180.
Phillips, P.C.B. and S.N. Durlauf (1986) Multiple time series regressions with integrated processes. *Review of Economic Studies* **53**, 473–95.
Phillips, P.C.B. and B.E. Hansen (1990) Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* **57**, 99–125.
Phillips, P.C.B. and M. Loretan (1991) Estimating long run economic equilibria. *Review of Economic Studies* **58**, 407–36.
Phillips, P.C.B. and K. Shimotsu (2004) Local Whittle estimation in nonstationary and unit root cases. *Annals of Statistics* **32**, 656–92.
Phillips, P.C.B. and K. Shimotsu (2005) Exact local Whittle estimation of fractional integration. *Annals of Statistics* **33**(4) 1890–933.
Porter-Hudak, S. (1990) An application of the seasonal fractionally differenced model to the monetary aggregate. *Journal of the American Statistical Association* **85**, 338–44.
Rainville, E.D. (1960) *Special Functions*. New York: Macmillan.
Rao, A.R. and D. Bhattacharya (1999) Hypothesis testing for long-term memory in hydrological series. *Journal of Hydrology* **216**, 183–96.
Ray, B. K. (1993) Long range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model. *International Journal of Forecasting* **9**, 255–69.
Robinson, P.M. (1978) Statistical inference for a random coefficient autoregressive model. *Scandinavian Journal of Statistics* **5**, 163–8.
Robinson, P.M. (1994a) Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association* **89**, 1420–37.
Robinson, P.M. (1994b) Semiparametric analysis of long memory time series. *Annals of Statistics* **22**, 515–39.
Robinson, P.M. (1995a) Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* **23**, 1630–61.
Robinson, P.M. (1995b) Log-periodogram regression of time series with long range dependence. *Annals of Statistics* **23**, 1048–72.
Robinson, P.M. (2003) Long memory time series. In P.M. Robinson (ed.), *Time Series with Long Memory*, pp. 1–48. Oxford: Oxford University Press.
Robinson, P.M. and J. Hualde (2003) Cointegration in fractional systems with unknown integration orders. *Econometrica* **71**, 1727–66.
Robinson, P.M. and F. Iacone (2005) Cointegration in fractional systems with deterministic trends. *Journal of Econometrics* **129**, 263–98.
Robinson, P.M. and D. Marinucci (2001) Narrow-band analysis of nonstationary processes. *Annals of Statistics* **29**, 947–86.
Robinson, P.M. and D. Marinucci (2003) Semiparametric frequency domain analysis of fractional cointegration. In P.M. Robinson (ed.), *Time Series with Long Memory*, pp. 334–73. Oxford: Oxford University Press.
Robinson, P.M. and Y. Yajima (2002) Determination of cointegrating rank in fractional systems. *Journal of Econometrics* **106**, 217–41.
Rudebusch, G.D. (1993) The uncertain unit root in real GNP. *American Economic Review* **83**, 264–72.
Sadek, N. and A. Khotanzad (2004) K-factor Gegenbauer ARMA process for network traffic simulation. *Computers and Communications* **2**, 963–8.
Sadique, S. and P. Silvapulle (2001) Long-term memory in stock market returns. International evidence. *International Journal of Finance and Economics* **6**, 59–67.

Saikkonen, P. (1991) Asymptotically efficient estimation of cointegrating regressions. *Econometric Theory* **7**, 1–21.
Sephton, P.S. and H.K. Larsen (1991) Tests of exchange market efficiency: fragile evidence from cointegration tests. *Journal of International Money and Finance* **10**, 561–70.
Shea, G. (1991) Uncertainty and implied variance bounds in long memory models of the interest rate term structure. *Empirical Economics* **16**, 287–312.
Smith, J., N. Taylor and S. Yadav (1997) Comparing the bias and misspecification in ARFIMA models. *Journal of Time Series Analysis* **18**, 507–27.
Sowell, F. (1990) The fractional unit root distribution. *Econometrica* **56**, 495–504.
Sowell, F. (1992a) Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics* **53**, 165–88.
Sowell, F. (1992b) Modelling long run behaviour with the fractional ARIMA model. *Journal of Monetary Economics* **29**, 277–302.
Stock, J.H. and M.W. Watson (1993) A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* **61**, 783–820.
Sutcliffe, A. (1994) Time series forecasting using fractional differencing. *Journal of Forecasting* **13**, 383–93.
Tanaka, K. (1999) The nonstationary fractional unit root. *Econometric Theory* **15**, 549–582.
Taqqu, M.S. (1975) Weak convergence to fractional motion and to the Rosenblatt process. *Z. Wahrscheinlichkeitstheorie verw., Geb*. **31**, 287–302.
Taqqu, M.S., W. Willinger and R. Sherman (1997) Proof of a fundamental result in self-similar traffic modelling. *Computer Communication Review* **27**, 5–23.
Taylor, M.P. (1988) An empirical estimation of the long run Purchasing Power Parity using cointegration techniques. *Applied Economics* **20**, 1369–81.
Tolvi, J. (2003) Long memory and outliers in stock market returns. *Applied Financial Economics* **13**, 495–502.
Tsay, W.J. (2000) The long memory story of the real interest rate. *Economics Letters* **67**, 325–30.
Tse, Y.K. (1998) The conditional heteroskedasticity of the Yen–Dollar exchange rate. *Journal of Applied Econometrics* **13**, 49–55.
Velasco, C. (1999a) Nonstationary log-periodogram regression. *Journal of Econometrics* **91**, 299–323.
Velasco, C. (1999b) Gaussian semiparametric estimation of nonstationary time series. *Journal of Time Series Analysis* **20**, 87–127.
Velasco, C. and P.M. Robinson (2000) Whittle pseudo maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association* **95**, 1229–43.
Wang, C. (2004) Futures trading activity and predictable foreign exchange movements. *Journal of Banking and Finance* **28**, 1023–41.
Wang, W., P.H.A., M. van Gelder, J.K. Vrijling and J. Ma (2005) Forecasting daily streamflow using hybrid ANN models. *Journal of Hydrology* **324**, 383–99.
Wang, W., P.H.A., M. van Gelder, J.K. Vrijling and X. Chen (2007) Detecting long memory: Monte Carlo simulations and application to daily streamflow processes. *Hydrology and Earth System Sciences* **11**, 851–62.

This page intentionally left blank

Part IV
Cross-section and Panel Data Applications

This page intentionally left blank

11
Discrete Choice Modeling

William Greene

Abstract

We detail the basic theory for models of discrete choice. This encompasses methods of estimation and analysis of models with discrete dependent variables. Entry level theory is presented for the practitioner. We then describe a few of the recent, frontier developments in theory and practice.

11.1 Introduction

This chapter will survey models for outcomes that arise through measurement of discrete consumer choices, such as whether to vote for a particular candidate, whether to purchase a car, how to get to work, whether to purchase insurance, where to shop, or whether to rent or buy a home or a car. Traditional economic theory for consumer choice – focused on utility maximization over bundles of continuous commodities – is relatively quiet on the subject of discrete choice among a set of specific alternatives. Econometric theory and applications, in contrast, contain a vast array of analyzes of discrete outcomes; discrete choice modeling has been one of the most fruitful areas of study in econometrics for several decades. There is a useful commonality in much of this. One can build an overview of models for discrete outcomes on a platform of individual maximizing behavior. Given that the literature is as vast as it is, and we have but a small number of pages within which to package it, this seems like a useful approach. In what follows, we will survey some of the techniques used to analyze individual random utility maximizing behavior.

We emphasize that we have chosen to focus on models for discrete *choice*, rather than models for discrete *dependent variables*. This provides us with several opportunities to focus and narrow this review. First, it allows us to limit the scope of the survey to a reasonably manageable few types of models. As noted, the literature on this topic is vast. We will use this framework to select a few classes of models that are used by analysts of individual choice. It also gives us license to make a few major omissions that might otherwise fall under the umbrella of discrete outcomes. One conspicuous case will be models for counts. Event counts are obviously discrete – models for them are used to study, e.g., traffic incidents, incidence of disease, health care system utilization, credit and financial markets, and an array of other settings. Models for counts can occupy an entire library of its own in this area – two excellent references are Cameron and Trivedi (1998) and Winkelmann (2003) – but this area will extend far beyond our reach. On the other hand, applications in health economics (system utilization) and industrial organization (patents and innovations) do lead to some settings in which individual or firm choice produces a count response. We will briefly consider models for counts from this standpoint. The reader will no doubt note other areas of discrete response analysis that are certainly important. Space limitations force us to consider a fairly small number of cases.

This chapter proceeds as follows. Section 11.2 details the estimation and inference tools used throughout the remainder of the survey, including the basic results in maximum likelihood estimation. Section 11.3 analyzes in detail the fundamental pillar of analysis of discrete choice, the model for binary choice – the choice between two alternatives. Most of the applications that follow are obtained by extending or building on the basic binary choice model. Thus we examine the binary choice model in greater detail than the others, as it also provides a convenient setting in which to develop the estimation and inferential concepts that carry over to the other models. Section 11.4 considers the immediate extension of the binary choice, bivariate and multivariate binary choice models. Section 11.5 returns to the single choice setting and examines ordered choice models. Models for count data are examined in section 11.6. Finally, section 11.7 turns to an area of literature in its own right, multinomial choice modeling. As before, but even more so here, we face the problem of surveying a huge literature in a few pages. We therefore describe the most fundamental elements of multinomial choice analysis, and point the reader toward more detailed sources in the literature. Section 11.8 concludes.

11.2 Specification, estimation and inference for discrete choice models

The classical theory of consumer behavior provides the departure point for economic models of discrete individual choice.[1] A representative consumer with preferences represented by a utility function defined over the consumption of a vector of goods, $U(\mathbf{d})$, is assumed to maximize this utility subject to a budget

constraint, $\mathbf{x}'\mathbf{d} \leq y$, where $\mathbf{x}$ is a vector of prices and y is income (or total expenditure). Assuming the necessary continuity and curvature conditions, a complete set of demand equations, $\mathbf{d}^* = \mathbf{d}(\mathbf{x}, y)$ results.[2] To extend the model of individual choice to observed market data, the demand system is assumed to hold at the aggregate level, and random elements (disturbances) are introduced to account for measurement error or optimization errors.

Since the 1960s, the availability of survey data on individual behavior has obviated the heroic assumption underlying the aggregate utility function or the (perhaps slightly less heroic) assumptions underlying the aggregate demand system. That progression has evolved to the contemporary literature with the appearance of large, detailed, high quality panel surveys, such as the German Socio-Economic Panel Survey (GSOEP) (see Hujer and Schneider, 1989) that we will use in this study and the British Household Panel Survey (BHPS) (http://www.iser.essex.ac.uk/ulsc/bhps), to name only two of many. The analysis of individual data to which the original theory applies has called for (at least) two more detailed developments of that theory.

First, the classical theory has relatively little to say about the *discrete* choices that consumers make. Individual data detail career choices, voting preferences, travel mode choices, discretized measures of the strength of preferences, and participation decisions of all sorts, such as labor supply behavior, whether to make a large purchase, whether to migrate, etc. The classical, calculus based theory of decisions made at the margins of consumption will comment on, e.g., how large a refrigerator a consumer will buy, but not whether they will buy a refrigerator instead of a car (this year), or what brand of car or refrigerator they will buy.

Second, the introduction of random elements in models of choice behavior as *disturbances* is much less comfortable at the individual level than in market demands. Researchers have considered more carefully the appropriate sources and form of random variation in individual models of discrete choice.

The *random utility model* of discrete choice provides the most general platform for the analysis of discrete choice. The extension of the classical theory of utility maximization to the choice among multiple discrete alternatives provides a straightforward framework for analyzing discrete choice in probabilistic, statistical, ultimately econometric, terms.

11.2.1 Discrete choice models and discrete dependent variables

Denote by "i" a consumer who is making a choice among a set of J_{it} choices in choice situation t. To put this in a context, which will help to secure the notation, envision a *stated choice experiment* in which individual i is offered the choice of several, J_{i1}, brands of automobiles with differing prices and characteristics and asked which they most prefer. In a second round of the experiment, the interviewer changes some of the features of some of the cars, and repeats the question. Denote by $A_{it,1}, \ldots, A_{it,Jit}$, $J_{it} \geq 2$, the set of alternatives available to the individual in choice situation t. It will be convenient to adopt the panel data notation, in which "t" denotes "*time*." The generality of the notation allows the choice set to vary from one individual to another, and across choice situations for the same individual. In

most of what follows, we will not need this level of generality, but the models to be developed will accommodate it.

We will formulate a model that describes the consumer choice in probabilistic terms. (A bit more of the underlying behavioral theory is presented in section 11.8.) The "model" will consist of a probability distribution defined over the set of choices:

$$P_{it,j} = \text{Prob(consumer } i \text{ makes choice } j \text{ at time } t \mid \text{choice set}), \quad j = 1, \ldots, J_{it}.$$

The manner in which the probabilities arise is an essential feature of the model. As noted earlier, choices are dependent on the environment in which they are made, which we characterize in terms of income, y, and prices, $\mathbf{x}$. Individual heterogeneity may be measured by such indicators as family size, gender, location, etc., which we collect in a set of variables, $\mathbf{z}$, and unmeasured, and therefore random from the point of view of the analyst, indicators, which we denote as $\mathbf{u}$. Common elements of the choice mechanism that constitute the interesting quantities that the analyst seeks to draw statistical inference about will be parameters, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, etc.[3] For purposes of translating the underlying choice process into an estimable econometric model, we define the choice indicators:

$$d_{it,j} = 1 \text{ if individual } i \text{ makes choice } j \text{ at time } t, \text{ and 0 otherwise.}$$

With all this in place, our discrete probability distribution will be defined by:

$$P_{it,j} = \text{Prob}\left(d_{it,j} = 1 | \mathbf{X}_{it}, \mathbf{z}_{it}, \mathbf{u}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \ldots\right), \quad j = 1, \ldots, J_{it}$$

where $\mathbf{X}_{it}$ is the set of attributes of all J_{it} choices in the choice set for individual i at time t. Note that being characteristics of the individual, and not the choices, $\mathbf{z}_{it}$ and $\mathbf{u}_{it}$ do not vary across the choices. Whether the preference parameters, $\boldsymbol{\beta}, \boldsymbol{\gamma}, \ldots$, should be allowed to vary (i.e., whether they do vary) across individuals – i.e., whether the parameters of the utility functions are heterogeneous – is a question that we will pursue at several points below. We will assume (not completely innocently) that in any choice situation, the individual actually makes a choice. It follows that:

$$\sum_{j=1}^{J_{it}} d_{it,j} = 1 \quad \text{and} \quad \sum_{j=1}^{J_{it}} P_{it,j} = 1.$$

The "model" consists of the interesting or useful features of $P_{it,j}$. The preceding discussion assumes that, at time t, the consumer makes a single decision. It will be necessary in section 11.4 to extend the model to cases of two or more decisions. This is straightforward, but requires a small change in notation and interpretation. We will defer that extension until we encounter it in the discussion in section 11.4.

We close this section with some definitions of terms that will be used throughout the text. The individual *characteristics*, such as gender or education, are denoted $\mathbf{z}_{it}$. Attributes of the choices, such as prices, are denoted $x_{it,j}$. We denote by *binomial* or *multinomial choice*, the single choice made between either two or more than two

choices. The term *binary choice* is often used interchangeably with the former. A *bivariate choice* or *multivariate choice* is the set of two or more choices made in a single choice situation. In one of our applications, an individual chooses not to visit a physician or to visit at least once; this is a binomial choice. This coupled with a second decision, whether to visit the hospital, constitutes a bivariate choice. In a different application, the choice of which of four modes to use for travel constitutes a multinomial choice.

11.2.2 Estimation and inference

"Estimation" in this setting is less clearly defined than in the familiar linear regression model. If the model is fully parametric, then the way that the parameters interact with the variables in the model, and the particular function that applies to the problem, are all fully specified. The model is then:

$$P_{it,j} = F_{it}\left(j, \mathbf{X}_{it}, \mathbf{z}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}_{it}\right) \quad j = 1, \ldots, J_u.$$

We will consider models that accommodate unobserved individual heterogeneity, $\mathbf{u}_{it}$, in sections 11.6 and 11.7. For the present, to avoid an inconvenience in the formulation, we consider a model involving only the observed data. Various approaches to estimation of parameters and derivative quantities in this model have been proposed, but the likelihood based estimator is by far the method of choice in the received literature. The log-likelihood for the model is:

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{j=1}^{J_{it}} d_{it,j} \ln F_{it}(j, \mathbf{X}_{it}, \mathbf{z}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}), \quad i = 1, \ldots, n \quad t = 1, \ldots, T_i.$$

The maximum likelihood estimator is that function of the data that maximizes $\ln L$.[4] (See, e.g., Greene, 2008a, Ch. 14, for discussion of maximum likelihood estimation.) The Bayesian estimator will be the mean of the posterior density:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{D}, \mathbf{X}, \mathbf{Z}) = \frac{L \times g(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\int_{\beta,\gamma} L \times g(\boldsymbol{\beta}, \boldsymbol{\gamma}) d\boldsymbol{\beta} d\boldsymbol{\gamma}}.$$

where $g(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the prior density for the model parameters and (**D**,**X**,**Z**) is the full sample of data on all variables in the model. (General discussions of Bayesian methods may be found in Koop, 2003; Lancaster, 2004; Geweke, 2005.) Semiparametric methods, generally in the index form, but without a specific distributional assumption, are common in the received literature, particularly in the analysis of binary choices and panel data. These will be considered briefly in sections 11.3.4 and 11.7.2. Nonparametric analysis of discrete choice data is on the frontier of the theory, and does not play much of a role in the empirical literature. We will note this segment of the development briefly in section 11.3.4.

Estimation and inference about model parameters is discussed in the sections to follow. Though the model is commonly formulated as an "index function" model, i.e.:

$$P_{it,j} = F_{it}\left(j, \mathbf{X}_{it}'\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}\right) \quad j = 1, \ldots, J_{it},$$

even in this form, it will generally bear little resemblance to the linear regression model. As in other nonlinear cases, the interpretation of the model coefficients is ambiguous. Partial effects on the probabilities associated with the choices for individual i at time t are defined as:

$$\boldsymbol{\delta}_{it}(j, \mathbf{X}_{it}'\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) = \partial F_{it}(j, \mathbf{X}_{it}'\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma})/\partial \begin{pmatrix} \mathbf{x}_{it,j} \\ \mathbf{z}_{it} \end{pmatrix} = F_{it}'(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}.$$

These are likely to be of interest for particular individuals, or averaged across individuals in the sample. A crucial implication for use of the model is that these partial effects may be quite different from the coefficients themselves. Since there is no "regression" model at work, this calls into question the interpretation of the model and its parts. No generality is possible at this point. We will return to the issue below.

A related exercise in marginal analysis of the sample and estimated model is to examine the aggregate outcomes predicted by the model:

$$\hat{n}_{t,j} = \sum_{i=1}^{n} \hat{F}_{it}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) = \sum_{i=1}^{n} \hat{d}_{it,j}.$$

where the "ˆ" indicates the estimate from the model. For example, if $x_{it,j,k}$ denotes a policy variable, a price or a tax, say, we might be interested in:

$$\Delta\hat{n}_{t,j} = \sum_{i=1}^{n} \hat{F}_{it}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}|x_{it,j,k}^{1}) - \sum_{i=1}^{n} \hat{F}_{it}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}|x_{it,j,k}^{0}).$$

Although the subject of the impact in the partial effect is already scaled – it is a probability between zero and one – it is still common for researchers to report elasticities of probabilities rather than partial effects. These are:

$$\eta_{it,j}(\text{variable}_{it,j,k}) = \frac{F_{it}'(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma})}{F_{it}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma})} \times \text{variable}_{it,j,k} \times \text{coefficient}_{k}.$$

This is prominently the case in the analysis of multinomial choice models, as we will explore in section 11.7.

Finally, again because the model does not correspond to a regression except in a very loose sense, the concept of fit measures is also ambiguous. There is no counterpart to "explained variation" or "total variation" in this class of models, so the idea behind the coefficient of determination (R^2) in linear regression has no meaning here. What is required to assess the fit of the model is, first, a specification of how the model will be used to predict the outcome (choice), then an assessment of how well the estimated model does in that regard.

11.2.3 Application

It will be helpful in the exposition below to illustrate the computations with a few concrete examples based on "live" data. We will use two familiar datasets. The RWM

Health Care data (our appelation) was used in Riphahn, Wambach and Million (2003) to analyze utilization of the German health care system. The dataset used is an unbalanced panel of 7,293 individual families observed over seven periods. It is part of the GSOEP, which can be downloaded from the archive site of the *Journal of Applied Econometrics* (http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/). We will use these to illustrate the single equation and panel data binary and ordered choice models and models for counts presented in sections 11.3–11.6. The second dataset is also widely used to illustrate multinomial choice models. These data, from Hensher and Greene (e.g., 2003), are a survey of 210 travelers between Sydney and Melbourne who chose among four modes, air, train, bus and car. We will use these data to illustrate a few multinomial choice models in section 11.7.

11.3 Binary choice

The second fundamental building block in the development of discrete choice models (after the model of random utility) is the basic model for choice between two alternatives. We would formulate this in a random utility framework with the utility of two choices:

$$U_{i,1} = \mathbf{x}_{i,1}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,1}$$
$$U_{i,2} = \mathbf{x}_{i,2}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,2}.$$

For convenience at this point, we assume there is a single choice made, so $T_i = 1$. The utility functions are in the index form, with characteristics and attributes and common (generic) coefficients. The random terms, $\varepsilon_{i,1}$ and $\varepsilon_{i,2}$, represent unmeasured influences on utility. (Looking forward, without these random terms, the model would imply that with sufficient data (and consistent parameter estimators), utility could be "observed" exactly, which seems improbable at best.) Consistent with the earlier description, the analyst observes the choice most preferred by the individual, that is, the one with the greater utility, say choice 1. Thus, the observed outcome reveals that:

$$U_{i,1} > U_{i,2},$$

or:

$$\mathbf{x}_{i,1}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,1} > \mathbf{x}_{i,2}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,2},$$

or:

$$(\mathbf{x}_{i,1}'\boldsymbol{\beta} - \mathbf{x}_{i,2}'\boldsymbol{\beta}) + (\mathbf{z}_i'\boldsymbol{\gamma} - \mathbf{z}_i'\boldsymbol{\gamma}) > (\varepsilon_{i,2} - \varepsilon_{i,1}), \tag{11.1}$$

or:

$$(\mathbf{x}_{i,1} - \mathbf{x}_{i,2})'\boldsymbol{\beta} > (\varepsilon_{i,2} - \varepsilon_{i,1}).$$

This exercise reveals several identification problems in the model as stated so far. First, we have implicitly assumed that, in the event that the two utilities are equal, the consumer chooses alternative 2. This is a normalization: recall that we assumed

earlier that the individual makes exactly one choice. Second, it is evident that, in describing the choice of process in this fashion, it is the relative values of the attributes of the choices that matter: the difference between $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$ is the determinant of the observed outcome, not the specific values of either. Third, note that the choice of invariant component, $\mathbf{z}_i$, has fallen out of the choice process. The implication is that, unless the characteristics influence the utilities differently, it is not possible to measure their impact on the choice process. Finally, $\varepsilon_{i,1}$ and $\varepsilon_{i,2}$ are random variables with so far unspecified means and variances. With respect to the means, if they are μ_1 and μ_2, only $\mu_2 - \mu_1$ enters the choice. As such, if the means were $\mu_1 + \phi$ and $\mu_2 + \phi$, the same outcome would be observed. These means cannot be measured with observed data, so at least one is normalized to zero. Finally, consider the outcome of scaling both utilities by an arbitrary constant, σ. The new random components would be $\sigma\varepsilon_{i,1} = \varepsilon^*_{i,1}$ and $\sigma\varepsilon_{i,2} = \varepsilon^*_{i,2}$, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ would be scaled likewise. However, this scaling of the model would have no impact on the observed outcome in the last line of equation (11.1). The same choice would be observed whatever positive value σ takes. Thus, there is yet one more indeterminacy in the model. This can be resolved in several ways. The most common expedient is to normalize the scaling of the random components to one.

Combining all of these, we obtain a conventional form of the model for the choice between two alternatives:

$$\begin{aligned}
&\Delta U_i = \mu + (\Delta \mathbf{x}_i)'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad E[\varepsilon_i|\mathbf{X}_i, \mathbf{z}_i] = 0, \quad Var[\varepsilon_i|\mathbf{X}_i, \mathbf{z}_i] = 1.\\
&d_{i1} = 1 \text{ if } \Delta U_i > 0 \text{ and } d_{i1} = 0 \text{ otherwise},\\
&d_{i2} = 1 - d_{i1}.
\end{aligned}$$

In a more familiar arrangement, we would have:

$$\begin{aligned}
d_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i\\
d_i &= 1 \text{ if } d_i^* > 0, \text{ and } d_i = 0 \text{ otherwise,}
\end{aligned} \tag{11.2}$$

where $d_i = 1$ indicates that choice 1 is selected and where the correspondence to the components of the more detailed model is direct.

11.3.1 Regression models

The preceding sections describe an underlying theoretical platform for a binary choice, based on a model of random utility. In order to translate it into an econometric model, we will add the assumptions behind the stochastic component of the specification, ε_i. To this point, the specification is semiparametric. We have not assumed anything specific about the underlying distribution, only that ε_i represents the random (from the point of view of the econometrician) element in the utility function of individual *i*. The restrictions imposed (zero mean, unit variance) are normalizations related to the identification issue and are not intended to be substantive restrictions on behavior. (Indeed, the unit variance assumption turns out to be unnecessary for some treatments. We will return to this below.)

We can approach the specification in equation (11.2) from a different viewpoint. The random utility approach specifies that d_i^* represents the strength of the individual's preference for alternative 1 relative to alternative 2. An alternative approach regards (11.2) as a *latent regression model*. The dependent variable is assumed to be unobservable; the observation is a censored variable that measures d_i^* relative to a benchmark, zero. For an example, consider a model of loan default. One would not typically think of loan default as a utility maximizing choice. On the other hand, in the context of (11.2), one might think of d_i^* as a latent measure of the financial distress of individual i. If d_i^* is high enough, the individual defaults, and we observe $d_i = 1$. By this construction, the appropriate model for d_i is a *censored regression*. Once we endow ε_i with a proper probability distribution, (11.2) can be construed as a regression model.

With the assumption of a specific distribution for ε_i, we obtain a statement of the choice probabilities:

$$\begin{aligned}\text{Prob}(d_i = 1|\mathbf{X}_i, \mathbf{z}_i) &= \text{Prob}(d_i^* > 0|\mathbf{X}_i, \mathbf{z}_i) \\ &= \text{Prob}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i > 0). \\ &= \text{Prob}[\varepsilon_i > -(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})] \\ &= 1 - \text{Prob}[\varepsilon_i \leq -(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})].\end{aligned}$$

It follows that:

$$\begin{aligned}E[d_i|\mathbf{X}_i, \mathbf{z}_i] &= 0 \times \text{Prob}(d_i = 0|\mathbf{X}_i, \mathbf{z}_i) + 1 \times \text{Prob}(d_i = 1|\mathbf{X}_i,\ \mathbf{z}_i) \\ &= \text{Prob}(d_i = 1|\mathbf{X}_i,\ \mathbf{z}_i),\end{aligned}$$

so we now have a regression model to manipulate as well. The implied probability endowed by our assumption of the distribution of ε_i becomes the regression of d_i on $\mathbf{X}_i$ and $\mathbf{z}_i$. By this construction, one might bypass the random utility apparatus, and simply embark on modeling:

$$\begin{aligned}d_i &= E[d_i|\mathbf{X}_i, \mathbf{z}_i] + a_i \\ &= \text{Prob}(d_i = 1|\mathbf{X}_i, \mathbf{z}_i) + a_i,\end{aligned}$$

where, by construction, a_i has zero mean, conditioned on the probability function. A remaining step is to construct the appropriate conditional mean function. This specification has suggested in some settings the *linear probability model*:

$$d_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + a_i.$$

(See, e.g., Aldrich and Nelson, 1984; Caudill, 1988; Heckman and Snyder, 1997; Angrist, 2001.) The linear probability model has some significant shortcomings, the most important of which is that the linear function cannot be constrained to lie between zero and one, so its interpretation as a probability model is suspect. With few exceptions, including those noted above, researchers have employed

proper probability models for the implied regressions. The logit and probit models described in the next section are the overwhelming choices in the received literature.

11.3.2 Estimation and inference in parametric binary choice models

A parametric model is completed by specifying a distribution for ε_i. Many candidates have been proposed, though there is little in the way of observable evidence that one can use to choose among the candidates.[5] For convenience, we will assume a symmetric distribution, such as the normal or logistic which are used in the overwhelming majority of studies. For a symmetric distribution:

$$\begin{aligned} 1 - \text{Prob}[\varepsilon_i \leq -(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})] &= \text{Prob}(\varepsilon_i \leq \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) \\ &= F(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}). \end{aligned}$$

Once again relying on the symmetry of the distribution, the probabilities associated with the two outcomes are:

$$\text{Prob}(d_i = 1|\mathbf{x}_i, \mathbf{z}_i) = F(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}),$$

and:

$$\text{Prob}(d_i = 0|\mathbf{x}_i, \mathbf{z}_i) = F[-(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})].$$

For the two outcomes $d_i = j, j = 0, 1$, these may be combined in the form suggested earlier:

$$F(j, \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) = F[(2j - 1)(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})],$$

where:

$$F(t) = \Lambda(t) = \frac{\exp(t)}{1 + \exp(t)} \text{ for the logistic distribution,}$$

and:

$$F(t) = \Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)dz \text{ for the normal distribution.}$$

The assumption of the logistic distribution gives rise to the logit model, while the normal distribution produces the probit model.

11.3.2.1 Parameter estimation

The model is now fully parameterized, so the analysis can proceed based either on the likelihood function or the posterior density. We consider the maximum likelihood estimator (MLE) first, and the Bayesian estimator in section 11.3.3.

The log-likelihood function for the observed data is:

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln \text{Prob}(d_i|\mathbf{x}_i, \mathbf{z}_i) \\ &= \sum_{d_i=1} \ln \text{Prob}(d_i = 1|\mathbf{x}_i, \mathbf{z}_i) + \sum_{d_i=0} \ln \text{Prob}(d_i = 0|\mathbf{x}_i, \mathbf{z}_i) \\ &= \sum_{i=1}^{n} \ln F[(2d_i - 1)(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})].\end{aligned}$$

Estimation by maximizing the log-likelihood is straightforward for this model. The gradient of the log-likelihood is:

$$\frac{\partial \ln L}{\partial \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}} = \sum_{i=1}^{n} (2d_i - 1) \frac{F'[(2d_i - 1)(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})]}{F[(2d_i - 1)(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})]} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} = \sum_{i=1}^{n} \mathbf{g}_i = \mathbf{g}.$$

The maximum likelihood estimators of the parameters are found by equating $\mathbf{g}$ to zero, an optimization problem that requires an iterative solution.[6] For convenience in what follows, we will define:

$$q_i = (2d_i - 1), \mathbf{w}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, t_i = q_i \mathbf{w}_i'\boldsymbol{\theta}, F_i = F(t_i), F_i' = dF_i/dt_i = f_i.$$

(Thus, F_i is the cumulative density function (c.d.f.) and f_i is the density for the assumed distribution.) It follows that:

$$\mathbf{g}_i = q_i F_i'(t_i)\mathbf{w}_i = q_i f_i \mathbf{w}_i.$$

Statistical inference about the parameters is made using one of the three conventional estimators of the asymptotic covariance matrix: the Berndt, Hall, Hall and Hausman (BHHH) (1974) estimator, based on the outer products of the first derivatives:

$$\mathbf{V}_{\text{BHHH}} = \left[\sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i'\right]^{-1},$$

the actual Hessian:

$$\mathbf{V}_{\text{H}} = \left[-\sum_{i=1}^{n} \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} = \left[-\sum_{i=1}^{n} \frac{F_i F_i'' - (F_i')^2}{F_i^2}, \mathbf{w}_i \mathbf{w}_i'\right]^{-1},$$

or the expected Hessian, which can be shown to equal:

$$\mathbf{V}_{\text{EH}} = \left[-\sum_{i=1}^{n} E_{d_i}\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)\right]^{-1} = \left[-\sum_{i=1}^{n} \frac{f(\mathbf{w}_i'\boldsymbol{\theta}) f(-\mathbf{w}_i'\boldsymbol{\theta})}{F_i(1 - F_i)} \mathbf{w}_i \mathbf{w}_i'\right]^{-1}.$$

It has become common, even *de rigueur*, to compute a "robust" covariance matrix for the MLE using $\mathbf{V}_{\text{H}} \times \mathbf{V}_{\text{BHHH}}^{-1} \times \mathbf{V}_H$, under the assumption that the MLE is robust to failures of the specification of the model. In fact, there is no obvious failure of

the assumptions of the model (distribution, omitted variables, heteroskedasticity, and correlation across observations) for which the MLE remains consistent, so the virtue of the "corrected" covariance matrix is questionable (see Freedman, 2006).

For the two distributions considered here, the derivatives are relatively simple. For the logistic:

$$F(t) = \Lambda(t), \quad f(t) = F'(t) = \Lambda(t)[1 - \Lambda(t)], \quad F''(t) = F'(t)[1 - 2\Lambda(t)].$$

For the normal distribution (probit model), the counterparts are:

$$F(t) = \Phi(t), \quad f(t) = F'(t) = \phi(t), \quad F''(t) = -t\phi(t).$$

In both cases, $f(t) = f(-t)$ and $F(-t) = 1 - F(t)$. For estimation and inference purposes, a further convenient result is, for the logistic distribution:

$$-[F(t)F''(t) - (F'(t))^2]/F(t)^2 = \Lambda(t)(1 - \Lambda(t)) > 0 \text{ for all } t,$$

while for the normal distribution:

$$-[F(t)F''(t) - (F'(t))^2]/F(t)^2 = t[\phi(t)/\Phi(t)] + [\phi(t)/\Phi(t)]^2 > 0 \text{ for all } t.^{7}$$

The implication is that both the second derivatives matrix and the expected second derivatives matrix are negative definite for all values of the parameters and data. Optimization using Newton's method or the method of scoring will always converge to the unique maximum of the log-likelihood function, so long as the weighting matrix ($\mathbf{V}_{\text{BHHH}}$, $\mathbf{V}_{\text{H}}$ or $\mathbf{V}_{\text{EH}}$) is not singular.[8]

11.3.2.2 Residuals and predictions

Two additional useful results are obtained from the necessary conditions for maximizing the log-likelihood function. First, the component of the score function that corresponds to the constant term is:

$$\sum_{i=1}^{n} q_i \frac{F'(q_i \mathbf{w}_i'\boldsymbol{\theta})}{F(q_i \mathbf{w}_i'\boldsymbol{\theta})} = 0.$$

The terms in this sum are the *generalized residuals* of the model. As do the ordinary residuals in the regression model, the generalized residuals sum to zero at the MLE. These terms have been used for specification testing in this model (see Chesher and Irish, 1987). For the logit model, it can be shown that the result above implies that:

$$\frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} F(\mathbf{w}_i'\boldsymbol{\theta}),$$

when F is evaluated at the MLEs of the parameters. The implication is that the average of the predicted probabilities from the logit model will equal the proportion of the observations that are equal to one. A similar (albeit inexact) outcome will be seen in empirical results for the probit model. The theoretical result has not been shown analytically.

11.3.2.3 Marginal effects

Partial effects in the binary choice model are computed for continuous variables using the general result:

$$\boldsymbol{\delta}_i = \frac{\partial \text{Prob}(d_i = 1|\mathbf{w}_i)}{\partial \mathbf{w}_i} = f(\mathbf{w}_i'\boldsymbol{\theta})\boldsymbol{\theta}.$$

For a binary variable, such as gender or degree attained, the counterpart would be:

$$\Delta_i = F(\mathbf{w}_i'\boldsymbol{\theta} + \gamma_k) - F(\mathbf{w}_i'\boldsymbol{\theta})$$

where γ_k is the coefficient on the dummy variable of interest (assumed to be a characteristic of the individual). These are typically evaluated for the average individual in the sample, though current practice somewhat favors the *average partial effect*:

$$\begin{aligned}\overline{\boldsymbol{\delta}} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \text{Prob}(d_i=1|\mathbf{w}_i)}{\partial \mathbf{w}_i} \\ &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}_i'\boldsymbol{\theta})\boldsymbol{\theta} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}_i'\boldsymbol{\theta})\right)\boldsymbol{\theta}.\end{aligned}$$

(The two estimators will typically not differ substantively.) Standard errors for partial effects are usually computed using the delta method. Let $\mathbf{V}$ denote the estimator of the asymptotic covariance matrix of the MLE of $\boldsymbol{\theta}$. For a particular vector, $\mathbf{w}_i$:

$$\boldsymbol{\Gamma}_i = \frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{\theta}'} = \left[f'(\mathbf{w}_i'\boldsymbol{\theta})\right] I + [f(\mathbf{w}_i'\boldsymbol{\theta})]\boldsymbol{\theta}\mathbf{w}_i'.$$

For a binary variable in the model, in addition to (or in) the $\mathbf{w}_i$, the corresponding row of $\boldsymbol{\Gamma}_i$ would be:

$$\boldsymbol{\Gamma}_{i,k} = \partial \Delta_{i,k}/\partial(\boldsymbol{\theta}', \gamma_k) = f(\mathbf{w}_i'\boldsymbol{\theta} + \gamma_k)[\mathbf{w}_i, 1] - f(\mathbf{w}_i'\boldsymbol{\theta})[\mathbf{w}_i, 0].$$

For the particular choice of $\mathbf{w}_i$, then, the estimator of the asymptotic covariance matrix for $\boldsymbol{\delta}_i$ would be $\boldsymbol{\Gamma}_i\mathbf{V}\boldsymbol{\Gamma}_i'$, computed at the maximum likelihood estimates. It is common to do this computation at the means of the data, $\overline{\mathbf{w}} = \frac{1}{n}\Sigma_{i=1}^{n}\mathbf{w}_i$. For the average partial effect, the computation is complicated a bit because the terms in $\overline{\boldsymbol{\delta}}$ are correlated – they use the same estimator of the parameters – so the variance of the mean is not $(1/n)$ times the sum of the variances. It can be shown (see Greene, 2008, Ch. 23) that the appropriate computation reduces to:

$$\text{Est.Asy.Var}[\overline{\boldsymbol{\delta}}] = \overline{\boldsymbol{\Gamma}}\mathbf{V}\overline{\boldsymbol{\Gamma}}', \text{ where } \overline{\boldsymbol{\Gamma}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}_i.$$

An alternative approach to computing standard errors for the marginal effects is the method of Krinsky and Robb (1986). A set of R random draws is taken from the estimated (asymptotic) normal population with mean $\hat{\boldsymbol{\theta}}_{MLE}$ and variance $\mathbf{V}$ and the empirical mean squared deviation of the estimated partial effects is computed using the MLE:

$$\text{Est.Asy.Var}[\overline{\boldsymbol{\delta}}] = \frac{1}{R}\sum_{r=1}^{R}\left(\overline{\boldsymbol{\delta}}_r - \overline{\boldsymbol{\delta}}\right)\left(\overline{\boldsymbol{\delta}}_r - \overline{\boldsymbol{\delta}}\right)',$$

where $\overline{\boldsymbol{\delta}}_r$ is computed at the random draw and $\overline{\boldsymbol{\delta}}$ is computed at $\hat{\boldsymbol{\theta}}_{MLE}$.

An empirical conundrum can arise when doing inference about partial effects rather than coefficients. For any particular variable, $\mathbf{w}_k$, the preceding theory does not guarantee that both the estimated coefficient, θ_k, and the associated partial effect, δ_k, will be "statistically significant," or statistically insignificant. In the event of a conflict, one is left with the uncomfortable problem of simultaneously rejecting and not rejecting the hypothesis that a variable should appear in the model. Opinions differ on how to proceed. Arguably, the inference should be about θ_k, not δ_k, since in the latter case, one is testing a hypothesis about a function of all the coefficients, not just the one of interest.

11.3.2.4 Hypothesis tests

Conventional hypothesis tests about restrictions on the model coefficients, $\boldsymbol{\theta}$, can be carried out using any of the three familiar procedures. Given the simplicity of the computations for the MLE, the likelihood ratio test is a natural candidate. The likelihood ratio statistic is:

$$\lambda_{\text{LR}} = 2[\ln L_1 - \ln L_0]$$

where "1" and "0" indicate the values of the log-likelihood computed at the unrestricted (alternative) estimator and the restricted (null) estimator, respectively. A hypothesis that is usually of interest in this setting is the null hypothesis that all coefficients save for the constant term are equal to zero. In this instance, it is simple to show that, regardless of the assumed distribution:

$$\ln L_0 = n[P_1 \ln P_1 + P_0 \ln P_0],$$

where P_1 is the proportion of observations for which d_i equals one, which is also $\overline{d} = \frac{1}{n}\Sigma_{i=1}^{n} d_i$, and $P_0 = 1 - P_1$. Wald statistics use the familiar results, all based on the unrestricted model. The general procedure assesses departures from the null hypothesis

$$\text{H}_0 : \mathbf{r}(\boldsymbol{\theta}, \mathbf{c}) = \mathbf{0},$$

where $\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})$ is a vector of J functionally independent restrictions on $\boldsymbol{\theta}$ and $\mathbf{c}$ is a vector of constants. The typical case is the set of linear restrictions, $\text{H}_0 : \mathbf{r}\boldsymbol{\theta} - \mathbf{c} = 0$, where $\mathbf{r}$ is a matrix of constants. The Wald statistic for testing the null hypothesis is constructed using the delta method to obtain an asymptotic covariance matrix for $\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})$. The statistic is:

$$\lambda_{\text{WALD}} = [\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})]'[\mathbf{R}(\boldsymbol{\theta}, \mathbf{c})\mathbf{V}\mathbf{R}(\boldsymbol{\theta}, \mathbf{c})']^{-1}[\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})],$$

where $\mathbf{R}(\boldsymbol{\theta}, \mathbf{c}) = \partial \mathbf{r}(\boldsymbol{\theta}, \mathbf{c})/\partial \boldsymbol{\theta}'$ and all computations are carried out using the unrestricted maximum likelihood estimator. The standard "t-test" of the significance of a coefficient is the most familiar example. The Lagrange multiplier (LM) statistic is:

$$\lambda_{\text{LM}} = \mathbf{g}^{0\prime}\mathbf{V}^0\mathbf{g}^0,$$

where "0" indicates that the computations are done using the restricted estimator and $\mathbf{V}$ is any of the estimators of the asymptotic covariance matrix of the MLE

mentioned earlier. Using $\mathbf{V}_{\text{BHHH}}$ produces a particularly convenient computation, as well as an interesting and surprisingly simple test of the null hypothesis that all coefficients save the constant are zero. Using $\mathbf{V}_{\text{BHHH}}$ and expanding the terms, we have:

$$\lambda_{\text{LM}} = \left(\sum_{i=1}^{n} q_i \mathbf{w}_i f_i^0\right)' \left(\sum_{i=1}^{n} q_i^2 (f_i^0)^2 \mathbf{w}_i \mathbf{w}_i'\right)^{-1} \left(\sum_{i=1}^{n} q_i \mathbf{w}_i f_i^0\right),$$

and an immediate simplification occurs because $q_i^2 = 1$. The density is computed at the restricted estimator, however obtained. If the null hypothesis is that all coefficients are zero save for the constant, then, for the logit model, $f_i^0 = f^0 = P_1(1-P_1)$. For the probit model, the estimator of the constant term will be $\Phi^{-1}(P_1)$ and $f^0 = \phi[\Phi^{-1}(P_1)]$. Taking this constant outside the summation in $\mathbf{g}$ leaves $\Sigma_{i=1}^{n} q_i \mathbf{w}_i = n[P_1 \overline{\mathbf{w}}_1 - P_0 \overline{\mathbf{w}}_0]$, where $\overline{\mathbf{w}}_1$ is the sample mean of the n_1 observations with d_i equal to one and $\overline{\mathbf{w}}_0$ is the mean of the n_0 remaining observations. Note that the constant f^0 falls out of the resulting statistic, and we are left with the LM statistic for testing this null hypothesis:

$$\lambda_{\text{LM}} = n^2 [P_1 \overline{\mathbf{w}}_1 - P_0 \overline{\mathbf{w}}_0]' (\mathbf{W}'\mathbf{W})^{-1} [P_1 \overline{\mathbf{w}}_1 - P_0 \overline{\mathbf{w}}_0],$$

where $\mathbf{W}$ is the data matrix with ith row equal to $\mathbf{w}_i'$. As in the case of the likelihood ratio (LR) statistic, the same computation is used for both the probit and logit models.

11.3.2.5 Specification tests

Two specification issues are typically addressed in the context of these parametric models, heteroskedasticity and the distributional assumption. For the former, since there are no useful "residuals" whose squares will reveal anything about scaling in the model, general approaches such as the Breusch and Pagan (1979, 1980) LM test or the White (1980) test are not available. Heteroskedasticity must be built into the model and tested parametrically. Moreover, there is no robust approach to estimation and inference that will accommodate heteroskedasticity without specifically making it part of the model. (In linear regression, the ordinary least squares (OLS) estimator and White's (1980) heteroskedasticity robust covariance matrix serve that purpose.) A common approach to modeling heteroskedasticity in parametric binary choice models is based on Harvey's (1976) exponential model:

$$d_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad E[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i] = 0, \quad \text{Var}[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i] = [\exp(\mathbf{v}_i'\boldsymbol{\tau})]^2$$
$$d_i = 1 \text{ if } d_i^* > 0, \text{ and } d_i = 0 \text{ otherwise,}$$

where $\mathbf{v}_i$ is a known set of variables (that does not include a constant term) and τ is a new parameter vector to be estimated. The adjustment of the log-likelihood is fairly straightforward; the terms are changed to accommodate

$$\text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i) = F[\mathbf{w}_i'\boldsymbol{\theta} / \exp(\mathbf{v}_i'\boldsymbol{\tau})].$$

Maximization of the likelihood function with respect to all the parameters is somewhat more complicated, as the function is no longer globally concave. The

complication arises in interpretation of the model. The partial effects in this augmented model are:

$$\boldsymbol{\delta}_i = \frac{\partial \text{Prob}(d_i = 1|\mathbf{w}_i, \mathbf{v}_i)}{\partial \begin{pmatrix} \mathbf{w}_i \\ \mathbf{v}_i \end{pmatrix}} = f\left(\frac{\mathbf{w}_i'\boldsymbol{\theta}}{\exp(\mathbf{v}_i'\boldsymbol{\tau})}\right)\begin{pmatrix} \boldsymbol{\theta} \\ [-(\mathbf{w}_i'\boldsymbol{\theta})/\exp(\mathbf{v}_i'\boldsymbol{\tau})]\boldsymbol{\tau} \end{pmatrix}.$$

If $\mathbf{w}_i$ and $\mathbf{v}_i$ have variables in common, then the two effects are added. Whether they do or not, this calls into question the interpretation of the original coefficients in the model. If $\mathbf{w}_i$ and $\mathbf{v}_i$ do share variables, then the partial effect may have sign and magnitude that both differ from those of the coefficients, $\boldsymbol{\theta}$. At a minimum, as before, the scales of the partial effects are different from those of the coefficients.

For testing for homoskedasticity, the same three statistics as before are useable. (This is a parametric restriction on the model; $H_0 : \boldsymbol{\tau} = \mathbf{0}$.) The derivatives of the log-likelihood function are presented in Greene (2008, Ch. 23). As usual, the LM test is the simplest to carry out. The term necessary to compute the LM statistic under the null hypothesis is:

$$\mathbf{g}_i = q_i f_i \begin{pmatrix} \mathbf{w}_i \\ (-\mathbf{w}_i'\boldsymbol{\theta})\mathbf{v}_i \end{pmatrix}.$$

A second specification test of interest concerns the distribution. Silva (2001) has suggested a score (LM) test that is based on adding a constructed variable to the logit or probit model. An alternative way of testing the two competing models could be based on Vuong's (1989) statistic. Vuong's test is computed using:

$$\lambda_{\text{Vuong}} = \frac{\sqrt{n}\overline{m}}{s_m}, \text{ where } \overline{m} = \frac{1}{n}\Sigma_{i=1}^{n}[\ln L_i(\text{probit}) - \ln L_i(\text{logit})],$$

and s_m is the sample standard deviation. Vuong shows that, under certain assumptions (likely to be met here for these two models), λ_{Vuong} has a limiting standard normal distribution. Large positive values (larger than +1.96) favor the probit model, while large negative values (less than −1.96) favor the logit model. The power of these two statistics for this setting remains to be investigated. As with all specification tests, the power depends crucially on the true but unknown underlying model, which may be unlike either candidate model.

11.3.2.6 The fit of the model

As noted earlier, in modeling binary (or other discrete) choices, there is no direct counterpart to the R^2 goodness-of-fit statistic. A common computation which, unfortunately in spite of its name, does not provide such a measure is the *likelihood ratio index*, which is also called the

$$\text{pseudo } R^2 = 1 - \ln L/\ln L_0,$$

where $\ln L$ is the log-likelihood for the estimated model (which must include a constant term) and $\ln L_0$ is the log-likelihood function for a model that only has a

constant. It is tempting to suggest that this measure measures the "contribution" of the variables to the fit of the model. It is a statistic that lies between zero and one, and it does rise unambiguously as variables are added to the model. However, the "fit" aspect of the statistic is ambiguous, since the likelihood function is not a fit measure. As a consequence, this measure can be distressingly small in a model that contains numerous precisely measured (highly significant) coefficients (see Wooldridge, 2002a, for discussion).

This does leave open the issue of how to assess the fit of the estimated model to the data. In order to address this question, the analyst must first decide what rule will be used to predict the observed outcome using the model, then determine how successful the model (and rule) are. A natural approach, since the model predicts probabilities of events, is to use the estimated probability, $F(\mathbf{w}_i'\boldsymbol{\theta})$. The prediction is based on the rule:

$$\text{Predict } d_i = 1 \text{ if the estimated Prob}(d_i = 1|\mathbf{w}_i) \text{ is greater than } P, \tag{11.3}$$

where P^* is to be chosen by the analyst. The usual choice of P^* is 0.5, reasoning that if the model predicts that the event is more likely to occur than not, we should predict that it will.[9] A summary 2×2 table of the number of cases in which the rule predicts correctly and incorrectly can be used to assess the fit of the model. Numerous single-valued functions of this tally have been suggested as counterparts to R^2. For example, Cramer (1999) proposed:

$$\lambda_C = (\text{average } \hat{P}_i|d_i = 1) - (\text{average } \hat{P}_i|d_i = 0).$$

This measure counts the correct predictions, and adds a penalty for incorrect predictions. Other modifications and similar alternatives have been suggested by Efron (1978), Kay and Little (1986), Ben-Akiva and Lerman (1985) and Zavoina and McKelvey (1975).

11.3.3 A Bayesian estimator

The preceding section has developed the classical MLE for binomial choice models. A Bayesian estimator for the probit model illustrates an intriguing technique for censored data models. The model framework is, as before:

$$d_i^* = \mathbf{w}_i'\boldsymbol{\theta} + \varepsilon_i, \varepsilon_i \sim N[0, 1] \tag{11.4}$$

$$d_i = 1 \text{ if } d_i^* > 0, \quad \text{otherwise } d_i = 0. \tag{11.5}$$

The data consist of $(\mathbf{d},\mathbf{W}) = (d_i,\mathbf{w}_i)$, $i = 1, \ldots, n$. The random variable d_i has a Bernoulli distribution with probabilities:

$$\text{Prob}[d_i = 1|\mathbf{w}_i] = \Phi(\mathbf{w}_i'\boldsymbol{\theta})$$

$$\text{Prob}[d_i = 0|\mathbf{w}_i] = 1 - \Phi(\mathbf{w}_i'\boldsymbol{\theta}).$$

The likelihood function for the observed data, $\mathbf{d}$, conditioned on $\mathbf{W}$ and $\boldsymbol{\theta}$, is:

$$L(\mathbf{d}|\mathbf{W}, \boldsymbol{\theta}) = \prod_{i=1}^{n} [\boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\theta})]^{d_i}[1 - \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\theta})]^{1-d_i}.$$

To obtain the posterior mean (Bayesian estimator), we assume a non-informative, flat (improper) prior for $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) \propto 1.$$

By Bayes' theorem, the posterior density would be:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{d}, \mathbf{W}) &= \frac{p(\mathbf{d}|\mathbf{W}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{d}|\mathbf{W}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{\prod_{i=1}^{n}[\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{d_i}[1-\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{1-d_i}(1)}{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n}[\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{d_i}[1-\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{1-d_i}(1)d\boldsymbol{\theta}}, \end{aligned}$$

and the estimator would be the posterior mean:

$$\hat{\boldsymbol{\theta}}_{BAYESIAN} = E[\boldsymbol{\theta}|\mathbf{d}, \mathbf{W}] = \frac{\int_{\boldsymbol{\theta}} \boldsymbol{\theta} \prod_{i=1}^{n}[\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{d_i}[1-\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{1-d_i}d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n}[\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{d_i}[1-\Phi(\mathbf{w}_i'\boldsymbol{\theta})]^{1-d_i}\, d\boldsymbol{\theta}}.$$

Evaluation of the integrals in $\hat{\boldsymbol{\theta}}_{\text{BAYESIAN}}$ is hopelessly complicated, but a solution using the Gibbs sampler and the technique of *data augmentation*, pioneered by Albert and Chib (1993), is surprisingly simple. We begin by treating the unobserved d_i^*s as unknowns to be estimated, along with $\boldsymbol{\theta}$. Thus, the $(K+n)\times 1$ parameter vector is $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \mathbf{d}^*)$. We now construct a Gibbs sampler. Consider, first, $p(\boldsymbol{\theta}|\, \mathbf{d}^*,\mathbf{d},\mathbf{W})$. If d_i^* is known, then d_i is known. It follows that:

$$p(\boldsymbol{\theta}|\mathbf{d}^*, \mathbf{d}, \mathbf{W}) = p(\boldsymbol{\theta}|\mathbf{d}^*, \mathbf{W}).$$

This posterior comes from a linear regression model with normally distributed disturbances and known $\sigma^2 = 1$ (see equation (11.4) above). This is the standard case for Bayesian analysis of the normal linear model with an uninformative prior for the slopes and known σ^2 (see, e.g., Koop, 2003; Greene, 2008a, sec. 18.3.1), with the additional simplification that $\sigma^2 = 1$. It follows that:

$$p(\boldsymbol{\theta}|\mathbf{d}^*, \mathbf{d}, \mathbf{W}) = N[\mathbf{q}^*, (\mathbf{W}'\mathbf{W})^{-1}],$$

where:

$$\mathbf{q}^* = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{d}^*.$$

For d_i^*, ignoring d_i for the moment, it would follow immediately from equation (11.4) that:

$$p(d_i^*|\boldsymbol{\theta}, \mathbf{W}) = N[\mathbf{w}_i'\boldsymbol{\theta}, 1].$$

However, d_i is informative about d_i^*. If d_i equals one, we know that $d_i^* > 0$ and, if d_i equals zero, then $d_i^* \leq 0$. The implication is that, conditioned on θ, $\mathbf{W}$, and $\mathbf{d}$, d_i^* has a truncated (above or below zero) normal distribution. The standard notation for this is:

$$p(d_i^*|\boldsymbol{\theta}, d_i = 1, \mathbf{w}_i) = N^+[\mathbf{w}_i'\boldsymbol{\theta}, 1]$$

$$p(d_i^*|\boldsymbol{\theta}, d_i = 0, \mathbf{w}_i) = N^-[\mathbf{w}_i'\boldsymbol{\theta}, 1].$$

These results set up the components for a Gibbs sampler that we can use to estimate the posterior means E[$\boldsymbol{\theta}$|**d**,**W**] and E[$\mathbf{d}^*$|**d**, **W**].

11.3.3.1 Gibbs sampler for the binomial probit model

1. Compute $\mathbf{W}'\mathbf{W}$ once at the outset and obtain $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}' = (\mathbf{W}'\mathbf{W})^{-1}$.
2. Start $\boldsymbol{\theta}$ at any value such as $\mathbf{0}$.
3. Obtain draws $U_{i,r}$ from the standard uniform distribution. Greene (2008a, p. 575, result (17–1)) shows how to transform a draw from $U[0,1]$ to a draw from the truncated normal with underlying mean μ and standard deviation σ. For this application, $\mu = \mathbf{w}_i'\boldsymbol{\theta}$ and $\sigma = 1$, so the draws from $p(\mathbf{d}^*|\boldsymbol{\theta},\mathbf{d},\mathbf{W})$ are obtained as:
$$d_{i,r}^*(r) = \mathbf{w}_i'\boldsymbol{\theta}_{r-1} + \boldsymbol{\Phi}^{-1}\left[1 - (1 - U_{i,r})\boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\theta}_{r-1})\right] \text{ if } d_i = 1$$
$$d_{i,r}^*(r) = \mathbf{w}_i'\boldsymbol{\theta}_{r-1} + \boldsymbol{\Phi}^{-1}\left[U_{i,r}\boldsymbol{\Phi}(-\mathbf{w}_i'\boldsymbol{\theta}_{r-1})\right] \text{ if } d_i = 0.$$
This step is used to draw the n observations on $d_{i,r}^*(r)$.
4. To draw an observation from the multivariate normal population of $p(\boldsymbol{\theta}|\mathbf{d}^*, \mathbf{d}, \mathbf{W}]$, we need to draw from the normal population with mean q_{r-1}^* and variance $(\mathbf{W}'\mathbf{W})^{-1}$. For this application, we use the results at step 3 to compute $\mathbf{q}^* = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{d}^*(r)$. We obtain a vector, $\mathbf{v}$, of K draws from the $N[0,1]$ population, and then compute $\boldsymbol{\theta}(r) = \mathbf{q}^* + \mathbf{L}\mathbf{v}$.

The iteration cycles between steps 3 and 4. This should be repeated several thousand times, discarding the burn-in draws, and then the estimator of $\boldsymbol{\theta}$ is the sample mean of the retained draws. The posterior variance is computed with the variance of the retained draws. Posterior estimates of d_i^* would typically not be useful.

This application of the Gibbs sampler demonstrates, in an uncomplicated case, how the algorithm can provide an alternative to actually maximizing the log-likelihood. The similarity of the method to the EM algorithm (Dempster, Laird and Rubin, 1977) is not coincidental. Both procedures use an estimate of the unobserved, censored data, and both estimate $\boldsymbol{\theta}$ by using OLS using the predicted data.

11.3.4 Semiparametric models

The fully parametric probit and logit models remain by far the mainstays of empirical research on binary choice. Fully nonparametric discrete choice models are fairly exotic and have made only limited inroads in the literature, most of which is theoretical (e.g., Matzkin, 1993). The middle ground is occupied by a few semiparametric models that have been proposed to relax the detailed assumptions of the probit and logit specifications. The single index model of Klein and Spady (1993) has been used in several applications, including Gerfin (1996), Horowitz (1993) and Fernandez and Rodriguez-Poo (1997), and provides the theoretical platform for a number of extensions.[10]

The single index formulation departs from a regression formulation:

$$E[d_i|\mathbf{w}_i] = E[d_i|\mathbf{w}_i'\boldsymbol{\theta}].$$

Then:

$$\text{Prob}(d_i = 1|\mathbf{w}_i) = F(\mathbf{w}_i'\boldsymbol{\theta}|\mathbf{w}_i) = G(\mathbf{w}_i'\boldsymbol{\theta}),$$

where G is an unknown continuous distribution function whose range is [0,1]. The function G is not specified *a priori*; it is estimated (pointwise) along with the parameters. (Since G as well as $\boldsymbol{\theta}$ is to be estimated, a constant term is not identified; essentially, G provides the location for the index that would otherwise be provided by a constant.) The criterion function for estimation, in which n subscripts denote estimators based on the sample of n observations of their unsubscripted counterparts, is:

$$\ln L_n = \frac{1}{n}\sum_{i=1}^{n}\{d_i \ln G_n(\mathbf{w}_i'\boldsymbol{\theta}_n) + (1-d_i)\ln[1-G_n(\mathbf{w}_i'\boldsymbol{\theta}_n)]\}.$$

The estimator of the probability function, G_n, is computed at each iteration using a nonparametric kernel estimator of the density of $\mathbf{w}_i'\boldsymbol{\theta}_n$. For the Klein and Spady estimator, the nonparametric regression estimator is:

$$G_n(z_i) = \frac{\overline{d}\,g_n(z_i|d_i=1)}{\overline{d}\,g_n(z_i|d_i=1) + (1-\overline{d})\,g_n(z_i|d_i=0)},$$

where $g_n(z_i|d_i)$ is the *kernel estimate of the density* of $z_i = \mathbf{w}_i'\boldsymbol{\theta}_n$. This result is:

$$g_n(z_i|d_i=1) = \frac{1}{n\overline{d}h_n}\sum_{j=1}^{n} d_j K\left(\frac{z_i - \mathbf{w}_j'\boldsymbol{\theta}_n}{h_n}\right);$$

$g_n(z_i|d_i=0)$ is obtained by replacing $\overline{d}$ with $1-\overline{d}$ in the leading scalar and d_j with $1-d_j$ in the summation. The scalar h_n is the bandwidth. There is no firm theory for choosing the kernel function or the bandwidth. Both Horowitz and Gerfin used the standard normal density. Two different methods for choosing the bandwidth are suggested by them. Klein and Spady provide theoretical background for computing asymptotic standard errors.

Manski's (1975, 1985, 1986, 1987) maximum score estimator is even less parameterized than Klein and Spady's model. The estimator is based on the fitting rule:

$$\text{Maximize}_{\theta} S_{N\alpha}(\theta) = \frac{1}{n}\sum_{i=1}^{n}[q_i - (1-2\alpha)]\text{sign}(\mathbf{w}_i'\boldsymbol{\theta}).^{11}$$

The parameter α is a preset quantile, and $q_i = 2d_i - 1$ as before. If α is set to 0.5, then the maximum score estimator chooses the $\boldsymbol{\theta}$ to maximize the number of times that the prediction has the same sign as z. This result matches our prediction rule in equation (11.3) with $P^* = 0.5$. So for $\alpha = 0.5$, the maximum score attempts to maximize the number of correct predictions. Since the sign of $\mathbf{w}'\boldsymbol{\theta}$ is the same for all positive multiples of $\boldsymbol{\theta}$, the estimator is computed subject to the constraint that $\boldsymbol{\theta}'\boldsymbol{\theta} = 1$. Variants of semiparametric estimators are discussed in Li and Racine (2007), including a modification by Horowitz (1992) and an estimator suggested by Lewbel (2000).

The semiparametric estimators of $\boldsymbol{\theta}$ are robust to variation in the distribution of the random elements in the model, and even to heteroskedasticity. Robustness

is an ambiguous virtue in this context. As we have seen, the raw coefficients are of questionable value in interpreting the model – in order to translate them into useful quantities we have computed partial effects and predicted probabilities. But the semiparametric models specifically program around the assumption of a fixed distribution and thus sacrifice the ability to compute partial effects or probabilities. What remains is the estimator of $\boldsymbol{\theta}$ and, in some cases, a covariance matrix that can be used to test the significance of coefficients or to test hypotheses about restrictions on structural coefficients.[12] Perhaps for these reasons, applied work in binary choice remains overwhelmingly dominated by the parametric models.

11.3.5 Endogenous right-hand-side variables

The presence of endogenous right-hand-side variables in a binary choice model presents familiar problems for estimation. The problem is made worse in nonlinear models because, even if one has an instrumental variable readily at hand, it may not be immediately clear what is to be done with it. The instrumental variable estimator for the linear model is based on moments of the data, the variances and covariances. In a binary choice setting, we are not using any form of least squares to estimate the parameters, so the instrumental variable (IV) method would appear not to apply. Generalized method of moments is a possibility. Consider the model:

$$
\begin{aligned}
d_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i + \varepsilon_i \\
d_i &= 1(d_i^* > 0) \\
E[\varepsilon_i | z_i] &= g(z_i) \neq 0.
\end{aligned}
$$

Thus z_i is endogenous in this model. The MLEs considered earlier will not consistently estimate $(\boldsymbol{\beta}, \gamma)$. (Without an additional specification that allows us to formalize $\text{Prob}(d_i = 1|\mathbf{x}_i, z_i)$, we cannot state what the MLE will, in fact, estimate.) Suppose that we have a relevant (not "weak") instrumental variable, w_i, such that:

$$
\begin{aligned}
E[\varepsilon_i | w_i, \mathbf{x}_i] &= 0 \\
E[w_i z_i] &\neq 0.
\end{aligned}
$$

A natural instrumental variable estimator would be based on the "moment" condition:

$$
E\left[(d_i^* - \mathbf{x}_i'\boldsymbol{\beta} - \gamma z_i) \begin{pmatrix} \mathbf{x}_i \\ w_i \end{pmatrix} \right] = \mathbf{0}.
$$

However, d_i^* is not observed: d_i is, but the "residual," $d_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma z_i$, would have no meaning even if the true parameters were known.[13] One approach that was used in Avery, Hansen and Hotz (1983), Butler and Chatterjee (1997) and Bertschek and Lechner (1998) is to assume that the instrumental variable is orthogonal to the residual $[d_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i)]$, i.e.:

$$
E\left[[d_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta} - \gamma z_i)] \begin{pmatrix} \mathbf{x}_i \\ w_i \end{pmatrix} \right] = \mathbf{0}.
$$

This form of the moment equation, based on observables, can form the basis of a straightforward two-step generalized method of moments (GMM) estimator.

The GMM estimator is not less parametric than the full information MLE described below, because the probit model based on the normal distribution is still invoked to specify the moment equation.[14] Nothing is gained in simplicity or robustness compared to full information maximum likelihood estimation, which we now consider. (As Bertschek and Lechner, 1998, argue, however, the gains might come in terms of practical implementation and computation time. The same considerations motivated Avery, Hansen and Hotz, 1983.)

The MLE requires a full specification of the model, including the assumption that underlies the endogeneity of z_i. This becomes essentially a simultaneous equations model. The model equations are:

$$\begin{aligned} &d_i^* = \mathbf{x}_i'\beta + \gamma z_i + \varepsilon_i,\ d_i = 1[d_i^* > 0], \\ &z_i = \mathbf{w}_i'\alpha + u_i, \\ &(\varepsilon_i, u_i) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2 \end{pmatrix}\right]. \end{aligned} \tag{11.6}$$

(We are assuming that there is a vector of instrumental variables, $\mathbf{w}_i$.) Probit estimation based on d_i and $(\mathbf{x}_i, z_i)$ will not consistently estimate $(\boldsymbol{\beta},\gamma)$ because of the correlation between z_i and ε_i induced by the correlation between u_i and ε_i. Several methods have been proposed for estimation of this model. One possibility is to use the partial reduced form obtained by inserting the second equation into the first. This becomes a probit model with probability $\text{Prob}(d_i = 1|\mathbf{x}_i, \mathbf{w}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}^* + \mathbf{w}_i'\boldsymbol{\alpha}^*)$. This will produce consistent estimates of $\boldsymbol{\beta}^* = \boldsymbol{\beta}/(1 + \gamma^2\sigma^2 + 2\gamma\sigma\rho)^{1/2}$ and $\boldsymbol{\alpha}^* = \gamma\boldsymbol{\alpha}/(1 + \gamma^2\sigma^2 + 2\gamma\sigma\rho)^{1/2}$ as the coefficients on $\mathbf{x}_i$ and $\mathbf{w}_i$, respectively. (The procedure will estimate a mixture of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ for any variable that appears in both $\mathbf{x}_i$ and $\mathbf{w}_i$.) In addition, linear regression of z_i on $\mathbf{w}_i$ produces estimates of $\boldsymbol{\alpha}$ and σ^2, but there is no method of moments estimator of ρ or γ produced by this procedure, so this estimator is incomplete. Newey (1987) suggested a "minimum chi-squared" estimator that does estimate all parameters. A more direct, and actually simpler, approach is full information maximum likelihood.

The log-likelihood is built up from the joint density of d_i and z_i, which we write as the product of the conditional and the marginal densities:

$$f(d_i, z_i) = f(d_i|z_i)f(z_i).$$

To derive the conditional distribution, we use results for the bivariate normal, and write:

$$\varepsilon_i|u_i = [(\rho\sigma)/\sigma^2]u_i + v_i,$$

where v_i is normally distributed with $\text{Var}[v_i] = (1 - \rho^2)$. Inserting this in the first equation of equation (11.6), we have:

$$d_i^*|z_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i + (\rho/\sigma)u_i + v_i.$$

Therefore:

$$\text{Prob}[d_i = 1|\mathbf{x}_i, z_i] = \Phi\left[\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i + (\rho/\sigma)u_i}{\sqrt{1-\rho^2}}\right].$$

Inserting the expression for $u_i = (z_i - \mathbf{w}_i'\boldsymbol{\alpha})$, and using the normal density for the marginal distribution of z_i in the second equation of (11.6), we obtain the log-likelihood function for the sample:

$$\ln L = \sum_{i=1}^{n} \ln \Phi\left[(2d_i - 1)\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u)(z_i - \mathbf{w}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)\right] + \ln\left[\frac{1}{\sigma_u}\phi\left(\frac{z_i - \mathbf{w}_i'\boldsymbol{\alpha}}{\sigma_u}\right)\right].$$

11.3.6 Panel data models

The ongoing development of large, rich panel data sets on individual and family market experiences, such as the GSOEP data we are using here, has brought attention to panel data approaches for discrete choice modeling. The extensions of familiar fixed and random effects models are not direct and bring statistical and computational issues that are not present in linear regression modeling. This section will detail the most widely used techniques. This area of research is one of the most active theoretical arenas as well. We will only have space to note the theoretical frontiers briefly in the conclusions.

11.3.6.1 Panel data modeling frameworks

The natural departure point for panel data analysis of binary choice is the extension of the familiar fixed and random effects linear regression models. Since the models considered here are nonlinear, however, the convenient least squares and feasible generalized least squares methods are unavailable. This proves to be more than an inconvenience in this setting, as it mandates consideration of some specification issues. We will begin by considering extensions of the fixed and random effects models, then turn to more general models of individual heterogeneity, the random parameters and latent class models. The various models described here all carry over to a range of specifications. However, in the applied literature, the binary choice model is the leading case.

11.3.6.2 Fixed effects model

The fixed effects model is:

$$d_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{it}, \quad t = 1, \ldots, T_i,\ i = 1, \ldots, n$$
$$d_{it} = 1 \text{ if } d_{it}^* > 0, \text{ and } d_{it} = 0 \text{ otherwise.}$$

We have made the distinction between time varying attributes and characteristics, $\mathbf{x}_{it}$, and time invariant characteristics, $\mathbf{z}_i$. The common effects, α_i, may be

correlated with the included variables, $\mathbf{x}_{it}$. Since the model is nonlinear, the least squares estimator is unuseable. The log-likelihood is:

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln F[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})].$$

In principle, direct (brute force) maximization of the function with respect to $(\alpha_1, \ldots, \alpha_n, \boldsymbol{\beta}, \boldsymbol{\gamma})$ can be used to obtain estimates of the parameters and of their asymptotic standard errors. However, several issues arise.

1. The number of individual intercept parameters may be excessive. In our application, e.g., there are 7,293 families. Direct maximization of the log-likelihood function for this many parameters is likely to be difficult. This purely practical issue does have a straightforward solution and is, in fact, not an obstacle to estimation (see Greene, 2001, 2008a, Ch. 23).
2. As in the case of the linear model, it is not possible to estimate the parameters that apply to time invariant variables, $\mathbf{z}_i$. In the linear case, the transformation to group mean deviations turns these variables into columns of zeros. A similar problem arises in this nonlinear model.
3. Groups of observations in which the outcome variable, d_{it}, is always one or always zero for $t = 1, \ldots, T_i$ must be dropped from the sample.
4. The full MLE for this model is inconsistent, a consequence of the *incidental parameters problem* (see Neyman and Scott, 1948; Lancaster, 2000). The problem arises because the number of α_i parameters in the model rises with n. With small T or T_i this produces a bias in the estimator of $\boldsymbol{\beta}$ that does not diminish with increases in n. The best known case, that of the logit model with $T = 2$, was documented by Andersen (1970), Hsiao (1986) and Abrevaya (1997), who showed analytically that, with $T = 2$, the MLE of $\boldsymbol{\theta}$ for the binary logit model in the presence of the fixed effects will converge to $2\boldsymbol{\theta}$. Results for other distributions and other values of T have not been obtained analytically, and are based on Monte Carlo studies. Table 11.1, extracted from Greene (2001, 2004a, 2004b), demonstrates the effect in the probit, logit, and ordered probit model discussed in section 11.5. (The conditional estimator is discussed below.) The model contains a continuous variable, x_{it1}, and a dummy variable, x_{it2}. The

Table 11.1 Means of empirical sampling distributions, $N = 1{,}000$ individuals based on 200 replications. Table entry is $\bar{\beta}_1, \bar{\beta}_2$.

	$T = 2$		$T = 3$		$T = 5$		$T = 8$		$T = 10$		$T = 20$	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
Logit	2.020,	2.027	1.698,	1.668	1.379,	1.323	1.217,	1.156	1.161,	1.135	1.069,	1.062
Logit-C[a]	0.994,	1.048	1.003,	0.999	0.996,	1.017	1.005,	0.988	1.002,	0.999	1.000,	1.004
Probit	2.083,	1.938	1.821,	1.777	1.589,	1.407	1.328,	1.243	1.247,	1.169	1.108,	1.068
Ord. probit	2.328,	2.605	1.592,	1.806	1.305,	1.415	1.166,	1.220	1.131,	1.158	1.058,	1.068

[a] Estimates obtained using the conditional likelihood function – fixed effects not estimated.

population values of both coefficients are 1.0. The results, which are consistent with other studies, e.g., Katz (2001), suggest the persistence of the "small T bias" out to fairly large T.

These problems, particularly the last, have made the full fixed effects approach unattractive. The specification, however, remains an attractive alternative to the random effects approach considered next. Two approaches have been taken to work around the incidental parameters problem in the fixed effects model. A variety of semiparametric models have been suggested, such as Honore and Kyriazidou (2000a, 2000b) and Honore (2002).[15] In a few cases, including the binomial logit (but not the probit), it is possible to condition the fixed effects out of the model. The operation is similar to the group mean deviations transformation in the linear regression model. For the binary logit model (omitting the time invariant variables), we have:

$$\text{Prob}(d_{it} = j_{it}|\mathbf{x}_{it}) = \frac{\exp[j_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]}{1 + \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})} \quad \text{where } j_{it} \text{ is the observed value.}$$

This is the term that enters the unconditional log-likelihood function. However, conditioning on $\Sigma_{t=1}^{T_i} j_{it} = S_i$, we have the joint probability:

$$\text{Prob}(d_{i1} = j_{i1},\ d_{i2} = j_{i2}, \ldots |\mathbf{x}_{it}, \Sigma_{t=1}^{T_i} d_{it} = S_i) = \frac{\exp(\Sigma_{t=1}^{T_i} j_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\Sigma_t d_{it}=S_i} \exp(\Sigma_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}.$$

(See Rasch, 1960; Andersen, 1970; Chamberlain, 1980.) The denominator of the conditional probability is the summation over the different realizations of $(d_{i1}, \ldots, d_{i,Ti})$ that can sum to S_i. Note that, in this formulation, if $S_i = 0$ or T_i, there is only one way for the realizations to sum to S_i, and the one term in the denominator equals the observed result in the numerator. The probability equals one and, as noted in point 3 above, this group falls out of the estimator. The conditional log-likelihood is the sum of the logs of the joint probabilities. The log-likelihood is free of the fixed effects, so the estimator has the usual properties, including consistency. This estimator was used by Cecchetti (1986) and Willis (2006) to analyze magazine price changes.

The conditional estimator is consistent, so it bypasses the incidental parameter problem. However, it does have a major shortcoming. By avoiding the estimation of the fixed effects, we have precluded computation of the partial effects or estimates of the probabilities for the outcomes. So, like the robust semiparametric estimators, this approach limits the analyst to simple inference about $\boldsymbol{\beta}$ itself. One approach that might provide some headway out of this constraint is to compute second-step estimates of α_i. Since we have in hand a consistent estimator of $\boldsymbol{\beta}$, we treat that as known, and return to the unconditional log-likelihood function. For individual i, the contribution to the log-likelihood is:

$$\ln L_i = \sum_{t=1}^{T_i} \ln F[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})].$$

For convenience, denote the "known" $\mathbf{x}_{it}'\boldsymbol{\beta}$ as b_{it}. The first-order condition for maximizing ln$L$ with respect to α_i, given known $\boldsymbol{\beta}$, is:

$$\partial \ln L_i/\partial b_{it} = \sum_{t=1}^{T_i}[d_{it} - F(\alpha_i + b_{it})] = 0.$$

This is one equation in one unknown that can be solved iteratively to provide an estimate of α_i. The resulting estimator is inconsistent, since T_i is fixed – the resulting estimates are also likely to be highly variable because of the small sample sizes. However, the inconsistency results not because it converges to something other than α_i. The estimator is inconsistent because its variance is $O(1/T_i)$. Consequently, an estimator of the average partial effects:

$$\bar{\hat{\boldsymbol{\delta}}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} f(\hat{\alpha}_i + \mathbf{x}_{it}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}},$$

may yet provide a useful estimate of the partial effects. This estimator remains to be examined empirically or theoretically.

The fixed effects model has the attractive aspect that it is a robust specification. The four shortcomings listed above, especially items 2 and 4, do reduce its appeal, however. The wisdom behind the linear model does not carry over to binary choice models because the estimation and inferential problems change substantively in nonlinear settings. The statistical aspects of the random effects model discussed next are more appealing. However, the model's assumption of orthogonality between the unobserved heterogeneity and the included variables is also unattractive. The Mundlak (1978) device is an intermediate step between these two that is sometimes used. This approach relies on a projection of the effects on the time invariant characteristics and group means of the time variables;

$$\alpha_i = \mathbf{z}_i'\boldsymbol{\gamma} + \pi_0 + \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \sigma_u u_i, \quad \text{where } E[u_i|\bar{\mathbf{x}}_i] = 0 \text{ and } \text{Var}[u_i|\bar{\mathbf{x}}_i] = 1.$$

(The location parameter π_0 accommodates a non-zero mean while the scale parameter, σ_u, picks up the variance of the effects, so the assumptions of zero mean and unit variance for u_i are just normalizations.) Inserting this into the fixed effects model produces a type of random effects model:

$$d_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \pi_0 + \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \sigma_u u_i + \varepsilon_{it}, \quad t = 1, \ldots, T_i, \quad i = 1, \ldots, n$$

$$d_{it} = 1 \text{ if } d_{it}^* > 0, \text{ and } d_{it} = 0 \text{ otherwise.}$$

If the presence of the projection on the group means successfully picks up the correlation between α_i and $\mathbf{x}_{it}$, then the parameters ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, π_0, $\boldsymbol{\pi}$, σ_u) can be estimated by maximum likelihood (ML) as a random effects model. The remaining assumptions (functional form and distribution) are assumed to hold (at least approximately), so that the random effects treatment is appropriate.

11.3.6.3 Random effects models and estimation

As suggested in the preceding section, the counterpart to a random effects model for binary choice would be:

$$d_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sigma_u u_i + \varepsilon_{it}, \quad t = 1, \ldots, T_i, \quad i = 1, \ldots, n,$$

where $E[u_i|\mathbf{x}_{it}] = 0$ and $\text{Var}[u_i|\mathbf{x}_{it}] = 1$ and:

$$d_{it} = 1 \text{ if } d_{it}^* > 0, \text{ and } d_{it} = 0 \text{ otherwise.}$$

(Since the random effects model can accommodate time invariant characteristics, we have reintroduced $\mathbf{z}_i$ into the model.) The random effects model is fitted by ML assuming normality for ε_{it} and u_i. (The most common application is the random effects probit model.)

To begin, suppose the common effect is ignored, and the "pooled" model is fitted by simple ML, ignoring the presence of the heterogeneity. The (incorrectly) assumed model is:

$$\text{Prob}(d_{it} = 1|\mathbf{x}_{it}) = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}).$$

In the presence of u_i, the correct model is:

$$\begin{aligned}\text{Prob}(d_{it} = 1|\mathbf{x}_{it}) &= \text{Prob}(\varepsilon_{it} + \sigma_u u_i < \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) \\ &= \text{Prob}\left(\frac{\varepsilon_{it} + \sigma_u u_i}{\sqrt{1+\sigma_u^2}} < \frac{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}}{\sqrt{1+\sigma_u^2}}\right) \\ &= \text{Prob}\left(v_{it} < \mathbf{x}_{it}'\boldsymbol{\beta}^u + \mathbf{z}_i'\boldsymbol{\gamma}^u\right), \quad v_{it} \sim N[0,1].\end{aligned}$$

Thus the marginal probability that d_{it} equals one obeys the assumptions of the familiar probit. However, the coefficient vector is not $\boldsymbol{\beta}$, but $\boldsymbol{\beta}^u = \boldsymbol{\beta}/(1+\sigma_u^2)^{1/2}$, and likewise for $\boldsymbol{\gamma}$. The upshot is that ignoring the heterogeneity (random effect) is not so benign here as in the linear regression model. In the regression case, ignoring a random effect that is uncorrelated with the included variables produces an inefficient, but consistent, estimator.

In spite of the preceding result, it has become common in the applied literature to report "robust," "cluster corrected" asymptotic covariance matrices for pooled estimators such as the MLE above. The underlying justification is that, while the MLE may be consistent (though it rarely is, as exemplified above), the asymptotic covariance matrix should account for the correlation across observations within a group. The corrected estimator is:

$$\begin{aligned}\text{Est.Asy.Var}\left[\hat{\boldsymbol{\theta}}_{MLE}\right] = &\left[\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{H}_{it}\right]^{-1}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\mathbf{g}_{it}\right)\left(\sum_{t=1}^{T_i}\mathbf{g}_{it}'\right)\right] \\ &\times\left[\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{H}_{it}\right]^{-1},\end{aligned}$$

where $\mathbf{H}_{it} = \partial^2 \ln F(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}))/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ and $\mathbf{g}_{it} = \partial \ln F(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}))/\partial\boldsymbol{\theta}$ and all terms are computed at the pooled MLE. The estimator has a passing resemblance to the White (1980) covariance estimator for the least squares coefficient estimator. However, the usefulness of this estimator rests on the assumption that the pooled estimator is consistent, which will generally not be the case.

Efficiency is a moot point for this estimator, since the probit MLE estimates $\boldsymbol{\beta}$ with a bias toward zero:

$$\begin{aligned}\text{plim } \hat{\boldsymbol{\beta}}_{MLE} &= \boldsymbol{\beta}^{u} \\ &= \boldsymbol{\beta}/(1+\sigma_u^2)^{1/2} \\ &= \boldsymbol{\beta}(1-\rho^2)^{1/2},\end{aligned}$$

where $\rho^2 = \text{Corr}^2[\varepsilon_{it}+u_i, \varepsilon_{is}+u_i]$ for $t \neq s$. Wooldridge (2002a) suggests that this may not be an issue here, since the real interest is in the partial effects, which are, for the correct model:

$$\boldsymbol{\delta}_{it} = \partial \text{Prob}[d_{it}=1|\mathbf{x}_{it}, \mathbf{z}_i]/\partial \mathbf{x}_{it} = \boldsymbol{\beta}^{u}\phi(\mathbf{x}_{it}'\boldsymbol{\beta}^{u} + \mathbf{z}_i'\boldsymbol{\gamma}^{u}).$$

These would then be averaged over the individuals in the sample. It follows, then, that the "pooled" estimator, which ignores the heterogeneity, does not estimate the structural parameters of the model correctly, but it does produce an appropriate estimator of the average partial effects.

In the random effects model, the observations are not statistically independent – because of the common u_i, the observations $(d_{i1}, \ldots, d_i, T_i, u_i)$ constitute a T_i+1 variate random vector. The contribution of observation i to the log-likelihood is the joint density:

$$f(d_{i1}, \ldots, d_{i,Ti}, u_i|\mathbf{X}_i) = f(d_{i1}, \ldots, d_{i,Ti}|\mathbf{X}_i, \mathbf{z}_i, u_i)f(u_i).$$

Conditioned on u_i, the T_i random outcomes, $d_{i1},\ldots,d_{i,Ti}$, are independent. This implies that (with the normality assumption now incorporated in the model) the contribution to the log-likelihood is:

$$\ln L_i = \ln\left\{\left[\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sigma_u u_i)\right)\right]\phi(u_i)\right\},$$

where $\phi(u_i)$ is the standard normal density. This joint density contains the unobserved u_i, which must be integrated out of the function to obtain the appropriate log-likelihood function in terms of the observed data. Combining all terms, we have the log-likelihood for the observed sample:

$$\ln L = \sum_{i=1}^{n}\ln\left[\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sigma_u u_i)\right)\right)\phi(u_i)du_i\right]. \tag{11.7}$$

Maximization of the log-likelihood with respect to $(\boldsymbol{\beta}, \sigma_u)$ requires evaluation of the integrals in equation (11.7). Since these do not exist in closed form, some method of approximation must be used. The most common approach is the Hermite quadrature method suggested by Butler and Moffitt (1982). The approximation is written:

$$\begin{aligned}&\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sigma_u u_i)\right)\right)\phi(u_i)du_i \approx \frac{1}{\sqrt{\pi}}\sum_{h=1}^{H} w_h \prod_{t=1}^{T_i} \\ &\quad \Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sqrt{2}\sigma_u z_h)\right),\end{aligned}$$

where w_h and z_h are the weights and nodes of the quadrature (see Abramovitz and Stegun, 1971) and H is the number of nodes chosen (typically 20, 32 or 64). An alternative approach to the approximation is suggested by noting that:

$$\left[\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\sigma_u u_i)\right)\right)\phi(u_i)du_i\right]$$
$$=E_{u_i}\left[\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\sigma_u u_i)\right)\right].$$

The expected value can be approximated satisfactorily by simulation by using a sufficiently large sample of random draws from the population of u_i:

$$\left[\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\sigma_u u_i)\right)\right)\phi(u_i)du_i\right]$$
$$\approx\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\sigma_u u_{ir})\right).$$

Sampling from the standard normal population is straightforward using modern software (see Greene, 2008a, Ch. 17). The right-hand side converges to the left-hand side as R increases (so long as $\sqrt{n}/R\rightarrow 0$; see Gourieroux and Monfort, 1996).[16] The *simulated log-likelihood* to be maximized is:

$$\ln L_S=\sum_{i=1}^{n}\ln\frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\Phi\left(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\sigma_u u_{ir})\right).$$

Recent research in numerical methods has revealed alternative approaches to random sampling to speed up the rate of convergence in the integration. Halton sequences (see Bhat, 1999) are often used to produce approximations which provide comparable accuracy with far fewer draws than the simulation approach.

11.3.6.4 Dynamic models

An important extension of the panel data treatment in the previous section is the dynamic model:

$$d_{it}^{*}=\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}+\lambda d_{i,t-1}+\alpha_i+\varepsilon_{it}$$
$$d_{it}=1\text{ if }d_{it}^{*}>0\text{ and 0 otherwise.}\qquad(11.8)$$

Recent applications include Hyslop's (1999) analysis of labor force participation, Wooldridge's (2005) study of union membership and Contoyannis, Jones and Rice's (2004) analysis of self-reported health status in the BHPS.[17] In these and other applications, the central feature is *state dependence*, or the *initial conditions problem*: individuals tend to "stick" with their previous position. Wooldridge (2002b) lays out conditions under which an appropriate treatment is to model the individual effect as being determined by the initial value in:

$$\alpha_i=\alpha_0+\alpha_1 d_{i0}+\overline{\mathbf{x}}_i'\boldsymbol{\pi}+\sigma_u u_i,\quad u_i\sim N[0,1].\qquad(11.9)$$

This is the Mundlak treatment suggested earlier with the addition of the initial state in the projection.[18] Inserting equation (11.9) in (11.8) produces an augmented

random effects model that can be estimated, as in the static case, by Hermite quadrature of maximum simulated likelihood (MSL).

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriadizou (2000a) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two-period panel with fixed effects. An extension of lagged effects to a parametric model is Chamberlain (1980), Jones and Landwehr (1988) and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits its usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In this framework, the narrow assumptions about the independent variables once again limit its practical applicability. Honore and Kyriazidou (2000b) have combined the logic of the conditional logit model and Manski's maximum score estimator. They specify:

$$\begin{aligned}
&\text{Prob}(d_{i0} = 1|\mathbf{X}_i, \mathbf{z}_i, \alpha_i) = F_0(\mathbf{X}_i, \mathbf{z}_i, \alpha_i), \text{ where } \mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}),\\
&\text{Prob}(d_{it} = 1|\mathbf{X}_i, \mathbf{z}_i, \alpha_i, d_{i0}, d_{i1}, \ldots, d_{i,t-1})\\
&\qquad = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \alpha_i + \lambda d_{i,t-1})\ t = 1, \ldots, T.
\end{aligned}$$

The analysis assumes a single regressor and focuses on the case of $T = 3$. The resulting estimator resembles Chamberlain's but relies on observations for which $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$, which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of $\mathbf{x}_{it}$ is a considerable restriction, and the authors propose a kernel density estimator for the difference, $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, instead, which does relax that restriction a bit. The end result is an estimator which converges (they conjecture) but to a non-normal distribution and at a rate slower than $n^{-1/3}$.

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and Macurdy (1981), Jakubson (1988), Keane (1993) and Beck, Epstein and Jackman (2001), to name just a few.[19]

11.3.6.5 Parameter heterogeneity: random parameters and latent class models

Among the central features of panel data treatments of individual data is the opportunity to model individual heterogeneity, both observed and unobserved. The preceding discussion develops a set of models in which latent heterogeneity is embodied in the additive effect, α_i. We can extend the model to allow heterogeneity in the other model parameters as well,. The resulting specification is:

$$d_{it}^* = \mathbf{w}_{it}'\boldsymbol{\theta}_i + \alpha_i + \varepsilon_i, \quad d_{it} = 1(d_{it}^* > 0). \tag{11.10}$$

The specification is completed by the assumptions about the process that generates the individual specific parameters. Note that, in this formulation, the "effect" α_i is now merely an individual specific constant term. It is thus convenient to absorb it into the rest of the parameter vector, $\boldsymbol{\theta}_i$, and assume that $\mathbf{w}_{it}$ contains a constant.

A random parameters model (or mixed model or hierarchical model), in which parameters are continuously distributed across individuals, can be written:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_i,$$

where $\mathbf{u}_i$ is a set of uncorrelated random variables with zero means (means are absorbed in $\boldsymbol{\theta}_0$) and unit variances (non-unit variances are contained in the parameter matrix $\boldsymbol{\Gamma}$). The random effects model examined earlier emerges if $\boldsymbol{\Delta} = \mathbf{0}$ and the only random component in $\boldsymbol{\theta}_i$ is the constant term, in which case $\boldsymbol{\Gamma}$ would have a single nonzero diagonal element equal to σ_u. For the more general case, we have a random parameters formulation in which:

$$E[\boldsymbol{\theta}_i|\mathbf{z}_i] = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}\mathbf{z}_i$$
$$\text{Var}[\boldsymbol{\theta}_i|\mathbf{z}_i] = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'.$$

A random parameters model of this sort can be estimated by Hermite quadrature (see Rabe-Hesketh, Skrondal and Pickles, 2005) or by MSL (see Train, 2003; Greene, 2008a, Ch. 17, 23). The simulated log-likelihood function for this model will be:

$$\ln L_S = \sum_{i=1}^{n} \ln \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \Phi\left(q_{it}(\mathbf{w}_{it}'(\boldsymbol{\theta}_0 + \Delta\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_{ir}))\right).$$

Partial effects in this model can be computed by averaging the partial effects at the population conditional means of the parameters, $E[\boldsymbol{\theta}_i|\mathbf{z}_i] = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}\mathbf{z}_i$.

11.3.7 Application

Riphahn, Wambach and Million (2003) were interested in counts of physician and hospital visits. In this application, they were particularly interested in the impact that the presence of private insurance had on utilization counts, i.e., whether the data contain evidence of moral hazard. The sample is an unbalanced panel of 7,293 households. The number of households varies over seven periods (1,525; 1,079; 825; 926; 1,051; 1,000; 887) with a total number of 27,326 observations. The variables in the data file are listed in Table 11.2. (Only a few of these were used in the applications.)

The model to be examined here (not the specification used in the original study) is:

$$\text{Prob}(Doctor_{it} = 1|\mathbf{x}_{it}) = F(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

Table 11.2 Variables in German health care data file

Variable		*Mean*	*Standard deviation*
Year	Calendar year of the observation	1987.82	3.17087
Age	Age in years	43.5257	11.3302
Female	Female = 1; male = 0	.478775	.499558
Married	Married = 1; else = 0	.758618	.427929
HhKids	Children under age 16 in the household = 1; else = 0	.402730	.490456
HhNInc	Household nominal monthly net income in German marks/10,000	.352084	.176908
Working	Employed = 1; else = 0	.677048	.467613
BlueC	Blue-collar employee = 1; else = 0	.243761	.429358
WhiteC	White-collar employee = 1; else = 0	.299605	.458093
Self	Self-employed = 1; else = 0	.0621752	.241478
Beamt	Civil servant = 1; else = 0	.0746908	.262897
Educ	Years of schooling	11.3206	2.32489
Haupts	Highest schooling degree is Hauptschul = 1; else = 0	.624277	.484318
Reals	Highest schooling degree is Realschul = 1; else = 0	.196809	.397594
Fachhs	Highest schooling degree is Polytechnic = 1; else = 0	.0408402	.197924
Abitur	Highest schooling degree is Abitur = 1; else = 0	.117031	.321464
Univ	Highest schooling degree is university = 1; else = 0	.0719461	.258403
Hsat	Health satisfaction, 0–10	6.78543	2.29372
Newhsat[a,b]	Health satisfaction, 0–10	6.78566	2.29373
Handdum	Handicapped = 1; else = 0	.214015	.410028
Handper	Degree of handicap in pct, 0–100	7.01229	19.2646
DocVis	Number of doctor visits in last three months	3.18352	5.68969
Doctor[b]	1 if Docvis > 0, 0 else	.629108	.483052
HospVis	Number of hospital visits in last calendar year	.138257	.884339
Hospital[b]	1 of Hospvis > 0, 0 else	.0876455	.282784
Public	Insured in public health insurance= 1; else = 0	.885713	.318165
AddOn	Insured by add-on insurance = 1; else = 0	.0188099	.135856

Data source: http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. From Riphahn, Wambach and Million (2003, pp. 387–405).
[a] *NEWHSAT = HSAT*; 40 observations on HSAT recorded between 6 and 7 were changed to 7.
[b] Transformed variable not in raw data file.

(In order to examine fixed effects models, we have not used any of the time invariant variables, such as gender.) Table 11.3 lists the maximum likelihood estimates and estimated asymptotic standard errors for several model specifications. Estimates of the logit model are shown first, followed by the probit estimates. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. The empirical regularity that the MLEs of the coefficients in the logit model are typically about 1.6 times their probit counterparts is strikingly evident in these results (e.g., the ratios are 1.613 and 1.597 for the coefficients on age and income, respectively). The apparent differences between the logit and probit results are resolved by a comparison of the partial effects also shown in Table 11.3. As anticipated, the results are essentially the same for

Table 11.3 Estimated parameters for panel data binary choice models

			Variable					
Model	*Estimate*	*lnL*	*Constant*	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit	β		**0.25112**	**0.020709**	**−0.18592**	**−0.22947**	**−0.045587**	**0.085293**
Pooled	(ME)	−17673.10		(0.00481)	(−0.0432)	(−0.0536)	(−0.0106)	(0.0199)
	[APE]			[0.00471]	[−0.0423]	[−0.0522]	[−0.0104]	[0.0194]
	St.Er.		0.091135	0.001285	0.075064	0.029537	0.005646	0.033286
	Rob.SE[e]		0.12827	0.001743	0.091546	0.038313	0.008075	0.045314
Logit R.E.	β		**−0.13460**	**0.039267**	**0.021914**	**−0.21598**	**−0.063578**	**0.025071**
$\rho^2 = 0.41607$	(ME)	−15261.90		(0.00642)	(0.00358)	(−0.0354)	(−0.0103)	(0.00410)
	St.Er.		0.17764	0.002465	0.11866	0.047738	0.011322	0.056282
Logit	β			**0.10475**	**−0.060973**	**−0.088407**	**−0.11671**	**−0.057318**
F.E.(U)[a]	(ME)	−9458.64		(0.0249)	(−0.0145)	(−0.0210)	(−0.0277)	(−0.0136)
	St.Er.			0.007255	0.17829	0.074399	0.066749	0.10609
Logit	β			**0.084760**	**−0.050383**	**−0.077764**	**−0.090816**	**−0.052072**
F.E.(C)[b]	(ME)	−6299.02		(0.00730)	(−0.00434)	(−0.00670)	(−0.00782)	(−0.00448)
	St.Er.			0.006502	0.15888	0.066282	0.056673	0.093044
Probit	β		**0.15500**	**0.012835**	**−0.11643**	**−0.14118**	**−0.028115**	**0.052260**
Pooled	(ME)	−17670.94		(0.00484)	(−0.0439)	(−0.0534)	(−0.0106)	(0.0198)
	St.Er.		0.056516	0.000790	0.046329	0.018218	0.003503	0.020462
	Rob.SE[e]		0.079591	0.001074	0.056543	0.023614	0.005014	0.027904
Bayesian	β (Mean)		**0.15729**	**0.012807**	**−0.11319**	**−0.14160**	**−0.028234**	**0.050943**
Pooled		N/A						
Probit	β (Var.)		0.057824	0.000784	0.048868	0.017385	0.003437	0.020729
Probit:RE[c]	β		**0.034113**	**0.020143**	**−0.003176**	**−0.15379**	**−0.033694**	**0.016325**
$\rho^2 = 0.44789$	(ME)	−16273.96		(0.00560)	(−0.00088)	(−0.0428)	(−0.00938)	(0.00454)
	St.Er.		0.096354	0.001319	0.066672	0.027043	0.006289	0.031347

Probit:RE[d]	β	−16279.97	**0.033290**	**0.020078**	**−0.002973**	**−0.153579**	**−0.033489**	**0.016826**
ρ^2 =0.44799	(ME)			(0.00715)	(−0.00106)	(−0.0547)	(−0.0119)	(0.00599)
	St.Er.		0.063229	0.000901	0.052012	0.020286	0.003931	0.022771
Probit	β	−9453.71		**0.062528**	**−0.034328**	**−0.048270**	**−0.072189**	**−0.032774**
F.E.(U)	(ME)			(0.0239)	(−0.0132)	(−0.0185)	(−0.0277)	(−0.0126)
	St.Er.			0.004322	0.10745	0.044559	0.040731	0.063627

[a] Unconditional fixed effects estimator.
[b] Conditional fixed effects estimator.
[c] Butler and Moffitt estimator.
[d] Maximum simulated likelihood estimator.
[e] Robust, "cluster" corrected standard error.

the two models. The first two rows of partial effects in Table 11.3 compare the partial effects computed at the means of the variables, shown in the first row, to the average partial effects, computed by averaging the individual partial effects, shown in the second row. As might be expected, the differences between them are inconsequential.

The log-likelihood for the probit model is slightly larger than for the logit: however, it is not possible to compare the two on this basis as the models are non-nested. The Vuong statistic, based on $v_i = \ln L_i(\text{logit}) - \ln L_i(\text{probit})$, equals -7.44, which favors the probit model. The aggregated prediction of the pooled logit model is shown in the following table, using the usual prediction rule, $P^* = 0.5$.

	Predicted	
Actual	0	1
0	378	9,757
1	394	16,797

Thus, we obtain correct prediction of $(378 + 16,797)/27,326 = 62.9\%$ of the observations. In spite of this apparently good model performance, the pseudo-R^2 is only $1 - (-17673.10)/(-18019.55) = 0.01923$. This suggests a disconnection between these two measures of model performance. As a final check on the model itself, we tested the null hypothesis that the five coefficients other than the constant term are zero in the probit specification. The likelihood ratio test is based on the statistic:

$$\lambda_{\text{LR}} = 2[-17670.94 - 27326(.37089 \ln .37089 + .62911 \ln .62911)] = 697.22.$$

The Wald statistic based on the full model is λ_{WALD} = 686.991. The LM statistic is computed as:

$$\lambda_{LM} = \mathbf{g}_0' \boldsymbol{X} (\mathbf{G}_0' \boldsymbol{G}_0)^{-1} \mathbf{X}' \mathbf{g}_0,$$

where $\mathbf{g}_0$ is the derivative of the log-likelihood when the model contains only a constant term. This is equal to $q_{it}\phi(q_{it}\beta_0)/\Phi(q_{it}\beta_0)$, where $\beta_0 = \Phi^{-1}(.62911) = .32949$. Then the ith row of $\mathbf{G}$ is $\mathbf{g}_{it,0}$ times the corresponding row of $\mathbf{X}$. The value of the LM statistic is 715.97. The 5% critical value from the chi-squared distribution with 5 degrees of freedom is 11.07 so, in all three cases, the null hypothesis that the slopes are zero is soundly rejected.

The second set of probit estimates was computed using the Gibbs sampler and a noninformative prior. We used only 500 replications, and discarded the first 100 for the burn-in. The similarity to the maximum likelihood estimates is what one would expect given the large sample size. We note, however, that, notwithstanding the striking similarity of the Gibbs sampler to the MLE, this is not an efficient method of estimating the parameters of a probit model. The estimator requires generation

of thousands of samples of potentially thousands of observations. We used only 500 replications to produce the results in Table 11.3. The computations took about five minutes. Using Newton's method to maximize the log-likelihood directly took less than five seconds. Unless one is wedded to the Bayesian paradigm then, on strictly practical grounds, the MLE would be the preferred estimator.

Table 11.3 also lists the probit and logit random and fixed effects estimators. The random effects estimators produce a reasonably large estimate of ρ^2, roughly 0.44. The high correlation across observations does cast some doubt on the validity of the pooled estimator. The pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The model marked "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note that, for all three fixed effects estimators, it is necessary to drop from the sample any groups that have Doctor_{it} equal to zero or one for every period. There were 3,046 such groups, which is about 42% of the sample. We also computed the probit random effects model in two ways: first by using the Butler and Moffitt method, then by using MSL estimation. In this case, the estimators are very similar, as might be expected. The estimated squared correlation coefficient is computed as $\rho^2 = \sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$. For the probit model, $\sigma_\varepsilon^2 = 1$. The MSL estimator computes $s_u = 0.9088376$, from which we obtained ρ^2. The estimated partial effects for the models are also shown in Table 11.3. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

Finally, we carried out two tests of the stability of the model. All of the estimators listed in Table 11.3 derive from a model in which it is assumed that the same coefficient vector applies in every period. To examine this assumption, we carried out a homogeneity test of the hypothesis:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_T,$$

for the $T = 7$ periods in the sample. The likelihood ratio statistic is:

$$\lambda = 2\left[\left(\Sigma_{t=1}^{T} \ln L_t\right) - \ln L_{POOLED}\right].$$

The first part of the statistic is obtained by dividing the sample into the seven years of data – the number of observations varies (3,874; 3,794; 3,792; 3,661; 4,483; 4,340; 3,377) – and then estimating the model separately for each year. The calculated statistic is 202.97. The 5% critical value from the chi squared distribution with $(T-1)6 = 36$ degrees of freedom is 50.998, so the homogeneity assumption is rejected by the data. As a second test, we separated the sample into men and women and once again tested for homogeneity. The likelihood ratio test statistic is:

$$\begin{aligned}\lambda &= 2[\ln L_{\text{FEMALE}} + \ln L_{\text{MALE}} - \ln L_{\text{POOLED}}] \\ &= 2[(-7855.219377) + (-9541.065897) - (-18019.55)] \\ &= 1246.529452.\end{aligned}$$

The 5% critical value from the chi-squared distribution with 6 degrees of freedom is 12.592, so this hypothesis is rejected as well.

11.4 Bivariate and multivariate binary choice

The health care data contain two binary variables, DOCTOR and HOSPITAL, which one would expect to be at least correlated if not jointly determined. The extension of the binary choice model to more than one choice is relatively uncomplicated, but does bring new statistical issues as well as new practical complications. We consider several two equation specifications first, as these are the leading cases, then consider the extension to an arbitrary number of binary choices.

11.4.1 Bivariate binary choice

A two-equation binary choice model would take the form of a seemingly unrelated regressions model:

$$d_{i,1}^* = \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0,$$
$$d_{i,2}^* = \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0,$$

where "1" and "2" distinguish the equations (and are distinct from the periods in a panel data case). The bivariate binary choice model arises when the two disturbances are correlated. There is no convenient approach for this model based on the logistic model, so we assume bivariate normality at the outset. The bivariate probit model has:

$$F(\varepsilon_{i,1}, \varepsilon_{i,2}) = N_2[(0,0),(1,1),\rho], \quad -1 < \rho < 1.$$

The probability associated with the joint event $d_{i,1} = d_{i,2} = 1$ is then:

$$\text{Prob}(d_{i,1} = 1, d_{i,2} = 1 | \mathbf{w}_{i,1}, \mathbf{w}_{i,2}) = \Phi_2\left[\mathbf{w}_{i,1}'\boldsymbol{\theta}_1, \mathbf{w}_{i,2}'\boldsymbol{\theta}_2, \rho\right],$$

where $\Phi_2[t_1, t_2, \rho]$ denotes the bivariate normal c.d.f. The log-likelihood function is the joint density for the observed outcomes. By extending the formulation of the univariate probit model in the preceding section, we obtain:

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2\left[\left(q_{i,1}\mathbf{w}_{i,1}'\boldsymbol{\theta}_1\right), \left(q_{i,2}\mathbf{w}_{i,2}'\boldsymbol{\theta}_2\right), \left(q_{i,1}q_{i,2}\rho\right)\right].$$

The bivariate normal integral does not exist in closed form, and must be approximated, typically with Hermite quadrature.

The model is otherwise conventional and the standard conditions for MLEs are obtained. Interpretation of the model does bring some complications, however. First, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are not the slopes of any recognizable conditional mean function and neither are the derivatives of the possibly interesting $\text{Prob}(d_{i,1} = 1, d_{i,2} = 1 | \mathbf{w}_{i,1}, \mathbf{w}_{i,2})$. Both of these are complicated functions of all the model parameters and both data vectors (see Greene, 2008a, sec. 23.8.3; Christofides, Stengos and Swidinsky, 1997; Christofides, Hardin and Stengos, 2000). Since this is a two-equation model, it is unclear what quantity should be analyzed when interpreting the coefficients in relation to partial effects. One possibility

is the joint probability, $\text{Prob}(d_{i,1} = 1, d_{i,2} = 1) = \Phi_2\left[\mathbf{w}_{i,1}'\boldsymbol{\theta}_1,\ \mathbf{w}_{i,2}'\boldsymbol{\theta}_2,\ \rho\right]$, that is analyzed by Christofides, Stengos and Swidinsky (1997). Greene (1996, 2008a) considers, instead, the conditional mean function $\text{E}[d_{i,1}|d_{i,2} = 1, \mathbf{w}_{i,1}, \mathbf{w}_{i,2}] = \Phi_2\left[\mathbf{w}_{i,1}'\boldsymbol{\theta}_1,\ \mathbf{w}_{i,2}'\boldsymbol{\theta}_2,\ \rho\right]/\Phi[\mathbf{w}_{i,2}'\boldsymbol{\theta}_2]$. In either case, the raw coefficients bear little resemblance to the partial effects.

For hypothesis testing about the coefficients, the standard results for Wald, LM and LR tests apply. The LM test is likely to be cumbersome because the derivatives of the log-likelihood function are complicated. The other two are straightforward. A hypothesis of interest is that the correlation is zero. For testing:

$$H_0 : \rho = 0,$$

all three likelihood-based procedures are straightforward, as the application below demonstrates. The LM statistic derived by Kiefer (1982) is:

$$\lambda_{\text{LM}} = \frac{\left\{\sum_{i=1}^{n}\left[q_i, 1q_i, 2\frac{\phi(\mathbf{w}_{i,1}'\boldsymbol{\theta}_1)\phi(\mathbf{w}_{i,2}'\boldsymbol{\theta}_2)}{\Phi(\mathbf{w}_{i,1}'\boldsymbol{\theta}_1)\Phi(\mathbf{w}_{i,2}'\boldsymbol{\theta}_2)}\right]\right\}^2}{\sum_{i=1}^{n}\left\{\frac{\left[\phi(\mathbf{w}_{i,1}'\boldsymbol{\theta}_1)\phi(\mathbf{w}_{i,2}'\boldsymbol{\theta}_2)\right]^2}{\Phi(\mathbf{w}_i', 1\boldsymbol{\theta}_1)\Phi(-\mathbf{w}_i', 1\boldsymbol{\theta}_1)\Phi(\mathbf{w}_i', 2\boldsymbol{\theta}_2)\Phi(-\mathbf{w}_i', 2\boldsymbol{\theta}_2)}\right\}},$$

where the two coefficient vectors are the MLEs from the univariate probit models estimated separately.

11.4.2 Recursive simultaneous equations

Section 11.3.5 considered a type of simultaneous equations model in which an endogenous regressor appears on the right-hand side of a probit model. Two other simultaneous equations specifications have attracted interest. Amemiya (1985) demonstrates that a fully simultaneous bivariate probit model,

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \gamma_1 d_{i,2} + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \gamma_2 d_{i,1} + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, \end{aligned}$$

is internally inconsistent and unidentified. However, a recursive model:

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, \qquad\qquad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \gamma_2 d_{i,1} + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, \\ (\varepsilon_{i,1}&\varepsilon_{i,2}) \sim N_2[(0,0),(1,1),\rho], \end{aligned}$$

is a straightforward extension of the bivariate model. For estimation of this model, we have the counterintuitive result that it can be fitted as an ordinary bivariate probit model with the additional right-hand-side variable in the second equation, ignoring the simultaneity. The recent literature provides a variety of applications of this model, including Greene (1998), Fabbri, Monfardini and Radice (2004), Kassouf and Hoffman (2006), White and Wolaver (2003), Gandelman (2005) and Greene, Rhine and Toussaint-Comeau (2006).

Interpretation of the components of this model is particularly complicated. Typically, interest will center on the second equation. In Greene (1998), the second equation concerned the presence of a gender economics course in a college curriculum, while the first equation specified the presence of a women's studies program on the campus. In Kassouf and Hoffman (2006), the authors were interested in the occurrence of work-related injuries while the first conditioning equation specified the use or non-use of protective equipment. Fabbri, Monfardini and Radice (2004) analyzed the choice of Cesarean delivery conditioned on hospital type (public or private). In Greene, Rhine and Toussaint-Comeau (2003), the main equation concerned use of a check cashing facility while the conditioning event in the first equation was whether or not the individual participated in the banking system. In all of these cases, the margin of interest is the impact of the variables in the model on the probability that $d_{i,2}$ equals one. Because $d_{i,1}$ appears in the second equation, there is (potentially) a direct effect (in $\mathbf{w}_{i,2}$) and an indirect effect transmitted to $d_{i,2}$ through the impact of the variable in question on the probability that $d_{i,1}$ equals one. Details on these computations appear in Greene (2008a) and Kassouf and Hoffmann (2006).

11.4.3 Sample selection in a bivariate probit model

Another bivariate probit model that is related to the recursive model of the preceding section is the *bivariate probit with sample selection*. The structural equations are

$$d_{i,1}^{*} = \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^{*} > 0, \text{ 0 otherwise,}$$

$$d_{i,2}^{*} = \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^{*} > 0, \text{ 0 otherwise, and if } d_{i,1} = 1,$$

$$d_{i,2}, \mathbf{w}_{i,2} \text{ are unobserved when } d_{i,1} = 0,$$

$$(\varepsilon_{i,1}\varepsilon_{i,2}) \sim N_2[(0,0),(1,1),\rho].$$

The first equation is a "selection equation." Presence in the sample of observations for the second equation is determined by the first. Like the recursive model, this framework has been used in a variety of applications. The first was a study of the choice of deductibles in insurance coverage by Wynand and van Praag (1981). Boyes, Hoffman and Low (1989) and Greene (1992) studied loan default in which the application is the selection rule. More recently, McQuestion (2000) has used the model to analyze health status (selection) and health behavior, and Lee, Lee and Eastwood (2003) have studied consumer adoption of computer banking technology.

Estimation of this sample selection model is done by maximum likelihood in one step.[20] The log-likelihood is:

$$\ln L = \sum_{d_{i,a}=0} \ln \Phi(-\mathbf{w}_{i,1}'\boldsymbol{\theta}_1) + \sum_{i=1,d_{i,1}=1}^{n} \ln \Phi_2\left[\mathbf{w}_{i,1}'\boldsymbol{\theta}_1, q_{i,2}\mathbf{w}_{i,2}'\boldsymbol{\theta}_2, q_{i,2}\rho\right].$$

As before, estimation and inference in this model follows the standard procedures.

11.4.4 Multivariate binary choice and the panel probit model

In principle, the bivariate probit model can be extended to an arbitrary number of equations, as:

$$
\begin{aligned}
&d_{i,1}^{*} = \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^{*} > 0,\\
&d_{i,2}^{*} = \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^{*} > 0\\
&\cdots\\
&d_{i,M}^{*} = \mathbf{w}_{i,M}'\boldsymbol{\theta}_M + \varepsilon_{i,M}, d_{i,M} = 1 \text{ if } d_{i,M}^{*} > 0,\\
&\begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \cdots \\ \varepsilon_{i,M} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{12} & 1 & \cdots & \rho_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{1M} & \rho_{2M} & \cdots & 1 \end{pmatrix}\right] \quad = \quad N_2[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

The obstacle to the use of this model is its computational burden. The log-likelihood is computed as follows. Let:

$$
\begin{aligned}
&\mathbf{Q}_i = \text{diag}(q_{i,1}, q_{i,2}, \ldots, q_{i,M})\\
&\mathbf{b}_i = (\mathbf{w}_{i,1}'\boldsymbol{\theta}_1, \mathbf{w}_{i,2}'\boldsymbol{\theta}_2, \ldots, \mathbf{w}_{i,M}'\boldsymbol{\theta}_M)'\\
&\mathbf{c}_i = \mathbf{Q}_i\mathbf{b}_i\\
&\mathbf{D}_i = \mathbf{Q}_i\mathbf{R}\mathbf{Q}_i.
\end{aligned}
$$

Then:

$$\ln L = \sum_{i=1}^{n} \ln \Phi_M[\mathbf{c}_i, \mathbf{D}_i].$$

Evaluation of the M-variate normal c.d.f. cannot be done analytically or with quadrature. It is done with simulation, typically using the GHK (Geweke, Hajivassilou and Keane) simulator.

This form of the model also generalizes the random effects probit model examined earlier. We can relax the assumption of equal cross-period correlations by writing:

$$
\begin{aligned}
&d_{it}^{*} = \mathbf{w}_{it}'\boldsymbol{\theta} + \varepsilon_{it}, \quad d_{it} = 1 \text{ if } d_{it}^{*} > 0, \ 0 \text{ otherwise},\\
&(\varepsilon_{i1}, \ldots, \varepsilon_{iT}) \sim N[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

This is precisely the model immediately above with the constraint that the coefficients in the equations are all the same. In this form, it is conventionally labeled the *panel probit model*.[21] Bertschek and Lechner (1998) devised a GMM estimator to circumvent the computational burden of this model. Greene (2004a) examined the same model, and considered alternative computational procedures as well as some variations of the model specification.

11.4.5 Application

Riphahn, Wambach and Million (2003) studied the joint determination of two counts, doctor visits and hospital visits. One would expect these to be highly correlated, so a bivariate probit model should apply to DOCTOR = 1 (DocVis > 0) and HOSPITAL = 1(HospVis > 0). The simple product moment correlation coefficient is inappropriate for binary variables. The tetrachoric correlation is used instead; this turns out to be the estimate of ρ in a bivariate probit model in which both equations contain only a constant term. The first estimated model in Table 11.4 reports a value of 0.311 with a standard error of only 0.0136, so the results are consistent with the conjecture. The second set of estimates assume $\rho = 0$; the estimates for the "Doctor" equation are reproduced from Table 11.3. As noted, there is evidence that ρ is positive. Kiefer's (1982) LM statistic equals 399.20. The limiting distribution is chi-squared with one degree of freedom – the 5% critical value is 3.84, so the hypothesis that the outcomes are uncorrelated is rejected. The Wald and likelihood ratio statistics based on the unrestricted model are $21.496^2 = 462.08$ and $2[17670.94 + 8084.465 - 25534.46] = 441.998$, respectively, so the hypothesis is rejected by all three tests. The third model shown in Table 11.4 is the unrestricted bivariate probit model, while the fourth is the recursive bivariate probit model with DOCTOR added to the right-hand side of the HOSPITAL equation. The results do not support this specification; the log-likelihood is almost unchanged. It is noteworthy that in this expanded specification, the estimate of ρ is no longer significant, as might have been expected.

Table 11.4 Estimated bivariate probit models (standard errors in parentheses)

	(1) *Tetrachoric corr.*		*(2)* *Uncorrelated*		*(3)* *Bivariate probit*		*(4)* *Recursive probit*	
	Doctor	*Hospital*	*Doctor*	*Hospital*	*Doctor*	*Hospital*	*Doctor*	*Hospital*
Constant	0.329	–1.355	0.155	–1.246	0.155	–1.249	0.155	–1.256
	(.0077)	(.0107)	(.0565)	(.0809)	(.0565)	(.0773)	(.0565)	(.481)
Age	.000	.000	.0128	.00488	.0128	.00489	.0128	.00486
	(.000)	(.000)	(.0008)	(.0011)	(.0008)	(.0011)	(.0008)	(.0025)
HhNInc	.000	.000	–.116	.0421	–.118	.0492	–.118	.0496
	(.000)	(.000)	(.0463)	(.0633)	(.0462)	(.0595)	(.0463)	(.0652)
HhKids	.000	.000	–.141	–.0147	–.141	–.0129	–.141	–.0125
	(.000)	(.000)	(.0182)	(.0256)	(.0181)	(.0257)	(.0181)	(.0386)
Educ	.000	.000	–.0281	–.026	–.028	–.026	–.028	–.026
	(.000)	(.000)	(.0035)	(.0052)	(.0035)	(.0051)	(.0035)	(.0066)
Married	.000	.000	.0522	–.0547	.0519	–.0546	.0519	–.0548
	(.000)	(.000)	(.0205)	(.0279)	(.0205)	(.0277)	(.0205)	(.0313)
Doctor								.00912
								(.663)
ρ	.311	(.0136)	.000		.303	(.0138)	.298	(.366)
Ln*L*	–25898.27		–1767.94	–8084.47	–25534.46		–25534.46	

11.5 Ordered choice

In the preceding sections, the consumer is assumed to maximize utility over a pair of alternatives. Models of ordered choice describe settings in which individuals reveal the strength of their utility with respect to a single outcome. For example, in a survey of voter preferences over a single issue (a new public facility or project, a political candidate, etc.), random utility is, as before:

$$U_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i.$$

The individual reveals a censored version of U_i^* through a discrete response, e.g.,

$$\begin{aligned} y_i = 0 &: \text{strongly dislike} \\ 1 &: \text{mildly dislike} \\ 2 &: \text{indifferent} \\ 3 &: \text{mildly prefer} \\ 4 &: \text{strongly prefer.} \end{aligned}$$

The translation between the underlying U_i^* and the observed y_i produces the ordered choice model:

$$\begin{aligned} y_i = 0 \quad & \text{if } U_i^* \le \mu_0 \\ 1 \quad & \text{if } \mu_0 < U_i^* \le \mu_1 \\ 2 \quad & \text{if } \mu_1 < U_i^* \le \mu_2 \\ & \cdots \\ J \quad & \text{if } \mu_{J-1} < U_i^* \le \mu_J, \end{aligned}$$

where $\mu_0, \ldots, \mu_J$ are threshold parameters that are to be estimated with the other model parameters subject to $\mu_j > \mu_{j-1}$ for all j. Assuming $\boldsymbol{\beta}$ contains a constant term, the distribution is located by the normalization $\mu_0 = 0$. At the upper tail, $\mu_J = +\infty$. Probabilities for the observed outcomes are derived from the laws of probability:

$$\text{Prob}(y_i = j | x_i, z_i) = \text{Prob}(\mu_{j-1} < U_i^* \le \mu_j), \text{ where } \mu_{-1} = -\infty.$$

As before, the observed data do not reveal information about the scaling of ε_i, so the variance is normalized to one. Two standard cases appear in the literature; if ε_i has a normal distribution, then the ordered probit model emerges, while if it has the standardized logistic distribution, the ordered logit model is produced. (Other distributions have been suggested as the model is internally consistent with any continuous distribution over the real line. However, these two overwhelmingly dominate the received applications.)

By the laws of probability:

$$\begin{aligned}\text{Prob}(y_i = j|\mathbf{x}_{i,},\ \mathbf{z}_i) &= \text{Prob}(U_i^* \le \mu_j) - \text{Prob}(U_i^* \le \mu_{j-1}) \\ &= F(\mu_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma}) - F(\mu_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma}),\end{aligned}$$

where $F(t)$ is the assumed c.d.f., either normal or logistic. These are the terms that enter the log-likelihood for a sample of n observations. The standard conditions for maximum likelihood estimation apply here. The results in Table 11.1 suggest that the force of the incidental parameters problem in the fixed effects case is similar to that for the binomial probit model.

As usual in discrete choice models, partial effects in this model differ substantively from the coefficients. Note, first, that there is no obvious regression at work. Since y_i is merely a labeling with no implicit scale, there is no conditional mean function to analyze. In order to analyze the impact of changes in a variable, say income, one can decompose the set of probabilities. For a continuous variable $x_{i,k}$, e.g.:

$$\begin{aligned}\delta_{i,k}(j) &= \partial\text{Prob}(y_i = j|x_i, z_i)/\partial x_{i,k} \\ &= -\boldsymbol{\beta}_k[f(\mu_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma}) - f(\mu_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma})], \quad j = 0, \ldots, J,\end{aligned}$$

where $f(t)$ is the density, $\text{d}F(t)/\text{d}t$. The sign of the partial effect is ambiguous, since the difference of the two densities can have either sign. Moreover, since $\Sigma_{j=0}^{J}\text{Prob}(y_i = 1|x_i, z_i) = 1$, it follows that $\Sigma_{j=1}^{J}\delta_{i,k}(j) = 0$. Since the c.d.f. is monotonic, there is one sign change in the set of partial effects as the example below demonstrates. For purposes of using and interpreting the model, it seems that the coefficients are of relatively little utility – neither the sign nor the magnitude directly indicates the effect of changes in a variable on the observed outcome.

Terza (1985) and Pudney and Shields (2000) suggested an extension of the ordered choice model that would accommodate heterogeneity in the threshold parameters. The extended model is:

$$\text{Prob}(y_i = j|x_i, z_i) = F(\mu_{i,j} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma}) - F(\mu_{i,j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\boldsymbol{\gamma}),$$

where:

$$\mu_{i,j} = \mathbf{v}_i'\boldsymbol{\pi}_j, \quad \text{where } \pi_0 = 0,$$

for a set of variables $\mathbf{v}_i$. The model as shown has two complications First, it is straightforward to constrain the fixed threshold parameters to preserve the ordering needed to ensure that all probabilities are positive.[22] When there are variables v_i in the construction, it is no longer possible to produce this result parametrically. The authors (apparently) did not find it necessary to confront this constraint. A second feature of the model (which was examined at length by the authors) is the unidentifiability of elements of $\boldsymbol{\pi}_j$ when v_i and $(\mathbf{x}_i, \mathbf{z}_i)$ contain the same variables. This is a result of the linear functional form assumed for $\mu_{i,j}$. Greene (2007a) and

Harris and Zhao (2007) suggested alternative parameterizations that circumvent these problems – a restricted version:

$$\mu_{i,j} = \exp(\mu_j + \mathbf{v}_i'\boldsymbol{\pi}_j),$$

and a counterpart to Pudney and Shields' (2000) formulation:

$$\mu_{i,j} = \exp(\mathbf{v}_i'\boldsymbol{\pi}_j).^{23}$$

11.5.1 Specification analysis

As in the binary choice case, the analysis of micro-level data is likely to encounter individual heterogeneity, not only in the means of utilities ($\mathbf{x}_i$,$\mathbf{z}_i$), but also in the scaling of U_i^*, i.e., in the variance of ε_i. Building heteroskedasticity into the model, as in the binary choice model shown earlier, is straightforward. If:

$$E[\varepsilon_i^2|\mathbf{v}_i] = [\exp(\mathbf{v}_i'\boldsymbol{\tau})]^2,$$

then the log-likelihood would become:

$$\ln L = \sum_{i=1}^{j} \ln\left[F\left(\frac{\mu_{y_i-1} - x_i'\boldsymbol{\beta} - z_i'\boldsymbol{\gamma}}{\exp(v_i'\boldsymbol{\tau})}\right) - F\left(\frac{\mu_{y_i} - x_i'\boldsymbol{\beta} - z_i'\boldsymbol{\gamma}}{\exp(v_i'\boldsymbol{\tau})}\right)\right].$$

As before, this complicates (even further) the interpretation of the model components and the partial effects.

There is no direct test for the distribution, since the alternatives are not nested. The Vuong test is a possibility, although the power of this test and its characteristics remain to be examined both analytically and empirically.

11.5.2 Bivariate ordered probit models

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 11.4. A direct analog to the base case two-equation model was used by Butler, Finegan and Siegfried (1998), who analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit in order to compress the description.) Step 1 involved a direct application of the ordered probit model to the level of calculus achievement, which is coded 0, 1, . . . , 6:

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i|\mathbf{x}_i \sim N[0,1], \\ m_i &= 0 \text{ if } -\infty < m_i^* \leq 0, \\ &\quad 1 \text{ if } 0 < m_i^* \leq \mu_1, \\ &\quad \cdots \\ &\quad 6 \text{ if } \mu_5 < m_i^* < +\infty. \end{aligned}$$

The authors argued that, although the various calculus courses can be ordered discretely by the material covered, the differences between the levels cannot be measured directly. Thus, this is an application of the ordered probit model. The independent variables in this first-step model included SAT (scholastic aptitude test) scores, foreign language proficiency, indicators of intended major, and several other variables related to areas of study.

The second step of the estimator involves regression analysis of the grade in the intermediate microeconomics or macroeconomics course. Grades in these courses were translated to a granular continuous scale (A = 4.0, A− = 3.7, etc.). A linear regression is then specified:

$$Grade_i = \mathbf{z}_i'\boldsymbol{\delta} + u_i, \quad \text{where } u_i|\mathbf{z}_i \sim N[0, \sigma_u^2].$$

Independent variables in this regression include, among others, (1) dummy variables for which the outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman grade point average, etc. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that $(\varepsilon_i, u_i|\mathbf{x}_i,\mathbf{z}_i) \sim N_2[(0,0),(1,\sigma_u^2),\rho\sigma_u]$. A non-zero ρ captures this "selection" effect. With this in place, the dummy variables now become endogenous. The solution is a "selection" correction where the modified equation becomes:

$$\begin{aligned} Grade_i|m_i &= \mathbf{z}_i'\boldsymbol{\delta} + E[u_i|m_i] + v_i \\ &= \mathbf{z}_i'\boldsymbol{\delta} + (\rho\sigma_u)[\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ldots, \mu_5)] + v_i. \end{aligned}$$

They thus adopt a "control function" approach to accommodate the endogeneity of the math attainment dummy variables. The term $\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ldots, \mu_5)$ is a generalized residual that is constructed using the estimates from the first-stage ordered probit model (a precise statement of the form of this variable is given in Tobias and Li, 2006). Linear regression of the course grade on $\mathbf{z}_i$ and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (1985) correction.

Tobias and Li (2006), in a replication of and comment on Butler, Finegan and Siegfried (1998), after roughly replicating the classical estimation results with a Bayesian estimator, observe that the *Grade* equation above could also be treated as an ordered probit model. The resulting bivariate ordered probit model would be:

$$\begin{array}{rlcrl} m_i^* = & \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, & \text{and} & g_i^* & = \mathbf{z}_i'\boldsymbol{\delta} + u_i, \\ m_i = & 0 \text{ if } -\infty < m_i^* \le 0, & & g_i & = 0 \text{ if } -\infty < g_i^* \le 0, \\ & 1 \text{ if } 0 < m_i^* \le \mu_1, & & & 1 \text{ if } 0 < g_i^* \le \alpha_1, \\ & \ldots & & & \ldots \\ & 6 \text{ if } \mu_5 < m_i^* < +\infty & & & 11 \text{ if } \mu_9 < g_i^* < +\infty, \end{array}$$

where $(\varepsilon_i,u_i|\mathbf{x}_i,\mathbf{z}_i) \sim N_2[(0,0),(1,\sigma_u^2),\rho\sigma_u]$.

Tobias and Li extended their analysis to this case simply by "transforming" the dependent variable in Butler, Finegan and Siegfried's second equation. Computing the log-likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model (see Greene, 2007b). However, the classical study of these data using the bivariate ordered approach remains to be done, so a side-by-side comparison to Tobias and Li's Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appear.

The bivariate ordered probit model has been applied in a number of settings in the recent empirical literature, including husband and wife's education levels (Magee, Burbidge and Robb, 2000), family size (Calhoun, 1991) and many others. In two early contributions to the field of pet econometrics, Butler and Chatterjee analyze ownership of cats and dogs (1995) and dogs and televisions (1997).

11.5.3 Panel data applications

11.5.3.1 Fixed effects

D'Addio, Eriksson and Frijters (2003), using methodology developed by Frijters, Haisken-DeNew and Shields (2004) and Ferrer-i-Carbonel and Frijters (2004), analyzed survey data on job satisfaction using the Danish component of the European Community Household Panel. Their estimator for an ordered logit model is built around the logic of Chamberlain's estimator for the binary logit model. The approach is robust to individual specific threshold parameters and allows time invariant variables, so it differs sharply from the fixed effects models we have considered thus far, as well as from the ordered probit model.[24] Unlike Chamberlain's estimator for the binary logit model, however, their conditional estimator is not a function of minimal sufficient statistics. As such, the incidental parameters problem remains an issue.

Das and van Soest (2000) proposed a somewhat simpler approach. (See, as well, Long's 1997 discussion of the "parallel regressions assumption," which employs this device in a cross-section framework.) Consider the base case ordered logit model with fixed effects:

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it}|\mathbf{X}_i \sim N[0,1]$$

$$y_{it} = j \text{ if } \mu_{j-1} < y_{it}^* < \mu_j, j = 0, 1, \ldots, J \text{ and } \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that:

$$\text{Prob}(y_{it} = j|\mathbf{X}_i) = \Lambda(\mu_j - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) - \Lambda(\mu_{j-1} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where $\Lambda(t)$ is the c.d.f. of the logistic distribution. Now, define a binary variable:

$$w_{it,j} = 1 \quad \text{if } y_{it} > j, \quad j = 0, \ldots, J-1.$$

It follows that:

$$\begin{aligned}\text{Prob}[w_{it,j} = 1|\mathbf{X}_i] &= \Lambda(\alpha_i - \mu_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= \Lambda(\theta_i + \mathbf{x}_{it}'\boldsymbol{\beta}).\end{aligned}$$

The "j" specific constant, which is the same for all individuals, is absorbed in θ_i. Thus a fixed effects binary logit model applies to each of the $J-1$ binary random variables, $w_{it,j}$. The method in section 11.3 can now be applied to each of the $J-1$ random samples. This provides $J-1$ estimators of the parameter vector $\boldsymbol{\beta}$ (but no estimator of the threshold parameters). The authors propose to reconcile these different estimators by using a minimum distance estimator of the common true $\boldsymbol{\beta}$. The minimum distance estimator at the second step is chosen to minimize:

$$q = \sum_{j=0}^{J-1} \sum_{m=0}^{J-1} \left(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}\right)' \left[\mathbf{V}_{jm}^{-1}\right] \left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right),$$

where $\left[\mathbf{V}_{jm}^{-1}\right]$ is the j,m block of the inverse of the $(J-1)K \times (J-1)K$ partitioned matrix $\mathbf{V}$ that contains Asy.Cov$\left[\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_m\right]$. The appropriate form of this matrix for a set of cross-section estimators is given in Brant (1990). Das and van Soest (2000) used the counterpart of Chamberlain's fixed effects estimator, but do not provide the specifics for computing the off diagonal blocks in $\mathbf{V}$.

The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Greene (2008a, sec. 16.9.6.c). The likelihood function is concave (see Pratt, 1981), so despite its superficial complexity, the estimation is straightforward. (In the application below, with over 27,000 observations and 7,293 individual effects, estimation of the full model required roughly five seconds of computation.) No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) results on the small T bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model. The Monte Carlo results in Table 11.1 suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. As in the binary choice case, the complication of the fixed effects model is the small sample bias, not the computation. The Das and van Soest (2000) approach finesses this problem, as their estimator is consistent, but at the cost of losing the information needed to compute partial effects or predicted probabilities.

11.5.3.2 Random effects

The random effects ordered probit model has been much more widely used than the fixed effects model. Applications include Groot and van den Brink (2003), who studied training levels of employees, with firm effects and gains to marriage, Winkelmann (2004), who examined subjective measures of well-being with individual and family effects, Contoyannis, Jones and Rice (2004), who analyzed self-reported measures of health status, and numerous others. In the simplest case, the quadrature method of Butler and Moffitt (1982) can be used.

11.5.4 Application

The GSOEP data that we have used earlier includes a self-reported measure of health satisfaction, *HSAT*, that takes values 0, 1, . . . , 10. This is a typical application

of a scale variable that reflects an underlying continuous variable, "health." The frequencies and sample proportions for the reported values are as follows:

NEWHSAT	0	1	2	3	4	5	6	7	8	9	10
Frequency	447	255	642	1173	1390	4233	2530	4231	6172	3061	3192
Proportion	1.6%	0.9%	2.3%	4.2%	5.0%	15.4%	9.2%	15.4%	22.5%	11.2%	11.6%

We have fitted pooled and panel data versions of the ordered probit model to these data. The model used is:

$$U_{it}^{*} = \beta_1 + \beta_2 Age_{it} + \beta_3\, Income_{it} + \beta_4\, Education_{it} + \beta_5\, Married_{it} + \beta_6\, Working_{it} + \varepsilon_{it} + c_i,$$

where c_i will be the common fixed or random effect. (We are interested in comparing the fixed and random effects estimators, so we have not included any time invariant variables such as gender in the equation.) Table 11.5 lists five estimated models. (Standard errors for the estimated threshold parameters are omitted.) The first is the pooled ordered probit model. The second and third are fixed effects. Column 2 shows the unconditional fixed effects estimates using the results in Greene (2008). Column 3 shows the Das and van Soest (2000) estimator. For the minimum distance estimator, we used an inefficient weighting matrix, the block diagonal matrix in which the *j*th block is the inverse of the *j*th asymptotic covariance matrix for the individual logit estimators. With this weighting matrix, the estimator is:

$$\hat{\beta}_{MDE} = \left[\sum_{j=0}^{9} \mathbf{V}_j^{-1}\right]^{-1} \sum_{j=0}^{9} \mathbf{V}_j^{-1} \hat{\beta}_j,$$

and the estimator of the asymptotic covariance matrix is approximately equal to the bracketed inverse matrix. The fourth set of results is the random effects estimator computed using the MSL method. This model can be estimated using Butler and Moffitt's quadrature method: however, we found that, even with a large number of nodes, the quadrature estimator converged to a point where the log-likelihood was far lower than the MSL estimator, and at parameter values that were implausibly different from the other estimates. Using different starting values and different numbers of quadrature points did not change this outcome. The MSL estimator for a random constant term is considerably slower, but produces more reasonable results. The fifth set of results is the Mundlak form of the random effects model, which includes the group means in the models as controls to accommodate possible correlation between the latent heterogeneity and the included variables. As noted earlier, the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 11.6 shows the partial effects for the pooled model to illustrate the computations.

Table 11.5 Estimated ordered probit models for health satisfaction

	(1) Pooled	*(2) Fixed effects uncon-ditional*	*(3) Fixed effects conditional*	*(4) Random effects*	*(5) Random effects Mundlak controls*	
Variable					Variables	Means
Constant	2.4739			3.8577	3.2603	
	(0.04669)			(0.05072)	(0.05323)	
Age	–0.01913	–0.07162	–0.1011	–0.03319	–0.06282	0.03940
	(0.00064)	(0.002743)	(0.002878)	(0.00065)	(0.00234)	(0.002442)
Income	0.1811	0.2992	0.4353	0.09436	0.2618	0.1461
	(0.03774)	(0.07058)	(0.07462)	(0.03632)	(0.06156)	(0.07695)
Kids	0.06081	–0.06385	–0.1170	0.01410	–0.05458	0.1854
	(0.01459)	(0.02837)	(0.03041)	(0.01421)	(0.02566)	(0.03129)
Education	0.03421	0.02590	0.06013	0.04728	0.02296	0.02257
	(0.002828)	(0.02677)	(0.02819)	(0.002863)	(0.02793)	(0.02807)
Married	0.02574	0.05157	0.08505	0.07327	0.04605	–0.04829
	(0.01623)	(0.04030)	(0.04181)	(0.01575)	(0.03506)	(0.03963)
Working	0.1292	–0.02659	–0.007969	0.07108	–0.02383	0.2702
	(0.01403)	(0.02758)	(0.02830)	(0.01338)	(0.02311)	(0.02856)
μ_1	0.1949	0.3249		0.2726	0.2752	
μ_2	0.5029	0.8449		0.7060	0.7119	
μ_3	0.8411	1.3940		1.1778	1.1867	
μ_4	1.111	1.8230		1.5512	1.5623	
μ_5	1.6700	2.6992		2.3244	2.3379	
μ_6	1.9350	3.1272		2.6957	2.7097	
μ_7	2.3468	3.7923		3.2757	3.2911	
μ_8	3.0023	4.8436		4.1967	4.2168	
μ_9	3.4615	5.5727		4.8308	4.8569	
σ_u	0.0000	0.0000		1.0078	0.9936	
ln$L$	–56813.52	–41875.63		–53215.54	–53070.43	

Winkelmann (2004) used the random effects approach to analyze the subjective well-being (SWB) question (also coded 0 to 10) in the GSOEP dataset. The ordered probit model in this study is based on the latent regression:

$$y^*_{imt} = \mathbf{x}'_{imt}\boldsymbol{\beta} + \varepsilon_{imt} + u_{im} + v_i.$$

The independent variables include age, gender, employment status, income, family size and an indicator for good health. An unusual feature of the model is the nested random effects, which include a family effect, v_i, as well as the individual family member (i in family m) effect, u_{im}. The MLE approach is unavailable in this nonlinear setting. Winkelmann instead employed a Hermite quadrature procedure to maximize the log-likelihood function.

Contoyannis, Jones and Rice (2004) analyzed a self-assessed health scale that ranged from 1 (very poor) to 5 (excellent) in the BHPS. Their model accommodated

Table 11.6 Estimated marginal effects: pooled model

HSAT	*Age*	*Income*	*Kids*	*Education*	*Married*	*Working*
0	0.0006	−0.0061	−0.0020	−0.0012	−0.0009	−0.0046
1	0.0003	−0.0031	−0.0010	−0.0006	−0.0004	−0.0023
2	0.0008	−0.0072	−0.0024	−0.0014	−0.0010	−0.0053
3	0.0012	−0.0113	−0.0038	−0.0021	−0.0016	−0.0083
4	0.0012	−0.0111	−0.0037	−0.0021	−0.0016	−0.0080
5	0.0024	−0.0231	−0.0078	−0.0044	−0.0033	−0.0163
6	0.0008	−0.0073	−0.0025	−0.0014	−0.0010	−0.0050
7	0.0003	−0.0024	−0.0009	−0.0005	−0.0003	−0.0012
8	−0.0019	0.0184	0.0061	0.0035	0.0026	0.0136
9	−0.0021	0.0198	0.0066	0.0037	0.0028	0.0141
10	−0.0035	0.0336	0.0114	0.0063	0.0047	0.0233

a variety of complications in survey data. The latent regression underlying their ordered probit model is:

$$h_{it}^{*} = \mathbf{x}_{it}^{\prime}\boldsymbol{\beta} + \mathbf{H}_{i,t-1}^{\prime}\boldsymbol{\gamma} + \alpha_i + \varepsilon_{it},$$

where $\mathbf{x}_{it}$ includes marital status, race, education, household size, age, income, and the number of children in the household. The lagged value, $\mathbf{H}_{i,t-1}$, is a set of binary variables for the observed health status in the previous period. In this case, the lagged values capture state dependence, as the assumption that the health outcome is redrawn randomly in each period is inconsistent with evident runs in the data. The initial formulation of the regression is a fixed effects model. To control for the possible correlation between the effects, α_i, and the regressors, and the initial conditions problem that helps to explain the state dependence, they use a hybrid of Mundlak's (1978) correction and a suggestion by Wooldridge (2002b) for modeling the initial conditions:

$$\alpha_i = \alpha_0 + \overline{\mathbf{x}}^{\prime}\alpha_1 + \mathbf{H}_{i,1}^{\prime}\boldsymbol{\delta} + u_i,$$

where u_i is exogenous. Inserting the second equation into the first produces a random effects model that can be fitted using Butler and Moffitt's (1982) quadrature method.

11.6 Models for counts

A model that is often used for interarrival times at such facilities as a telephone switch, an ATM machine, or at the service window of a bank or gasoline station, is the *exponential model*:

$$f(t) = \theta \exp(-\theta t), \quad t \geq 0, \theta > 0,$$

where the continuous variable, t, is the time between arrivals. The expected interarrival time in this distribution is $E[t] = 1/\theta$. Consider the number of arrivals, y,

that occur *per unit of time*. It can be shown that this discrete random variable has the *Poisson probability distribution*:

$$f(y) = \exp(-\lambda)\lambda^{y}/y!, \quad \lambda = 1/\theta > 0, y = 0, 1, \ldots$$

The expected value of this discrete random variable is $E[y] = 1/\theta$. The *Poisson regression model* arises from the specification:

$$E[y_i|\mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}).$$

The log-linear form is used to ensure that the mean is positive. Estimation of the Poisson model by ML is straightforward owing to the simplicity of the log-likelihood and its derivatives:

$$\begin{array}{ll} \ln L & = \sum_{i=1}^{n} -\lambda_i + y_i(\mathbf{x}_i'\boldsymbol{\beta}) - \ln\Gamma(y_i + 1) \\ \partial \ln L/\partial\boldsymbol{\beta} & = \sum_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i \\ \partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' & = \sum_{i=1}^{n} -\lambda_i x_i \mathbf{x}_i'. \end{array}$$

Inference about parameters is based on either the actual (and expected) Hessian:

$$V = \left[\sum_{i=1}^{n} \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} = \left[\mathbf{X}'\hat{\Lambda}\mathbf{X}\right]^{-1},$$

or the BHHH estimator, which is:

$$\mathbf{V}_{\text{BHHH}} = \left[\sum_{i=1}^{n}(y_i - \hat{\lambda}_i)^2 x_i \mathbf{x}_i'\right]^{-1} = \left[\sum_{i=1}^{n} \varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'\right]^{-1} = \left[\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}\right]^{-1}.$$

Hypothesis tests about the parameters may be based on the likelihood ratio or Wald statistics, or the LM statistic, which is particularly convenient here:

$$\lambda_{LM} = \left[\sum_{i=1}^{n} \hat{\varepsilon}_i^0 x\right]' V_{\text{BHHH}} \left[\sum_{i=1}^{n} \hat{\varepsilon}_i^0 x\right],$$

where the residuals are computed at the restricted estimates. For example, under the null hypothesis that all coefficients are zero save for the constant term, $\hat{\lambda}_i^0 = \overline{y}$, $\hat{\varepsilon}_i^0 = y_i - \overline{y}$ and:

$$\lambda_{LM} = \left[\sum_{i=1}^{n}(y_i - \overline{y})\mathbf{x}_i\right]' \left[\sum_{i=1}^{n}(y_i - \overline{y})^2 \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_{i=1}^{n}(y_i - \overline{y})\mathbf{x}_i\right].$$

The Poisson model is one in which the MLE is robust to certain misspecifications of the model, such as the failure to incorporate latent heterogeneity into the mean (i.e., one fits the Poisson model when the negative binomial is appropriate.) In this case, the robust (sandwich) covariance matrix:

$$\text{Robust Est.Asy.Var}\left[\hat{\boldsymbol{\beta}}\right] = \left[\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}\right]^{-1} \left[\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}\right] \left[\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}\right]^{-1},$$

is appropriate to accommodate this failure of the model. It has become common to employ this estimator with all specifications, including the negative binomial. One might question the virtue of this. Since the negative binomial model already accounts for the latent heterogeneity, it is unclear what *additional* failure of the assumptions of the model this estimator would be robust to.

11.6.1 Heterogeneity and the negative binomial model

The Poisson model is typically only the departure point for the analysis of count data. The simple model has (at least) two shortcomings that arise from heterogeneity that is not explicitly incorporated in the model.

One easily remedied minor issue concerns the units of measurement of the data. In the Poisson model (and negative binomial model below), the parameter λ_i is the expected number of events *per unit of time*. Thus there is a presumption in the model formulation, e.g., the Poisson, that the same amount of time is observed for each *i*. In a spatial context, such as measurements of the incidence of a disease per group of N_i persons, or the number of bomb craters per square mile in London in 1940, the assumption would be that the same physical area or the same size of population applies to each observation. Where this differs by individual, it will introduce a type of heteroskedasticity in the model. The simple remedy is to modify the model to account for the *exposure,* T_i, of the observation as follows:

$$\text{Prob}(y_i = j|\mathbf{x}_i, T_i) = \frac{\exp(-T_i\phi_i)(T_i\phi_i)^j}{j!},\ \phi_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),\ j = 0, 1, \ldots.$$

The original model is returned if we write $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \ln T_i)$. Thus, when the exposure differs by observation, the appropriate accommodation is to include the log of exposure in the regression part of the model with a coefficient of 1.0. (For less than obvious reasons, the term "*offset variable*" is commonly associated with the exposure variable T_i.) Note that if T_i is the same for all *i*, ln *T* will simply vanish into the constant term of the model (assuming one is included in $\mathbf{x}_i$).

The less straightforward restriction of the Poisson model is that $E[y_i|x_i] = \text{Var}[y_i|x_i]$. This equidispersion assumption is a major shortcoming. Observed data rarely, if ever, display this feature. The very large amount of research activity on functional forms for count models is often focused on testing for equidispersion and building functional forms that relax this assumption.

The overdispersion found in observed data can be attributed to omitted heterogeneity in the Poisson model. A more complete regression specification would be:

$$E[y_i|\mathbf{x}_i] = \lambda_i = h_i \exp(\mathbf{x}_i'\boldsymbol{\beta}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i),$$

where the heterogeneity, h_i, has mean one and non-zero variance. Two candidates for the distribution of ε_i have dominated the literature, the log-normal model discussed later and the log-gamma model. The most common specification is the log-gamma model, which derives from the gamma variable:

$$f[h_i] = [\theta^\theta / \Gamma(\theta)] \exp(-\theta h_i) h_i^{\theta-1}, \quad h_i \geq 0.^{25}$$

This gamma distributed random variable has mean 1.0 and variance $1/\theta$. (A separate variance parameter is not identified – the scaling in the model is, once again, absorbed by the coefficient vector.) If we write the Poisson–log-gamma model as:

$$f(y_i|\mathbf{x}_i, h_i) = \exp(-h_i\lambda_i)(h_i\lambda_i)^{y_i} / \Gamma(y_i + 1),$$

then the unconditional distribution is:

$$f(y_i|\mathbf{x}_i) = \int_0^\infty f(y_i, v_i|\mathbf{x}_i)dv_i = \int_0^\infty f(y_i|\mathbf{x}_i, v_i)f(v_i)dv_i.$$

The integral can be obtained in closed form; the result is the *negative binomial model*:

$$\begin{aligned}\text{Prob}(Y = y_i|\mathbf{x}_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^\theta,\\ \lambda_i &= \exp(\mathbf{x}_i'\beta),\\ r_i &= \lambda_i/(\theta + \lambda_i).\end{aligned}$$

The recent literature, mostly associating the result with Cameron and Trivedi (1986, 1998), defines this form of the negative binomial model as the *Negbin 2* (NB2) form of the probability. This is the default form of the model in the received econometrics packages that provide an estimator for this model. The *Negbin 1* (NB1) form of the model results if θ is replaced with $\theta_i = \theta\lambda_i$. Then, r_i reduces to $r = 1/(1 + \theta)$, and the density becomes:

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i)} r^{y_i}(1 - r)^{\theta\lambda_i}.$$

This is not a simple reparameterization of the model. The results in the example below demonstrate that the log-likelihood functions of the two forms are not equal at the maxima, and their parameters are not simple transformations of each other. We are not aware of a theory that justifies using one form or the other for the negative binomial model. Neither is a restricted version of the other, so we cannot carry out a nested likelihood ratio test. The more general *Negbin P* (NBP) family (Greene, 2008b) does nest both of them, so this may provide a more general, encompassing approach to finding the right specification. The NBP model is obtained by replacing θ in the NB2 form with $\theta\lambda_i^{2-P}$. We have examined the cases of $P = 1$ and $P = 2$ above and, for general P:

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i^Q + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i^Q)} \left(\frac{\lambda}{\theta\lambda_i^Q + \lambda_i}\right)^{y_i} \left(\frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i}\right)^{\theta\lambda_i^Q}, \quad Q = 2 - P.$$

The conditional mean function for the three cases considered is:

$$E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}) \times \theta^{2-P} = \alpha^{P-2}\lambda_i, \quad \text{where } \alpha = 1/\theta.$$

The parameter P is picking up the scaling. A general result is that, for all three variants of the model:

$$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + \alpha\lambda_i^{P-1}).$$

Thus, the NB2 form has a variance function that is quadratic in the mean, while the NB1 form's variance is a simple multiple of the mean. There have been many

other functional forms proposed for count data models, including the generalized Poisson, gamma, and Polya–Aeppli forms described in Winkelmann (2003) and Greene (2007a, Ch. 24).

The heteroskedasticity in the count models is induced by the relationship between the variance and the mean. The single parameter θ picks up an implicit overall scaling, so it does not contribute to this aspect of the model. As in the linear model, microeconomic data are likely to induce heterogeneity in both the mean and variance of the response variable. A specification that allows independent variation of both will be of some virtue. The result:

$$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + (1/\theta)\lambda_i^{P-1}),$$

suggests that a natural platform for separately modeling heteroskedasticity will be the dispersion parameter, θ, which we now parameterize as:

$$\theta_i = \theta \exp(\mathbf{z}_i'\boldsymbol{\delta}).$$

Operationally, this is a relatively minor extension of the model. But it is likely to introduce a quite substantial increase in the flexibility of the specification. Indeed, a heterogeneous NBP model is likely to be sufficiently parameterized to accommodate the behavior of most datasets. (Of course, the specialized models discussed below, e.g., the zero inflation models, may yet be more appropriate for a given situation.)

11.6.2 Extended models for counts: two-part, zero inflation, sample selection, bivariate

"Non-Poissonness" arises from a variety of sources in addition to the latent heterogeneity modeled in the previous section. A variety of *two-part models* have been proposed to accommodate elements of the decision process.

11.6.2.1 Hurdle model

The hurdle model (Mullahy, 1986; Gurmu, 1997) consists of a participation equation and a conditional Poisson or negative binomial model. The structural equations are:

$$\begin{aligned}
&\text{Prob}(y_i > 0|\mathbf{z}_i) && = \text{a binary choice mechanism, such as probit or logit}\\
&\text{Prob}(y_i = j|y_i > 0, \mathbf{x}_i) && = \text{truncated Poisson or negative binomial.}
\end{aligned}$$

(See Shaw, 1988.) For a logit participation equation and a Poisson count, the probabilities for the observed data that enter the log-likelihood function would be:

$$\text{Prob}(y_i = 0|\mathbf{z}_i) = \frac{1}{1 + \exp(\mathbf{z}_i'\alpha)}$$

$$\begin{aligned}
\text{Prob}(y_i = j|\mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(y_i > 0|\mathbf{z}_i) \times \text{Prob}(y_i = j|y_i > 0, \mathbf{x}_i)\\
&= \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\alpha})} \frac{\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]}.
\end{aligned}$$

This model might apply for on-site counts of the use of certain facilities such as recreational sites. The expectation in the hurdle model is easily found using the rules of probability:

$$\begin{aligned} E[y_i|\mathbf{x}_i, \mathbf{z}_i] &= \frac{\exp(\mathbf{z}_i'\alpha)}{1+\exp(\mathbf{z}_i'\alpha)} E[y_i|y_i > 0, \mathbf{x}_i, \mathbf{z}_i] \\ &= \frac{\exp(\mathbf{z}_i'\alpha)}{1+\exp(\mathbf{z}_i'\alpha)} \frac{\lambda_i}{[1-\exp(-\lambda_i)]}. \end{aligned}$$

As usual, the intricacy of the function mandates some caution in interpreting the model coefficients. In particular:

$$\begin{aligned} \boldsymbol{\delta}_i(\mathbf{x}_i) &= \frac{\partial E[y_i|\mathbf{x}_i,\mathbf{z}_i]}{\partial \mathbf{x}_i} \\ &= \left\{ \frac{\exp(\mathbf{z}_i'\alpha)}{1+\exp(\mathbf{z}_i'\alpha)} \frac{\lambda_i}{[1-\exp(-\lambda_i)]} \left(1 - \frac{\lambda_i \exp(-\lambda_i)}{[1-\exp(-\lambda_i)]}\right) \right\} \boldsymbol{\beta}. \end{aligned}$$

The complication of the partial effects is compounded if $\mathbf{z}_i$ contains any of the variables that also appear in $\mathbf{x}_i$. The other part of the partial effect is:

$$\boldsymbol{\delta}_i(\mathbf{z}_i) = \left\{ \frac{\exp(\mathbf{z}_i'\alpha)}{[1+\exp(\mathbf{z}_i'\alpha)]^2} \frac{\lambda_i}{[1-\exp(-\lambda_i)]} \right\} \alpha.$$

11.6.2.2 Zero inflation models

A related formulation is the *zero inflation model*, which is a type of *latent class model*. The model accommodates a situation in which the zero outcome can arise in either of two mechanisms. In one regime, the outcome is always zero; in the other, the outcome is generated by the Poisson or negative binomial process that might also produce a zero. The example suggested in Lambert's (1992) pioneering application is a manufacturing process that produces a number of defective parts, y_i, equal to zero if the process is under control, or equal to a Poisson outcome if the process is not under control. The applicable distribution is:

$$\begin{aligned} \text{Prob}(y_i = 0|\mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(\text{regime } 0|\mathbf{z}_i) + \text{Prob}(\text{regime } 1|\mathbf{z}_i) \\ &\quad \text{Prob}(y_i = 0|\text{regime1}, \mathbf{x}_i) \\ &= F(r_i|\mathbf{z}_i) + [1 - F(r_i|\mathbf{z}_i)]\text{Prob}(y_i = 0|\mathbf{x}_i) \\ \text{Prob}(y_i = j|y_i > 0, \mathbf{x}_i, \mathbf{z}_i) &= [1 - F(r_i|\mathbf{z}_i)]\text{Prob}(y_i = j|\mathbf{x}_i). \end{aligned}$$

The density governing the count process may be the Poisson or negative binomial model. The regime process is typically specified as a logit model, though the probit model is often used as well. Finally, two forms are used for the regime model, the standard probit or logit model with covariate vector, $\mathbf{z}_i$, and the zero inflated poisson, ZIP(τ) form, which takes the form (for the logit–Poisson model):

$$\begin{aligned} \text{Prob}(y_i = 0|\mathbf{x}_i) &= \Lambda(\tau \mathbf{x}_i'\boldsymbol{\beta}) + [1 - \Lambda(\tau \mathbf{x}_i'\boldsymbol{\beta})] \exp(-\lambda_i) \\ \text{Prob}(y_i = j|y_i > 0, \mathbf{x}_i) &= [1 - \Lambda(\tau \mathbf{x}_i'\boldsymbol{\beta})] \exp(\lambda_i)\lambda_i^j/j!, \end{aligned}$$

where $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ and τ is a single new, free parameter to be estimated. (Researchers usually find that the τ form of the model is more restrictive than desired.) The conditional mean function is:

$$E[y_i|\mathbf{x}_i, \mathbf{z}_i] = [1 - F(r_i|\mathbf{z}_i)]\lambda_i.$$

11.6.2.3 Sample selection

We consider an extension of the classic model of sample selection (Heckman, 1979) to the models for count outcomes. In the context of the applications considered here, for example, we might consider a sample based on only those individuals who have health insurance. The generic model will take the form:

$s_i^* = \mathbf{z}_i'\boldsymbol{\alpha} + u_i, \quad u_i \sim N[0,1],$	
$s_i = \mathbf{1}(\mathbf{s}_i^* > 0)$	(probit selection equation)
$\lambda_i\|\varepsilon_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i + \sigma\varepsilon_i), \varepsilon_i \sim N[0,1]$	(index function with heterogeneity)[26]
$y_i\|\mathbf{x}_i, \varepsilon_i \sim \text{Poisson}(y_i\|\mathbf{x}_i, \varepsilon_i)$	(Poisson model for outcome)
$[u_i, \varepsilon_i] \sim N[(0,1),(1,\rho,1)]$	
$y_i, \mathbf{x}_i$ are observed only when $s_i = 1$.	

The count model is the heterogeneity model suggested earlier with log-normal rather than log-gamma heterogeneity. The conventional approach of fitting the probit selection equation, computing an inverse Mills ratio, and adding it as an extra regressor in the Poisson model, is inappropriate here (see Greene, 1995, 1997, 2006). A formal approach for this model is developed in Terza (1994, 1998) and Greene (1995, 2006, 2007b). Formal results are collected in Greene (2006). The generic result for the count model (which can be adapted to the negative binomial or other models) is:

$$f(y_i, s_i|\mathbf{x}_i, \mathbf{z}_i) = \int_{-\infty}^{\infty} [(1 - s_i) + s_i f(y_i|\mathbf{x}_i, \varepsilon_i)] \Phi\left((2s_i - 1)[\mathbf{z}_i'\boldsymbol{\alpha}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right)\phi(\varepsilon_i)d\varepsilon_i,$$

with:

$$f(y_i|\mathbf{x}_i, \varepsilon_i) = \frac{\exp(-\lambda_i|\mathbf{x}_i, \varepsilon_i)(\lambda_i|\mathbf{x}_i, \varepsilon_i)^{y_i}}{\Gamma(y_i + 1)}, \ \lambda_i|\mathbf{x}_i, \varepsilon_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i + \sigma\varepsilon_i).$$

The integral does not exist in closed form, but the model can be fitted by approximating the integrals with Hermite quadrature:

$$\ln L_Q = \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{\pi}}\sum_{h=1}^{H} \omega_h\left[(1 - s_i) + s_i f(y_i|\mathbf{x}_i, \nu_h)\right]\Phi\left[(2s_i - 1)\left(\mathbf{z}_i'\boldsymbol{\gamma}_i + \tau\nu_h\right)\right]\right],$$

or simulation, for which the simulated log-likelihood is:

$$\ln L_S = \sum_{i=1}^{N} \log\frac{1}{R}\sum_{r=1}^{R}[(1 - s_i) + s_i f(y_i|\mathbf{x}_i, \sigma\varepsilon_{ir})]\Phi[(2s_i - 1)\left(\mathbf{z}_i'\boldsymbol{\gamma}_i + \tau\varepsilon_{ir}\right)],$$

where $\gamma = \boldsymbol{\alpha}/(1-\rho^2)^{1/2}$ and $\tau = \rho/(1-\rho^2)^{1/2}$. There is a minor extension of this model that might be interesting for the health care application examined in this study. The count variables and all the covariates in both equations would be observed for all observations. Thus, to use the full sample of data, the appropriate log-likelihood would be:

$$f(y_i, z_i|\mathbf{x}_i, \mathbf{z}_i) = \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i, \varepsilon_i)]\Phi\left((2s_i - 1)[\mathbf{z}_i'\boldsymbol{\gamma}_i + \tau\varepsilon_i]\right)\phi(\varepsilon_i)d\varepsilon_i.$$

11.6.2.4 Bivariate Poisson model

The application from which our examples are drawn was a study of the two count variables, DocVis (visits to the doctor) and HospVis (visits to the hospital). Riphahn, Wambach and Million (2003) were interested in a bivariate count model for the two outcomes. One approach to formulating a two-equation Poisson model is to treat the correlation as arising from the intervention of a latent common Poisson process. The model is:

$$y_1 = y_1^* + U$$
$$y_2 = y_2^* + U,$$

where y_1^*, y_2^* and U are three independent Poisson processes. This model is analogous to the seemingly unrelated regressions model (see King, 1989). The major shortcoming of this approach is that it forces the two variables to be positively correlated. For the application considered here, it is at least possible that the preventive motivation for physician visits could result in a negative correlation between physician and inpatient hospital visits. The approach proposed by Riphahn, Wambach and A. Million (2003) adapted for a random effects panel data model, is $y_{it,j} \sim$ Poisson $(\lambda_{it,j})$, where:

$$\lambda_{it,j} = \exp(\mathbf{x}_{it,j}'\boldsymbol{\beta} + u_{i,j} + \varepsilon_{it,j}), \qquad j = 1, 2.$$

and where the unique heterogeneity, $(\varepsilon_{it,1}, \varepsilon_{it,2})$, has a bivariate normal distribution with correlation ρ, and the random effects, which are constant through time, have independent normal distributions. Thus the correlation between the conditional means is that induced by the two log-normal variables, $\exp(\varepsilon_{it,1})$ and $\exp(\varepsilon_{it,2})$. The implied correlation between $y_{it,1}$ and $y_{it,2}$ was not derived. This would be weaker than ρ, since both variables have additional variation around the correlated conditional mean functions.

In order to formulate the log-likelihood function, the random components must be integrated out. There are no closed forms for the integrals based on the normal distribution – the problem is similar to that in the sample selection model. The authors used a quadrature procedure to approximate the integrals. The log-likelihood could also be maximized by using simulation. Separate models were fitted to men and women in the sample. The pooling hypothesis was rejected for all specifications considered.

11.6.3 Panel data models

Hausman, Hall and Griliches (1984) (HHG) report the following conditional density for the fixed effects negative binomial (FENB) model:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i} | \sum_{t=1}^{T_i} y_{it}\right) = \frac{\Gamma(1+\Sigma_{t=1}^{T_i} y_{it})\Gamma(\Sigma_{t=1}^{T_i}\lambda_{it})}{\Gamma(\Sigma_{t=1}^{T_i} y_{it} + \Sigma_{t=1}^{T_i}\lambda_{it})} \prod_{t=1}^{T_i} \frac{\Gamma(y_{it}+\lambda_{it})}{\Gamma(1+y_{it})\Gamma(\lambda_{it})},$$

which is free of the fixed effects. This is the default FENB formulation used in popular software packages such as SAS, Stata and LIMDEP. Researchers accustomed to the admonishments that fixed effects models cannot contain overall constants or time invariant covariates are sometimes surprised to find (perhaps accidentally) that this fixed effects model allows both. (This issue is explored at length in Allison, 2000; Allison and Waterman, 2002.) The resolution of this apparent contradiction is that the HHG FENB model is not obtained by shifting the conditional mean function by the fixed effect, $\ln \lambda_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i$, as it is in the Poisson model. Rather, the HHG model is obtained by building the fixed effect into the model as an individual specific θ_i in the NB1 form. In the negative binomial models, the conditional mean functions are:

$$\text{NB1} : E[y_{it}|\mathbf{x}_{it}] = \theta_i \phi_{it} = \theta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) = exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \ln\theta_i),$$

$$\text{NB2} : E[y_{it}|\mathbf{x}_{it}] = \exp(\alpha_i)\phi_{it} = \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i),$$

so, superficially, the formulations do produce the same interpretation. However, the parameter θ_i in the NB1 model enters the variance function in a different manner:

$$\text{NB1} : \quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \theta_i\phi_{it}[1+\theta_i],$$

$$\text{NB2} : \quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \lambda_{it}[1+\theta\lambda_{it}].$$

The relationship between the mean and the variance is different for the two models. For estimation purposes, one can explain the apparent contradiction noted earlier by observing that, in the NB1 formulation, the individual effect is identified separately from the mean in the skedastic (scaling) function. This is not true for the FENB2 form. In order to obtain a counterpart to the HHG model, we would replace θ with θ_i (and λ_i with λ_{it}). Greene (2007a) analyzes the more familiar FENB2 form with the same treatment of λ_{it}. Estimates for both models appear below. Comparison of the suggested NB2 model to the HHG model remains for future investigation.

Once again, theory does not provide a reason to prefer the NB1 formulation over the more familiar NB2 model. The NB1 form does extend beyond the interpretation of the fixed effect as carrying only the sum of all the time invariant effects in the conditional mean function. The appearance of $\ln\theta_i$ in the conditional mean is an artifact of the exponential mean form; θ_i is a scaling parameter in this model. In its favor, the HHG model, being conditionally independent of the fixed effects,

finesses the incidental parameters problem – the estimator of $\boldsymbol{\beta}$ in this model is consistent. This is not the case for the FENB2 form, where:

$$Q_i = \frac{\theta}{\theta + \Sigma_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for $Y_i = \Sigma_t y_{it}$ with mean $\Lambda_i = \Sigma_t \lambda_{it}$.

Like the fixed effects model, introducing random effects into the negative binomial model adds some additional complexity. We do note, since the negative binomial model derives from the Poisson model by adding latent heterogeneity to the conditional mean, that adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time. However, one might prefer to interpret the negative binomial as the density for y_{it} in its own right, and treat the common effects in the familiar fashion. Hausman, Hall and Griliches' (1984) random effects negative binomial model is a hierarchical model that is constructed as follows. The heterogeneity is assumed to enter λ_{it} additively with a gamma distribution with mean 1, $\Gamma(\theta_i,\theta_i)$. Then, $\theta_i/(1+\theta_i)$ is assumed to have a beta distribution with parameters a and b. The resulting unconditional density after the heterogeneity is integrated out is:

$$p(y_{i1}, y_{i2}, \ldots, y_{iT_i}) = \frac{\Gamma(a+b)\Gamma\left(a + \Sigma_{t=1}^{T_i} \lambda_{it}\right) \Gamma\left(b + \Sigma_{t=1}^{T_i} y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a + \Sigma_{t=1}^{T_i} \lambda_{it} + b + \Sigma_{t=1}^{T_i} y_{it}\right)}.$$

As before, the relationship between the heterogeneity and the conditional mean function is unclear, since the random effect impacts upon the parameter of the skedastic function. An alternative approach that maintains the essential flavor of the Poisson model (and other random effects models) is to augment the NB2 form with the random effect:

$$\text{Prob}(Y = y_{it} | \mathbf{x}_{it}, \varepsilon_i) = \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it}+1)\Gamma(\theta)} r_{it}^{y_{it}} (1 - r_{it})^{\theta},$$

$$\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_i),$$

$$r_{it} = \lambda_{it}/(\theta + \lambda_{it}).$$

We then estimate the parameters by forming the conditional (on ε_i) log-likelihood and integrating ε_i out either by quadrature or simulation. The parameters are simpler to interpret by this construction. Estimates of the two forms of the random effects model are presented below for comparison.

11.6.4 Application

The study by Ripahn, Wambach and Million (2003) that provided the data we have used in numerous earlier examples analyzed the two count variables DocVis and HospVis. The authors were interested in the joint determination of these two

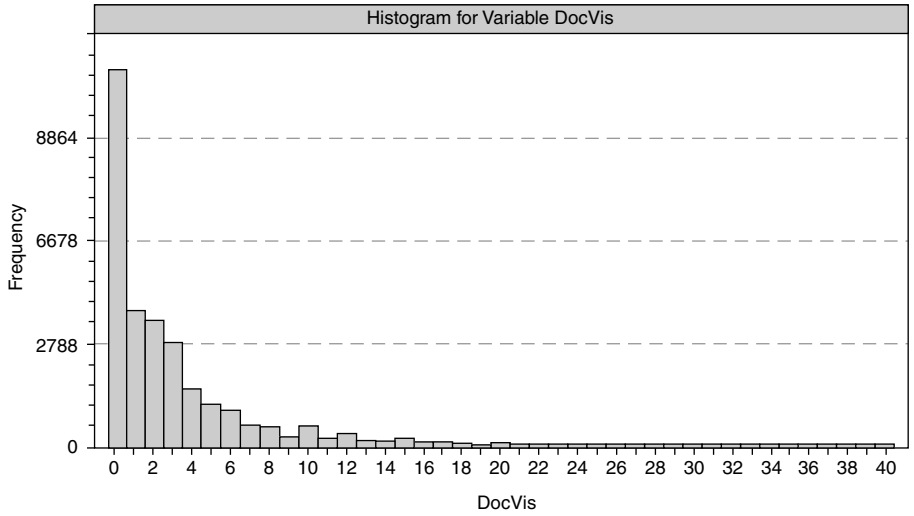


Figure 11.1 Histogram of count variable DocVis

count variables. One of the issues considered in the study was whether the data contained evidence of moral hazard, that is, whether health care utilization as measured by these two outcomes was influenced by the subscription to health insurance. The data contain indicators of two levels of insurance coverage: *Public*, which is the main source of insurance, and *Addon*, which is a secondary optional insurance. In the sample of 27,326 observations (family/years), 24,203 individuals held the public insurance. (There is quite a lot of within group variation in this, as individuals did not routinely obtain the insurance for all periods). Of these 24,203, 23,689 had only public insurance and 514 had both types. (One could not have only the addon insurance.) To explore the issue, we have analyzed the DocVis variable with the count data models described above. Figure 11.1 shows a histogram for this count variable. (There is a very long tail of extreme observations in these data, extending up to 121. The histogram omits the 91 observations with DocVis greater than 40. All observations are included in the sample used to estimate the models.) The exogenous variables in our model are:

$$\mathbf{x}_{it} = (1, \textit{Age, Education, Income, Kids, Public}).$$

(Variables are described in Table 11.2. Those listed are a small subset of those used in the original study, chosen here only for a convenient example.)

Table 11.7 presents the estimates of several count models. In all specifications, the coefficient on *Public* is positive, large, and highly statistically significant, which is consistent with the results in Riphahn, Wambach and Million (2003). The large spike at zero in the histogram casts some doubt on the Poisson specification. As a first step in extending the model, we estimated an alternative model that has a

Table 11.7 Estimated pooled models for DocVis (standard errors in parentheses)

Variable	*Poisson*	*Geometric*	*NB2*	*NB2 heterogeneous*	*NB1*	*NBP*
Constant	0.7162	0.7579	0.7628	0.7928	0.6848	0.6517
	(0.03287)	(0.06314)	(0.07247)	(0.07459)	(0.06807)	(0.07759)
Age	0.01844	0.01809	0.01803	0.01704	0.01585	0.01907
	(0.000332)	(0.00669)	(0.000792)	(0.000815)	(0.00070)	(0.0008078)
Education	–0.03429	–0.03799	–0.03839	–0.03581	–0.02381	–0.03388
	(0.00180)	(0.00343)	(0.003965)	(0.004034	(0.00370)	(0.004308)
Income	–0.4751	–0.4278	–0.4206	–0.4108	–0.1892	–0.3337
	(0.02198)	(0.04137)	(0.04700)	(0.04752)	(0.04452)	(0.05161)
Kids	–0.1582	–0.1520	–0.1513	–0.1568	–0.1342	–0.1622
	(0.00796)	(0.01561)	(0.01738)	(0.01773)	(0.01647)	(0.01856)
Public	0.2364	0.2327	0.2324	0.2411	0.1616	0.2195
	(0.0133)	(0.02443)	(0.02900)	(0.03006)	(0.02678)	(0.03155)
P	0.0000	0.0000	2.0000	2.0000	1.0000	1.5473
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.03444)
$\boldsymbol{\theta}$	0.0000	0.0000	1.9242	2.6060	6.1865	3.2470
	(0.0000)	(0.0000)	(0.02008)	(0.05954)	(0.06861)	(0.1346)
δ (Female)	0.0000	0.0000	0.0000	–0.3838	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.02046)	(0.0000)	(0.0000)
δ (Married)	0.0000	0.0000	0.0000	–0.1359	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.02307)	(0.0000)	(0.0000)
ln*L*	–104440.3	–61873.55	–60265.49	–60121.77	–60260.68	–60197.15

distribution that appears more like that in the figure, a *geometric regression model*:

$$\text{Prob}(y_i = j|\mathbf{x}_i) = \pi_i(1-\pi_i)^j, \qquad \pi_i = 1/(1+\lambda_i), \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad j = 0, 1, \ldots$$

This is the distribution for the number of failures before the first success in independent trials with success probability equal to π_i. It is suggested here simply as an alternative functional form for the model. The two models are similarly parameterized. The geometric model also has conditional mean equal to $(1-\pi_i)/\pi_i = \lambda_i$, like the Poisson. The variance is equal to $(1/\pi_i)\lambda_i > \lambda_i$, so the geometric distribution is overdispersed – it allocates more mass to the zero outcome. Based on the log-likelihoods, the Poisson model would be overwhelmingly rejected. However, since the models are not nested, this is not a valid test. Using, instead, the Vuong statistic based on $v_i = \ln L_i(\text{geometric}) - \ln L_i(\text{Poisson})$, we obtain a statistic of +37.89, which, as expected, strongly rejects the Poisson model.

The various formal test statistics strongly reject the hypothesis of equidispersion. Cameron and Trivedi's (1986) semiparametric tests from the Poisson model have *t*-statistics of 22.147 for $g_i = \mu_i$ and 22.504 for $g_i = \mu_i^2$. Both of these are far larger than the critical value of 1.96. The LM statistic (see Winkelmann, 2003) is 972,714.48, which is also larger than any critical value. On any of these bases, we would reject the hypothesis of equidispersion. The Wald and likelihood ratio tests based on the negative binomial models produce the same conclusion. For

comparing the different negative binomial models, note that NB2 is the worst of the three by the likelihood function, though NB1 and NB2 are not directly comparable. On the other hand, in the NBP model, the estimate of P is more than 10 standard errors from 1 or 2, so both NB1 and NB2 would be rejected in favor of the unrestricted NBP form of the model. The NBP and the heterogeneous NB2 model are not nested either, but on comparing the log-likelihoods, it does appear that the heterogeneous model is substantially superior. We computed the Vuong statistic based on the individual contributions to the log-likelihoods, with $v_i = \ln L_i(\text{NBP}) - \ln L_i(\text{NB2 – H})$. The value of the statistic is -3.27. On this basis, we would reject NBP in favor of NB2–H. Finally, with regard to the original question, the coefficient on PUBLIC is larger than 10 times its estimated standard error in every specification. We would conclude that the results are consistent with the proposition that there is evidence of moral hazard.

Estimates of the two two-part models, zero inflated and hurdle, are presented in Table 11.8. The regime equation for both is assumed to be a logit binary choice model with:

$$\mathbf{z}_{it} = (1, \textit{Age, Female, Married, Kids, Income, Self-employed}).$$

There is little theoretical basis for choosing between the two models. The interpretation of the data-generating process is quite similar in both cases. Each posits a regime in which the individual chooses whether or not to "participate" in the

Table 11.8 Two-part models for DocVis

	Poisson	*Poisson/logit zero inflation*		*Poisson/logit hurdle*	
Variable	*Count*	*Count*	*Regime*	*Count*	*Regime*
Constant	0.7162	1.3689	0.4789	1.4187	–0.5105
	(0.03287)	(0.01338)	(0.0651)	(0.0128)	(0.0637)
Age	0.01844	0.01067	–0.01984	0.01059	0.02068
	(0.000332)	(0.00013)	(0.00133)	(0.00012)	(0.00131)
Education	–0.03429	–0.02038	0.0000	–0.02215	0.0000
	(0.00180)	(0.00075)	(0.0000)	(0.00072)	(0.0000)
Income	–0.4751	–0.4131	0.1663	–0.4560	–0.2499
	(0.02198)	(0.00869)	(0.0758)	(0.00831)	(0.0724)
Kids	–0.1582	–0.08639	0.2306	–0.08862	–0.2378
	(0.00796)	(0.00316)	(0.0303)	(0.00297)	(0.0297)
Public	0.2364	0.1573	0.0000	0.1547	0.0000
	(0.0133)	(0.00604)	(0.0000)	(0.006037)	(0.0000)
Female	0.0000	0.0000	–0.58789	0.0000	0.5812
	(0.0000)	(0.0000)	(0.0265)	(0.0000)	(0.0260)
Married	0.0000	0.0000	–0.1257	0.0000	0.1271
	(0.0000)	(0.0000)	(0.0342)	(0.0000)	(0.0336)
Self-employed	0.0000	0.0000	0.4172	0.0000	–0.4137
	(0.0000)	(0.0000)	(0.0521)	(0.0000)	(0.0513)
log-likelihood	–104440.3	–83648.75		–83988.80	

health care system and a process that generates the count when they do. Nonetheless, there is little doubt that both are improvements on the Poisson regression. The average predicted probability of the zero outcome is 0.04826, so the Poisson model predicts $n\bar{\hat{P}}_0 = 1,319$ zero observations. The frequency in the sample is 10,135. The counterparts for the ZIP model are 0.36340 and 9,930. The Poisson model is not nested in the ZIP model – setting the ZIP coefficients to zero forces the regime probability to 0.5, not to 1.0. Thus the models cannot be compared by log-likelihoods. The Vuong statistic strongly supports the zero inflation model, being +47.05. Similar results are obtained for the hurdle model with the same specification.

The German health care panel data set contains 7,293 individuals with group sizes ranging from 1 to 7 and Table 11.9 presents the fixed and random effects estimates of the equation for DocVis. The pooled estimates are also shown for comparison. Overall, the panel data treatments bring large changes in the estimates compared to the pooled estimates. There is also a considerable amount of variation across the specifications. With respect to the parameter of interest, *Public*, we find that the size of the coefficient falls substantially with all panel data treatments. Whether using the pooled, fixed or random effects specifications, the test statistics (Wald, LR) all reject the Poisson model in favor of the negative binomial. Similarly, either common effects specification is preferred to the pooled estimator. There is no simple basis for choosing between the fixed and random effects models, and we have further blurred the distinction by suggesting two formulations for each of them. We do note that the two random effects estimators are producing similar results, which one might hope for, but the two fixed effects estimators are producing very different estimates. The NB1 estimates include two coefficients, on *Income* and *Education*, that are positive, but negative in every other case.

Moreover, the coefficient on *Public*, which is large and significant throughout the table, becomes small and less significant with the fixed effects estimators.

11.7 Multinomial unordered choices

We now extend the random utility, discrete choice model of sections 11.2–11.4 to a setting in which the individual chooses among multiple alternatives (see Hensher, Rose and Greene, 2005). The random utility model is:

$$U_{it,j} = \mathbf{x}_{it,j}'\boldsymbol{\beta} + \mathbf{z}_{it}'\boldsymbol{\gamma} + \varepsilon_{it,j}, \quad j = 1, \ldots, J_{it}, \quad t = 1, \ldots, T_i,$$

where, as before, we consider individual i in choice situation t, choosing among a possibly variable number of choices, J_{it}, and a possibly individual specific number of choice situations. For the present, for convenience, we assume $T_i = 1$ – a single-choice situation. This will be generalized later. The extension to variable choice set sizes, J_{it}, turns out to be essentially a minor modification of the mathematics, so it will also prove convenient to assume J_{it} is fixed at J. The random utility model is thus:

$$U_{i,j} = \mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,j}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n.$$

Table 11.9 Estimated panel data models for doctor visits (standard errors in parentheses)

Variable	*Poisson*			*Negative binomial*				
	Pooled (robust s.e.)	*Fixed effects*	*Random effects*	*Pooled NB2*	*Fixed effects*		*Random effects*	
					FE-NB1	*FE-NB2*	*HHG-gamma*	*Normal*
Constant	0.7162 (0.1319)	0.0000	0.4957 (0.05463)	0.7628 (0.07247)	–1.2354 (0.1079)	0.0000	–0.6343 (0.07328)	0.1169 (0.06612)
Age	0.01844 (0.001336)	0.03115 (0.001443)	0.02329 (0.0004458)	0.01803 (0.0007916)	0.02389 (0.001188)	0.04479 (0.002769)	0.01899 (0.0007820)	0.02231 (0.0006969)
Educ	–0.03429 (0.007255)	–0.03803 (0.01733)	–0.03427 (0.004352)	–0.03839 (0.003965)	0.01652 (0.006501)	–0.04589 (0.02967)	–0.01779 (0.004056)	–0.03773 (0.003595)
Income	–0.4751 (.08212)	–0.3030 (0.04104)	–0.2646 (0.01520)	–0.4206 (0.04700)	0.02373 (0.05530)	–0.1968 (0.07320)	–0.08126 (0.04565)	–0.1743 (0.04273)
Kids	–0.1582 (0.03115)	–0.001927 (0.01546)	–0.03854 (0.005272)	–0.1513 (0.01738)	–0.03381 (0.02116)	–0.001274 (0.02920)	–0.1103 (0.01675)	–0.1187 (0.01582)
Public	0.2365 (0.04307)	0.1015 (0.02980)	0.1535 (0.01268)	0.2324 (0.02900)	0.05837 (0.03896)	0.09700 (0.05334)	0.1486 (0.02834)	0.1940 (0.02574)
θ	0.0000	0.0000	1.1646 (0.01940)	1.9242 (0.02008)	0.0000	1.9199 (0.02994)	0.0000	1.0808 (0.01203)
a	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.1463 (0.05955)	0.0000
b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.8011 (0.1145)	0.0000
σ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9737 (0.008235)
ln*L*	–104440.3	–60337.13	–71763.13	–60265.49	–34016.16	–49476.36	–58182.52	–58177.66

The earlier assumptions are extended as well. The axioms of choice will imply that preferences are transitive, reflexive and complete. Thus, in any choice situation, the individual will make a choice, and that choice, j_i, will be such that:

$$U_{i,ji} > U_{i,m} \text{ for all } m = 1, \ldots, J \text{ and } m \neq j_i.$$

Reverting back to the optimization problem, utility maximization over continuous choices subject to a budget constraint produces the complete set of demands, $\mathbf{d}_i$ (prices, income). Inserting the demands back into the utility function produces the indirect utility function:

$$U_i^* = U_i[\mathbf{x}(\text{prices, income})].$$

This formulation is convenient for discrete choice modeling, as the data typically observed on the right-hand sides of the model equations will be income, prices, other characteristics of the individual such as age and sex, and attributes of the choices, such as model or type. The random utility model for multinomial unordered choices is then taken to be defined over the indirect utilities.

11.7.1 Multinomial logit and multinomial probit models

Not all stochastic specifications for $\varepsilon_{i,j}$ are consistent with utility maximization. McFadden (1981) showed that the i.i.d. Type 1 extreme value distribution:

$$F(\varepsilon_{i,j}) = \exp(-\exp(-\varepsilon_{i,j})), \quad j = 1, \ldots, J, \quad i = 1, \ldots, n,$$

produces a probabilistic choice model that is consistent with utility maximization. The resulting choice probabilities are:

$$\text{Prob}(d_{i,j} = 1|\mathbf{X}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})}{\Sigma_{m=1}^{J} \exp(\mathbf{x}_{i,m}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma})},$$

$$d_{i,j} = 1 \text{ if } U_{i,ji} > U_{i,m}, , m = 1, \ldots, J \text{ and } m \neq j.$$

This is the *multinomial logit model*. The components, $\mathbf{x}_{i,j}$, are the attributes of the choices (prices, features, etc.), while the $\mathbf{z}_i$ are the characteristics of the individual (income, age, sex). We noted at the outset of section 11.2 that identification of the model parameters requires that $\boldsymbol{\gamma}$ varies across the choices. Thus, the full model is:

$$\text{Prob}(d_{i,j} = 1|\mathbf{X}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_j)}{\Sigma_{m=1}^{J} \exp(\mathbf{x}_{i,m}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_m)}, \boldsymbol{\gamma}_J = 0,$$

$$d_{i,j} = 1 \text{ if } U_{i,ji} > U_{i,m}, , m = 1, \ldots, J \text{ and } m \neq j.$$

The log-likelihood function is:

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{i,j} \ln \left[\frac{\exp(\mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_j)}{\Sigma_{m=1}^{J} \exp(\mathbf{x}_{i,m}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_m)} \right].$$

The multinomial logit specification implies the peculiar restriction that

$$\frac{\partial \ln \text{Prob}(choice = j)}{\partial \mathbf{x}_{i.m}} = [1(j = m) - \text{Prob}(choice = m)]\boldsymbol{\beta}.$$

Thus the impact of a change in an attribute of a particular choice on the set of choice probabilities is the same for all (other) choices. For example, in our application,

$$\frac{\partial \ln P_{TRAIN}}{\partial Cost_{AIR}} = \frac{\partial \ln P_{BUS}}{\partial Cost_{AIR}} = \frac{\partial \ln P_{CAR}}{\partial Cost_{AIR}} = (-P_{AIR})\boldsymbol{\beta}_{Cost}.$$

This striking result, termed *independence from irrelevant alternatives* (IIA), follows from the initial assumptions of independent and identical distributions for $\varepsilon_{i,j}$. This is a major shortcoming of the model, and has motivated much of the research on specification of discrete choice models. Many model extensions have been proposed, including a heteroskedastic extreme value model (Bhat, 1995), the Dogit (dodging the logit model, Gaudry and Dagenais, 1979), and a host of others. The major extensions of the canonical multinomial logit (MNL) model have been the multinomial probit (MNP) model, the nested logit model and the current frontier, the mixed logit model. We consider each of these in turn.

11.7.1.1 Multinomial probit model

The MNP model (Daganzo, 1979) replaces the i.i.d. assumptions of the MNL model with a multivariate normality assumption:

$$\boldsymbol{\varepsilon}_i \sim N_J[\mathbf{0}, \boldsymbol{\Sigma}].$$

This specification relaxes the independence assumption. In principle, it can also relax the assumption of identical (marginal) distributions as well. Recall that, since only the most preferred choice is revealed, information about utilities is obtained in the form of differences, $U_{i,j} - U_{i,m}$. It follows that identification restrictions are required, as only some, or certain combinations of, elements of $\boldsymbol{\Sigma}$ are estimable. The simplest approach to securing identification that is used in practice is to impose that the last row of $\boldsymbol{\Sigma}$ be equal to $(0, 0, \ldots, 1)$, and one other diagonal element also equals 1. The remaining elements of $\boldsymbol{\Sigma}$ may be unrestricted, subject to the requirement that the matrix be positive definite. This can be done by a Cholesky decomposition, $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$, where $\mathbf{C}$ is a lower triangular matrix.

The MNP model relaxes the IIA assumptions. The shortcoming of the model is its computational demands. The relevant probabilities that enter the log-likelihood function and its derivatives must be approximated by simulation. The GHK simulator (Manski and Lerman, 1977; Geweke, Keane and Runkle, 1994) is commonly used. The Gibbs sampler with non-informative priors (Allenby and Rossi, 1999) has also proved useful for estimating the model parameters. Even with the GHK simulator, however, computation of the probabilities by simulation is time consuming.

11.7.2 Nested logit models

The nested logit model allows for the grouping of alternatives into "nests" with correlation across elements in a group. The natural analogy is to a "tree structure"; e.g., Figure 11.2 suggests an elaborate, three-level treatment of an eight-alternative-choice set.

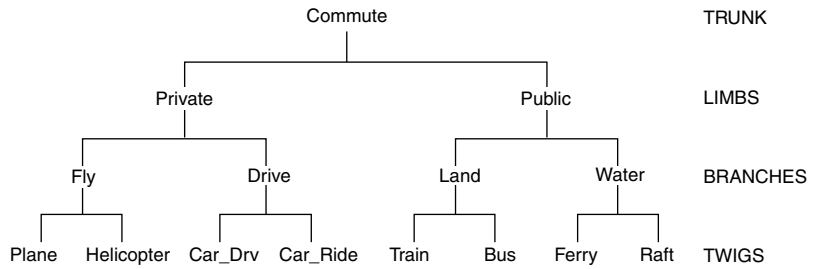


Figure 11.2 Nested choice set

The specific choice probabilities are redefined to be the conditional probability of alternative twig j in branch b, limb l, and trunk r, $j|b,l,r$, At the next level up the tree, we define the conditional probability of choosing a particular branch in limb l, trunk r, $b|l,r$, the conditional probability of choosing a limb in trunk r, $l|r$, and, finally, the probability of choosing a trunk r. By the laws of probability, the unconditional probability of the observed choices made by an individual is:

$$P(j,b,l,r) = P(j|b,l,r) \times P(b|l,r) \times P(l|r) \times P(r).$$

This is the contribution of an individual observation to the likelihood function for the sample. (Note that in our example, there is only one trunk, so $P(r) = 1$.)

The two-level nested logit model is the leading case, and occupies most of the received applications. In this instance, a common specification places the individual specific characteristics, such as demographic variables, in the branch probabilities. For this basic model, then:

$$P(j|b) = \frac{\exp(\mathbf{x}'_{j|b}\boldsymbol{\beta})}{\sum_{q|b}\exp(\mathbf{x}'_{q|b}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_{j|b}\boldsymbol{\beta})}{\exp(J_b)},$$

where J_b is the inclusive value for branch b,

$$J_b = \log \sum_{q|b} \exp(\mathbf{x}'_{q|b}\boldsymbol{\beta}).$$

At the next level up the tree, we define the conditional probability of choosing a particular branch:

$$P(b) = \frac{\exp\left[\lambda_b(\mathbf{z}'_i\boldsymbol{\gamma}_b + J_b)\right]}{\sum_s \exp\left[\lambda_s(\mathbf{z}'_i\boldsymbol{\gamma}_s + J_s)\right]} = \frac{\exp\left[\lambda_b(\mathbf{z}'_i\boldsymbol{\gamma}_b + J_b)\right]}{\exp(I)},$$

where I is the inclusive value for the model:

$$I = \log \sum_s \exp[\lambda_s(\mathbf{z}_i'\boldsymbol{\delta}_s + J_s)].$$

The original MNL results if the inclusive parameters, λ_s, are all equal to one.

Alternative normalizations and aspects of the nested logit model are discussed in Hensher and Greene (2002) and Hunt (2000). A second form moves the scaling down to the twig level, rather than at the branch level. Here it is made explicit that, within a branch, the scaling must be the same for all alternatives, but it can differ between the branches:

$$P(j|b) = \frac{\exp\left[\mu_b(\mathbf{x}_{j|b}'\boldsymbol{\beta})\right]}{\sum_{q|b}\exp\left[\mu_b(\mathbf{x}_{q|b}'\boldsymbol{\beta})\right]} = \frac{\exp\left[\mu_b(\mathbf{x}_{j|b}'\boldsymbol{\beta})\right]}{\exp(J_b)}.$$

Note that, in the summation in the inclusive value I, the scaling parameter is not varying with the summation index. It is the same for all twigs in the branch.

At the next level up the tree, we define the conditional probability of choosing the particular branch:

$$P(b|l) = \frac{\exp\left[\left(\mathbf{z}_i'\boldsymbol{\gamma}_s + (1/\mu_b)J_b\right)\right]}{\sum_s \exp\left[\left(\mathbf{z}_i'\boldsymbol{\gamma}_s + (1/\mu_s)J_s\right)\right]} = \frac{\exp\left[\left(\mathbf{z}_i'\boldsymbol{\gamma}_s + (1/\mu_b)J_b\right)\right]}{\exp(I)},$$

where I_l is the inclusive value for limb l:

$$I_l = \log \sum_{s|l} \exp\left[\gamma_l\left(\boldsymbol{\alpha}'\mathbf{y}_{s|l} + (1/\mu_{s|l})J_{s|l}\right)\right].$$

In the nested logit model with $P(j,b,l,r) = P(j|b,l,r)\times P(b|l,r)\times P(l|r)\times P(r)$, the marginal effect of a change in attribute "k" in the utility function for alternative "J" in branch "B" of limb "L" of trunk "R" on the probability of choice "j" in branch "b" of limb "l" of trunk "r" is computed using the following result: lower-case letters indicate the twig, branch, limb and trunk of the outcome upon which the effect is being exerted. Upper-case letters indicate the twig, branch, limb and trunk which contain the outcome whose attribute is being changed:

$$\frac{\partial \log P(alt = j, limb = l, branch = b, trunk = r)}{\partial x(k)|alt = J, limb = L, branch = B, trunk = r} = D(k|J,B,L,R) = \Delta(k)\times F,$$

where $\Delta(k)$ = coefficient on $x(k)$ in $U(J|B,L,R)$ and:

$$\begin{aligned}
F &= \mathbf{1}(r=R)\times\mathbf{1}(l=L)\times\mathbf{1}(b=B)\times[\mathbf{1}(j=J)\,P(J|BLR)] && \text{(trunk effect)},\\
&= 1(r=R)\times\mathbf{1}(l=L)\times[\mathbf{1}(b=B) - P(B|LR)]\times P(J|BLR)]\times\tau_{B|LR} && \text{(limb effect)},\\
&= 1(r=R)\times[\mathbf{1}(l=L) - P(L|R)]\times P(B|LR)\times P(J|BLR)\times\tau_{B|LR}\times\sigma_{L|R} && \text{(branch effect)},\\
&= [\mathbf{1}(r=R)P(R)]\times P(L|R)\times P(B|LR)\times P(J|BLR)\times\tau_{B|LR}\times\sigma_{L|R}\times\phi_R && \text{(twig effect)},
\end{aligned}$$

where $\tau_{B|LR,}$ $\sigma_{L|R}$ and ϕ_R are parameters in the MNL probabilities. The marginal effect is:

$$\partial P(j,b,l,r)/\partial x(k)|J,B,L,R = P(j,b,l,r)\Delta(k)F.$$

A marginal effect has four components: an effect on the probability of the particular trunk, one on the probability for the limb, one for the branch, and one for the probability for the twig. (Note that with one trunk, $P(l) = P(1) = 1$, and likewise for limbs and branches.) For continuous variables, such as cost, it is common to report, instead:

$$Elasticity = x(k)|J,B,L,R \times \Delta(k|J,B,L,R) \times F.$$

The formulation of the nested logit model imposes no restrictions on the inclusive value parameters. However, the assumption of utility maximization and the stochastic underpinnings of the model do imply certain restrictions. For the former, in principle, the inclusive value parameters must lie between zero and one. For the latter, the restrictions are implied by the way the random terms in the utility functions are constructed. In particular, the nesting aspect of the model is obtained by writing:

$$\varepsilon_{j|b,l,r} = u_{j|b,l,r} + v_{b|l,r}.$$

That is, within a branch, the random terms are viewed as the sum of a unique component and a common component. This has certain implications for the structure of the scale parameters in the model. In particular, it is the source of the oft cited (and oft violated) constraint that the IV parameters must lie between zero and one. These are explored in Hunt (2000) and Hensher and Greene (2002).

11.7.3 Mixed logit and error component models

This model is somewhat similar to the random coefficients model for linear regressions (see Bhat, 1996; Jain, Vilcassim and Chintagunta, 1994; Revelt and Train, 1998; Train, 2003; Berry, Levinsohn and Pakes, 1995). The model formulation is a one-level MNL for individuals $i = 1,\ldots,n$ in choice setting t. We begin with the basic form of the MNL model, with alternative specific constants α_{ji} and attributes $\mathbf{x}_{ji}$:

$$\text{Prob}(y_{it} = j|\mathbf{X}_{it}) = \frac{\exp\left(\alpha_{ji} +_i \mathbf{x}'_{it,j}\boldsymbol{\beta}_i\right)}{\sum_{q=1}^{J_{it}} \exp\left(\alpha_{qi} +_i \mathbf{x}'_{it,q}\boldsymbol{\beta}_i\right)}.$$

The random parameters model emerges as the form of the individual specific parameter vector, $\boldsymbol{\beta}_i$, is developed. The most familiar, simplest version of the model specifies:

$$\beta_{ki} = \beta_k + \sigma_k v_{ki},$$
$$\alpha_{ji} = \alpha_j + \sigma_j v_{ji},$$

where β_k is the population mean, v_{ki} is the individual specific heterogeneity, with mean zero and standard deviation one, and σ_k is the standard deviation of the distribution of the β_{ki}s around β_k. The term "mixed logit" is often used in the literature

(e.g., Revelt and Train, 1998, and, especially, McFadden and Train, 2000) for this model. The choice-specific constants, α_{ji}, and the elements of β_i are distributed randomly across individuals with fixed means. A refinement of the model is to allow the means of the parameter distributions to be heterogeneous with observed data $\mathbf{z}_i$ (which does not include a constant). This would be a set of choice invariant characteristics that produce individual heterogeneity in the means of the randomly distributed coefficients so that:

$$\beta_{ki} = \beta_k + \mathbf{z}_i'\boldsymbol{\delta}_k + \sigma_k v_{ki},$$

and likewise for the constants. The model is not limited to the normal distribution. One important variation is the log-normal model,

$$\beta_{ki} = \exp(\beta_k + \mathbf{z}_i'\boldsymbol{\delta}_k + \sigma_k v_{ki}).$$

The v_{ki}s are individual and choice specific, unobserved random disturbances – the source of the heterogeneity. Thus, as stated above, in the population, if the random terms are normally distributed:

$$\beta_{ki} \sim \text{Normal or Lognormal } [\beta_k + \mathbf{z}_i'\boldsymbol{\delta}_k, \sigma_k^2].$$

(Other distributions may be specified.) For the full vector of K random coefficients in the model, we may write the full set of random parameters as:

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta} z_i + \boldsymbol{\Gamma} v_i,$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix which contains σ_k on its diagonal. For convenience at this point, we will simply gather the parameters or choice specific constants under the subscript "k."

Greene and Hensher (2006) have developed a counterpart to the random effects model that essentially generalizes the mixed logit model to a stochastic form of the nested logit model. The general notation is fairly cumbersome, but an example suffices to develop the model structure. Consider a four outcome-choice set: Air, Train, Bus, Car. The utility functions in an MNL or mixed logit model could be:

$$\begin{array}{llllll}
U_{it,Air} & = \alpha_{Air} & +\mathbf{x}_{it,Air}'\boldsymbol{\beta}_i & +\varepsilon_{it,Air} & +\theta_1 E_{i,Private} \\
U_{it,Train} & = \alpha_{Train} & +\mathbf{x}_{it,Train}'\boldsymbol{\beta}_{ii} & +\varepsilon_{it,Train} & +\theta_2 E_{i,Public} \\
U_{it,Bus} & = \alpha_{Bus} & +\mathbf{x}_{it,Bus}'\boldsymbol{\beta}_{ii} & +\varepsilon_{it,Bus} & +\theta_2 E_{i,Public} \\
U_{it,Car} & = \mathbf{x}_{it,Car}'\boldsymbol{\beta}_i & +\varepsilon_{it,Car} & +\theta_1 E_{i,Private}, &
\end{array}$$

where the components $E_{i,Private}$ and $E_{i,Public}$ are independent, normally distributed random elements of the utility functions. Thus this is a two-level nested logit model.

The probabilities defined above are conditioned on the random terms, $\mathbf{v}_i$, and the error components, $\mathbf{E}_i$. The unconditional probabilities are obtained by integrating v_{ik} and E_{im} out of the conditional probabilities: $P_j = E_{v,E}[P(j|\mathbf{v}_i,\mathbf{E}_i)]$. This is a multiple integral which does not exist in closed form. The integral is approximated by simulation (see Greene and Hensher, 2006, and Greene, 2007a, for discussion). Parameters are estimated by maximizing the simulated log-likelihood.

11.7.4 Application

The multinomial choice models are illustrated with a well-known data survey of commuters between Sydney and Melbourne (see Greene, 2007a, and references cited). A sample of 210 travelers between Sydney and Melbourne were asked which of four travel modes they chose air, train, bus or car. The variables used in the models are:

TTME = Terminal time, in minutes, zero for car,
INVT = In-vehicle time for the journey
GC = generalized cost = in-vehicle cost + a wage times INVT
HINC = household income
PSIZE = traveling party size:

Table 11.10 lists descriptive statistics for the variables in the model. The left-hand side for each panel lists the means and standard deviations for the variables

Table 11.10 Descriptive statistics for variables

Variable	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*
AIR	All 210 obs.		58 observations that chose AIR	
TTME	61.010	15.719	46.534	24.389
INVT	133.710	48.521	124.828	50.288
GC	102.648	30.575	113.552	33.198
PSIZE	1.743	1.012	1.569	.819
HINC	34.548	19.711	41.724	19.115
TRAIN	All 210 obs.		63 observations that chose TRAIN	
TTME	35.690	12.279	28.524	19.354
INVT	608.286	251.797	532.667	249.360
GC	130.200	58.235	106.619	49.601
PSIZE	1.743	1.012	1.667	.898
HINC	34.548	19.711	23.063	17.287
BUS	All 210 obs.		30 observations that chose BUS	
TTME	41.657	12.077	25.200	14.919
INVT	629.462	235.408	618.833	273.610
GC	115.257	44.934	108.133	43.244
PSIZE	1.743	1.012	1.333	.661
HINC	34.548	19.711	29.700	16.851
CAR	All 210 obs.		59 observations that chose CAR	
TTME	.000	.000	.000	.000
INVT	573.205	274.855	527.373	301.131
GC	95.414	46.827	89.085	49.833
PSIZE	1.743	1.012	2.203	1.270
HINC	34.548	19.711	42.220	17.685

for all observations, in each of the four choices. The right-hand side reports the same statistics for the observations that made the particular choices. Thus, for example, the average terminal time for all 210 observations for the air choice is 61.01 minutes. For the 58 individuals who chose air, the average terminal time for air is 46.534 minutes. We note before beginning that the sample proportions for the four travel modes in this sample are 0.27619, 0.30000, 0.14286 and 0.28095, respectively. Long study of this market revealed that the population values of these proportions should be closer to 0.14, 0.13, 0.09 and 0.64, respectively. The sample observations were deliberately drawn so that the car alternative received fewer observations than random sampling would predict. The sample is *choice based*. A general adjustment for that phenomenon is the Manski–Lerman (1977) weighted endogenous sampling maximum likelihood (WESML) correction, which consists of two parts. First, we would fit a weighted log-likelihood:

$$\ln L(\text{WESML}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\pi_j}{p_j} d_{ij} \ln \Pi_{ij},$$

where $d_{ij} = 1$ if individual i chooses alternative j and 0 otherwise, π_j is the true population proportion, p_j is the sample proportion, and Π_{ij} is the probability for outcome j implied by the model. The second aspect of the correction is to use a sandwich style corrected estimator for the asymptotic covariance matrix of the MLE:

$$\text{V(WESML)} = \mathbf{H}^{-1}(\mathbf{G}'\mathbf{G})\mathbf{H}^{-1},$$

where $\mathbf{H}$ is the inverse of the (weighted) Hessian and $(\mathbf{G}'\mathbf{G})^{-1}$ would be the BHHH estimator based on first derivatives. The results to follow do not include this correction – the results in the example would change slightly if they were incorporated.

We fit a variety of models. The same utility functions were specified for all:

$$\text{U}_{i,\text{AIR}} = \alpha_{\text{AIR}} + \beta_{tt}\text{TTME}_{i,\text{AIR}} + \beta_{it}\text{INVT}_{i,\text{AIR}} + \beta_{\text{gc}}\text{GC}_{i,\text{AIR}} + \gamma_{\text{A}}\text{HINC}_i + \varepsilon_{i,\text{AIR}},$$

$$\text{U}_{i,\text{TRAIN}} = \alpha_{\text{TRAIN}} + \beta_{tt}\text{TTME}_{i,\text{TRAIN}} + \beta_{it}\text{INVT}_{i,\text{TRAIN}} + \beta_{\text{gc}}\text{GC}_{i,\text{TRAIN}} + \varepsilon_{i,\text{TRAIN}},$$

$$\text{U}_{i,\text{BUS}} = \alpha_{\text{BUS}} + \beta_{tt}\text{TTME}_{i,\text{BUS}} + \beta_{it}\text{INVT}_{i,\text{BUS}} + \beta_{\text{gc}}\text{GC}_{i,\text{BUS}} + \varepsilon_{i,\text{BUS}},$$

$$\text{U}_{i,\text{CAR}} = \beta_{tt}\text{TTME}_{i,\text{CAR}} + \beta_{it}\text{INVT}_{i,\text{CAR}} + \beta_{\text{gc}}GC_{i,\text{CAR}} + \varepsilon_{i,\text{CAR}}.$$

The estimated parameters for the several specifications are given in Table 11.11. Model MNL is the base case multinomial logit model. Model MNP is the multinomial probit model. The three nested logit (NL) models are nested logit models with different tree structures:

NL(1) = Private (air, car), Public (train, bus).
NL(2) = Fly (air), Ground (train, bus, car)
NL(3) = Fly (air), Rail (train), Drive (car), Autobus (bus).

For the third of these, one of the inclusive value parameters, μ_j, must be constrained to equal one. Model HEV (heteroskedastic extreme value) is the extreme value

Table 11.11 Estimated multinomial choice models (standard errors in parentheses)

	MNL	*MNP*[a]	*NL (1)*[b]	*NL (2)*[b]	*NL (3)*[b]	*HEV*[c]	*RPL*[d]
α_{AIR}	3.139	–2.769	1.110	3.261	1.825	2.405	6.930
	(.984)	(1.997)	(.877)	(.879)	(.621)	(2.692)	(4.053)
α_{TRAIN}	3.558	3.137	1.468	3.039	2.113	6.701	17.994
	(.443)	(1.0599)	(.452)	(.601)	(.493)	(2.852)	(4.745)
α_{BUS}	3.134	2.581	.971	2.721	1.877	6.150	16.556
	(.452)	(.419)	(.475)	(.604)	(.746)	(2.483)	(4.585)
α_{CAR}	.000	.000	.000	.000	.000	.000	.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Term	–.0963	–.0548	–.0655	–.0742	–.0415	–.164	–.385
time	(.0103)	(.0227)	(.0116)	(.00134)	(.0148)	(.0799)	(.0857)
Inv. time	–.00379	–.00447	–.00422	–.0167	–.00767	–.00744	–.0241
	(.00118)	(.00139)	(.000919)	(.00142)	(.00197)	(.00300)	(.00589)
Gen. cost	–.00139	–.0183	–.000449	.00639	–.00051	–.0299	–.0397
	(.00623)	(.00827)	(.00467)	(.00679)	(.00340)	(.0185)	(.0238)
Income	.0185	.0702	.0169	.0195	.00868	.0604	.156
	(.0108)	(.0398)	(.00691)	(.00878)	(.00389)	(.0456)	(.0715)
Scale (1)		5.073	3.097	1.278	3.400	.386	.261
		(2.172)	(.627)	(.289)	(1.238)	(.189)	(.0794)
Scale (2)		1.221	1.989	.197	1.0839	.745	.0176
		(.911)	(.423)	(.0679)	(.109)	(.376)	(.00564)
Scale (3)		1.000			1.130	.964	.0369
		(0.000)			(.144)	(.587)	(.0350)
Scale (4)		1.000			1.000	1.000	
		(0.000)			(0.000)	(0.000)	
P. size						–.208	
						(.0739)	
ρ (air, train)		.736					
		(.323)					
ρ (air, bus)		.649					
		(.475)					
ρ (train, bus)		.655					
		(.292)					
ln*L*	–193.4981	–191.8264	–178.7135	–166.3662	–190.9303	–186.1741	–168.1089

[a] Scale parameters are standard deviations.
[b] Scale parameters are IV parameters.
[c] Scale parameters are σ_j.
[d] Scale parameters are standard deviations of random parameters.

model with the variances allowed to differ across utility functions. In addition, we introduced heteroskedasticity in the model, so that:

$$\text{Var}[\varepsilon_{i,j}] = \sigma_j^2 \times \exp(\theta \text{ party size}).$$

Finally, the last model, RPL (random parameters logit), is a random parameters specification in which the parameters on TTME, INVT and GC are allowed to vary randomly across individuals.

It is difficult to obtain a precise interpretation of the coefficients in the utility functions. The elasticities of the probabilities of specific outcomes with respect to changes in the attributes of those and other outcomes are more informative. We write these:

$$\delta(m)_{j,l} = \frac{\partial \log P_j}{\partial \log x_{m,l}},$$

where m indicates the mth attribute, j is the choice probability affected and l is the choice utility function in which $x_{m,l}$ changes. These are given in Table 11.12. (specific forms of these appear in Greene, 2008a; Hensher *et al.*, 2005). The values given are averaged over the observations. Note that for the MNL form, $\delta(m)_{j,l}$ is the same for all l. This is the force of the IIA assumptions made at the outset. This is a motivation for the more elaborate functions considered.

Table 11.12 Estimated elasticities with respect to changes in GC

Effect is on choice of:	*CG changes in choices:*							
	Air		*Train*		*Bus*		*Car*	
	MNL	–.0994	MNL	.0455	MNL	.0210	MNL	.0346
Air	MNP	–.5230	MNP	.3060	MNP	.1179	MNP	.1006
	NL(2)	.595	NL(2)	–.0310	NL(2)	–.0200	NL(2)	–.0430
	HEV	–.9158	HEV	.3771	HEV	.2339	HEV	.2144
	RPL	–.4808	RPL	.2361	RPL	.1440	RPL	.0663
	MNL	.0435	MNL	–.1357	MNL	.0210	MNL	.0346
Train	MNP	.3889	MNP	–3.4650	MNP	1.1148	MNP	.9416
	NL(2)	–.2440	NL(2)	–.2160	NL(2)	–.127	NL(2)	.5420
	HEV	.3443	HEV	–1.7389	HEV	.4105	HEV	.4621
	RPL	.3167	RPL	–1.4151	RPL	.5715	RPL	.2360
	MNL	.0435	MNL	.0455	MNL	–.1394	MNL	.0346
Bus	MNP	.2859	MNP	2.454	MNP	–4.4750	MNP	1.2686
	NL(2)	–.2440	NL(2)	–.2160	NL(2)	.6100	NL(2)	–.2900
	HEV	.4744	HEV	1.2723	HEV	–3.1008	HEV	.8358
	RPL	.7109	RPL	1.8434	RPL	–2.9242	RPL	.3246
	MNL	.0435	MNL	.0455	MNL	.0210	MNL	–.0982
Car	MNP	.1113	MNP	.8592	MNP	.5587	MNP	–1.4023
	NL(2)	–.2440	NL(2)	.3940	NL(2)	–.1270	NL(2)	–.2900
	HEV	.4133	HEV	.8108	HEV	.6190	HEV	–1.7829
	RPL	.2489	RPL	.6300	RPL	.2973	RPL	–1.0332

11.8 Summary and conclusions

This chapter has outlined the basic modeling frameworks that are used in analyzing microeconomic data when the response variable corresponds to a discrete choice. The essential binary choice model is the foundation for a vast array of applications and theoretical developments. The full set of results for the fully parametric models based on the normal distribution, as well as many non- and semiparametric models,

are well established. Ongoing contemporary theoretical research is largely focused on less parametric approaches and on panel data. The parametric models developed here still overwhelmingly dominate the received applications.

Notes

1. For a lengthy and detailed development of these ideas, see Daniel McFadden's Nobel Prize Lecture (McFadden, 2001).
2. See, as well, Samuelson (1947) and Goldberger (1987).
3. Some formulations of the models, such as models of heteroskedasticity and the random parameters, will also involve additional parameters. These will be introduced later. They are omitted at this point to avoid cluttering the notation.
4. The formulation assumes that the T_i choices made by individual i are unconditionally independent. This assumption may be inappropriate. In one of our applications, the assumption is testable.
5. See, e.g., the documentation for LIMDEP (Econometric Software, Inc., 2007) or Stata (Stata, Inc., 2007).
6. We are assuming that the data are "well behaved" so that the conditions underlying the standard optimality properties of MLEs are met here. The conditions and the properties are discussed in Greene (2008a). We will take them as given in what follows.
7. The sign of the result for the logistic distribution is obvious. See, e.g., Maddala (1983, p. 366) for a proof of the result for the normal distribution.
8. There are data configurations, in addition to simple multicollinarity, that can produce singularities. Another possibility is that of a variable in $\mathbf{x}_i$ or $\mathbf{z}_i$ that can predict d_i perfectly based on a specific cut point in the range of that variable.
9. Recall that the average predicted probability, $\overline{\hat{P}}$, equals the average outcome in the binary choice model, P_1. To a fair approximation, the standard deviation of the predicted probabilities will equal $[P_1(1 - P_1)]^{0.5}$. If the sample is highly unbalanced, say $P_1 < 0.05$ or $P_1 > 0.95$, then a predicted probability as large as (or as small as) 0.5 may become unlikely. It is common in unbalanced panels for the simple prediction rule always to predict the same value.
10. A symposium on the subject is Hardle and Manski (1993).
11. See Manski (1975, 1985, 1986) and Manski and Thompson (1986). For extensions of this model, see Horowitz (1992), Charlier, Melenberg and van Soest (1995) and Kyriazidou (1997).
12. Bootstrapping has been used to estimate the asymptotic covariance matrix for the maximum score estimator. However, Abrevaya and Huang (2005) have recently cast doubt on the validity of that approach. No other strategy is available for statistical inference in this model.
13. One would proceed in precisely this fashion if the central specification were a linear probability model (LPM) to begin with. See, e.g., Eisenberg and Rowe (2006) or Angrist (2001) for an application and some analysis of this case.
14. This is precisely the platform that underlies the generalized linear models/generalized estimating equations (GLIM/GEE) treatment of binary choice models in, e.g., the widely used programs SAS and Stata.
15. Much of the recent research in semiparametric and nonparametric analysis of discrete choice and limited dependent variable models has focused on how to accommodate individual heterogeneity in panel data models while avoiding the incidental parameters problem.
16. The requirement does not state how large R must be, only that it "increase" faster than $n^{1/2}$. In practice, analysts typically use several hundred, perhaps up to 1,000, random draws for their simulations.

17. See, as well, Hsiao (2003) for a survey of dynamic panel data models and other applications by van Doorslaer and Nonneman (1987), Wagstaff (1993) and Vella and Verbeek (1999).
18. This is the formulation used by Contoyannis *et al.* (2004). Wooldridge (2006) suggested, instead, that the projection be upon all of the data, $(\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots)$. Two major practical problems with this approach are that, in a model with a large number of regressors, which is common when using large, elaborate panel data sets, the number of variables in the resulting model will become excessive. Second, this approach breaks down if the panel is unbalanced, as it was in the Contoyannis *et al.* study.
19. Beck *et al.* (2001) is a bit different from the others mentioned in that, in their study of "state failure," they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in such a way that makes the asymptotics with respect to T appropriate. They can analyze the data essentially in a time series framework. Sepanski (2000) is another application which combines state dependence and the random coefficient specification of Akin, Guilkey and Sickles (1979).
20. Wynand and van Praag (1981) used a two-step procedure similar to Heckman's (1979) procedure for the linear model. Applications since then have used the MLE.
21. Since the coefficient vectors are assumed to be the same in every period, it is only necessary to normalize one of the diagonal elements in $\mathbf{R}$ to 1.0. See Greene (2004a) for discussion.
22. For example, the parameters can be written in terms of a set of latent parameters so that $\mu_1 = \tau_1^2$, $\mu_2 = \tau_1^2 + \tau_2^2$, etc. Typically, the explicit reparameterization is unnecessary.
23. One could argue that this reformulation achieves identification purely "through functional form," rather than through the theoretical underpinnings of the model. Of course, this assertion elevates the linear specification to a default position of prominence, which seems unwarranted. Moreover, arguably the underlying theory (as, in fact, suggested in passing by Pudney and Shields, 2000) *is* that there are different effects of the regressors on the thresholds and on the underlying utility.
24. Cross-section versions of the ordered probit model with individual specific thresholds appear in Terza (1985), Pudney and Shields (2000) and in Greene (2007a).
25. No theory justifies the choice of the log-gamma density. It is essentially the same as a conjugate prior in Bayesian analysis, chosen for its mathematical convenience.
26. The use of the linear index form is a convenience. The random component, ε, could enter the model in some other form, with no change in the general approach.

References

Abramovitz, M. and I. Stegun (1971) *Handbook of Mathematical Functions*. New York: Dover Press.

Abrevaya, J. (1997) The equivalence of two estimators of the fixed effects logit model. *Economics Letters* **55**(1), 41–4.

Abrevaya, J. and J. Huang (2005) On the bootstrap of the maximum score estimator. *Econometrica* **73**(4), 1175–204.

Akin, J., D. Guilkey and R. Sickles (1979) A random coefficient probit model with an application to a study of migration. *Journal of Econometrics* **11**, 233–46.

Albert, J. and S. Chib (1993) Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association* **88**, 669–79.

Aldrich, J. and F. Nelson (1984) *Linear Probability, Logit, and Probit Models*. Beverly Hills: Sage Publications.

Allenby, G.M. and P.E. Rossi (1999) Marketing models of consumer heterogeneity. *Journal of Econometrics* **89**, 57–78.

Allison, P. (2000) Problems with fixed-effects negative binomial models. Manuscript, Department of Sociology, University of Pennsylvania.

Allison, P. and R. Waterman (2002) Fixed-effects negative binomial regression models. Manuscript, Department of Sociology, University of Pennsylvania.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Andersen, E. (1970) Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32**, 283–301.

Angrist, J. (2001) Estimation of limited dependent variable models with binary endogenous regressors: simple strategies for empirical practice. *Journal of Business and Economic Statistics* **19**(1), 1–14.

Avery, R., L. Hansen, and J. Hotz (1983) Multiperiod probit models and orthogonality condition estimation. *International Economic Review* **24**, 21–35.

Beck, N., D. Epstein and S. Jackman (2001) Estimating dynamic time series cross section models with a binary dependent variable. Manuscript, Department of Political Science, University of California, San Diego.

Ben-Akiva, M. and S. Lerman (1985) *Discrete Choice Analysis*. London: MIT Press.

Berndt, E., B. Hall, R. Hall and J. Hausman (1974) Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3/4**, 653–65.

Berry, S., J. Levinsohn and A. Pakes (1995) Automobile prices in market equilibrium. *Econometrica* **63**(4), 841–90.

Bertschek, I. and M. Lechner (1998) Convenient estimators for the panel probit model. *Journal of Econometrics* **87**(2), 329–72.

Bhat, C. (1995) A heteroscedastic extreme value model of intercity mode choice. *Transportation Research* **30**(1), 16–29.

Bhat, C. (1996) Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. Department of Civil Engineering, University of Massachusetts, Amherst, Working Paper.

Bhat, C. (1999) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Manuscript, Department of Civil Engineering, University of Texas, Austin.

Breusch, T. and A. Pagan (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–94.

Breusch, T. and A. Pagan (1980) The LM test and its applications to model specification in econometrics. *Review of Economic Studies* **47**, 239–54.

Boyes, W., D. Hoffman and S. Low (1989) An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* **40**, 3–14.

Butler, J. and P. Chatterjee (1995) Pet econometrics: ownership of cats and dogs. Department of Economics, Vanderbilt University, Working Paper 95-WP1.

Butler, J. and P. Chatterjee (1997) Tests of the specification of univariate and bivariate ordered probit. *Review of Economics and Statistics* **79**, 343–7.

Butler, J., T. Finegan and J. Siegfried (1998) Does more calculus improve student learning in intermediate micro- and macroeconomic theory? *Journal of Applied Econometrics* **13**(2), 185–202.

Butler, J. and R. Moffitt (1982) A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* **50**, 761–64.

Calhoun, C. (1991) Desired and excess fertility in Europe and the United States: indirect estimates from world fertility survey data. *European Journal of Population* **7**, 29–57.

Cameron, A. and P. Trivedi (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**, 29–54.

Cameron, C. and P. Trivedi (1998) *Regression Analysis of Count Data*. New York: Cambridge University Press.

Caudill, S. (1988) An advantage of the linear probability model over probit or logit. *Oxford Bulletin of Economics and Statistics* **50**, 425–7.

Cecchetti, S. (1986) The frequency of price adjustment: a study of the newsstand prices of magazines. *Journal of Econometrics* **31**(3), 255–74.
Chamberlain, G. (1980) Analysis of covariance with qualitative data. *Review of Economic Studies* **47**, 225–38.
Charlier, E., B. Melenberg and A. van Soest (1995) A smoothed maximum score estimator for the binary choice panel data model with an application to labor force participation. *Statistica Neerlander* **49**, 324–43.
Chesher, A. and M. Irish (1987) Residual analysis in the grouped data and censored normal linear model. *Journal of Econometrics* **34**, 33–62.
Christofides, L., T. Hardin and R. Stengos (2000) On the calculation of marginal effects in the bivariate probit model: corrigendum. *Economics Letters* **68**, 339–40.
Christofides, L., T. Stengos and R. Swidinsky (1997) On the calculation of marginal effects in the bivariate probit model. *Economics Letters* **54**(3), 203–8.
Contoyannis, C., A. Jones and N. Rice (2004) The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics* **19**(4), 473–503.
Cramer, J. (1999) Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society, Series D (The Statistician)* **48**, 85–94.
D'Addio, A.C., T. Eriksson and P. Frijters (2003) An analysis of the determinants of job satisfaction when individuals' baseline satisfaction levels may differ. Center for Applied Microeconometrics, University of Copenhagen, Working Paper 2003–16.
Daganzo, C. (1979) *The Multinomial Probit Model: The Theory and Its Application to Demand Forecasting*. New York: Academic Press.
Das M. and A. van Soest (2000) A panel data model for subjective information on household income growth. *Journal of Economic Behavior and Organization* **40**, 409–26.
Dempster, A., N. Laird and D. Rubin (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
Econometric Software, Inc. (2007) LIMDEP, Version 9.0. Plainview, New York: Econometric Software, Inc.
Efron, B. (1978) Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association* **73**, 113–212.
Eisenberg, D. and B. Rowe (2006) The effect of serving in the Vietnam war on smoking behavior later in life. Manuscript, School of Public Health, University of Michigan.
Fabbri, D., C. Monfardini and R. Radice (2004) Testing exogeneity in the bivariate probit model: Monte Carlo evidence and an application to health economics. Manuscript, Department of Economics, University of Bologna.
Ferrer-i-Carbonel, A. and P. Frijter (2004) The effect of methodology on the determinants of happiness. *Economic Journal* **114**, 715–19.
Fernandez, A. and J. Rodriguez-Poo (1997) Estimation and testing in female labor participation models: parametric and semiparametric models. *Econometric Reviews* **16**, 229–48.
Freedman, D. (2006) On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors." *American Statistician* **60**(4), 299–302.
Frijters P., J. Haisken-DeNew and M. Shields (2004) The value of reunification in Germany: an analysis of changes in life satisfaction. *Journal of Human Resources* **39**(3), 649–74.
Gandelman, N. (2005) Homeownership and gender. Manuscript, Universidad ORT, Uruguay.
Gaudry, M. and M. Dagenais (1979) The dogit model. *Transportation Research, Series B* **13**, 105–11.
Gerfin, M. (1996) Parametric and semi-parametric estimation of the binary response model. *Journal of Applied Econometrics* **11**, 321–40.
Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics*. New York: John Wiley and Sons.
Geweke, J., M. Keane and D. Runkle (1994) Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics* **76**, 609–32.

Goldberger, A. (1987) *Functional Form and Utility: A Review of Consumer Demand Theory*. Boulder, Colo: Westview Press.
Gourieroux, C. and A. Monfort (1996) *Simulation-Based Methods Econometric Methods*. Oxford: Oxford University Press.
Greene, W. (1992) A statistical model for credit scoring. Department of Economics, Stern School of Business, New York University, Working Paper 92–29.
Greene, W. (1995) Sample selection in the poisson regression model. Working Paper No. EC-95–6, Department of Economics, Stern School of Business, New York University.
Greene, W. (1996) Marginal effects in the bivariate probit model. Working Paper No. 96–11, Department of Economics, Stern School of Business, New York University.
Greene, W. (1997) FIML estimation of sample selection models for count data. Working Paper No. 97–02, Department of Economics, Stern School of Business, New York University.
Greene, W. (1998) Gender economics courses in liberal arts colleges: further results. *Journal of Economic Education* **29**(4), 291–300.
Greene, W. (2001) Fixed and random effects in nonlinear models. Working Paper No. EC-01–01, Department of Economics, Stern School of Business, New York University.
Greene, W. (2004a) Convenient estimators for the panel probit model. *Empirical Economics* **29**(1), 21–47.
Greene, W. (2004b) Fixed effects and bias due to the incidental parameters problem in the Tobit model. *Econometric Reviews* **23**(2), 125–47.
Greene, W. (2006) Censored data and truncated distributions. In T. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics*, *Volume 1*. Basingstoke: Palgrave Macmillan.
Greene, W. (2007a) *LIMDEP/NLOGIT manual*. Plainview, New York: Econometric Software, Inc.
Greene, W. (2007b) A method of incorporating sample selection in a nonlinear model. Working Paper No. 07–16, Department of Economics, Stern School of Business, New York University.
Greene, W. (2008a) *Econometric Analysis* (sixth edition). Upper Saddle River: Prentice Hall.
Greene, W. (2008b) Functional forms for the negative binomial model for count data. *Economics Letters* **99**, 585–90.
Greene, W. and D. Hensher (2006) Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. *Transportation Research, B: Methodology* **40**(1), 75–92.
Greene, W., S. Rhine and M. Toussaint-Comeau (2006) The importance of check-cashing businesses to the unbanked: a look at racial/ethnic differences. *Review of Economics and Statistics* **88**(1), 146–57.
Groot, W. and H.M. Van den Brink (2003) Firm-related training tracks: a random effects ordered probit model. University of Amsterdam, http://www1.fee.uva.nl/scholar/wp/wp23-01.pdf.
Gurmu, S. (1997) Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics* **12**(3), 225–42.
Hardle, W. and C. Manski (1993) Nonparametric and semiparametric approaches to discrete response analysis. *Journal of Econometrics* **58**, 1–274.
Harris, M. and X. Zhao (2007) Modelling tobacco consumption with a zero inflated ordered probit model. School of Business and Economics, Monash University, Working Paper 14/04.
Hausman, J., B. Hall and Z. Griliches (1984) Economic models for count data with an application to the patents–R&D relationship. *Econometrica* **52**, 909–38.
Heckman, J. (1978) State dependence against the hypothesis of spurious state dependence. *Annalse de l'INSEE* **30**, 227–69.
Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica* **47**, 153–61.
Heckman, J. (1981a) Statistical models for discrete panel data. In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass.: MIT Press.

Heckman, J. (1981b) Heterogeneity and state dependence. In S. Rosen (ed.), *Studies of Labor Markets*. Chicago: University of Chicago Press.
Heckman, J. and T. MaCurdy (1981) A life cycle model of female labor supply. *Review of Economic Studies* **47**, 247–83.
Heckman, J. and J. Snyder (1997) Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *Rand Journal of Economics* **28**(0).
Hensher, D. and W. Greene (2002) Specification and estimation of the nested logit model: alternative normalizations. *Transportation Research B* **36**, 1–17.
Hensher, D. and W. Greene (2003) The mixed logit model: the state of practice. *Transportation Research B* **30**, 133–76.
Hensher, D., J. Rose and W. Greene (2005) *Applied Choice Analysis*. Cambridge: Cambridge University Press.
Honore, B. (2002) Non-linear models with panel data. Institute For Fiscal Studies, CEMMAP, Working Paper CWP13/02.
Honore, B. and E. Kyriazidou (2000a) Panel data discrete choice models with lagged dependent variables. *Econometrica* **68**(4), 839–74.
Honore, B. and E. Kyriazidou (2000b) Estimation of tobit-type models with individual specific effects. *Econometric Reviews* **19**(3), 341–66.
Horowitz, J. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica* **60**, 505–31.
Horowitz, J. (1993) Semiparametric Estimation of a work-trip mode choice model. *Journal of Econometrics* **58**, 49–70.
Hsiao, C. (1986) *Analysis of Panel Data*. Cambridge: Cambridge University Press.
Hsiao, C. (2003) *Analysis of Panel Data* (second edition). Cambridge: Cambridge University Press.
Hujer, R. and H. Schneider (1989) The analysis of labor market mobility using panel data. *European Economic Review* **33**, 530–6.
Hunt, G. (2000) Alternative nested logit model structures and the special case of partial degeneracy. *Journal of Regional Science* **40** (February), 89–113.
Hyslop, D. (1999) State dependence, serial correlation, and heterogeneity in labor force participation of married women. *Econometrica* **67**(6), 1255–94.
Jain, D.C., N.J. Vilcassim and K.D. Chintagunta (1994) A random-coefficients logit brand-choice model applied to panel data. *Journal of Business and Economic Statistics* **12**, 317–28.
Jones, J. and J. Landwehr (1988) Removing heterogeneity bias from logit model estimation. *Marketing Science* **7**(1), 41–59.
Kassouf, A. and R. Hoffmann (2006) Work related injuries involving children and adolescents: application of a recursive bivariate probit model. *Brazilian Review of Econometrics* **26**(1), 105–26.
Katz, E. (2001) Bias in conditional and unconditional fixed effects logit estimation. *Political Analysis* **9**(4), 379–84.
Kay, R. and S. Little (1986) Assessing the fit of the logistic model: a case study of children with hemolytic uremic syndrome. *Applied Statistics* **35**, 16–30.
Kiefer, N. (1982) Testing for independence in multivariate probit models. *Biometrika* **69**, 161–6.
King, G. (1989) A seemingly unrelated poisson regression model. *Sociological Methods and Research* **17**(3), 235–55.
Klein, R. and R. Spady (1993) An efficient semiparametric estimator for discrete choice. *Econometrica* **61**, 387–421.
Koop, G. (2003) *Bayesian Econometrics*. New York: John Wiley and Sons.
Krinsky, I. and L. Robb (1986) On approximating the statistical properties of elasticities. *Review of Economics and Statistics* **68**(4), 715–19.

Kyriazidou, E. (1997) Estimation of a panel data sample selection model. *Econometrica* **65**(6), 1335–64.
Lambert, D. (1992) Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* **34**(1), 1–14.
Lancaster, T. (2000) The incidental parameters problem since 1948. *Journal of Econometrics* **95**, 391–414.
Lancaster, T. (2004) *An Introduction to Modern Bayesian Inference*. Oxford: Oxford University Press.
Lee, E., J. Lee and D. Eastwood (2003) A two step estimation of consumer adoption of technology based service innovations. *Journal of Consumer Affairs* **37**(2), 37–62.
Lerman, S. and C. Manski (1981) On the use of simulated frequencies to approximate choice probabilities. In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass.: MIT Press.
Lewbel, A. (2000) Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics* **97**(1), 145–77.
Li, Q and J. Racine (2007) *Nonparametric Econometrics*. Princeton: Princeton University Press.
Long, S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.
Maddala, G. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.
Magee, L., J. Burbidge and L. Robb (2000) The correlation between husband's and wife's education: Canada. 1971–1996. *Social and Economic Dimensions of an Aging Population Research Papers* **24**, McMaster University.
Magnac, T. (1997) State dependence and heterogeneity in youth unemployment histories. INRA and CREST, Paris, Working Paper.
Manski, C. (1975) The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* **3**, 205–28.
Manski, C. (1985) Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313–33.
Manski, C. (1986) Operational characteristics of the maximum score estimator. *Journal of Econometrics* **32**, 85–100.
Manski, C. (1987) Semiparametric analysis of the random effects linear model from binary response data. *Econometrica* **55**, 357–62.
Manski, C. and S. Lerman (1977) The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977–88.
Manski, C. and S. Thompson (1986) MSCORE: a program for maximum score estimation of linear quantile regressions from binary response data. Mimeo, Department of Economics, University of Wisconsin, Madison.
Matzkin, R. (1993) Nonparametric identification and estimation of polytomous choice models. *Journal of Econometrics* **58**, 137–68.
McFadden, D. (1981) Econometric Models of Probabilistic Choice. In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass.: MIT Press.
McFadden, D. (2001) Economic choices. *American Economic Review* **93**(3), 351–78.
McFadden, D. and K. Train (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics* **15**, 447–70.
McQuestion, M. (2000) A bivariate probit analysis of social interaction and treatment effects. Center for Demography and Ecology, University of Wisconsin, Working Paper 2000–05.
Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–65.
Mundlak, Y. (1978) On the pooling of time series and cross sectional data. *Econometrica* **56**, 69–86.

Murphy, K. and R. Topel (1985) Estimation and inference in two step econometric models. *Journal of Business and Economic Statistics* **3**, 370–9.
Newey, W. (1987) Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* **36**, 231–50.
Neyman, J. and E. Scott (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.
Pudney, S. and M. Shields (2000) Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics* **15**(4), 367–99.
Pratt, J. (1981) Concavity of the log likelihood. *Journal of the American Statistical Association* **76**, 103–6.
Rabe-Hesketh, S., A. Skrondal and A. Pickles (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**(2), 301–23.
Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Paedogiska.
Revelt, D. and K. Train (1998) Mixed logit with repeated choices: households' choice of appliance efficiency level. *Review of Economics and Statistics* **80**(4), 647–57.
Riphahn, R., A. Wambach and A. Million (2003) Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics* **18**(4), 387–405.
Samuelson, P. (1947) *Foundations of Economic Analysis*. Boston: Atheneum Press.
Sepanski, J. (2000) On a random coefficients probit model. *Communications in Statistics – Theory and Methods* **29**, 2493–2505.
Shaw, D. (1988) "On-site samples" regression problems of nonnegative integers, truncation, and endogenous stratification. *Journal of Econometrics* **37**, 211–23.
Silva, J. (2001) A score test for non-nested hypotheses with applications to discrete response models. *Journal of Applied Econometrics* **16**(5), 577–98.
Stata, Inc. (2006) *Stata User's Guide*, Version 9.0. College Station, Texas: Stata Press.
Terza, J. (1985) Ordinal probit: a generalization. *Communications in Statistics* **14**, 1–12.
Terza, J. (1994) Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. Working paper IPRE 94–14. Department of Economics, Pennsylvania State University.
Terza, J. (1998) Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. *Journal of Econometrics* **84**(1), 129–54.
Tobias, J. and M. Li (2006) Calculus attainment and grades received in intermediate economic theory. *Journal of Applied Econometrics* **21**(6), 893–6.
Train, K. (2003) *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
van Doorslaer, E. and W. Nonneman (1987) Economic incentives in the health care industry: implications for health policy making. *Health Policy* **7**(2), 109–14.
Vella, F. and M. Verbeek (1999) Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics* **90**, 239–63.
Vuong, Q. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–34.
Wagstaff, A. (1993) The demand for health: an empirical reformulation of the Grossman model. *Health Economics* **2**, 189–98.
White, H. (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* **48**, 817–38.
White, N. and A. Wolaver (2003) Occupation choice, information and migration. *Review of Regional Studies* **33**(2), 142–63.
Willis, J. (2006) Magazine prices revisited. *Journal of Applied Econometrics* **21**(3), 337–44.
Winkelmann, R. (2003) *Econometric Analysis of Count Data* (fourth edition). Heidelberg: Springer Verlag.

Winkelmann, R. (2004) Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics* **19**(4) 455–72.
Wooldridge, J. (1995) Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* **68**(1), 115–32.
Wooldridge, J. (2002a) *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.
Wooldridge, J. (2002b) Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. CEMMAP, IFS and University College, London, Working Paper CWP18/02.
Wooldridge, J. (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**(1), 39–54.
Wynand, P. and B. van Praag (1981) The demand for deductibles in private health insurance. *Journal of Econometrics* **17**, 229–52.
Zavoina, R. and W. McKelvey (1975) A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, Summer, 103–20.

12

Panel Data Methods and Applications to Health Economics

Andrew M. Jones

Abstract

Much of the empirical analysis done by health economists seeks to estimate the impact of specific health policies, and the greatest challenge for successful applied work is to find appropriate sources of variation to identify the treatment effects of interest. Estimation can be prone to selection bias when the assignment to treatments is associated with the potential outcomes of the treatment. Overcoming this bias requires variation in the assignment of treatments that is independent of the outcomes. One source of independent variation comes from randomized controlled experiments. But, in practice, most economic studies have to draw on non-experimental data. Many studies seek to use variation across time and events that takes the form of a quasi-experimental design, or "natural experiment," that mimics the features of a genuine experiment. This chapter reviews the data and methods that are used in applied health economics with a particular emphasis on the use of panel data. The focus is on nonlinear models and methods that can accommodate unobserved heterogeneity. These include conditional estimators, maximum simulated likelihood, Bayesian MCMC, finite mixtures and copulas.

12.1 Introduction

A common thread that runs through this chapter is the "evaluation problem": is it possible to identify the impact of policies from empirical data? The focus of the chapter is on individual-level longitudinal data, so consider an "outcome" y_{it}, for individual i at time t. The treatment effect of interest is:

$$TE_{it} = \Delta_{it} = y_{it}^1 - y_{it}^0, \tag{12.1}$$

where 1 denotes treatment and 0 denotes control.[1] The pure treatment effect cannot be identified because the counterfactual can never be observed: each individual is either treated or untreated at a particular point in time so only one of the potential outcomes can be observed. The outcome that is actually observed can be written in terms of the potential outcomes:

$$y_{it} = y_{it}^0 + d_{it}(y_{it}^1 - y_{it}^0), \tag{12.2}$$

where d_{it} is an indicator of treatment.

One response to the problem of defining a counterfactual is to concentrate on the average treatment effect (ATE), comparing the average outcomes between the treated and controls:

$$ATE = E\left(y_{it}^1 - y_{it}^0\right). \tag{12.3}$$

When there is heterogeneity in individual responses to the treatment that may influence the assignment of treatment, for example, when doctors select patients on the basis of their capacity to benefit, attention is likely to focus on the average treatment effect on the treated (ATET) rather than the ATE:

$$ATET = E\left(y_{it}^1 - y_{it}^0 \mid d_{it} = 1\right). \tag{12.4}$$

This is the average effect of treatment for those individuals who would actually select into treatment.

Moving towards a regression framework, assume that the observed outcome under the two treatment regimes is given by the general regression model:

$$y_{it} = f_j(x_{it}, u_{jit}), j = 0, 1. \tag{12.5}$$

The vector x includes observable factors that influence the outcome and may influence the assignment of treatment (reflecting "selection on observables"). The u are unobservable factors that influence the outcomes and may influence the assignment of treatment ("selection on unobservables"). Formulating the problem in this way requires the SUTVA (stable unit-treatment value assumption) to hold – an individual's potential outcomes and treatments are independent of others in the population, ruling out spillover and general equilibrium effects. These spillovers may be important in some health economics applications and the evaluation of treatment effects would then have to be designed to accommodate them (see Chandra and Staiger, 2007; Miguel and Kremer, 2004). Using linear functions for $f(.)$ gives a switching regression model:

$$y_{it} = x_{it}'\beta_j + u_{jit}, j = 0, 1. \tag{12.6}$$

A simplification of this model, which assumes a homogeneous treatment effect so that only the intercept varies with treatment, gives the regression function:

$$y_{it} = x_{it}'\beta + d_{it}\,\delta + u_{it}\,. \tag{12.7}$$

In this case ATE = ATET = δ.

If unobserved factors (u) influence whether an individual is selected into the treatment group or how they respond to the treatment, this will lead to biased estimates of the treatment effect. A randomized experimental design may achieve the desired orthogonality of measured covariates (x, d) and unobservables (u). However, econometric studies typically rely on observational data gathered in a non-experimental setting. One strategy is to rely on selection on observables: finding a sufficiently rich set of observable characteristics so that unobservables can be assumed to have no systematic influence on treatments. This approach includes matching estimators and inverse probability weighted estimators. In contrast, the selection on unobservables strategy looks for factors that predict treatment, but have no direct effect on outcomes and which can therefore be used to mimic random assignment of treatment. This approach includes using within-individual variation to allow for time invariant individual heterogeneity in panel data models (fixed effects) as well as conventional instrumental variables (IVs) estimators. It also includes multiple equation models in which equations for the treatment and outcome are estimated jointly by full information maximum likelihood (FIML). "Natural experiments" often lead to the use of difference-in-differences estimators, which combine selection on observables (by including x in the regression models) with selection on unobservables (by using differencing to control for time invariant heterogeneity).

Natural experiments are often also linked to IV estimation, which relies on instruments (z) that predict the assignment of treatment, but do not have a direct effect on the outcome. When there is heterogeneity in the response to treatment the IV estimator identifies a local average treatment effect, or LATE (Imbens and Angrist, 1994; McClellan *et al.,* 1994). This is the average treatment effect over the sub-group of the population that are induced to participate in the treatment by variation in the instrument. The fact that IV estimates only identify the LATE and that the results are therefore contingent on the set of instruments explains why different empirical studies can produce quite different estimates, even though they examine the same outcomes and treatments. Heterogeneity in treatment effects is likely to be widespread: for example, Auld (2006a) finds considerable heterogeneity in the treatment effect of local HIV infection prevalence on risky sexual behavior among gay men in the San Francisco Men's Health Study (SFMHS), with HIV prevalence having less impact among those at high risk.

Recent work by Heckman and Vytlacil has extended the analysis of local treatment effects by specifying a model for the assignment of treatment and using it to identify those individuals who are indifferent between treatments, given x and z (see, e.g., Heckman and Vitlacyl, 1999, 2007; see Basu *et al.,* 2007, for an application to health data). This approach defines the marginal treatment effect (MTE): the treatment effect among those individuals at the margin. The MTE provides a building block for the LATE, ATET and ATE. It can be identified using local instrumental variables (LIVs) methods or by specifying multiple equation models with a common factor structure (see, e.g., Aakvik *et al.,* 2005; Basu *et al.,* 2007).

For example, in Aakvik *et al.* (2005) the treatment is a Norwegian vocational rehabilitation (VR) program and the outcome is a binary measure of employment. Analysis is based on a 10% sample of all those who applied for VR in 1989. To define the treatment effects of interest, Aakvik *et al.*, specify a discrete choice model with a common factor structure. There is a switching regression for the binary indicator of employment under the two treatment regimes:

$$\begin{aligned} y_i^1 &= f_1(x_i, u_{1i}) = 1(x_i' \beta_1 \geq u_{1i}) \\ y_i^0 &= f_0(x_i, u_{0i}) = 1(x_i' \beta_0 \geq u_{0i}), \end{aligned} \tag{12.8}$$

along with a latent variable model for the assignment of treatment:

$$\begin{aligned} &d_i = 1 \\ &\text{if} \\ &d_i^* = z_i' \beta_d - u_{di} > 0. \end{aligned} \tag{12.9}$$

The error terms are assumed to have a common factor structure:

$$\begin{aligned} u_{di} &= -\eta_i + \varepsilon_{di} \\ u_{1i} &= -\alpha_1 \eta_i + \varepsilon_{1i} \\ u_{0i} &= -\alpha_0 \eta_i + \varepsilon_{0i}\,. \end{aligned} \tag{12.10}$$

Estimation is by FIML, assuming that the error components are jointly normal.

Given this set-up, the treatment effects of interest can be defined as follows:

$$MTE(x,u) = E\left(\Delta | x, d^* = 0\right) = E\left(\Delta | x, u_d = z_i' \beta_d\right) \tag{12.11}$$

$$ATE(x) = \int_u MTE(x,u) dF(u) = E\left(\Delta | x\right) \tag{12.12}$$

$$ATE = E(ATE(x)) = \int_x E\left(\Delta | x\right) dF(x) \tag{12.13}$$

$$ATT(x,u) = E\left(\Delta | x, d = 1\right) = E\left(\Delta | x, u_d < z_i' \beta_d\right), \tag{12.14}$$

where $F(u)$ is the distribution of u and $F(x)$ is the distribution of the xs. Aakvik *et al.* (2005) do not use the concept of the LATE in their study but, based on the notation of their model, it could be expressed as:

$$LATE(x, z, \tilde{z}) = E\left(\Delta | x, z_i' \beta_d < u_d < \tilde{z}_i' \beta_d\right), \tag{12.15}$$

where, for illustration, it is assumed that assignment to treatment is monotonically related to a single instrument that takes two values z and $\tilde{z}$, where $z_i' \beta_d < \tilde{z}_i' \beta_d$. The LATE defines the treatment effect for all those individuals who are induced into the treatment by the change in the instrument (see, e.g., Basu *et al.*, 2007).

The nonlinear model is identified by functional form, but an exclusion restriction is also imposed by including an instrument – the degree of rationing of VR places in the individual's locality – in z, but not in x. The apparent positive impact of

the VR program is reversed when selection bias is taken into account and there is evidence of perverse cream-skimming, with those most likely to benefit being the least likely to be selected by the program administrators.

A note on the scope of the chapter

This chapter takes the identification of treatment effects as its starting point and concentrates on microeconometric methods that can be used with longitudinal and other complex and multilevel datasets. Although the methods described in the chapter are widely used throughout applied econometrics, the applications reviewed here all relate to one specific area: health economics. The chapter follows an earlier review of the literature on "health econometrics" (Jones, 2000) and concentrates on studies that have appeared as peer-reviewed publications from 2000 onwards. The emphasis is on applications that use health and health care as outcomes. Less attention is devoted to the large number of studies of health-related behaviors, such as diet, smoking, drinking and illicit drugs (see, e.g., Cowell, 2006; Dee *et al.,* 2005; Forster and Jones, 2001; Harris *et al.,* 2006; Terza, 2002; Van Ours, 2006) and to those that investigate the impact of health, health care and health insurance on labor market outcomes (see, e.g., Askildsen *et al.,* 2005; Au *et al.,* 2005; Auld, 2002; Bradley *et al.,* 2005; Contoyannis and Rice, 2001; Disney *et al.,* 2006; French, 2005; Hogelund and Holm, 2006; Morris, 2006, 2007; Royalty and Abraham, 2006; Stewart, 2001; Van Ours, 2004) or labor outcomes for health care professionals (see, e.g., Arulampalam *et al.,* 2004; Frijters *et al.,* 2006; Holmas, 2002). The scope does not include studies that use econometric techniques in the context of contingent valuation and discrete choice experiments, where random effects models are often applied (see, e.g., Ryan *et al.,* 2006); multinomial models of the choice of insurance plans or health care providers (see, e.g., Deb and Trivedi, 2006; Ho, 2006; Sahn *et al.,* 2003); productivity analysis based on models of cost and production functions and estimation of stochastic frontier models (see, e.g., Bradford *et al.,* 2001; Burgess, 2006; Dranove and Lindrooth, 2003; Smith and Street, 2005; Wilson and Carey, 2004); and in the context of cost-benefit and cost-effectiveness analysis, where econometric methods are starting to be used alongside methods from biostatistics and epidemiology (see, e.g., Briggs, 2006; Hoch *et al.,* 2002; Willan *et al.,* 2004).

The focus is primarily on studies that use micro-level data derived from longitudinal, multilevel and other complex data structures. Relatively few cross-section studies are discussed and the chapter does not attempt to review studies that use aggregate time series or panels and that apply pure time series methods (see, e.g., Aakvik and Holmas, 2006; Abadie and Gay, 2006; Chou, 2007; García-Ferrer *et al.,* 2007; Leigh and Jencks, 2007; Or *et al.,* 2005; Paton, 2002; Ruhm, 2003; Wang and Rettenmaier, 2007). Analysis of longitudinal data often makes use of the methods of survival analysis (see, e.g., Arulampalam *et al.,* 2004; Chou, 2002; Disney *et al.,* 2006; Farsi and Ridder, 2006; Forster and Jones, 2001; Frijters *et al.,* 2006; Harrison, 2007; Holmas, 2002; Kyle, 2007; Picone *et al.,* 2003a; Stewart, 2001; Van Ours, 2004, 2006), but these methods are not discussed in detail here.

The empirical findings of many of the studies are discussed, but no attempt is made to provide a systematic synthesis of the empirical results. It is notable that meta-analyses of regression results are beginning to appear in the health economics literature. For example, Gallet and List (2003) present a meta-analysis of the tobacco price elasticity and Gemmill *et al.* (2007) carry out a meta-regression of estimates of the price elasticity of prescription drugs.

On the whole, the original sources for the econometric methods are not cited. These are reviewed in Volume 1 of this Handbook (particularly the chapters by Badi Baltagi, William Greene and Lung-fei Lee) and in other chapters in this second volume, in particular those by Colin Cameron, William Greene, and David Jacho-Chávez and Pravin Trivedi.

12.2 Identification strategies: finding relevant variation

The success of applied work depends on finding appropriate sources of variation to identify the effects of interest. Estimation of treatment effects can be prone to selection bias, where the assignment to treatments is associated with the potential outcomes of the treatment. Overcoming this selection bias requires variation in the assignment of treatments that is independent of the outcomes. One source of independent variation comes from randomized controlled experiments. While these are the norm in the evaluation of new clinical therapies, their use for the evaluation of social programs remains rare (Gertler, 2004; Kremer, 2003; Miguel and Kremer, 2004). Most economic studies have to draw on non-experimental, or observational, data. This section presents a series of case studies from the recent literature and describes how these have sought out relevant identifying information.

12.2.1 Randomized experiments

The "gold standard" methodology that is used to identify the efficacy and effectiveness of new medical technologies is the randomized clinical trial (RCT). Much of the work done by health economists to measure the cost-effectiveness of these technologies draws on data collected within RCTs to perform statistical analyses (Briggs, 2006) or to calibrate decision analytic models (Claxton *et al.,* 2006). Econometric methods are sometimes used in secondary analysis of such data to model costs and outcomes as functions of observable covariates (Willan *et al.,* 2004).

Broader randomized social experiments are far less prevalent. One exception, which has played a highly influential role in the development of health economics and has driven many of the early developments in the use of econometrics in the field, is the RAND Health Insurance Experiment (Manning *et al.,* 1987). The RAND experiment was designed to address the problem of self-selection in the choice of insurance plans. Participants were randomized between a health maintenance organization (HMO) and reimbursement plans and across plans with different levels of co-payments and a plan with deductibles. The RAND study has had a strong influence, especially in the US. It has focused attention on the use of two-part or multi-part models to model health care utilization and expenditures and on the choice of functional form to deal with heavily skewed data and the consequent

problems of retransformation back to the "natural scale" in the presence of individual heterogeneity (Manning, 2006). The RAND data is available over the internet and has been used to test more recent developments of econometric methods (Bago d'Uva, 2006; Deb and Trivedi, 2002; Gilleskie and Mroz, 2004; Vera-Hernandez, 2003).

More recently, randomized experiments have begun to play an influential role in research and policy in developing countries. This is exemplified by the studies of Gertler (2004) and Miguel and Kremer (2004). The Mexican government's PROGRESA program, which was initiated in 1997, has received considerable attention and has influenced policy throughout Latin America. The program relies on conditional cash transfers that are designed to influence the use of health and welfare services for children in poor families. It covers 2.6 million families in 50,000 rural villages. The program focuses on health, hygiene and nutrition. It links substantial cash transfers, on average amounting to 20–30% of household income, to the use of prenatal care, well-baby care and immunization, nutrition monitoring and supplementation, preventive check-ups and participation in educational programs. PROGRESA works by first selecting whole communities to participate in the scheme and then selecting households within those communities that satisfy the eligibility criteria to receive the benefits of the scheme. Financial constraints on the implementation of PROGRESA meant that its introduction was phased. To make the implementation equitable, communities were selected randomly to receive the benefits either immediately or with a delay. The random phasing provides researchers with an ideal opportunity to use a randomized design in the evaluation of the impact of the program. Of the communities selected for the program, 320 were randomly selected to receive the intervention in August–September 1998 with the remaining 185 delayed for two years. The communities in the control group were not informed that they would eventually receive the program, reducing the scope for anticipation of treatment to influence the outcome.

Gertler (2004) focuses on health outcomes among children. These include self-reported morbidity, measured by illnesses in the past month, as reported by the child's mother, and objective measures including anthropometric measures of height and stunting and a biomarker for anaemia (haemoglobin levels). The analysis is restricted to those households, in both the treatment and control groups, that satisfy the eligibility criteria for PROGRESA. Although the data is randomized, multivariate regression models are used to control for observed covariates and to reduce idiosyncratic variation. Individual and village random effects are included, the latter to allow for the clustered sampling. The results show significant improvements in both self-reported and objective measures of health and the impact increases with length of exposure to the program. Gertler (2004) is careful to note that the comparison of treated and controls does not explain the mechanism behind this effect: for example, it is not possible to say whether an unconditional transfer would have had the same effect as the conditional one.

In their "worms" paper, Miguel and Kremer (2004) analyze a randomized experiment to evaluate the impact of the Kenyan Primary School Deworming Project

(PSDP) on hookworm infection rates and on school attendance. The program included drug therapy and public health education on avoiding hookworm infection, with the assignment of treatment randomly phased. Randomization was done at the level of schools rather than individuals: one group of schools received treatment in 1998 and 1999, another group only in 1999 and a third group only in 2003. Data was collected in 1998 and 1999 so, in the Miguel and Kremer study, the first group are the treated and the second and third groups make up the controls. Miguel and Kremer argue that randomization at the levels of schools is crucial in this context as it avoids biases created by spillover effects of the deworming program in reducing infection rates. They argue that an ideal prospective study would randomize treatments across pupils within schools, across schools within clusters and across these clusters. This multilevel variation in the assignment of treatments could then be used to estimate different levels of the treatment effect in the case where spillovers are important.

12.2.2 Natural experiments

12.2.2.1 Health shocks

Almond (2006) makes inventive use of the 1918 influenza pandemic as a natural experiment to provide evidence in favour of the "fetal origins hypothesis." Cohorts that were in utero during the pandemic, between the fall of 1918 and January 1919, are shown to have poorer outcomes: lower educational attainment, more disability, lower income, lower socioeconomic status and higher transfer payments. The pandemic has the potential to be used as a natural experiment: it was unanticipated, the period of exposure was short and the impact varied systematically across states. The study uses discontinuity across birth cohorts to identify the long-term effects, drawing on data from the 1960, 1970 and 1980 US Census microdata (which identify quarter of birth). Geographic variation is also exploited, based on the "laggard" states where the epidemic had less pronounced long-term effects, although this does reduce the sample size available. This is a paper where simple graphical analysis tells the main story, although it is backed up by thorough statistical modeling.

Doyle (2005) makes innovative use of data on severe traffic accidents to measure variation in unanticipated health shocks and finds that, in the United States, the uninsured receive 20% less treatment and have a substantially higher mortality rate. The Crash Outcome Data Evaluation System (CODES) links police accident reports to hospital discharge data. This study uses data for Wisconsin covering 1992–97, with a sample of 28,236 individuals, 10% of whom were uninsured. Severe traffic accidents are assumed to be unanticipated at the time that insurance is taken out and the consequent use of health care is non-discretionary. Descriptive evidence suggests that the uninsured are riskier drivers and have worse health problems, creating a problem of selection bias. To deal with selection a control group is selected from those with medical insurance, but without car insurance. Within-hospital variation and time effects are controlled for. The robustness of the findings is checked by using the sub-sample where both insured and uninsured individuals

are injured in the same crash, allowing for different severities of accident (accident fixed effects). Also the sample of passengers is used to abstract from differences in the quality of driving between the treated and control groups. Robustness is assessed by doing separate analyses by diagnoses and by medical procedures. In the light of these results, the lower levels of treatment for the uninsured are attributed to decisions made by providers in response to insurance status rather than differences in background characteristics of the patients. Similar issues are faced by Levitt and Porter (2001), who address the problem of selection bias in the US Fatality Analysis Reporting System (FARS), which they use for an analysis of the effectiveness of seat belts and air bags. The problem arises because data is only included for fatal crashes. The use of safety devices influences the probability of survival and hence of inclusion in the sample. The identification strategy adopted to get around this problem is to use a sample based on crashes where someone in a different car dies. The aim is to make the sample selection independent of the observation's own treatment status and outcomes. In doing so they find that seat belts are more effective and air bags are less effective than previous evidence had suggested.

12.2.2.2 Economic shocks

Evans and Lien (2005) make use of the 1992 Port Authority Transit (PAT) strike in Allegheny County, Pennsylvania, as a source of independent variation in access to prenatal care. Prenatal visits were affected most for black women and city residents (in Pittsburgh) and the results show that, for these groups, missing visits early in pregnancy had a detrimental effect, but missing those later in the pregnancy did not. The main source of information is observational data from the 1990–94 US Natality Detail Files, which contain a census of births in each given year, taken from birth records. This is augmented by survey data that is used to assess the impact of the strike on access to prenatal care. A control group of counties that were not affected by the strike are selected on the basis of regression analyses. The use of prenatal care by women who were pregnant at the time of the strike is included in regression equations for birth weight, gestation, maternal weight gain and maternal smoking. Models are estimated by ordinary least squares (OLS) and two-stage least squares (2SLS), the latter using the strike as an instrument. These produce similar results, suggesting that selection bias is not a problem. The clearest effect of prenatal care is on maternal smoking. The robustness of the findings is tested by checking for a general decline in earnings or employment coincident with the strike and for evidence of increases in abortions or "unwanted" births.

In Frijters *et al.* (2005) the reunification of Germany in 1990 provides a natural experiment to assess the causal effect of income on self-reported health satisfaction. A positive and statistically significant effect is found, but the effect is small. The increase in incomes for those in East Germany is used as a source of independent variation in income that is not contaminated by reverse causality from health. The suitability of this setting as a natural experiment is justified by the fact that the changes in income associated with the fall of the Berlin Wall are assumed

to have been unanticipated, that the income transfers were large in magnitude (affecting the real value of savings, collectively bargained wages and pay in general) and that there was individual variation in the impact, with civil servants experiencing an immediate effect. This variation is exploited in an econometric framework that also allows for entry and attrition from the panel dataset and for inherent individual heterogeneity. The analysis uses longitudinal data from the German Socio-economic Panel (GSOEP) from 1984 to 2002 for West Germans and from 1990 to 2002 for East Germans. Data for East Germans is not available prior to reunification, so separate models are estimated for East and West Germans and the natural experiment has to be used indirectly.

In 1996 a crisis in the public pension system in Russia meant that 14 million of the 39 million state pensioners faced substantial arrears in their payments. Jensen and Richter (2004) exploit this pensions crisis as a natural experiment. Their findings show a doubling of poverty rates, significant declines in calorie and protein intake, and reductions in the use of health services and medications. They also show evidence of attempts to mitigate the loss of pension income through work, sales of assets, borrowing and private transfers. Data from the Russian Longitudinal Monitoring Survey (RLMS) for 1995 and 1996 is used to assess the impact of the crisis. Identification stems from geographic variation in arrears, which arises because decisions were regionally decentralized across administrative areas (oblasts) and discretion was exercised within oblasts. The control group is made up of households who continued to receive their pensions. Estimation uses a difference-in-differences design, with the policy effect measured by an interaction between the post-1996 period and whether an individual's pension was in arrears. The identification strategy relies on the assignment of arrears not being associated with outcomes prior to the crisis. The paper attempts to assess the validity of this assumption and presents evidence of a common trend for treated and controls prior to the crisis.

In contrast to Jensen and Richter (2004), Duflo (2000) uses a positive economic shock associated with public pensions as a source of exogenous variation in income in her study of child health in South Africa. The end of the Apartheid era in the early 1990s led to large increases in benefits for black Africans within the South African Old Age Pension System. Duflo's study uses cross-section data collected during 1993 and faces a selection problem, as children living in households with pension recipients are more likely to be disadvantaged and to live in rural areas. Her identification strategy compares eligible and non-eligible households and those children exposed to the increased household pension income for all of their lives or for only a fraction of their lives. Outcomes are measured using height-for-age z-scores and there is evidence of an effect on child health and nutrition. This effect is entirely attributable to pensions received by women and the effect is strongest for girls.

Chay and Greenstone (2003) bring together a comprehensive set of data sources within a quasi-experimental research design to investigate the impact of atmospheric pollution on infant health in the US. They find significant effects of total suspended particles (TSPs) on infant mortality, mostly driven by deaths within one

month of birth. There is heterogeneity in this effect, with the impact on infant mortality rates being twice as large among blacks. Identification is based on geographic variation in the impact of the 1981–82 recession on levels of TSP, which is treated as a source of random variation. County-level data from various sources are merged for the period 1978–84. First differenced (fixed effects) models and, unusually, differenced models with fixed effects for the trend (double differencing) are used. The latter allows for heterogenity in trends. The analysis goes a step further by selecting neighboring counties as controls. This uses non-manufacturing counties that either are or are not neighbors to manufacturing counties to try and isolate the effects of pollution from the socioeconomic effects of the recession.

Lindahl (2005) shows how lottery winnings can provide one source of exogenous variation in income, in an attempt to overcome the selection biases inherent in disentangling the socioeconomic gradient in health. There is a statistically significant effect of income on morbidity and mortality and the magnitude of this effect is largely unchanged when lottery winnings are used as an instrument, although the estimates are less precise. This income effect is not apparent for the sub-sample aged over 60. Data from the Swedish Level of Living Surveys (SLLS) for 1968, 1974 and 1981 are matched with register data on income and deaths up to 1997. Morbidity is measured by combining 48 symptoms into a standardized measure and mortality is measured as death within five or ten years of the surveys. Lottery winnings are treated as a source of exogenous variation in income: assuming that the variation is independent of health. Models are estimated with lottery winnings included directly. Then OLS and instrumental variable estimates (using winnings as the instrument) are compared for the sample of individuals who are identified as "players." The magnitudes of the income effects are similar although standard errors are inflated when IV is used. A similar strategy is adopted by Gardner and Oswald (2007). They use data on the General Health Questionnaire (GHQ-12) measure of psychological well-being from the British Household Panel Survey (BHPS) for 1996–2003 and compare those who received lottery winnings of between £1,000 and £120,000 to two control groups, those with smaller wins and those with no wins. The study finds a statistically significant effect of 1.4 GHQ-12 points after two years (compared to the average drop of 5 points associated with widowhood). An important caveat is the small number of treated cases: there are only 137 observations with large lottery wins.

In Van Den Berg *et al.* (2006) the state of the economy during infancy is shown to have long-term consequences for mortality rates in this inventive study of those born in the nineteenth century in the Netherlands. The analysis finds a significant effect of the stage of the business cycle (boom or bust) at the time of birth on individuals' subsequent age of death. Data from the Historical Sample of the Netherlands (HSN), drawn from registers of births, marriages and deaths, covers 14,000 individuals born between 1812 and 1912 with follow-up to 2000. This data is merged with macroeconomic time series that are used to identify the phases of the business cycle. Macroeconomic conditions early in life are used as

an instrument for socioeconomic conditions in infancy in order to avoid the problems of unobservable heterogeneity bias that plague cross-section comparisons. The impact of early life conditions is analyzed nonparametrically, comparing those born in booms and recessions, and through duration analysis of the individual mortality data.

12.2.2.3 Educational reforms

Lleras-Muney (2005) shows how historical changes in the US educational system can be used as a natural experiment, based on a discontinuity design, to identify the effect of education on adult health. The estimated effect is larger than previous studies have suggested, with the magnitude of the instrumental variables estimate of the local average treatment effect three times larger than the OLS estimate. The natural experiment is based on changes across states in compulsory schooling and child labor laws between 1915 and 1939. Identification stems from variation over states and across time in the age at which children had to enter school, the age at which they could leave school and get a work permit, and whether those with work permits had to continue in school part time. Estimation uses a regression discontinuity design, which attributes any jumps associated with school leaving age to the policy effect. This is applied using a linear probability specification for deaths as a function of years of schooling. Synthetic cohorts are constructed from successive US Censuses (1960, 1970 and 1980) to select those who were aged 14 between 1915 and 1939 and to follow-up subsequent mortality rates. These are synthetic in the sense that they do not follow the same individuals and are based on gender, birth cohort and state of birth.

Educational reforms are also used as a natural experiment in Arendt (2005). In this case the analysis focuses on Denmark and reforms in 1958 that removed formal tests before middle school, and 1975, that increased the compulsory minimum school leaving age. Data are taken from the Danish National Work Environment Cohort Study (WECS), with two waves in 1990 and 1995, and covers workers aged 18–59 in 1990. The impact of years of schooling on outcomes later in life is estimated using two-stage conditional maximum likelihood (2SCML) estimates, allowing for a random individual effect, for self-reported health. Models are also estimated for body mass index and for an indicator of never having smoked. The latter is included for comparison as, for most people, it is determined while they are still in education. The impact of education on health is amplified when instruments are used, but, at the same time, the standard errors are inflated so that exogeneity is not rejected. However, tests suggest that there may be a problem of weak instruments, as the reforms have low explanatory power in the reduced form equations.

12.2.2.4 Health policies and reforms

Bleakley (2007) is a good example of combining a natural experiment with a long-term follow-up to explore the economic consequences of a public health

intervention aimed at children. A program aimed at the eradication of hookworm is shown to lead to a long-term gain in the income of beneficiaries. Areas with greater scope for benefiting, due to higher levels of hookworm infection, show greater contemporaneous increases in school enrollment and attendance and in literacy among children. The natural experiment is the Rockefeller Sanitary Commission's (RSC's) funding of treatment and education programs to eradicate hookworm in the Southern US, which took place around 1910–15. This policy intervention was implemented over a well-defined and relatively short period. Geographic differences in infection rates prior to the intervention can be used to formulate and identify a treatment/control design, estimated using difference-in-differences. Data on long-term consequences for the cohorts exposed to the eradication program are obtained from the US Census available through the Integrated Public Use Micro Sample (IPUMS).

In 1966 the Ceauşescu regime in Romania banned abortion and family planning. Birth rates doubled the following year. In Pop-Eleches (2006) this provides an interesting contrast to studies that have examined moves in the opposite direction in the United States. The raw data show an improvement in educational attainment and labor market outcomes associated with the ban, but these results are reversed by allowing for compositional changes in the type of families having children. The findings are explained by the fact that affluent urban women were more likely to have abortions and use contraception before the ban. Data are drawn from a 15% sample of the 1992 Romanian census and focus on children born between January and October 1967. There is a spike in births between July and October due to the ban, but all of these children entered school in the same year and experienced the same overcrowding effect. Although it is not labeled as such, the paper uses a discontinuity design estimating a simple difference equation that includes a dummy variable for the period after the policy. Additional covariates are included, but there is no control group and identification relies on any sudden changes in outcomes for those born just before or just after the ban.

Lakdawalla *et al.* (2006) present a careful application of the method of instrumental variables, based on state-level variation in Medicaid eligibility in the US, which shows that an unintended consequence of highly active antiretroviral therapy (HAART) is to increase risky sexual behavior among patients who are HIV+. Simple correlations show lower sexual activity among those who are HIV+, but this is because of the debilitating effects of the disease and does not show the causal effect of treatment. Panel data on HIV+ patients in care are taken from the HIV Costs and Sevices Utilization Study (HCSUS) for the period 1996–98. The outcome of interest is the number of sex partners of the previous six months and the treatment is HAART, which is inferred from records of medications. Simple, unconditional estimates do not show a difference in sexual activity. But when treatment is instrumented by variation in the eligibility rules for Medicaid across states a positive effect emerges. The validity of these instruments is checked by examining the reduced form association between Medicaid eligibility and sexual activity prior to

the introduction of HAART in 1996. There is no association pre-1996, but there is post-1996.

12.2.3 Natural controls

12.2.3.1 Families

Auld and Sidhu (2005) present evidence that around a quarter of the association between schooling and health is attributable to variation in cognitive ability and that the causal effect of schooling on health is concentrated among those with low levels of education. Estimates that allow for both schooling and health to be influenced by a common "third factor" diminish the effect of schooling on health except for those individuals with no greater than high school education. The models are estimated using the 1979 and 2000 US National Longitudinal Survey of Youth (NLSY). The validity of the estimated causal effects relies on the use of parental education as a source of independent variation, using variation in the individual's own education that is associated with their parents' educational achievements.

Siblings who are brought up together share common background characteristics which may be unobserved and also influence the treatments and outcomes of interest to researchers. Using within-sibling variation can control for these factors. Holmlund (2005) shows how variation within biological sisters can be used to assess the long-term consequences of teenage pregnancy for educational outcomes. The siblings approach and standard cross-section methods produce similar results providing heterogeneity within the family is controlled for. The potential for selection bias is that teenage mothers may have family backgrounds that would lead to poorer outcomes irrespective of an early pregnancy. Variation within biological sisters can be used to control for these "family effects." However, within-sibling variation will not deal with heterogeneity within the family and the study controls for observable pre-motherhood school performance, measured by the grade point average (GPA) from primary school, to try and control for this. Data are taken from a 20% sample of each cohort born in Sweden between 1974 and 1977, with the population register used to identify siblings.

Sibling fixed effects play a role in Currie and Stabile's (2006) study of the impact of Attention Deficit Hyperactivity Disorder (ADHD) on educational outcomes. Within-sibling variation is used to control for omitted variables at the level of the family. Data from the Canadian National Longitudinal Survey of Children and Youth (NLSCY) and the US NLSY are used. ADHD symptoms are based on parental reports and are recorded in 1994 in the Canadian data and between 1990 and 1994 in the US data. Educational outcomes include the repetition of grades, enrollment in special education, reading and math tests and delinquency. These are measured in 1998 for Canada and 1998–2000 for the US. The study finds large effects, relative to chronic physical conditions, and for low levels of ADHD symptoms in cases that would not usually receive treatment. The results for Canada and the US are similar to each other.

12.2.3.2 Twin studies

Within-sibling variation can control for common factors relating to family background, upbringing and environment. But siblings are born at different times and they have different genes. Twin studies take the notion of natural controls a step further by removing the genetic variation (at least for monozygotic twins). In the context of research on birth outcomes, using twins means that the siblings share the same pregnancy and are born at the same time. This controls for unobservable characteristics of their mother and her behavior and environment during the pregnancy. Almond *et al.* (2005) show that using variation in birth weight between twins leads to lower estimates of the impact of low birth weight (LBW), defined as less than 2,500 g, on short-run outcomes than is typically found in cross-section studies. They find heterogeneity in the effects of LBW, suggesting a highly nonlinear relationship. Two identification strategies are adopted. The first exploits variation "within mothers" by comparing outcomes for heavier and lighter infants for all twins born in the US between 1983 and 2000. Using this within-variation should control for all observed and unobserved characteristics of the mother. The second exploits variation "between mothers" in a complementary analysis of maternal smoking and singleton births. The strategy here is to attribute the whole effect of smoking to LBW and compare it with the twins estimates. Data are drawn from two sources: linked birth and infant deaths data from the US National Center for Health Statistics (NCHS), covering the population of US twins, and data from hospital discharge abstracts from the Healthcare Cost and Utilization Project (HCUP) state inpatient database. There are some caveats to bear in mind with this study. Some useful descriptive analysis presented in the paper highlights the inherent differences between twins and singletons (the latter are more healthy). This raises questions about external validity of analysis based on samples of twins rather than the general population. The study only uses short-run outcomes and may miss long-term consequences (see the studies by Behrman and Rosenzweig, 2004, and Black *et al.,* 2007, below). For fraternal twins, genetic differences may mean that changes in birth weight may be associated with changes in unobservables (there is evidence of a negative correlation of birth weight with congenital defects) and the fixed effects approach may overestimate the impact of birth weight. Also the data do not distinguish between monozygotic (identical) and dizygotic (fraternal) twins.

In Behrman and Rosenzweig (2004), variation between monozygotic twins provides a way of identifying the impact of birth weight on long-term outcomes, such as measures of adult health, anthropometric measures and adult schooling and earnings. Increased birth weight, as measured on the birth certificate, increases schooling among adults and this effect is underestimated by 50% when cross-section variation is used to identify the effect. Data were collected through a survey mailed to monozygotic twins on the Minnesota Twins Registry, the largest birth certificate based registry in the US. The identification strategy assumes that difference in birth weight reflects random differences in nutrition in the womb that are uncorrelated with individual endowments and therefore avoids selection

bias. The estimates allow for heterogeneous treatment effects and show an impact on labor market outcomes for low birth weight, but not for high. The implications of the US results for worldwide health inequalities are explored at the end of the paper.

Black *et al.* (2007) use Norwegian registry data and, like Behrman and Rosenzweig (2004), they use twins to investigate the impact of low birth weight on long-term socioeconomic outcomes rather than just short-run outcomes. Within-twins fixed effects estimates are shown to be significant and similar to standard least squares estimates for long-run outcomes, such as height, IQ , earnings and education, while the estimates for short-run outcomes are smaller for the twins data, as suggested by Almond *et al.* (2005). The analysis is made possible by the richness of the data, which use personal identifiers to link all Norwegian births between 1967 and 1997, as recorded in the birth registry, with other registry data for those aged 16–74 in the period 1986–2002. The register data is augmented with military records and a survey of twins that identifies zygocity. Within-twin variation is used to capture unobservable socioeconomic and genetic factors that may confound the causal effect of birth weight. This means that identification stems from differences in nutrition in utero (resulting from different placentas for fraternal twins and different positioning on the placenta for monozygotic). Birth order is included as a control. The robustness of the findings is assessed by separate analyses for mothers who have more than one singleton birth, allowing for mother fixed effects rather than pregnancy fixed effects. To assess the role of zygocity the sample is restricted to same-sex twins. The sub-sample where there is survey data on zygocity is also used. The findings are robust, but reveal interesting evidence that those who participate in twins studies are a self-selected sample. Also it should be borne in mind that selection into the sample of registry data for long-run outcomes may be affected by infant mortality. Finally, there are substantial differences between twins and singletons in terms of factors, such as gestation and the age of their mothers, and twins usually appear in the lower part of the distribution of birth weights.

12.2.3.3 Communities

Many studies use variation within groups, communities or geographic areas to control for unobservable factors that are common to all those within the community or locality. For example, Wagstaff (2007) controls for village effects in a study of the impact of health shocks, such as the death of a working-age member of the household, on incomes of urban and rural households in Vietnam based on the Vietnam Living Standards Survey (VLSS). Arcidiacono and Nicholson (2005) find that adding fixed effects for individual medical schools eliminates the positive peer effects that appear to exist when selection bias is not taken into account. The inclusion of school effects means that the impact of peer effects on a student's achievements and on their choice of specialty are identified by variations over time within schools in the ability and preferences of students. The aim is to separate correlated effects from exogenous peer effects. The study relies on data for graduates from US Medical Schools over a relatively short period, 1996–98, so identification may be limited by a lack of variation over time. Currie and Neidell (2005) use

variation within Californian zip code areas to identify the impact of air pollution on infant mortality. They find a statistically significant effect even at low levels of air pollution. The impact of this effect is quantified: it is estimated that reduction in pollution in California over the 1990s saved around 1,000 infant lives. The study takes data from the California birth cohort files and matches it to EPA data on air quality - specifically measures of carbon monoxide, ozone and particulate matter (PM10) – and information on weather patterns from the National Climatic Data Center. A linear model is used to approximate the discrete hazard function for infant deaths and month, year and zip code fixed effects are included in the model, this relies on variation within cells of observations defined by month, year and locality. The study complements the natural experiment presented by Chay and Greenstone (2003) that is described above.

12.2.4 Anti-tests

One way to assess the robustness of an identification strategy is to find an anti-test (or placebo test). Anti-tests provide counter-evidence by applying a model or identification strategy in a context where no effect should be detected. If an apparent "effect" is found then the validity of the identification strategy must be called into question.

For many years the standard empirical strategy to test for the phenomenon of supplier-induced demand (SID) in medical care has been to include a measure of the supply of doctors – usually the physician density, measuring the number of doctors in a locality per head of population – in empirical models of health care utilization or expenditure. This strategy is plagued by omitted variable bias and identification problems. To assess the robustness of the approach, Dranove and Wehner (1994) apply the physician density strategy using the obstetrician/population ratio and the volume of births as the measure of utilization. The physician density test shows evidence that the number of births (and hence pregnancies) is "supplier induced": casting obvious doubt on the reliability of the approach. However, failure of the methodology does not imply rejection of SID. For example, Gruber and Owings (1996) find evidence of increased C-section rates in response to a fall in fertility in the US between 1970 and 1982: a shift by obstetricians to more lucrative procedures in response to economic pressures.

In their "addiction to milk" paper, Auld and Grootendorst (2004) use non-addictive substances, such as milk, eggs and oranges, to construct an anti-test and demonstrate that evidence for the rational addiction hypothesis based on aggregate data may be spurious. Numerous studies have applied the canonical rational addiction equation of Becker *et al.* (1994) to substances such as alcohol, cigarettes and cocaine, and claim to have found support for rational addiction; but Auld and Grootendorst (2004) show that these findings are mimicked when the model is applied to Canadian aggregate data for non-addictive substances. Monte Carlo simulations show that spurious evidence is likely when the time series data exhibit high serial correlation, when prices are poor instruments, when overidentified instrumental variable estimators are used, or when theoretical restrictions are imposed by fixing the implied discount rate in the model.

The idea of an anti-test may provide a useful strategy as part of a robustness/sensitivity analysis. A good example of this is Galiani *et al.*'s (2005) evaluation of the impact of the privatization of local water services on child mortality in Argentina. They adopt two strategies for assessing the reliability of their difference-in-differences approach that can both be interpreted as anti- or placebo tests. The first, which is a good practice to adopt in any difference-in-differences analysis, is to estimate a placebo regression: the model of interest is estimated using only data from the pre-treatment period, but including an indicator of those cases that will go on to be treated. If this indicator of hypothetical treatment is significant it is a sign that the treated and controls are not comparable and that the "parallel trends" assumption required for difference-in-differences analysis is not valid. The second strategy adopted by Galiani *et al.* (2005) is that, as well as measuring deaths from infectious and parasitic diseases, they include measures of deaths from causes unrelated to water quality. The fact that they detect a reduction for the former but not for the latter creates confidence in their difference-in-differences identification strategy.

12.3 Data and measurement issues

12.3.1 Administrative data or sample surveys

Much of the applied work done by health economists uses social surveys. These are often designed to provide representative random samples of the underlying population. Most often the sampling follows a multi-stage design with clustered and/or stratified sampling (see, e.g., Jones *et al.,* 2007b). Data may be collected by face-to-face interviews or postal, telephone or web-based questionnaires, and in health surveys this is often supplemented by clinical tests and measurements. Many surveys are one-off cross-sections, but increasingly researchers have turned to longitudinal, or panel, surveys which give repeated observations on the units of interest, whether they be individuals, households or organizations. Sample surveys are the mainstay of microeconometric research and some of the more popular datasets are summarized in Table 12.1.

In health economics, administrative datasets often prove more useful and reliable than social surveys. Administrative datasets include sources, such as tax records, reimbursement and claims databases, and population registers of births, deaths, cancer cases, HIV/AIDS cases, unemployment, etc. (see, e.g., Aakvik *et al.,* 2003; Aakvik *et al.,* 2005; Atella *et al.,* 2006; Black *et al.,* 2007; Chalkley and Tilley, 2006; Dano, 2005; Dranove *et al.,* 2003; Dusheiko *et al.,* 2004, 2006, 2007; Farsi and Ridder, 2006; Gravelle *et al.,* 2003; Ho, 2002; Lee and Jones, 2004, 2006; Marini *et al.,* 2008; Martin *et al.,* 2007; Propper *et al.,* 2002, 2004, 2005; Rice *et al.,* 2000; Seshamani and Gray, 2004). These datasets are collected primarily for administrative purposes and are made available to researchers for secondary analysis. Some countries allow comprehensive linkage of different sources of administrative data based on personal identification numbers (see, e.g., Black *et al.,* 2007). Administrative datasets are typically large, often with millions rather than thousands of

Table 12.1 Key datasets cited in the review

Acronym	*Full title, origin*	*Format*	*Homepage*
AddHealth	National Longitudinal Study of Adolescent Health, US	Panel survey	http://www.cpc.unc.edu/projects/addhealth/
AHEAD	Assets and Health Dynamics Among the Oldest-Old, US	Panel survey	http://hrsonline.isr.umich.edu/
BHPS	British Household Panel Survey, UK	Panel survey	http://www.data-archive. ac.uk/
BRFSS	Behavioral Risk Factor Surveillance System, US	Telephone survey	http://www.cdc.gov/brfss/
CHNS	China Health and Nutrition Surveys	Panel survey	http://www.cpc.unc.edu/projects/china
ECHP	European Community Household Panel, EC-15	Panel survey	http://forum.europa.eu.int/Public/irc/dsis/echpanel/home
ELSA	English Longitudinal Survey of Ageing	Panel survey	http://www.ifs.org.uk/elsa/
GSCF	Gansu Survey of Children and Families, China	Longitudinal multilevel survey	http://china.pop.upenn.edu/Gansu/intro.htm
GSOEP	German Socioeconomic Panel	Panel survey	http://www.diw.de/english/sop/
HALS	Health and Lifestyle Survey, GB	Panel survey	http://www.data-archive.ac.uk/
HCSUS	HIV Cost and Services Utilization Study, US	Panel survey	http://www.rand.org/health/projects/hcsus/
HCUP	Healthcare Cost and Utilization Project, US	Administrative	http://www.hcup-us.ahrq.gov/overview.jsp
HES	Hospital Episode Statistics, England and Wales	Administrative	http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/HospitalEpisodeStatistics/index.htm
HRS	Health and Retirement Survey, US	Panel survey	http://hrsonline.isr.umich.edu/
HSE	Health Survey for England, Welsh Health Survey, Scottish Health Survey	Repeated cross-sections	http://www.data-archive.ac.uk/
HSN	Historical Sample of the Netherlands	Longitudinal sample from census and registers	http://www.iisg.nl/~hsn/
LASA	Longitudinal Aging Study Amsterdam	Panel survey	http://www.lasa-vu.nl
LSMS	Living Standards Measurement Study, World Bank	Repeated cross-sections	http://www.worldbank.org/html/prdph/lsms/lsmshome.html

Continued

Table 12.1 Continued

Acronym	*Full title, origin*	*Format*	*Homepage*
MTR	Minnesota Twin Registry, US	Longitudinal Register	http://www.psych.umn.edu/psylabs/mtfs/default.htm
NDF	Natality Detail Files, US	Census of births	http://www.cdc.gov/nchs/products/elec_prods/subject/natality.htm
MEPS	National Medical Expenditure Panel Survey, US	Panel Survey	http://www.ahrq.gov/data/mepsweb.htm
NCDS	National Child Development Survey, UK	Cohort study	http://www.data-archive.ac.uk/
NHANES	National Health and Nutrition Examination Surveys, US	Repeated cross-sections	http://www.cdc.gov/nchs/nhanes.htm
NLSCY	National Longitudinal Survey of Children and Youth, Canada	Panel survey	http://www.statcan.ca/
NLSY	National Longitudinal Survey of Youth, US	Panel survey	http://www.bls.gov/nls/
NLTCS	National Long Term Care Survey, US	Panel survey	http://www.nltcs.aas.duke.edu/index.htm
NPHS	National Population Health Survey, Canada	Panel survey	http://www.statcan.ca/english/survey/household/health/health.htm
PSBH	Panel Study of Belgian Households	Panel survey	http://www.psbh.be/
PSID	Panel Study of Income Dynamics, US	Panel Survey	http://psidonline.isr.umich.edu/
RAND HIE	RAND Health Insurance Experiment, US	Panel, randomized experiment	http://www.icpsr.umich.edu/ICPSR/access/index.html
RLMS	Russian Longitudinal Monitoring Study	Panel survey	http://www.cpc.unc.edu/rlms/
SHARE	Survey of Health, Ageing and Retirement in Europe	Panel Survey	http://www.share-project.org/
SLID	Survey of Labour and Income Dynamics, Canada	Panel survey	http://www.statcan.ca/
US Census	United States census	Population census	http://www.census.gov/
WHS	World Health Survey, World Health Organization	Repeated cross-sections	http://www.who.int/healthinfo/survey/en/index.html

Continued

observations, and are comprehensive, often providing observations on a complete population rather than a random sample. They tend to be less prone to unit and item non-response than survey data and may give better coverage of hard-to-reach groups of the population and the socially disadvantaged. Also they tend to be less affected by reporting bias, but are still vulnerable to data input and coding errors. Given their primary purpose, administrative datasets are not designed by and for researchers. This means they may not contain all of the variables that are of interest to researchers, such as socioeconomic characteristics, and many different sources may have to be combined to produce a usable dataset. In some cases, sources of administrative data may be combined and made available with researchers in mind. For example, the Oxford Record Linkage Study (ORLS), used by Seshamani and Gray (2004), is a longitudinal dataset that links statistical abstracts for hospital inpatient and day cases to birth and death certificates for people living in the Oxford region of England. It provides 10 million records for over 5 million people between 1963 and 1999.

Dusheiko *et al.*'s (2004) study of the impact of practice budgets for GPs on hospital waiting times in the English National Health Service (NHS) provides an example of the complex and painstaking process that is often required to link administrative data. Information on waiting times was obtained from the Hospital Episode Statistics (HES) for 1997/98 to 2000/01. HES is an annual database of hospital inpatient activity, including day cases, with more than 10 million records per year. Dusheiko *et al.* extracted information on the waiting times for over 5 million finished consultant episodes and linked average waiting times to GP practices. Information on practice populations was obtained from the Primary Care Trust (PCT) database at the National Primary Care Research and Development Centre (NPCRDC: http://www.primary-care-db.org.uk). Practice characteristics, such as the GP's age and sex, qualifications, size of practice, etc., were obtained from the Prescription Pricing Authority, the Department of Health's Organisational Codes Service and their General Medical Statistics, along with the NPCRDC database. Patient characteristics for each practice were obtained from the 1991 Census and components of the Index of Multiple Deprivation, with these small area data mapped to GP practices. Finally, supply side factors, such as distances to hospitals, were obtained from the Department of Health's Allocation of Resources to English Areas (AREA) project.

Some of the key administrative datasets that have been used in health economics are summarized in Table 12.1.

12.3.1.1 Non-response and attrition

Non-response and attrition are a common feature of longitudinal survey data. Nicoletti and Peracchi (2005) list possible reasons for non-response: these include demographic events, such as death; movement out of the scope of the survey, such as institutionalization or emigration; refusal to respond at subsequent waves; absence of the person at the address, along with other types of non-contact. Jones *et al.* (2006) investigate health-related non-response in the first 11 waves of the BHPS and the full eight waves of the European Community Household Panel

(ECHP). They explore its consequences for dynamic models of the association between socioeconomic status and self-assessed health (SAH). Descriptive evidence shows that there is health-related non-response in the data, with those in very poor initial health more likely to drop out, and variable addition tests provide evidence of non-response bias in the panel data models of SAH. Nevertheless a comparison of estimates – based on the balanced sample, the unbalanced sample and corrected for non-response using inverse probability weights – shows that, on the whole, there are not substantive differences in the average partial effects of the variables of interest.

Inverse probablity weights are used to attempt to control for attrition: this works by estimating separate probit equations for whether an individual responds or does not respond at each of the waves of the panel. Then the inverse of the predicted probabilities of response from these models are used to weight the contributions to the log likelihood function in the pooled probit models for SAH. The rationale for this approach is that a type of individual who has a low probability of responding represents more individuals in the underlying population and therefore should be given a higher weight. The appropriateness of this approach relies on the assumption that non-response is ignorable conditional on the variables that are included in the models for non-response ("selection on observables"). If this assumption holds then inverse probability estimates give consistent estimates. The findings in Jones *et al.* (2006) and the earlier work by Contoyannis *et al.* (2004b) suggest that, while health-related non-response clearly exists, on the whole it does not appear to distort the magnitudes of the estimated dynamics of SAH and the relationship between socioeconomic status and SAH. Similar findings have been reported concerning the limited influence of non-response bias in models of income dynamics and various labor market outcomes and on measures of social exclusion, such as poverty rates and income inequality indices.

12.3.2 Health outcomes

12.3.2.1 Self-reported data

Self-assessed health is often included in general social surveys. For example, in the BHPS, SAH is an ordered categorical variable based on the question: "Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?" The validity of self-reported measures of health has caused considerable debate. As a self-reported subjective measure of health, SAH may be prone to measurement error. General evidence of non-random measurement error in self-reported health is reviewed in Currie and Madrian (1999) and Lindeboom (2006).

Self-assessed health is not the only source of concern with self-reported data. Baker *et al.* (2004) use careful record linkage to check for flaws in self-reported data on specific chronic conditions. Survey data from the Canadian National Population Health Survey for 1996–97 are linked to ICD-9 (International Classification of Diseases – ninth revision) codes for Ontario residents from administrative data on

utilization of services for the Ontario Health Insurance Plan (OHIP) for the survey year and the five previous years. Linear probability models are used to analyze the probabilities of false negatives and false positives in the self-reported data. This shows that reporting errors are associated with individual characteristics. The data reveal a large number of false negatives, although the probability declines for those with more recorded medical treatments, suggesting that this reflects undiagnosed conditions and lack of information among respondents. The number of false positives is much smaller, but there is some evidence of "justification bias": those not in work are more likely to report false positives for conditions, such as hypertension, ulcers and bronchitis.

A more favorable view of self-reported measures emerges in the work of Benitez-Silva *et al.* (2004). They take the relatively small sub-sample of respondents to the first three waves of the US Health and Retirement Survey (HRS) who had applied for disability benefit from the Social Security Administration (SSA) and compare their self-reported disability to the outcome of the SSA decision. In this case the SSA decision to award benefits is used as an objective indicator to assess the reliability of the self-reported data on limitations that prevent work. Conditional moment tests for whether self-reported disability is an unbiased indicator of the SSA decision suggest that a large fraction of this population report their health accurately. Unlike Baker *et al.* (2004), this study relies on information collected within the HRS rather than matching the survey data with administrative records. McGarry (2004) adopts another strategy to get around the problem of "justification bias." Rather than using data on actual retirement, she uses information from the HRS on the subjective expected probability of retirement by age 62, which is collected while people are still in work. Using this measure she finds strong effects of health on expected retirement age.

It is sometimes argued that the mapping of health into SAH categories may vary with respondent characteristics. This source of measurement error has been termed "state-dependent reporting bias" (Kerkhofs and Lindeboom, 1995), "scale of reference bias" (Groot, 2000) and "response category cut-point shift" (Murray *et al.* 2001; Sadana *et al.* 2000). Regression analysis of SAH is often done by specifying an ordered probability model, such as the ordered probit or logit. Then the symptoms of measurement error can be captured by making the cut-points dependent on some or all of the exogenous variables used in the model and estimating a generalized ordered model. This requires strong *a priori* restrictions on which variables affect health and which affect reporting in order to separately identify the influence of variables on latent health and on measurement error. Attempts to surmount this fundamental identification problem include modeling the reporting bias based on more "objective" indicators of true health (Kerkhofs and Lindeboom, 1995; Lindeboom and Van Doorslaer, 2004) and the use of "vignettes" to fix the scale (Das and Hammer, 2005; Murray *et al.,* 2001). Lindeboom and Van Doorslaer (2004) analyze SAH in the Canadian National Population Health Survey and use the McMaster Health Utility Index (HUI-3) as their objective measure of health. They find evidence of reporting bias with respect to age and gender, but not for income, education or linguistic group.

A similar identification strategy to Lindeboom and Van Doorslaer (2004), that relies on objective measures capturing all of the genuine variation in health, is adopted by Etilé and Milcent (2006), who estimate generalized ordered probit models with a four-category measure of SAH. They construct synthetic measures of objective health from a latent class analysis of a set of self-reported indicators, such as activities of daily living (ADLs) and body mass index (BMI). The latent class analysis is used to condense the sample into six classes and indicators for these classes are included in the ordered probit model. This works by constructing classes such that the underlying health indicators are independent of each other, conditional on class membership. Latent class models are presented as an alternative to the grade of membership approach that has been used in some earlier work (see, e.g., Lindeboom *et al.,* 2002). Estimates on a sample of 2,956 individuals aged under 65 from the French Enquête Permanente sur les Conditions de Vie des Menages (EPCV) survey show evidence of reporting bias and, unlike Lindeboom and Van Doorslaer (2004), there is evidence that this is related to income. Etilé and Milcent's (2006) findings suggest a concave relationship between health and income in terms of health production and convexity with respect to reporting, with overoptimism among the rich and overpessimism among the poor. They conclude that the problems of reporting bias can be minimized by collapsing the four-point scale into a binary measure of poor health.

Jurges (2007) focuses on cross-country differences in reporting of SAH as measured for those aged over 50 in the ten countries covered by the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). Generalized ordered probit models are used to regress SAH on a set of objective measures, such as grip strength, walking speed and BMI, to get a set of disability weights. The average thresholds across the SHARE countries are then used to reclassify the reported data, assuming that the disability weights are constant across countries. The variation in SAH is decomposed into the component that is explained by the objective measures and the component attributed to reporting bias. The findings suggest that those in the Danish and Swedish samples overrate their health, while those in Germany underrate their health. For Austria and Greece there is little bias.

12.3.2.2 Anthropometric measures

Anthropometric measures have long played a role in studies of developing countries, especially those focused on child health issues. With the growing problem of adult and childhood obesity in more affluent nations, they are increasingly being used in that context as well. Typical anthropometric measures are height and weight, which may be self-reported or measured by a professional; infant length for children aged under two; demi-span, which is based on the length of an outstretched arm and is used among older populations who may have difficulties standing straight; the waist-to-hip ratio; and BMI, which is the most commonly used indicator of obesity. BMI is calculated as weight in kilograms divided by height in meters squared, with a BMI of 30 or greater indicating obesity and 25–30 indicating overweight. The confounding effect of levels of muscle development means

that measures of body fat are sometimes used instead of BMI. The height-for-age z-score standardizes a child's measured height using the median (or mean) and standard deviation for children of the same age and sex from a reference population, such as the US National Center for Health Statistics reference population of well-nourished American children (e.g., Duflo, 2000). Height is also compared to a standard distribution to construct measures of stunting (e.g., Gertler, 2004). For example, Chen and Zhou (2007) use height to measure the long-term health consequences of childhood exposure to the 1959–61 famine in China, which is estimated to have caused 15–30 million excess deaths. They adopt a difference-in-differences approach that exploits regional differences in exposure to the famine.

Anthropometric measures may also play a role as biomarkers (discussed in more detail below). For example, height and weight can be used as predictors of mortality, stroke and cardiovascular disease and the waist-to-hip ratio is a predictor for hypertension, late-onset diabetes, cardiovascular disease, stroke and some forms of cancer.

12.3.2.3 Biomarkers

Biological markers, or biomarkers, are likely to play an increasing role in future research in health economics as they become incorporated into an increasing range of datasets, including longitudinal datasets, such as the US Health and Retirement Survey (HRS), the English Longitudinal Survey of Ageing (ELSA) and the planned UK Longitudinal Household Study. This trend is likely to be enhanced by the availability of DNA information and genetic screening, which provide greater potential to control for individual heterogeneity. Biomarkers are biological or physiological measures that indicate the presence of a disease or the propensity to develop a disease. They can be used to identify risk factors and as objective measures of health that avoid contamination by reporting bias (see, e.g., Adda and Cornaglia, 2006; Banks *et al.*, 2006; Currie *et al.*, 2007).

Biomarkers for cardiovascular disease include elevated blood pressure and variability in the heart rate. Metabolic biomarkers include serum HDL (high-density lipoprotein) and total cholesterol and triglycerides, which are predictors of heart disease; fibrinogen, which is linked to blood clotting and the risk of heart disease; and glycated haemoglobin, which is a proxy indicator for diabetes. Biomarkers linked to the immune system include interleukin-6 (IL-6), which is a predictor of Alzheimer's disease, arthritis, diabetes and osteoporosis; C-reactive protein (CRP), which indicates lupus, pneumonia, rheumatoid arthritis, rheumatic fever and tuberculosis; ferritin and hemoglobin, which indicate iron deficiency; serum retinol, which indicates vitamin A deficiency. Biomarkers linked to hormonal indicators of stress (HPA axis) include cortisol, adrenocorticotropic hormone (ACTH) and dehydroepiandrosterone-sulphate (DHEA-S). Biomarkers linked to the sympathetic nervous system include norepiphrenine, which is associated with longevity, and epiphrenine (adrenaline), which is linked to cognitive decline and longevity. Biomarkers may also be used as objective indicators of physical functioning and

health limitations; these include lung function tests, such as forced expiratory volume (FEV) and forced vital capacity (FVC), as measured by a spirometer, and grip strength, measured by a gripometer.

Datasets which contain biomarkers and that have been used in economic research include ELSA, the UK Health and Lifestyle Survey (HALS), the US Health and Retirement Survey (HRS), the Health Survey of England (HSE), the UK National Child Development Study (NCDS) and the more recent cohort studies, the US National Health and Nutrition Examination Surveys (NHANES), SHARE, and the Whitehall Study of English civil servants. To take ELSA as an example: participants in the study are visited by a registered nurse who takes measurements of blood pressure; lung function; height, weight and the waist-to-hip ratio; grip strength, as a measure of upper body strength; a measure of lower body strength, based on standing up from a chair without using the arms; a saliva sample, that is used to measure cortisol, which is a marker for stress; and a blood sample, which is used to test total cholesterol, HDL cholesterol, fibrinogen, CRP, ferritin, glycated haemoglobin and haemoglobin. The blood samples from ELSA and from wave 7 of the NCDS will allow DNA to be extracted. DNA can also be collected using mouth or cheek swabs (as in the US AddHealth Survey).

Banks *et al.* (2006) study the socioeconomic gradient in health in the UK and the US and they compare both self-reported outcomes and objective outcomes based on biomarkers. To do this they use the ELSA data for the UK and the HRS and NHANES data for the US. The self-reported measures include the general question on SAH as well as self-reported indicators of chronic conditions, such as diabetes, hypertension and cancer. The biomarkers are glycosated hemoglobin levels above 6.5%, as a marker for diabetes; systolic blood pressure over 140 mm Hg, and diastolic blood pressure over 90 mm Hg, as a measure of hypertension; CRP greater than 3 mg/L, as a marker of high risk of arteriosclerosis; fibrinogen over 400 mg/dl, as a marker for cardiovascular disease, and HDL cholesterol over 40 mg/dl, as an indicator of reduced risk of coronary heart disease. Banks *et al.* (2006) find that, on average, respondents in the US reported better SAH, but the opposite holds true for the biomarkers. Their results show a strong socioeconomic gradient in self-assessed health, self-reported diseases and in the biomarkers. The gradient appears strongest for the biomarkers. Comparing the self-reported data with the biomarkers allows a measure of the socioeconomic gradient in undiagnosed cases. A gradient is apparent for diabetes, but not for hypertension.

A novel feature of Adda and Cornaglia (2006) is the use of biomarkers, in this case cotinine, within an economic study of smoking. Cotinine is a metabolite of nicotine and can be used as a biomarker for levels of tobacco consumption that is not contaminated by problems of measurement error, such as recall bias and deliberate deception, that may affect self-reported consumption. The study shows that smokers engage in compensatory behavior, increasing their intensity of smoking and offsetting the impact of tobacco tax increases. Data on cotinine is collected from saliva samples as part of the repeated cross-section data in the NHANES for 1999–2000. Evidence based on the biomarker is contrasted with

self-reported consumption. Cotinine has another advantage in studies of smoking and health as it provides a way of measuring passive smoking, especially among children.

12.3.3 Modeling costs and expenditure

Individual-level data on medical expenditures and costs of treatment is typically distinguished by a spike at zero, if there are non-users in the data, and a strongly skewed distribution with heavy tails. This kind of data is most often used in two areas of application: risk adjustment and cost-effectiveness analysis. In risk adjustment the emphasis is on predicting the treatment costs for particular types of patient, often with very large datasets. Cost-effectiveness analyses tend to work with smaller datasets and the scope for parametric modeling may be more limited (Briggs *et al.,* 2005). In the context of clinical trials, attention has focused on methods to deal with censoring of cost data due to limited follow-up (e.g., Baser *et al.,* 2006; Raikou and McGuire, 2004, 2006).

The presence of a substantial proportion of zeros in the data has typically been handled by using a two-part model, which distinguishes between a binary indicator, used to model the probability of any costs, and a conditional regression model for the positive costs. OLS applied to the level of costs (y) can perform poorly, due to the high degree of skewness and excess kurtosis, and the positive observations are often transformed prior to estimation. The most common transformation is the logarithm of y, although the square root is sometimes used as well. As the policy interest typically focuses on modeling costs on the original scale, the regression results have to be retransformed back to that scale. This weakens the case for working with transformed data and, in particular, problems arise with the retransformation if there is heteroskedasticity in the data on the transformed scale (Manning, 1998; Manning and Mullahy, 2001; Mullahy, 1998). Ai and Norton (2000) provide standard errors for the retransformed estimates when there is heteroskedasicity.

More recently, attention has shifted to other estimators. Basu *et al.* (2004) compare log-transformed models to the Cox proportional hazard model. Gilleskie and Mroz (2004) propose a flexible approach that divides the data into discrete intervals then applies discrete hazard models, implemented as sequential logits. Conway and Deb (2005) use a finite mixture model. Cooper *et al.* (2007) use hierarchical regressions implemented using Bayesian Markov chain Monte Carlo (MCMC) estimation. But the dominant approach in the recent literature has been the use of generalized linear models (GLMs) (e.g., Buntin and Zaslavsky, 2004; Manning, 2006; Manning *et al.,* 2005; Manning and Mullahy, 2001). The GLM specifies a link function for the relationship between the conditional mean, $\mu = E(y|x)$, and a linear function of the covariates and specifies the form of the conditional variance, $V(y|x)$, usually assuming that it can be specified as a simple function of the mean. The models are estimated using a quasi-likelihood approach derived from the quasi-score or "estimating equations." The most popular specification of the GLM for costs has been the log-link with a gamma error (Blough *et al.* 1999; Manning *et al.,* 2005; Manning and Mullahy, 2001, 2005). Cantoni and Ronchetti (2006) propose a

robust variant of GLM that is less sensitive to outliers. In response to the problem of selecting the appropriate link and variance functions, Basu and Rathouz (2005) suggest a flexible semiparametric extension of the GLM model. Their model incorporates a Box–Cox transformation into the link function which includes the log-link as a special case along with other power functions of y. The model, which is labeled the extended estimating equations (EEE) approach, also allows for flexible specifications of the variance using the power variance and quadratic variance families to nest common distributions, such as the Poisson, gamma, inverse Gaussian and negative binomial. Basu *et al.* (2006) apply the EEE method to claims data on the incremental costs associated with heart failure.

12.4 Methods for dealing with unobserved heterogeneity and dependence

12.4.1 Deviations and conditional estimates

Consider a linear panel data regression model with repeated measurements ($t = 1, \ldots, T_i$) for a sample of n individuals ($i = 1, \ldots, n$):

$$y_{it} = x_{it}' \beta + u_i + \varepsilon_{it} . \tag{12.16}$$

Correlation between the unobservable individual effects (u) and the regressors (x) will lead to an omitted variable bias and inconsistent estimates of the βs. The individual effects can be swept from the equation by transforming variables into deviations from their within-group means, or by using orthogonal deviations, based on the mean of the future values of the variables. Applying least squares to the mean deviations gives the covariance or within-groups estimator of β. Similarly, the model could be estimated in first differences to eliminate the individual effects. Identification of β rests on there being sufficient variation over time so the estimators may perform poorly when there is insufficient variation.

Many of the outcomes used in health economics are binary or ordered categorical measures, such as SAH. Fixed effects panel data methods, that allow for a correlation between the individual effect and the regressors of the model, are not, however, readily available for categorical data due to the incidental parameter problem, which means that the individual effect cannot in general be swept out of the model by taking deviations. For binary data the problem can be surmounted by using the conditional fixed effects logit, which uses a sufficient statistic to eliminate the individual effect from the log-likelihood function (Chamberlain, 1980). In the case of the logistic regression, the within-individual sum of y_{it} is a sufficient statistic and conditional maximum likelihood (ML) estimates are consistent. Although the conditional logit provides consistent parameter estimates, the approach has practical drawbacks for the researcher. First, by only using observations that have within-individual variation in the outcome and in the regressors, the method often leads to a substantial reduction in sample size. Second, it is hard to calculate partial effects of a variable of interest due to the inherent lack of information on

the distribution of the individual heterogeneity, which is conditioned out of the model.

The conditional logit can be applied to ordered data by choosing a particular threshold value and collapsing the data into a binary measure. A recent extension of Chamberlain's model, the conditional ordered fixed effects logit, proposed by Ferrer-i-Carbonel and Frijters (2004), and applied to data on SAH by Frijters *et al.* (2005), suggests a method to reduce the drastic loss in the number of observations by identifying individual-specific threshold values to collapse the ordered dependent variable into a binary format. Das and van Soest (1999) combine adjacent categories so that the dependent variable is summarized as a binary variable, and then use conditional logits. They repeat this for all the possible combinations of adjacent categories to get a set of estimates of the parameters of interest. They then define a linear combination of these estimates, with the optimal weighting matrix used to compute the final estimate obtained from a minimum distance approach. Ferrer-i-Carbonel and Frijters (2004) also propose an estimator that collapses the ordered variable into a binary format, but they use an individual specific threshold value. To find this individual threshold, the authors maximize a weighted sum of log-likelihood functions, similar to Das and van Soest (1999), subject to the constraint that the sum of squared weights across all possible threshold values across all individuals must be equal to the number of individuals in the sample. The threshold is selected for which the analytical expected Hessian is minimized. However, this formulation of the estimator is highly computation intensive, a fact which makes its wider application less attractive. In a simplification of this estimator, one can simply use the within-individual means as a cut-off criterion.

12.4.1.1 Dynamic models

Even for linear models the within-groups estimator breaks down in dynamic models, such as:

$$y_{it} = \alpha y_{it-1} + u_i + \varepsilon_{it}\,. \tag{12.17}$$

This is because the group mean is a function of $\varepsilon_{it}$and ε_{it-1}. An alternative is to use the differenced equation:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \varepsilon_{it}, \tag{12.18}$$

in which case both y_{it-2} and Δy_{it-2} are valid instruments for Δy_{it-1} as long as the error term (ε_{it}) does not exhibit autocorrelation. Arellano and Bond (1991) proposed generalized method of moments (GMM) estimators for dynamic panel data models: linear models that can include leads and lags of the dependent variable as well as a fixed effect. Instruments are created within the model by first taking differences of the equation to sweep out the individual effect and then using lagged levels or differences of the regressors as instruments.

Bover and Arellano (1997) extend the use of GMM to dynamic specifications for categorical and limited dependent variable models, where it is not possible to take first differences or orthogonal deviations as the latent variable y^* is unobserved.

One of the advantages of using panel data is the possibility of accounting for the correlation amongst the effects and the explanatory variables. To allow for this correlation, Chamberlain (1984) suggested using a random effects approach and specifying a distribution for the individual effects conditional on the values of the explanatory variables at each wave of the panel. This specification may contain polynomial terms and interactions in the *x*s as well. Combining this with assumptions about the conditional expectation of the initial and final values of the latent variable allows the dynamic model to be solved out to give linear reduced forms for the latent variables at each wave of the panel. Estimates of the reduced forms will be sensitive to assumptions about the distribution of the error terms, the linearity of the expected value, and the conditional mean independence assumption. However, these hypotheses can be checked by specification tests at the level of the reduced form, which is easier to do than testing the dynamic specification. At the second stage, on the basis of the reduced form coefficients, the parameters of the underlying dynamic structural model can be derived using various estimators. The simplest is to apply the within-groups transformation to the dynamic model after replacing the latent variables by their predicted counterparts (Bover and Arellano, 1997). This two-step within-groups procedure is simple to apply, but provides inefficient parameter estimates. Chamberlain (1984) proposed a fully efficient minimum distance (MD) estimator. Instead of using Chamberlain's approach, Bover and Arellano (1997) propose a three-step within-groups GMM, which also facilitates tests of the over identifying restrictions.

12.4.2 Numerical integration and classical simulation-based inference

In panel data specifications, unobserved heterogeneity is often modeled as a random effect and "integrated out" of the log-likelihood function. Monte Carlo simulation techniques can be used to deal with the computational intractability of nonlinear models, such as panel and multinomial probit models. Popular methods of simulation-based inference include classical maximum simulated likelihood (MSL) estimation, and Bayesian MCMC estimation. This section introduces the classical approach (for a review of the methods and applications to health economics, see Contoyannis *et al.,* 2004a).

Numerical integration by quadrature works well with low dimensions, but computational problems arise with higher dimensions. Instead, Monte Carlo simulation can be used to approximate integrals that are numerically intractable. This includes numerous models derived from the multivariate normal distribution. Simulation approaches use pseudo-random draws of the evaluation points and computational cost rises less rapidly than with quadrature.

The principle behind simulation-based estimation is to replace a population value by a sample analogue. This means that laws of large numbers and central limit theorems can be used to derive the statistical properties of the estimators. The basic problem is to evaluate an integral of the form:

$$\int_u [h(u)]dF(u) = E_u [h(u)], \tag{12.19}$$

where $h(u)$ is a nonlinear function of the random vector u, which has a multivariate density $f(u)$and distribution function $F(u)$. This kind of expression arises in panel data models with random effects specifications and with autocorrelated errors and in multiple equation models with correlated unobservables. The integral can be approximated using draws from $f(u), u_r, r = 1, \ldots, R$, such that:

$$\int [h(u)]dF(u) \approx \frac{1}{R}\sum_{r=1}^{R}\left[h(u_r)\right]. \tag{12.20}$$

MSL is a simple extension of classical ML estimation and is useful in many cases where the log-likelihood function involves high dimensional integrals. The idea is to replace individual contributions to the sample likelihood function (L_i) with an average over R random draws:

$$l_i = \frac{1}{R}\sum_{r=1}^{R}\left[l(u_{ir})\right], \tag{12.21}$$

where $l(u_{ir})$ is an unbiased simulator of L_i.The MSL estimates are the parameter values that maximize:

$$Lnl_{=}\sum_{i=1}^{n}\left[Lnl_i\right]. \tag{12.22}$$

For likelihoods derived from the multivariate normal the Geweke–Hajivassilou–Keane (GHK) simulator is often used. In practice, Halton sequences or antithetics can be used to reduce the variance of the simulator (see Contoyannis *et al.,* 2004a, for details).

12.4.3 Bayesian MCMC

In Bayesian analysis a prior density of the parameters of interest, say $\pi(\theta)$, is updated using information from sample data. Given a specified sample likelihood for the observed data, $l(y|\theta)$, the posterior density of θ is given by Bayes' theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)l(y|\theta)}{\pi(y)}, \tag{12.23}$$

where:

$$\pi(y) = \int \pi(\theta)l(y|\theta)d\theta. \tag{12.24}$$

The scaling factor $\pi(y)$ is known as the predictive likelihood and is used to compare models. It determines the probability that the specified model is correct. The posterior density $\pi(\theta|y)$ reflects updated beliefs about the parameters. Given the posterior distribution, a 95% credible interval can be constructed that contains the

true parameter with probability equal to 95%. Point estimates for the parameters can be computed using the posterior mean:

$$E(\theta|y) = \int \theta\pi(\theta|y)d\theta. \tag{12.25}$$

Bayesian estimates can be difficult to compute directly. For instance, the posterior mean is an integral with dimension equal to the number of parameters in the model. In order to overcome the difficulties in obtaining the characteristics of the posterior density, MCMC simulation methods are often used. The methods provide a sample from the posterior distribution and posterior moments and credible intervals are obtained from this sample (see Contoyannis *et al.,* 2004a, for details).

Bayesian MCMC simulation is built on the Gibbs sampling algorithm. To implement Gibbs sampling the vector of parameters is sub-divided into groups. For example, with two groups let $\theta = (\theta_1, \theta_2)$. Then, a draw from the joint distribution $\pi(\theta_1, \theta_2)$ can be obtained in two steps: first, draw θ_1 from the marginal distribution $\pi(\theta_1)$; then draw θ_2 from the conditional distribution $\pi(\theta_2|\theta_1)$. However, in many situations it is possible to sample from the conditional distribution, but it is not obvious how to sample from the marginal. The Gibbs sampling algorithm solves this problem by sampling iteratively from the full set of conditional distributions. Even though the Gibbs sampling algorithm never actually draws from the marginal, after a sufficiently large number of iterations the draws can be regarded as a sample from the joint distribution. There are situations in which it is not possible to sample from a conditional density, and hence Gibbs sampling cannot be applied directly. In these situations, Gibbs sampling can be combined with a so-called Metropolis step as part of a Metropolis–Hastings algorithm. In the Metropolis step, values for the parameters are drawn from an arbitrary density, and accepted or rejected with some probability. An attraction of MCMC is that latent or missing data can be treated as parameters to be estimated. Although this data augmentation method introduces many more parameters into the model, the conditional densities often belong to well-known families and there are simple methods to sample from them. This makes the use of MCMC especially convenient in nonlinear models, where the latent variables (y^*) can be treated as parameters to be estimated. Once the y^*s have been simulated, the estimation step involves the estimation of normal-linear models for y^*.

12.4.4 Finite mixture models

12.4.4.1 Latent class models

Recently the latent class framework has been used in models for health care utilization with individual data. Deb and Trivedi (2002) note that this framework "provides a natural representation of the individuals in a finite number of latent classes, that can be regarded as types or groups."[2] The segmentation can represent individual unobserved characteristics, such as unmeasured health status. The latent class (or finite mixture) framework offers a representation of heterogeneity, where individuals are drawn from a finite number of latent classes. For example, Conway and Deb (2005) show that allowing for heterogeneity between

"normal" and "complicated" pregnancies leads to evidence that early prenatal care is effective: on average, bringing the onset of prenatal care forward by one week increases birth weights by 30–60 g in normal pregnancies. Using a finite mixture model to capture the bimodality of the distribution of birth weights counteracts evidence from the standard 2SLS approach that the effects of prenatal care are weak or nonexistent. Estimates of the mixture model use observational data from the 1988 US National Maternal and Infant Health Survey, and the empirical findings are augmented by simulation results that show that the conventional findings could be attributable to the existence of a relatively small proportion (10–15%) of "complicated" pregnancies in the population.

To specify a finite mixture model, consider a vector of outcomes y_i that are observed for individual i: these may be repeated observations in a panel data model or related outcomes in a multiple equation model and they are linked by common unobservable heterogeneity. Then assume that each individual belongs to one of a set of latent classes $j = 1, \ldots, C$, and that individuals are heterogeneous across classes. Conditional on the observed covariates, there is homogeneity within a given class j. Given the class that individual i belongs to, the outcomes have a joint density $f_j(y_i|x_i;\theta_j)$ where the θ_j are vectors of parameters that are specific to each class. The probability of belonging to class j is π_{ij}, where $0 < \pi_{ij} < 1$ and $\sum_{j=1}^{C} \pi_{ij} = 1$. Unconditionally on the latent class the individual belongs to, the joint density of y_i is given by:

$$f\left(y_i|x_i;\pi_{i1},\ldots,\pi_{iC};\theta_1,\ldots,\theta_C\right) = \sum_{j=1}^{C} \pi_{ij} f_j\left(y_i|x_i;\theta_j\right). \tag{12.26}$$

The discrete distribution of the heterogeneity has C mass points and the πs need to be estimated along with the θ_js.

In many empirical applications of finite mixture models the class membership probabilities are treated as fixed parameters $\pi_{ij} = \pi_j, j = 1, \ldots, C$, (e.g., Atella *et al.*, 2004; Bago d'Uva, 2006; Deb, 2001; Deb and Holmes, 2000; Deb and Trivedi, 1997, 2002; Jiménez-Martin *et al.*, 2002). A more general approach is to parameterize the heterogeneity as a function of individual characteristics. To implement this approach in the case of the latent class model, class membership can be modeled as a multinomial logit (as in, e.g., Clark *et al.*, 2005; Etilé, 2006):

$$\pi_{ij} = \frac{\exp\left(z_i'\gamma_j\right)}{\sum_{k=1}^{C} \exp\left(z_i'\gamma_k\right)}, \; j = 1, \ldots, C, \tag{12.27}$$

with the normalization $\gamma_C = 0$. This approach specifies the determinants of class membership. In a panel data context, this parameterization provides a way to account for the possibility that the observed regressors may be correlated with the individual heterogeneity. Letting $z_i = \bar{x}_i$ be the average over the observed panel of the observations on the covariates is in line with what has been done in recent studies to allow for the correlation between covariates and random effects,

following the suggestion of Mundlak (1978) and Chamberlain (1984). The vectors of parameters $\theta_1, \ldots, \theta_C, \gamma_1, \ldots, \gamma_{C-1}$ are estimated jointly by ML.

After estimating the model, it is possible to calculate the posterior probability that each individual belongs to a given class. The posterior probability of membership of class j depends on the relative contribution of that class to the individual's likelihood function. This is given by:

$$P[i \in j] = \frac{\pi_{ij} f_j\left(y_i | x_i; \theta_j\right)}{\sum_{k=1}^{C} \pi_{ik} f_k\left(y_i | x_i; \theta_k\right)}. \tag{12.28}$$

Each individual can then be assigned to the class that has the highest posterior probability for them.

12.4.4.2 Finite density estimators and discrete factor models

In LCMs class membership has a discrete distribution with a fixed number of mass points. The models are very flexible in that all of the parameters can be allowed to vary across classes. A special case of this general model assumes that the slope coefficients are fixed across classes and that only the intercepts vary. This case is widely used to model unobserved heterogeneity, without imposing parametric assumptions on the distribution of the heterogeneity. The specification has a dual interpretation: the population may truly fall into a discrete set of classes or types or, alternatively, the mass points can be viewed as an approximation of some underlying continuous distribution – the finite density estimator. The finite density estimator was introduced into the econometrics literature by Heckman and Singer (1984) in the context of hazard models (see, e.g., Van Ours, 2004, 2006). More recently, the finite density estimator has been widely used in multiple equation models where a common factor structure is assumed, as in equation (12.9) above. This is often called the discrete factor model (DFM).

Since the DFM includes an intercept for each equation, the location of the distribution of the common factor η is arbitrary; also the scale of η is arbitrary and undetermined (Mroz, 1999). Therefore, identification of the DFM requires some normalizations. The existing literature on the DFM offers a range of equivalent strategies to identify the additional parameters of the discrete distribution by fixing the scale and the location of the distribution. If both are fixed, one of the factor loadings is set to 1 and either one of the η_j is set to 0 (see Mroz, 1999) or the mean of the discrete distribution is restricted to be 0, so that one of the η_j can be expressed as a function of the others (Kan *et al.,* 2003). If only the location is fixed, the first and last mass points are set to 0 and 1 (this strategy is used by Mroz, 1999, when $C > 2$). Other applications also impose that the remaining mass points follow a logistic distribution such that $\eta_k \in (0,1)$ (see Mello *et al.*, 2002, Picone *et al.*, 2003b). The π_k can be parameterized using various distributions, such as the logistic, normal or the sine function, such that each π_k is between 0 and 1 and they sum to 1.

12.4.5 Copulas

The presence of common unobservables leads to multiple equation models and the need to specify multivariate distributions. But the menu of parametric forms available for bivariate and, more generally, multivariate distributions is limited. In many applications multivariate normality may be unappealing: for example, with heavily skewed and long-tailed data on costs of care or on quality adjusted life-years (QALYs) (Quinn, 2005) or for rare events. Copulas provide an alternative and are a method of constructing multivariate distributions from univariate marginal distributions (see Trivedi and Zimmer, 2005). A copula is a function that can be interpreted as a joint probability whose arguments are the univariate cumulative distribution functions (CDFs) of the marginal distributions. The fact that the CDF is used means that the marginal distributions are fixed and invariant to transformations of the random variable. The functional form selected for the copula uniquely determines the form of the dependence, independently of the functional forms of the marginal distributions. The attractions of copulas are that they are flexible – they can mix together marginal distributions of different types, whether they be continuous, integer valued counts or categorical; they allow for richer concepts of dependence than the standard linear measure, including measures of tail dependence; and they are computationally tractable and avoid the need for numerical integration or simulation.

A key result in the theory of copulas is Sklar's theorem, which shows that all multivariate distributions can be represented by a copula. So in the bivariate case, if two random variables have a joint distribution $F(x_1, x_2)$ and marginal distributions $F_1(x_1)$ and $F_2(x_2)$ then Sklar's theorem establishes that there exists a copula function $C(\cdot)$ such that:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)). \tag{12.29}$$

In practice the unique copula that characterizes the true joint distribution is unknown. So particular functional forms have to be selected and compared in terms of their goodness of fit. There is a long list of copulas to choose from. Common choices include the Frank copula and the Farlie–Gumbel–Morgenstern (FGM) copula, which is a first-order approximation of the Frank copula and is tractable to use in applied work (Prieger, 2002; Smith, 2003; Zimmer and Trivedi, 2006). The bivariate form of the Frank copula is:

$$C_\theta(u, v) = -\theta^{-1} \log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right), \tag{12.30}$$

where the parameter θ captures dependence. The bivariate form of the FGM copula is:

$$C_\theta(u, v) = uv(1 + \theta(1 - u)(1 - v)). \tag{12.31}$$

Other common choices include elliptical copulas, such as the Gaussian and Student's t, and the Clayton copulas. The Gaussian copula is an example of a copula derived by the method of inversion and takes the form:

$$C_\theta(u, v) = \Phi_2(\Phi_1^{-1}(u), \Phi_1^{-1}(v)), \tag{12.32}$$

where Φ_2 and Φ_1 denote bivariate and univariate standard normal distribution functions.

12.5 Models for longitudinal data

12.5.1 Applications of linear models

12.5.1.1 Models for longitudinal and spatial panels

In Lindeboom *et al.* (2002) models for cognitive status and emotional well-being are estimated using linear fixed effects specifications, estimated in first differences. The data are taken from three waves of the Dutch Longitudinal Ageing Study Amsterdam (LASA) and exhibit considerable attrition. Lindeboom *et al.* note that the fixed effects specification is robust to selection associated with time invariant unobservables, but they also include indicators of patterns of response in their model. They find a large impact of life events, such as bereavement, on mental health among the elderly. Carey (2000) estimates the impact of length of stay (LOS) on total hospital costs using a panel of 2,792 US hospitals for the period 1987–92. The individual effect in these models captures factors, such as the quality of care provided in each hospital, which is likely to be correlated with both costs and LOS. To allow for these correlated effects, she uses the Chamberlain minimum distance estimator and finds that the elasticity of total costs with respect to LOS is low.

The rapid growth in dental care expenditure in Taiwan, after the inauguration of national health insurance (NHI) in 1995, led the government to reform the payment system and introduce global budgeting for outpatient dental care in July 1998. In response to the introduction of global budgets, dentists might alter their supply behavior, changing the number of visits, the amount of expenditure, and the type of services provided. Lee and Jones (2004) develop two-way fixed effects models to estimate these effects using panel data constructed from outpatient dental care expenditures claims from the Taiwanese National Health Insurance system. The availability of a long panel, with up to 48 monthly observations, allows them to estimate a policy effect for each dentist in the panel, using within-dentist variation and effectively treating each dentist as their own control group. The individual effects are an important component in the panel data model to investigate dentists' responses to the introduction of global budgeting. The magnitude of the dentist effects measures individual heterogeneity in activities that could not be captured by observable factors in the regression. This allows them to estimate individual-specific responses to the payment reform as well as calculating the average policy effect. They model the individual response to the policy changes by estimating the differences of individual fixed effects between two separate models: pre- and post-global budget. This policy effect can be interpreted as, holding other observable variables constant, the extent to which each individual dentist's activity changed after the introduction of global budgeting. They use OLS to analyze the factors influencing variation in the policy effects across different dentists. These factors include information on the dentist's demographic characteristics, such as age and gender; the type and ownership of affiliated medical institutions (public or

private and hospital or clinic); environmental characteristics, such as the dentist-population ratio and annual household income; and whether the dentist practices in a deprived area. The overall effect of global budgets is to constrain costs, but there is evidence of a change in the mix of services. Male and younger dentists have higher policy effects than female and older dentists. Global budgets favor dentists in deprived areas and there is some evidence of increases in the expenditure per visit and the volume of composite resin fillings.

Applications of linear models are not confined to longitudinal datasets. Moscone *et al.* (2007) take a spatial panel approach to data on mental health expenditure by local authorities in England. They compare specifications that allow for a spatial autoregressive process in the error term; a random effects model with spatially lagged dependent variables; and a random effects model with spatial autocorrelation. The random effects model with lagged dependent variables proves to be the preferred specification.

12.5.1.2 Dynamic panel data models: GMM estimators

Brown *et al.* (2005) apply dynamic panel data models to assess the impact of HMOs on the supply of doctors at the local level. They construct a panel of Californian counties for the years 1988–98 using data from the American Medical Association and specify reduced-form models for the supply of doctors per 100,000 inhabitants for both specialists and primary care. First differencing is used to deal with omitted variables and dynamics are included to allow for inertia in supply responses. The models are estimated by the systems version of the Arellano–Bond estimator (Arellano and Bond, 1991; Arellano and Bover, 1995; Blundell and Bond, 1998), with both one-step and two-step estimates of the standard errors. The internal instruments, drawn from lagged variables, are augmented by some external instruments. The tests for autocorrelation are consistent with the assumptions of the Arellano–Bond model: there is evidence of first order autocorrelation in the residuals, but not of higher order autocorrelation. The results show that the supply of specialists is responsive to changes in the relative market penetration of HMOs, but the supply of primary care doctors is not.

The Arellano–Bond approach is also adopted by Tamm *et al.* (2007) to estimate price effects on insurers' market shares. They use an unbalanced panel of insurers, who were active in the German social health insurance system between January 2001 and April 2004. This issue is important as price sensitivity among consumers is a precursor for the success of reforms based on the notion of managed competition. They adopt a dynamic panel data model for the logarithm of each insurer's market share. This aggregate model is motivated by an individual-level multinomial logit model for the choice of insurer and dynamics are introduced to capture the fact that only a fraction of consumers will switch companies during a given year. The model is estimated using the Arellano–Bond estimator and the systems-GMM estimator. The standard Arellano–Bond GMM estimator, which relies on lagged levels to instrument lagged differences, may perform poorly. This suggests using the systems approach, but this approach requires stronger restrictions on the initial conditions: the first-period error term and the first differences of the regressors

have to be uncorrelated with the individual effect. In this case the orthogonality conditions for the systems approach are rejected. The hypothesis of a unit root is not rejected and models are also estimated for differences in log-market share as a function of levels of the regressors. The estimated short-run price elasticities are small but, due to the dynamics, the long-run impact of prices is substantial.

Other applications of the GMM and systems-GMM approaches include Baltagi *et al.* (2005), who estimate dynamic equations for the log of hours worked by Norwegian doctors, using panel data from the personnel register of the Norwegian Association of Local and Regional Authorities. Clark and Etilé (2002) use seven waves of BHPS to estimate dynamic models for cigarette consumption. This creates some problems when the data are first differenced as there is considerable heaping of the self-reported data around focal values, such as 20 cigarettes per day. Health shocks are shown to influence levels of smoking. Hauck and Rice (2004) use 11 waves of the BHPS to estimate models for mental health, measured by the GHQ-12 score. They compare static variance components models and dynamic GMM estimators and find greater persistence of mental health problems among those with lower socioeconomic status. Windmeijer *et al.* (2005) use the systems-GMM estimator in panel data models of the demand for outpatient visits by GP practices in England.

12.5.2 Applications with categorical outcomes

12.5.2.1 Pooled and random effects specifications

Contoyannis *et al.* (2003) consider the determinants of a binary indicator for functional limitations using seven waves (1991–97) of the BHPS. Their models allow for persistence in the observed outcomes due to state dependence (a direct effect of previous health status), unobservable individual effects (heterogeneity which is due to unobserved factors that are fixed over time) and persistence in the transitory error component. Allowing for persistence is important: a comparison of the observed outcomes with those predicted by a simple binomial model shows that persistence is substantial in the data. They estimate models for the repeatedly observed binary health indicator, with and without state dependence, using panel probit models. These are estimated by MSL using the GHK simulator with antithetic acceleration. They also implement a test for the existence of asymptotic bias due to simulation which is used to select the number of replications required for use in MSL.

In related work Contoyannis *et al.* (2004b) explore the dynamics of SAH in the BHPS. The variable of interest is an ordered measure and the BHPS reveals evidence of considerable persistence in individual's health status. As SAH is measured at each wave of the panel there are repeated measurements for a sample of individuals. SAH is modeled using a latent variable specification, which is estimated using pooled ordered probits (with robust inference) and random effects ordered probit models. The presence of lagged health is designed to capture state dependence, the influence of previous health history on current health. The error term is split into two components; the first captures time invariant individual heterogeneity; the second is a time varying idiosyncratic component.

In this kind of application it is quite likely that the unobserved individual effect will be correlated with the observed regressors, such as household income. To allow for this possibility Contoyannis *et al.* (2004b) parameterize the individual effect (Chamberlain, 1984; Mundlak, 1978; Wooldridge, 2005). This allows for correlation between the individual effects and the means of the regressors. In addition, because they are estimating dynamic models, they need to take account of the problem of initial conditions. It is well known that in dynamic specifications the individual effect will be correlated with the lagged dependent variable, which gives rise to what is known as the *initial conditions problem*, that an individual's health at the start of the panel is not randomly distributed and will reflect the individual's previous experience and be influenced by the unobservable individual heterogeneity. To deal with the initial conditions an attractively simple approach suggested by Wooldridge (2005) is used. This involves parameterizing the distribution of the individual effects as a linear function of initial health at the first wave of the panel and of the within-individual means of the regressors, and assuming that it has a conditional normal distribution. As long as the correlation between the individual effect and initial health and the regressors is captured by this equation it will control for the problem of correlated effects. Its ease of implementation stems from the fact that the equation for u_i can be substituted back into the main equation and the model can then be estimated as a pooled ordered probit or a random effects ordered probit using standard software to retrieve the parameters of interest. Contoyannis *et al.* (2004b) find that SAH is characterized by substantial positive state dependence and unobserved permanent heterogeneity. Including state dependence dramatically reduces the impact of individual heterogeneity. Conditioning on the initial period health outcomes and within-individual averages of the exogenous variables reduces the impact of heterogeneity and state dependence. Unobservable heterogeneity accounts for around 30% of the unexplained variation in health.

Similar dynamic panel probit models are used by Gannon (2005). In her case the outcome of interest is a binary measure of labor force participation, which is assumed to be a function of past labor force participation and health limitations. This gives dynamic panel probit models that are estimated in pooled and random effects versions using the Wooldridge (2005) approach to deal with the initial conditions problem. The models are estimated with the Living in Ireland Survey (LIS), which is the Irish component of the ECHP. Nolan (2007) also uses the LIS, but uses a dynamic random effects Poisson specification to model GP visits. She adopts the Wooldridge approach to model the initial conditions. In contrast, Arulampalam and Bhalotra (2006) use Heckman's approach to specify the initial conditions in a Markov model of infant deaths among Indian families.

12.5.2.2 GMM estimators

In Jones and Labeaga (2003) a panel of Spanish households is used to test the empirical formulation of the rational addiction model (Becker *et al.,* 1994). This dataset raises problems of measurement errors, censoring, and unobservable heterogeneity. Jones and Labeaga (2003) use sample separation information to exclude those

households who never purchase tobacco. To deal with the remaining zeros, they compare specifications based on infrequency of purchase and on censoring. GMM and systems-GMM are used to deal with errors-in-variables and unobservable heterogeneity (Arellano and Bond, 1991; Bover and Arellano, 1997). Within-groups two-step, within-groups three-step GMM and MD methods are used to allow for censoring. To reduce the influence of distributional assumptions they adopt a semiparametric approach to estimate each of the T cross-section equations using Powell (1986) symmetrically censored least squares (SCLS). There is evidence that the rational addiction specification is sensitive to unobservable heterogeneity and censoring and the results suggest that failure to account for heterogeneity may lead to overestimates of the impact of addiction. The panel data estimators imply that behavior is more forward-looking than suggested by the results that fail to correct for heterogeneity.

12.5.2.3 Finite mixture models

Deb (2001) applies a random effects probit model in which the distribution of the individual effect is approximated by a discrete density. This is an example of a finite mixture model and it relaxes the normality assumption for the distribution of the random effects. Deb uses Monte Carlo experiments to assess the small sample properties of the estimator. These show that only 3–4 points of support are required for the discrete density to mimic normal and chi-square densities and to provide approximately unbiased estimates of the structural parameters and the variance of the individual effects. Deb applies the model to a cross-section of individuals clustered in families, where the random effect represents unobserved family effects. It is assumed therefore that all individuals in each family belong to the same latent class. This approach aims to approximate the distribution of the random (family) intercepts, whereas the responses to the explanatory variables are not allowed to vary across latent classes. Clark *et al.* (2005) develop a latent class ordered probit model for reported well-being, in which individual time invariant heterogeneity is allowed both in the intercept and in the income effect.

12.5.3 Applications with count data

12.5.3.1 Poisson/log-normal mixtures

Winkelmann (2004a) proposes an alternative two-part, or hurdle, model based on a Poisson/log-normal mixture rather than the usual negative binomial (Negbin) variants. Hurdle, or two-part, models make a distinction between the decision to seek care, modeled as a binary choice, and the conditional number of visits, modeled as a truncated count data regression. They are the most widely used specification in the recent applied literature (see, e.g., Alvarez and Delgado, 2002; Chang and Trivedi, 2003; Sarma and Simpson, 2006; Yen *et al.,* 2001), although Santos Silva and Windmeijer (2001) have pointed out that they may be problematic in applications where it cannot be assumed that there is a single spell of illness for each period of observation in the data. Winkelmann's (2004a) model leads to a probit equation for the first hurdle and a truncated Poisson/log-normal model for the second. Unlike the Negbin model, the latter does not have a closed form and is

estimated using Gauss–Hermite quadrature. An application to the 1997 reform of co-payments for prescription drugs in Germany uses data on quarterly doctor visits in the GSOEP. This confirms Deb and Trivedi's (2002) result that finite mixture models outperform Negbin hurdle models, but the results show that the normal hurdle model fits better than both of these specifications.

The innovation in Van Ourti (2004) is to include a Gaussian random effect in the two-part model. The time invariant individual effect appears as a common factor in both parts of the model, with the factor loading in the first equation normalized to one for identification. The individual effect is then integrated out with the resulting integral evaluated by Gauss–Hermite quadrature. In an empirical application to GP and specialist visits and to nights in hospital in the Panel Study of Belgian Households (PSBH), the panel version of the two-part model (2PM-PA) is compared to a one part-model with a Gaussian individual effect (1PM-PA) and to pooled versions of the one-part and two-part models (1PM-PO, 2PM-PO). On the basis of the log-likehoods and the Akaike information criterion (AIC) and Bayesian information criterion (BIC), the 2PM-PA specification is preferred to the simpler specifications.

While Van Ourti (2004) extends the Gaussian random effects model to the two-part specification for count data, Munkin and Trivedi (1999) and Riphahn *et al.* (2003) do the same for a bivariate count data model, dealing with the case where there are two dependent variables both measured as integer counts. Munkin and Trivedi (1999) propose a model that is designed for cross-section data and that is applied to the number of emergency room visits and of hospitalizations in data from the US National Medical Expenditure Survey, 1987–88 (as used by Deb and Trivedi, 1997). They construct the bivariate model by specifying the marginal distributions of the counts and then adding a correlated heterogeneity term to give a Poisson-log-normal mixture. This has conditional mean functions:

$$\begin{aligned} \lambda_{i1} &= \exp(x_{i1}^{'}\beta_1 + \varepsilon_{i1}) \\ \lambda_{i2} &= \exp(x_{i2}^{'}\beta_2 + \varepsilon_{i2}) \\ (\varepsilon_{i1}, \varepsilon_{i2}) &\sim N[(0,0), (\sigma_1^2, \rho\,\sigma_1\,\sigma_2, \sigma_2^2)]. \end{aligned} \tag{12.33}$$

The log-likelihood function for the model involves a two-dimensional integral and is estimated by MSL with antithetics and a correction for first order simulation bias. Using a bivariate model does not lead to substantial changes in the estimated effects of the regressors, but the overall fit of the model does improve.

The model proposed by Riphahn *et al.* (2003) is designed for panel data and they apply it to separate measures of doctor visits and hospital inpatient visits from 12 waves of the GSOEP for 1984–95, focusing on West Germans aged between 25 and 65. Their specification extends the single equation model of Geil *et al.* (1997) by adding two Gaussian error components to each equation; one, a time invariant individual effect (u_{ij}); the other, a time-varying idiosyncratic error term (ε_{itj}):

$$\begin{aligned} \lambda_{it1} &= \exp(x_{it1}^{'}\beta_1 + u_{i1} + \varepsilon_{it1}) \\ \lambda_{it2} &= \exp(x_{it2}^{'}\beta_2 + u_{i2} + \varepsilon_{it2}). \end{aligned} \tag{12.34}$$

Correlation between the two equations is introduced by assuming that the error terms are drawn from a bivariate normal distribution. Although it is not clear from the paper, it appears that the individual effects (u_{ij}) are assumed to be independent of each other, which seems rather restrictive. Computation is based on a combination of Gauss–Hermite quadrature, to integrate over the time invariant individual effect, and Gauss–Legendre quadrature, to integrate over the bivariate distribution within each period.

12.5.3.2 Finite mixtures

Deb and Trivedi (1997, 2002), Deb and Holmes (2000), Jiménez–Martin *et al.* (2002) and Sarma and Simpson (2006) estimate finite mixture models for count measures of health care use, in which a Negbin distribution is assumed within each latent class. Lourenço and Ferreira (2005) extend the application of the Negbin finite mixture model to a truncated sample from the 2003–04 Europep survey for Portugal, where data are only collected for those who visit health centers. This means that the data are drawn from an endogenous sampling scheme and are truncated at zero, and raises the question of whether the distribution of unobserved heterogeneity should be defined over the whole population or only the truncated sample.

Jochmann and Leon-Gonzalez (2004) propose a specification that uses a semiparametric Bayesian approach, which can be seen as an extension of Deb and Trivedi (1997). They start with a parametric "random coefficients" specification of the Poisson model as a benchmark. In this model the random slopes (b_i) are assumed to be drawn from a multivariate normal distribution, so the conditional mean function takes the form:

$$\lambda_{it} = \exp(x_{it}' \beta + w_{it}' b_i + \varepsilon_{it}). \tag{12.35}$$

The semiparametric element of the model is introduced by using a Dirichlet process mixture for the prior on the random effects. This gives a mixture model with a random number of components and extends the usual treatment of LCMs that fix the number of components. The Dirichlet process specifies a base distribution, in this case assumed to be normal, and a fixed number of mass points, in this case set equal to 10. Then, new draws of the random effects are a mixture of draws from the base distribution and draws from existing clusters of values. The end product is a discrete distribution where the number of mass points is random. Estimation of the model is done by MCMC, based on Gibbs sampling, with data augmentation (the random effects and latent variables are treated as parameters to be estimated) and incorporating a Metropolis–Hastings step where the Gibbs sampling cannot be used. The MCMC algorithm was run for 30,000 iterations, discarding the first 5,000 for the burn-in period. The application uses data on the number of visits to the doctor in the previous quarter from the 1997–2001 waves of the GSOEP. The aim is to test for horizontal equity in the delivery of care by seeing whether non-need factors play a significant role in explaining variation in utilization. This is tested using Bayes factors and horizontal equity is not rejected with these data.

The two dominant strands in the recent literature, hurdle models and finite mixture models, are brought together in the latent class hurdle model developed

by Bago d'Uva (2006). Her model uses a panel of individuals across time: individuals i are observed T_i times. Let y_{it} represent the number of doctor visits in year t. The joint density of $y_i = \left(y_{i1}, \ldots, y_{iT_i}\right)$ is given by:

$$g\left(y_i|x_i;\pi_{i1},\ldots,\pi_{iC};\theta_1,\ldots,\theta_C\right) = \sum_{j=1}^{C}\pi_{ij}\prod_{t=1}^{T_i} f_j\left(y_{it}|x_{it};\theta_j\right), \tag{12.36}$$

where x_i is a vector of covariates, including a constant, the θ_j are vectors of parameters, and $0 < \pi_{ij} < 1$ and $\sum_{j=1}^{C}\pi_{ij} = 1$. Conditional on the class that the individual belongs to, the number of visits in period t, y_{it}, is assumed to be determined by a hurdle model. The underlying distribution for the two stages of the hurdle model is the Negbin. Formally, for each component $j = 1, \ldots C$, it is assumed that the probability of zero visits and the probability of observing y_{it} visits, given that y_{it} is positive, are given by the following expressions:

$$\begin{aligned} f_j\left(0|x_{it};\beta_{j1}\right) &= P\left[y_{it}=0|x_{it},\beta_{j1}\right] = \left(\lambda_{j1,it}^{1-k}+1\right)^{-\lambda_{j1,it}^{k}} \\ f_j\left(y_{it}|y_{it}>0,x_{it};\beta_{j2}\right) &= \frac{\Gamma\left(y_{it}+\frac{\lambda_{j2,it}^{k}}{\alpha_j}\right)\left(\alpha_j\lambda_{j2,it}^{1-k}+1\right)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}}\left(1+\frac{\lambda_{j2,it}^{k-1}}{\alpha_j}\right)^{-y_{it}}}{\Gamma\left(\frac{\lambda_{j2,it}^{k}}{\alpha_j}\right)\Gamma(y_{it}+1)\left[1-\left(\alpha_j\lambda_{j2,it}^{1-k}+1\right)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}}\right]} \end{aligned}, \tag{12.37}$$

where $\lambda_{j1,it} = \exp\left(x_{it}'\beta_{j1}\right)$, $\lambda_{j2,it} = \exp\left(x_{it}'\beta_{j2}\right)$, α_j are overdispersion parameters and k is an arbitrary constant.

As in the standard hurdle model, β_{j1} can be different from β_{j2}, reflecting the fact that the determinants of care are allowed to have different effects on the probability of seeking care and on the conditional number of visits. On the other hand, having $\left[\beta_{j1},\beta_{j2}\right] \neq \left[\beta_{l1},\beta_{l2}\right]$ for $j \neq l$ reflects the differences between the latent classes. Various special cases are nested within the general model. It can be assumed that all the slopes are the same, varying only the constant terms, $\beta_{j1,0}$ and $\beta_{j2,0}$, and the overdispersion parameters α_j. This represents a case where there is unobserved individual heterogeneity, but not in the responses to the covariates. The most flexible version allows α_j and all elements of β_{j1} and β_{j2} to vary across classes. The finite mixture hurdle model also accommodates a mixture of sub-populations for which health care use is determined by a Negbin model (the two decision processes are indistinguishable) and sub-populations for which utilization is determined by a hurdle model. This is obtained by setting $\beta_{j1} = \beta_{j2}$ for some classes. Setting them equal for all of the classes gives a panel data version of Deb and Trivedi's (1997) latent class Negbin model. Bago d'Uva (2006) applies the latent class hurdle model to panel data from the RAND Health Insurance Experiment and finds a higher price effect on health care utilization for the latent class of "low users." This is mostly attributable to the impact of price on the probability of seeing a doctor rather than the conditional number of visits.

12.5.4 Applications of quantile regression and other semiparametric methods

Quantile regression is of particular value when there is interest in the full conditional distribution of an outcome rather than just the conditional expectation of the variable. The method provides one way to allow for heterogeneity in treatment effects over the range of the conditional distribution. It is a semiparametric approach that avoids distributional assumptions about the error term. It also has the attractive property of invariance to monotone transformations of y and therefore avoids the retransformation problem.

Kan and Tsai (2004) apply quantile regression to the conditional distribution of BMI using data from the Cardiovascular Disease Risk Factors Two-Township Study (CVDFACTS) for Taiwan. Quantile regression is well suited to an analysis of obesity as interest focuses on the upper tail of the distribution of BMI rather than the area around the mean. Other studies have tended to create an indicator variable for obesity using the published clinical thresholds, but the quantile approach makes better use of the available variation in BMI in the upper tail of the distribution.

Lee and Jones (2006) use the same dataset for Taiwan dentistry as Lee and Jones (2004) to provide evidence on heterogeneity in dentists' activity. The heterogeneity is of particular importance because dentists' responses are likely to differ widely from high- to low-activity dentists. Quantile regressions provide a useful method to investigate the differential responses of dentists to various observable variables. It is shown that time trends for dentists at higher quantiles have greater fluctuations than those at lower quantiles. The clinic–hospital gap in activity at higher quantiles is greater than at low quantiles. Clinic dentists at higher quantiles have much higher numbers of visits and numbers of treatments than those at lower quantiles, but they provide less intensity of care. Dentists in deprived areas have higher activity than those in non-deprived areas, but only when they are located at higher quantiles.

Winkelmann (2006) applies quantile regression to count data on doctor visits in the GSOEP. This allows an evaluation of the impact of the 1997 reform of copayments for medicines on the full conditional distribution and not just the mean. The quantiles of a count are integer valued and cannot be represented by a continuous function of the covariates, such as $\exp(x'\beta)$, so Winkelmann (2006) adopts the method proposed by Machado and Santos Silva (2005). This transforms the data by "jittering": adding a uniform random variable to the counts and then applying quantile regression to the resulting continuous variable. The results show a greater impact of the reform on the lower quantiles, which is consistent with the earlier evidence from a hurdle model in Winkelmann (2004) that showed a larger effect in the first part of the model.

Applications of other semiparametric regression methods are relatively sparse in the health economics literature. Askildsen *et al.* (2003) use Kyriazidou's (1997) semiparametric estimator for a panel data sample selection model in order to estimate nurses' labor supply in Norway. This allows for individual effects in the selection equation and the hours equation that may be correlated with each other

and with the observed covariates. A conditional logit is used to estimate the selection equation. Then the hours equation is estimated in first differences by weighted least squares, with kernel weights applied to the difference in the linear index from the selection equation between different periods. The aim is to difference observations across periods for which the probability of selection is (approximately) the same. Rettenmaier and Wang (2006) use a semiparametric estimator for the Tobit model to model persistence in Medicare reimbursements. The model allows for fixed effects and a lagged dependent variable, but assumes that initial conditions are fixed.

It is widely believed that income has a direct effect on health, but it is often argued that indirect income effects due to relative deprivation may be equally important. Wildman and Jones (2007) investigate these relationships using parametric and semiparametric panel data models. By allowing for a flexible functional form for income, they seek to ensure that coefficients on relative deprivation variables are not an artifact of a highly non-linear relationship between health and income. Parametric estimation may lead to biased coefficients if the parameterization of the explanatory variables is incorrect. Semiparametric partially linear estimation overcomes this problem by allowing an unspecified relationship between health and income (Robinson, 1998). The results provide strong evidence for the impact of income on self-reported measures of health for men and women. These results are robust across a range of techniques and are resilient to the inclusion of measures of relative deprivation. The parametric results for relative deprivation largely reject its influence on health, although there is some evidence of an effect in the semiparametric models.

12.6 Multiple equation models

12.6.1 Applications using MSL

Balia and Jones (2008) use the first wave of the British HALS from 1984 to 1985 along with the longitudinal follow-up from May 2003 to model the determinants of premature mortality and to assess the relative contribution of lifestyle factors to the gradient in mortality in Britain. A behavioral model, that relates mortality to observable and unobservable factors, is used to motivate the empirical specification. Death, SAH and a range of health-related behaviors are measured as binary outcomes and, to capture the effect of lifestyles on mortality and morbidity in the presence of common unobservable factors, the model is estimated as a recursive multivariate probit. The full system is estimated by MSL and health inequality is explored using a decomposition analysis of the Gini coefficient. The results contradict the view that lifestyles only play a minor part in health inequalities.

Deb and Trivedi (2006) and Deb *et al.* (2006a) use a latent factor specification to model selection into treatment in nonlinear models and adopt a MSL estimator. The aim is to estimate causal effects using a structural model, motivated by a selection on unobservables approach, in which the parametric distribution of

the unobservables is specified and the full model is estimated by ML. In this case the outcomes of interest are binary and count measures of health care utilization. The treatment variables reflect the individual's choice of insurance plan, which is modeled using a random utility framework. The options are categorized as HMOs, other-managed care and non-managed care plans, with the choice of options specified using a multinomial logit specification. In Deb and Trivedi (2006) the data are taken from the 1996 US Medical Expenditure Panel Survey (MEPS), while in Deb *et al.* (2006a) the MEPS data are augmented by the 1996 Community Tracking Survey (CTS) as a test for the external validity of the findings. Utilization and insurance plan are modeled simultaneously to take account of the possibility of selection by patients, insurers and providers. Identification of the latent factor model rests on the assumption that each of the multinomial choices depends on a unique latent factor, based on independent and identically distributed (i.i.d.) normal draws, and these are allowed to be freely correlated with the error in the outcome equation. The parametric specification is identified by functional form, but exclusion restrictions are also imposed, using employment status and occupational sector as predictors of insurance plans, while excluding them from the outcome equations. Simple linear models are used to provide informal checks for the validity of these instruments. Estimation uses MSL, accelerated by using quasi-random draws from Halton sequences. The results do not show evidence of favorable selection into HMO plans, but the average treatment effects on the use of care are much larger when selection is taken into account and there is considerable heterogeneity in the effects. Monte Carlo simulation is also used to compute the standard errors of these treatment effects.

Other applications of MSL include Lindeboom *et al.* (2002), who use the LASA panel to estimate a five-equation model for the use of long-term care services among elderly residents of Amsterdam. They take draws from a multivariate normal distribution and use antithetics to accelerate the estimation. The results show strong effects of health status, sex, socioeconomic status and prices on the use of institutional care. Two papers by Pudney and Shields (2000a, 2000b) use MSL to estimate a system of equations comprised of a generalized ordered probit model for British nurses' pay grades along with auxiliary equations for training, career breaks, work outside the NHS, and part-time work. A common factor structure leads to a log-likelihood based on the multivariate normal and hence the need for simulation estimation. The MSL estimation only uses 50 replications, without acceleration, which is relatively few for this kind of application. Pudney and Shields (2000b) make a case for identification based on functional form rather than exclusion restrictions. Vera-Hernandez (2003) makes use of MSL to estimate a structural model of insurance coverage and health care use applied to data from the RAND Health Insurance Experiment.

12.6.2 Applications using Bayesian MCMC

Patient selection makes it hard to identify the impact of the size and characteristics of hospitals on the quality of their outcomes. The problem arises when there is

selective admission that is influenced by unobservables, such as unmeasured severity, that are also associated with the quality of outcomes. Geweke *et al.* (2003) find evidence of patient selection among 78,848 Medicare patients treated in 114 hospitals in Los Angeles county between 1989 and 1992. They focus on patients aged over 65 with a diagnosis of pneumonia taken from administrative data on hospital discharges collected by the State of California Office of Statewide Health Planning and Development. The quality of the clinical outcomes is measured by deaths in hospital within ten days of admission. A structural probit model for deaths is coupled with a reduced form multinomial probit model for the patient's choice of hospital, allowing correlation in the error terms to capture patient selection. The system of equations is estimated by a Bayesian approach using the MCMC simulator of the posterior distribution. Gibbs sampling with data augmentation breaks the estimation into steps, first simulating the latent dependent variables and then estimating the linear simultaneous equations system. The model is identified by using distances from the patients' homes to the hospitals as an instrument. The raw data and simple probits do not show a relationship between hospital size and mortality rates, but the MCMC results reveal a U-shaped relationship, with better quality in the smallest and largest hospitals.

Deb *et al.* (2006b) start with a conventional two-part model for medical expenditure. This is applied to data on ambulatory care, which has 17% zero observations, and hospital care, which has 94% zero observations and which exhibits positive skewness and excess kurtosis. The data are drawn from the US MEPS for 1996–2001, giving six repeated cross-sections and 20,460 observations. The standard two-part set-up is used with a binary choice equation and a conditional regression for the logarithm of expenditure. However, this is augmented by a multinomial probit model to allow for the endogenous selection of insurance plans, which fall into three categories: HMO, PPO (preferred provider organization) and FFS (fee-for-service). To capture the possibility of selection bias the error terms from the insurance equations (u) are assumed to be linearly associated with the error terms in the two parts of the model for expenditure:

$$\begin{aligned} \varepsilon_1 &= u'\delta + \upsilon \\ \varepsilon_2 &= u'\pi + \tau. \end{aligned} \tag{12.38}$$

This assumes that the εs are only conditionally independent given u and relaxes the usual assumption that the two parts of the model are independent. Like Geweke *et al.* (2003), the full system of equations is estimated by Bayesian MCMC and Bayes factors are used to construct a test for the exogeneity of the choice of insurance plan. This test shows evidence of substantial selection bias. Having estimated the model, the authors show how to define and compute estimated treatment effects for the impact of insurance plan on expenditure, using data augmentation to impute the latent variables. It should be noted that the approach used to compute the treatment effects involves a standard retransformation for log-scale data and therefore relies on a strong assumption about the absence of heteroskedasticity in the expenditure data (Manning and Mullahy, 2001).

12.6.3 Applications using finite mixtures

The common factor model was introduced earlier in this chapter in the context of Aakvik *et al.*'s (2005) evaluation of a Norwegian VR program (see equations (12.8)–(12.10) above). Aakvik *et al.*'s application takes a parametric approach and assumes that the common factor is normally distributed. An alternative, semiparametric approach is to use a finite density and estimate a DFM. The DFM has two main advantages over MSL. First, it is semiparametric and therefore more robust than the parametric approach, which relies on strong distributional assumptions. Also, in general, it is easier to compute, involving standard numerical methods for maximum likelihood estimation or the use of the expectation maximization (EM) algorithm, rather than computationally intensive Monte Carlo simulation (see, e.g., Arcidiancono *et al.,* 2007). However, in practice there can be problems with identification, manifested in failure of convergence and problems with multiple optima.

Like Aakvik *et al.* (2005), Aakvik *et al.* (2003) use a latent variable framework to specify the impact of multidisciplinary treatment for back pain on the probability of leaving sickness benefits. Using a structural model, based on latent variables, allows them to define the ATE, the ATET and to allow for heterogeneous MTEs. In this application the unobservables are modeled using a discrete factor structure with distance to the nearest hospital used to identify the model. The estimates show a positive effect of around 6 percentage points on the probability of leaving sickness benefits.

Rous and Hotchkiss (2003) use data from 254 Texas counties on all reported births in 1993 to explore the impact of prenatal care on birth weight. They estimate a discrete factor model with three equations: a logit model for whether the pregnancy is carried to full-term and linear regressions for a measure of prenatal care visits and for birth weight. Travel distance to the nearest provider of abortions is used to identify the first equation, the availability of obstetricians is used for the second and the gender of the child for the third. The factor model shows evidence of adverse selection effects. Picone *et al.* (2003b) merge panel data from the US National Long-Term Care Survey (NLTCS) for 1984–95 with the National Death Index to investigate the impact of treatment intensity on survival rates and other health outcomes. Their model involves a system of three equations for treatment intensity, length of stay and health outcomes. These are assumed to have a common factor structure which is estimated using a one-factor model. The selection of models is based on the likelihood ratio (LR) statistic (Mroz, 1999) and 1,000 bootstrap replications are used to avoid the problem of local optima. The model is identified by excluding area data on the cost of capital, the Herfindahl index for hospital concentration and a wage index from the mortality equation. The results suggest that treatment intensity has a beneficial effect. Hamilton *et al.* (2000) compare US and Canadian data to see whether waiting time for surgery for hip fractures influences the outcomes of the treatment, measured by length of stay and inpatient mortality. They use discharge data for 20,995 patients admitted to acute hospitals in Quebec and Massachusetts between 1990 and 1992 and estimate a competing

risks model for delay before surgery and post-surgery length of stay. They use the Heckman–Singer specification for the common heterogeneity with two mass points (Heckman and Singer, 1984). Day of week of admission is used as an instrument for delay until surgery. The raw data show a strong relationship between delays and outcomes, but this disappears once unobserved frailty is taken into account. The higher observed inpatient mortality in Quebec is attributed to the longer length of stay rather than poorer outcomes: the longer the patients remain in hospital, the more likely they are to die there rather than at home.

Other applications of the discrete factor model include Bhattacharya *et al.* (2003), who model the impact of public and private insurance on HIV-related mortality. Mello *et al.* (2002) use a discrete factor specification for a two-equation model of health plan choice, whether to join a Medicare HMO, and various measures of subsequent health care utilization applied to data from the Medicare Current Beneficiary Survey for 1993–96. A similar model of HMO enrollment and subsequent hospital use by Kan *et al.* (2003) finds evidence of strong selection effects when a discrete factor specification is used to deal with unobservables. Rous and Hotchkiss (2003) use the Nepal Living Standards Survey for 1996 in a model of choice of health care provider, levels of expenditure and health outcomes with a two-factor specification to allow for community and household effects. Holmes (2005) combines a multinomial logit model with a discrete factor specification to evaluate the US National Health Service Corps (NHSC) program that was designed to encourage doctors to locate in underserviced areas.

The discrete factor model allows for a common factor in the intercept of each equation. In contrast, the LCM allows all of the parameters to vary across the latent classes. Clark and Etilé (2006) use the latent class framework to approximate the continuous distribution of the individual effects in a dynamic random effects bivariate probit model. They apply the model to data on smoking among couples in the BHPS and use the simulated annealing version of the EM algorithm to estimate the model. Atella *et al.* (2004) develop a latent class model for the joint decisions of consulting three types of physician. The authors assume that, within a latent class, each decision can be modeled by an independent probit, so the joint distribution of the three binary outcomes is a product of probits.

12.6.4 Applications using copulas

The copula approach leads to closed forms and avoids the need for numerical integration. It also circumvents the problem of the limited menu of parametric specifications of multivariate distributions that are available, especially when normality is an unsuitable assumption, for example, when dealing with highly skewed data.

Although copulas are not mentioned explicitly, Prieger (2002) proposes an extension of the sample selection model that is built around the FGM copula. This is applied to data from the 1996 wave of the MEPS. The outcome equation is a duration model for hospital length of stay and the selection equation measures whether the individual had an inpatient stay or not during the survey period.

Prieger's approach builds on Lee's (1983) selection model that uses a bivariate normal copula. In both cases the use of copulas to model the joint distribution allows for the marginal distributions to be non-normal. As the outcome is a measure of duration, exponential, gamma, log-logistic, log-normal and Weibull specifications are considered and the gamma is selected as the preferred model. This reflects an attractive feature of the copula approach, that model selection can focus on finding an acceptable specification of the marginal distributions before turning to the joint distribution. The specifications of the full selection model are compared using the AIC, BIC, and the consistent Akaike information criterion (CAIC) as well as the Voung test for non-nested models. These tests favor the FGM over the bivariate normal copula in terms of goodness-of-fit. Smith (2003) uses the same data as Prieger (2002), but a different copula, the Frank copula from the Archimedean class, that improves the fit of the model and changes the estimates of average length of stay.

Copulas are used by Zimmer and Trivedi (2006) to model dependence in a system of nonlinear reduced form equations. Their application focuses on couples' decisions about insurance coverage and health care use and consists of three equations: one for the husband's utilization, one for the wife's and one for whether or not the couple take out separate health insurance policies. The marginal distributions for utilization are assumed to be Negbin (NB2) and a probit is used for the insurance equation. The equations are assumed to be linked by common unobservable factors and the joint distribution is modeled using a trivariate Frank copula, which is derived using the mixture-of-powers approach. The use of copulas avoids having to select a parametric specification for the unobservable heterogeneity and is computationally tractable. To emphasize these points, Zimmer and Trivedi compare their results to those derived from an MSL approach, based on multivariate normality. The model is applied to data from four waves of the US MEPS. An apparent limitation of the copula approach is that it works by specifying marginal distributions and then modeling dependence, so that the emphasis is on a system of reduced form equations rather than on conditional distributions. However, Zimmer and Trivedi show how the estimates can be used to derive the conditional distribution and hence compute the ATE of insurance coverage on health care use.

12.7 Evaluation of treatment effects

12.7.1 Matching

Matching provides a general approach to deal with selection on observables. It addresses the problem that, in the observed data, confounding factors may be non-randomly distributed over the treated and controls. Rosenbaum and Rubin (1983) showed that, rather than matching on an entire set of observable characteristics (x), the dimensions of the problem can be reduced by matching on the basis of their probability of receiving treatment, $P\left(d = 1|x\right)$, known as the propensity score. In practice, propensity score matching (PSM) estimators do not rely on

exact matching and instead weight observations by their proximity, in terms of the propensity score.

An important requirement is that the model for treatment, used to construct the propensity score, should only include variables that are unaffected by participation in the treatment, or the anticipation of participation. The matching variables should be either time invariant characteristics or variables that are measured before participation in the treatment and that are not affected by anticipation of participation. The crucial condition for identification of treatment effects using the matching approach is that the selection into treatment should be ignorable, conditional on the observed covariates.[3]

Jones *et al*. (2007a) use the ECHP to estimate the impact of private health insurance coverage on the use of specialist visits in four European countries that allow supplementary coverage. The results show that the probability of having private insurance increases with income and with better reported health. Private insurance has a positive effect on the probability of specialist visits in all countries, although the magnitude is sensitive to the choice of estimator. They match treated individuals with non-treated individuals inversely weighted for the distance in terms of estimated propensity scores, with weights constructed using kernel smoothed distance weighting. They ensure that all cases are supported by controls. The quality of the matching can be assessed by computing the reduction of the pseudo R^2 of the insurance regression before and after matching. To evaluate the extent to which matching on propensity scores balances the distribution of the *x*s between the insured and the uninsured group, they compute the bias reduction due to matching for each of the *x*s.

Dano (2005) uses a 10% sample of the Danish population, drawn from register data, to give a panel for 1981–2000. She estimates the impact of injuries sustained in road traffic accidents on economic outcomes. Although these accidents are unanticipated health shocks, their incidence varies with observable and unobservable characteristics that can be associated with the outcomes and the estimates of the treatment effect need to be adjusted for this. Due to the large sample size, one-to-one matching without replacement is used, with matching on the linear index from the propensity score. The matching is combined with a difference-in-differences approach to control for time invariant unobservables. The study finds an impact of injuries on earned income for older and low-income individuals, but also shows the compensating effects of public transfers in the Danish system.

García-Gómez and López-Nicolás (2006) adopt a matched panel data difference-in-differences estimator. They use panel data from the Spanish component of the ECHP to explore the impact of health shocks on employment and of employment shocks on health. Their strategy is to match exactly on pre-treatment outcomes, in order to control for time invariant observables. This means that controls are restricted to those individuals who were identical to the treated in terms of their pre-treatment outcomes. In order to define the treated and the controls, a three-year window is adopted. In the case of health shocks, the treated are those who are in good health in the first year and then move to bad health in the next two years, and who are in employment for the first two years. The controls remain in good

health throughout the period and are also in employment in the first two years. The outcome is employment status in the third year. In addition to matching on pre-treatment outcomes, PSM is used to make the treated and controls comparable in terms of their observed characteristics. Four methods of matching are compared: nearest neighbor matching on the propensity score; kernel matching on the propensity score; matching of the four nearest neighbors on a set of explanatory variables; and simple matching on the four nearest neighbors according to the propensity score.

Other studies that use matching are summarized in Table 12.2.

12.7.2 Regression discontinuity

The regression discontinuity (RD) design exploits situations where the assignment to treatment changes discontinuously with respect to a threshold value of one or more exogenous variables. The contrast between individuals on either side of the discontinuity is used to identify the treatment effect. In a sharp regression discontinuity design, passing the threshold completely determines the allocation of treatment. In a fuzzy design, which is more likely in practice, the allocation of treatment is stochastic and the threshold creates a discontinuity in the probability of treatment. The discontinuity design relies on a comparison of observations "before and after" the threshold and does not have a separate control group. For this reason, applications typically use a narrowly defined neighborhood around the discontinuity to try and ensure that the treated and untreated observations are comparable in other respects. Studies that use a discontinuity design were described in section 12.2 above (Almond, 2006; Lleras-Muney, 2005; Pop-Eleches, 2006).

12.7.3 Difference-in-differences

The difference-in-differences, or diff-in-diff (DD) approach to evaluation with non-experimental data has been applied extensively in the health economics literature. The method is based on a before and after (pre–post) design with a control group. It can be used with both panel data and repeated cross-sections and requires treatment and control groups to be specified.

The basic form of the DD estimator of the average treatment effect compares mean outcomes for the treated (1) and controls (0) before (B) and after (A) the treatment:

$$ATE_{DD} = (\bar{y}_A^1 - \bar{y}_B^1) - (\bar{y}_A^0 - \bar{y}_B^0). \tag{12.39}$$

With individual panel data, the DD estimator can be computed using a two-way fixed effects specification:

$$y_{it} = \gamma(T_i P_t) + x_{it}'\beta + \nu_t + u_i + \varepsilon_{it}. \tag{12.40}$$

The treatment effect is identified by the parameter (γ) on the interaction term between the indicator for whether the individual is in the treated group (T) and the indicator for the post-treatment period (P).[4] The observed regressors (x) control for any observable differences between the treated and controls and the individual effects (u) control for any time invariant unobservable differences that may be

Table 12.1 Studies that use matching estimators

Study	*Outcome*	*Treatment*	*Method*	*Comment*
Dano (2005)	Earnings, annual employment rate, disposable income, public transfer income	Road traffic accidents	Uses difference-in-differences matching estimator with panel constructed from register data. One-to-one matching without replacement based on linear index from propensity score. Checks for balancing	Finds an impact on earned income for older and low income individuals. Also shows compensating effects of public transfers in the Danish system
Frolich *et al*. (2004)	Labor market outcomes: reintegration into the labor force	Vocational rehabilitation program in western Sweden	Multiple treatments with matching based on multivariate balancing scores computed from multinomial probit models. Uses nearest neighbours with replacement. Checks for balancing of covariates after matching	Finds a negative effect of rehabilitation. Many of the effects are insignificant: many controls are used repeatedly in the matching
García-Gómez and López-Nicolás (2006)	Employment, income, SAH	Health shocks and employment shocks	Use matched difference-in-differences estimator. Combine exact matching on pre-treatment outcomes with propensity score matching. Use both nearest neighbor and kernel-smoothed matching	Find effects of health shocks on employment and activity. Also find an effect of transition to unemployment on SAH

Girma and Paton (2006)	Teenage pregnancy rates	Free over-the-counter access to emergency birth control for teenagers at pharmacies in England	Use difference-in-differences matching estimator with panel data Matching based on nearest neighbors. Impose common support and test for balancing	No evidence that policy leads to lower teenage pregnancy rates
Jalan and Ravallion (2003)	Prevalence and duration of diarrhea in children aged under five in rural India	Piped water supply	Propensity score matching using logit, matching to five nearest neighbors	Finds that matching on household- as well as village-level controls makes a difference. Average effects are misleading as there is a lot of heterogeneity in the treatment effects, with little impact on poorer and less-educated mothers

correlated with the outcome and with the allocation of treatment. In this sense, the DD estimator combines selection on unobservables with selection on observables, so long as the unobservables are time invariant. The time effects (ν) take account of any time trend in the data that is common to both the treatment and control groups. This implies a "parallel trend" assumption. When the DD estimator is applied to repeated cross-section data a further assumption is required: that the composition of the treatment and control groups remains stable over time.

The DD estimator in equation (12.40) is defined above by using the interaction between the post-treatment period (P) and belonging to the treatment group (T). In some applications exposure to treatment may be defined by the interaction between more than two factors. For example, in Schmidt's (2007) evaluation of the impact of infertility insurance mandates on birth rates in the US, an indicator of whether states have mandates is interacted with whether or not women are over 35, as those over 35 are most likely to suffer infertility. Multiple interactions can be used to define exactly who is exposed to treatment and also to allow for heterogeneity in the size of the treatment effect. This approach is often labeled difference-in-difference-in-differences (DDD).

Chalkley and Tilley (2006) show how economic incentives can influence dental practice. This study exploits the comparison between self-employed and salaried dentists working for the NHS in Scotland to show that the financial incentives of FFS increase the intensity of treatment by around 21%. Using a DD approach, the paper finds that self-employed dentists treat exempt patients, who are assumed to be more likely to be influenced by supplier inducement, more intensively than non-exempt patients, relative to salaried dentists who do not face the financial incentive of FFS. These findings are based on an administrative database, the Management Information and Dental Accounting System (MIDAS), that records claims for self-employed and salaried dentists. The database provides a panel of dentists and patients and can be used to control for the practice style of individual dentists as well as measures of patient need.

In January 1994 the health authorities in Belgium increased co-payment rates for home and office visits to GPs and for visits to specialists. The increases were substantial: 35% for GP home visits, 45% for GP office visits and 60% for specialist visits. Cockx and Brasseur (2003) use this change in prices as a natural experiment. To create a control group they use those who were exempt from charges due to low income among widows, orphans, the disabled and retired. This means that identification relies on the treatment and control groups being comparable and the authors note that identification can only be achieved for low-income groups. A DD estimator is applied to the logarithm of utilization and the model is extended to a Rotterdam demand system to accommodate substitution effects induced by the change in relative prices. Interaction terms are used to allow for heterogeneity in the treatment effects. A similar set of reforms in Germany is used as a natural experiment by Winkelmann (2004b). In this case co-payments for prescription drugs increased by 6 DM on July 1, 1997, leading to relative price increases of up to 200% depending on the pack size. The policy is evaluated using data from the GSOEP for the years around the reform, 1995–96 and 1998–99. The control group

here is those exempt from these charges, made up of those with private insurance, co-insured children and low-income households. A DD strategy is adopted, with the effect of the co-payments on doctor visits identified by the interaction between the timing of the reform and exemption from charges. Winkelmann argues that the assumption of a common trend, which is essential for identification, is plausible in this context.

The use of a control group in DD methods help avoids the spurious inferences that can arise in a simple before and after comparison. For example, in Wagstaff and Yu's (2007) evaluation of the impact of the World Bank's Health VIII project in Gansu province in China they find only a small reduction in out-of-pocket spending on health care in counties exposed to the program. A before-and-after comparison would suggest that there was no significant improvement in this outcome. However, the trend in the control group of counties shows a rise in out-of-pocket payments, so the DD estimate reveals better outcomes among the treated relative to the controls. An important caveat is that the validity of the DD estimates relies on the comparability of the treated and controls in terms of the underlying trend in the outcomes. The comparability of treatment and control groups can be enhanced by combining the DD approach with matching estimators (as in Dawson *et al.,* 2007; Galiani *et al.,* 2005; Wagstaff and Yu, 2007). In Wagstaff and Yu (2007), the unmatched DD estimate for the impact of Health VIII on the availability of medical equipment in township health centers does not show a significant effect, but when the controls are matched with treated counties an effect is revealed. This is because the counties selected for the project tended to be poorer than the average among the controls so that, on average, the funds available to invest in equipment in the control counties lead to higher rates of increase. Matching with control counties that face the same sort of financial constraints allows a reliable comparison to be made.

The careful use of matching estimators, which should include tests of whether the treated cases and selected controls are balanced in terms of their observed characteristics, provides a link to strategies for testing the robustness of the identification assumptions that are built into the DD approach. The comparability of the treatment and control groups can be assessed by comparing their observed characteristics prior to treatment and, in particular, by testing the parallel trend assumption prior to treatment. A good example of this is Galiani *et al.* (2005). They use a DD strategy applied to panel data on municipalities in Argentina. The treatment of interest is the privatization of local water services that took place in the 1990s and the outcomes are measures of general and cause-specific child mortality. There is sufficient data for the pre-treatment period to do graphical analysis of the trends in the treatment and control groups. More importantly, it is possible to estimate the two-way fixed effects specification for mortality rates only using the data from the pre-treatment period, but including an indicator of which municipalities would go on to be treated. Evidence that this indicator is significant would undermine the parallel trends assumption and mean that areas that privatized their water supply were systematically different in terms of (trends in) mortality. In fact, Galiani *et al.* find that the common trend assumption is not rejected. Their DD

estimators are refined using a PSM approach. The results show a significant reduction in deaths from infectious and parasitic diseases and suggest that privatization helped to reduce health inequalities. Other studies that combine DD with matching are Dano (2005), García-Gómez and López-Nicolás (2006), Girma and Paton (2006) and Marini *et al.* (2008).

Numerous studies in health economics use the DD strategy. Some of these are summarized in Table 12.3.

12.7.4 Instrumental variables

The DD design is often applied in the context of natural experiments. Natural experiments and natural controls also form the basis for the IV approach, which is intended to capture "selection on unobservables" (see, e.g., Auld, 2006b). This approach relies on the variation in treatment that can be attributed to variation in an exogenous variable, or instrument; assigning individuals to treatments on the basis that the instrument mimics the random assignment of an experimental design. This approach is often hard to apply in practice as instruments should be both powerful predictors of treatment and have no direct effect on outcomes. The search for convincing instruments is therefore fraught with difficulty.

There are two broad estimation strategies. The FIML approach specifies a complete system of equations for the outcomes and treatments and estimates them jointly, allowing for common unobservable factors and identifying the model through exclusion restrictions. Estimation can be by MLE, MSL, MCMC, DFM or copulas, as described in section 12.6 above. The more commonly used approach is the limited-information or single equation approach, using IV estimators, such as 2SLS, GMM and 2SCML. Some studies that use instrumental variables were described in section 12.2 above (Arendt, 2005; Auld and Sidhu, 2005; Evans and Lien, 2005; Gardner and Oswald, 2007; Lakdawalla *et al.,* 2006; Lindahl, 2005). Other applications are too numerous to describe in detail here (some examples are Cawley *et al.* 2006; Contoyannis *et al.,* 2005; Dubay *et al.,* 2001; Dusheiko *et al.,* 2004, 2007; Elliott *et al.,* 2007; Guaraglia and Rossi 2004; Hadley *et al.,* 2003; Jewell and Triunfo, 2006; Kessler and McClellan, 2002; Lindrooth and Weisbrod, 2007; Meer *et al.,* 2003; Sasso and Buchmueller, 2004; Schellhorn, 2001; Sloan *et al.,* 2001; Van Houtven and Norton, 2004; Yelowitz, 2000).

In section 12.1 it was emphasized that, when treatment effects are heterogeneous, the IV estimator identifies a local average treatment effect (LATE) and that this estimate is conditional on the set of instruments that are used. In a recent paper, Basu *et al.* (2007) apply Heckman and Vitlacyl's (1999) LIV estimator which identifies MTEs over the support of the propensity score $p(d = 1|x, z)$. Computation of the LIV estimator involves regressing the outcome y on the observed regressors $x$and on a flexible function of the propensity score, which is estimated using x and the instruments z. The model could be estimated semiparametrically, for example, by using a partially linear model, or, as in Basu *et al.*, by adding polynomial and interaction terms between x and $p(d = 1|x, z)$. The LIV estimator of the $MTE(x,u_d)$

Table 12.3 Studies that use DD

Study	*Outcome*	*Treatment*	*Treated/controls*	*Comments*
Adams (2007)	Relative wages for older workers	Introduction of pure community rating among small group health insurers in New York in 1993	(i) Data for small and large firms in New York only: time interacted with older workers and employment in small firm (ii) Data for small firms in all states: time interacted with older workers and New York	Uses two sets of DDD estimates. Finds an increase in relative wages of older workers in small firms
Carpenter (2004)	Self-reported alcohol use and drunk driving	Zero tolerance drunk driving laws	In DDD estimators zero tolerance laws interacted with age 18–20, the group affected by the laws, and age 22–24	Finds an effect on episodic drinking among males, but for other outcomes
Chalkley and Tilley (2006)	Intensity of dental treatment	Remuneration of dentists, FFS or salary	Interacts salaried versus self-employed dentists with exempt versus non-exempt patients	Find that self-employed dentists treat exempt patients more intensively
Chen *et al*. (2007)	Use of outpatient and inpatient care and health status among the elderly	Introduction of NHI in Taiwan in 1995	Those who became insured under NHI/those already insured	Find an impact on use of care, which is larger for those on low incomes, but not on health
Chen and Zhou (2007)	Height, labor supply, earnings	1959–61 famine in China	Birth cohorts interacted with regional differences in severity of famine (proxied by excess mortality)	Find worse outcomes for those exposed to the famine
Cockx and Brasseur (2003)	Demand for physician services	Increase in co-payments	Non-exempt/exempt	DD applied to a Rotterdam demand system

Continued

Table 12.3 Continued

Study	*Outcome*	*Treatment*	*Treated/controls*	*Comments*
Currie and Hotz (2004)	Childhood accident requiring medical attention	Child care regulations	Aged 0–4/aged 5–9	
Davidoff *et al.* (2005)	Employer-sponsored insurance coverage	Reforms of state regulations of small group insurance market in US	Interacts indicators for reform with high/low risk and small/large firms	Uses DD and DDD estimates with repeated cross-sections
Dawson *et al.* (2007)	Waiting times in ophthalmology	London patient choice experiment (LPCP)	LCPC hospitals/three comparator groups: all hospitals in rest of England; a matched control group; all hospitals from four large metropolitan areas	DD combined with PSM. The results show a small reduction in average waiting times and reduced dispersion within the LPCP hospitals
Dranove *et al.* (2003)	Health care expenditures and health outcomes for patients receiving coronary artery bypass graft (CABG)	Mandatory CABG report card laws adopted in New York State in 1991 and Pennsylvania in 1993	Hospitals affected by reform/hospitals in other states	Use DD estimators in both hospital level and patient level analyses. Find an adverse effect of report cards on outcomes, due to the patient selection they induce. The validity of the DD strategy is assessed by using a cohort of acute myocardial infarction (AMI) patients who, as emergency admissions, should not be subject to patient selection
Dusheiko *et al.* (2006)	Admissions to hospital (chargeable elective, non-chargeable elective and emergency)	Abolition of GP fundholding in England	(Ex-)fundholders/ non-fundholders	Find that abolition of fundholding increased admissions

Finkelstein (2004)	Private insurance coverage for prescription drug expenditure	Medicare coverage	Those aged 63–64 at start of HRS panel survey, who became eligible for Medicare/(younger controls) those aged 60–62 who were never eligible, (older controls) those aged 65–67 who were eligible throughout	Finds no evidence that Medicare is associated with levels of private coverage for drug expenditures
Galiani *et al*. (2005)	Child mortality (general and cause-specific)	Privatization of local water services in Argentina	Privately provided water services/publically provided	Includes careful analysis of robustness of results to "parallel" trends assumption, matching methods and use of cause-specific mortality to assess validity of findings
Jensen and Richter (2004)	Poverty, nutrition, use of medical care	1996 pensions crisis in Russia	In arrears/those who received pensions	
Liu *et al*. (2004)	Hospital length of stay and charges	Postpartum discharge laws in US states	States that enacted the laws/states that did not	Find heterogeneity in the effects depending on how the laws were implemented

Continued

Table 12.3 Continued

Study	*Outcome*	*Treatment*	*Treated/controls*	*Comments*
Marini *et al.* (2008)	Retained surplus/deficit and Reference Cost Index (RCI)	Introduction of Foundation Trusts in England in 2004	Foundation Trusts/three comparator groups: all hospitals in rest of England; a matched control group; all hospitals eligible for Foundation status	Difference in differences combined with PSM. The results show a small increase in surplus and a small decrease in the RCI within Foundation Trusts
Propper *et al.* (2002)	Waiting times for hospital admissions in England (North West Anglia region)	GP fundholding	Fundholders' patients/non-fundholders' patients	Effects on waiting only applied to a limited set of patients and of procedures
Schmidt (2007)	Infertility (first birth rates)	Infertility insurance mandates in US states	Uses interaction between time, whether states have a mandate and age of women (over 35)	Uses DDD estimates. Finds that mandates significantly increase birth rates for women over 35
Wagstaff and Yu (2007)	Use of services, catastrophic expenses, health outcomes	World Bank's Health VIII project, Gansu province, China	Project/non-project counties	DD combined with PSM
Winkelmann (2004a)	Doctor visits	Increase in co-payments for prescription drugs	Non-exempt/exempt	
Wolfe *et al.* (2006)	Public health care coverage of "welfare leavers"	Wisconsin BadgerCare Program (expanded public health insurance eligibility)	(i) Between cohort strategy: 1997 cohort/1995 cohort of leavers (ii) Within cohorts: newly eligible/continuously eligible	DD estimates suggest that BadgerCare increased public coverage by 17–25 percentage points

is then:

$$LIV = \left[\frac{\partial E(y|x, p(x,z))}{\partial p(x,z)} \right]_{1-p(x,z)=u_d} . \tag{12.41}$$

This can be used to test for heterogeneity in the treatment effect and to construct estimates of the other treatment effects of interest, such as the ATE and ATET.

12.8 Future prospects

Like other areas that are rooted in applied microeconomics, such as development, environmental and labor economics, modern empirical analysis in health economics is dominated by the tools of microeconometrics and the use of individual-level data drawn from social surveys and administrative sources. The trend is towards complex longitudinal and multilevel data structures, with increasing reliance on linkage of a variety of sources. Health economists have exploited the full range of microeconometric techniques, and applications to health data have driven methodological innovations in the context of variables with skewed and heavy-tailed distributions, multinomial choices, count data and mixture models.

Much of the empirical analysis done by health economists seeks to estimate causal effects and fits within the treatment–outcomes framework. A challenge for successful applied work is to find appropriate sources of variation to identify the treatment effects of interest. Estimation of causal effects can be prone to selection bias, when the assignment to treatments is associated with the potential outcomes of the treatment. Overcoming this selection bias requires variation in the assignment of treatments that is independent of the outcomes. One source of independent variation comes from randomized controlled experiments. These are the norm in the evaluation of new clinical therapies, but their use for the evaluation of broader health and social programs remains relatively rare.

Applied researchers in health economics face a twin challenge. The first is to make the best possible use of the available non-experimental data by combining robust econometric methods, such as those presented in this chapter, with imaginative and convincing sources of identification. The second is to seek opportunities to encourage the agencies responsible for designing and funding health and social programs to collect new and comprehensive longitudinal datasets, to facilitate the linkage of different datasets and to make greater use of randomized designs in the evaluation of new initiatives.

Acknowledgments

This chapter draws on joint work with Teresa Bago d'Uva, Silvia Balia, Paul Contoyannis, Cristina Hernández Quevedo, Xander Koolman, José Labeaga, Miaw-chwen Lee, Roberto Leon Gonzalez, Nigel Rice, Stefanie Schurer, Eddy van Doorslaer and John Wildman, and by Casey Quinn. I am grateful for comments from Giorgia Marini, Edward Norton, Pedro Rosa Dias and João Santos Silva.

Notes

1. This chapter uses the "treatment-outcome" terminology that is commonplace in the evaluation literature. In practice many treatments are broad policy reforms associated with the financing and delivery of health care rather than specific clinical interventions. Treatment effects are defined here in terms of a binary treatment with just two regimes – the treated and the controls. In practice there may be multiple treatments and varying intensities of treatment.
2. In health economics, latent class models (LCMs) have typically been applied in the context of nonlinear regression models, to allow for the role of unobserved heterogeneity in the relationship between an observed outcome and a set of regressors. In the statistics literature, LCMs are more commonly applied in the context of latent structural variables and a set of observed indicators, such that the indicators are orthogonal conditional on class membership.
3. Comparisons of treatment effects estimated using randomized experiments with those estimated by matching methods in the labor economics literature cast considerable doubt on the validity of such ignorability conditions and hence on the matching approach (e.g., Agodini and Dynarski, 2004; LaLonde, 1986; Smith and Todd, 2001).
4. Bertrand *et al*. (2004) highlight the risk of making misleading inferences using the standard DD estimator if there is serial correlation in the outcomes and the standard errors are not adjusted to take account of it.

References

Aakvik, A., J.J. Heckman and E.J. Vytlacil (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* **125**, 15–51.

Aakvik, A. and T.H. Holmas (2006) Access to primary health care and health outcomes: the relationship between GP characteristics and mortality rates. *Journal of Health Economics* **25**, 1139–53.

Aakvik, A., T.H. Holmas and E. Kjerstad (2003) A low-key social insurance reform-effects of multidisciplinary outpatient treatment for back pain patients in Norway. *Journal of Health Economics* **22**, 747–62.

Abadie, A. and S. Gay (2006) The impact of presumed consent legislation on cadaveric organ donation: a cross-country study. *Journal of Health Economics* **25**, 599–620.

Adams, S. (2007) Health Insurance market reform and employee compensation: the case of pure community rating in New York. *Journal of Public Economics* **91**, 1119–33.

Adda, J. and F. Cornaglia (2006) Taxes, cigarette consumption and smoking intensity. *American Economic Review* **96**, 1013–28.

Agodini, R. and M. Dynarski (2004) Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics* **86**, 180–94.

Ai, C. and E.C. Norton (2000) Standard errors for the retransformation problem with heteroscedasticity. *Journal of Health Economics* **19**, 697–718.

Almond, D. (2006) Is the 1918 influenza pandemic over? Long term effects of in utero influenza exposure in the post 1940 US. *Journal of Political Economy* **114**, 672–712.

Almond, D., K.Y. Chay and D.S. Lee (2005) The costs of low birth weight. *Quarterly Journal of Economics* **120**, 1031–83.

Alvarez, B. and M.A. Delgado (2002) Goodness-of-fit techniques for count data models: an application to the demand for dental care in Spain. *Empirical Economics* **27**, 543–67.

Arcidiacono, P. and S. Nicholson (2005) Peer effects in medical school. *Journal of Public Economics* **89**, 327–50.

Arcidiancono, P., H. Sieg and F. Sloan (2007) Living rationally under the volcano? An empirical analysis of heavy drinking and smoking. *International Economic Review* **48**, 37–65.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277–97.

Arellano, M. and O. Bover (1995) Another look at the instrumental variables estimation of error component models. *Journal of Econometrics* **68**, 29–52.

Arendt, J.N. (2005) Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education Review* **24**, 149–60.

Arulampalam, W. and S. Bhalotra (2006) Sibling death clustering in India: state dependence versus unobserved heterogeneity. *Journal of the Royal Statistical Society Series A (Statistics in Society)* **169**, 829–48.

Arulampalam, W., R.A. Naylor and J.P. Smith (2004) A hazard model of the probability of medical school drop-out in the UK. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **167**, 157–78.

Askildsen, J.E., B.H. Baltagi and T.H. Holmas (2003) Wage policy in the health care sector: a panel data analysis of nurses' labour supply. *Health Economics* **12**, 705–19.

Askildsen, J.E., E. Bratberg and O.A. Nilsen (2005) Unemployment, labour force composition and sickness absence: a panel data study. *Health Economics* **14**, 1087–101.

Atella, V., F. Brindisi, P. Deb and F.C. Rosati (2004) Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics* **13**, 657–68.

Atella, V., F. Peracchi, D. Depalo and C. Rossetti (2006) Drug compliance, co-payment and health outcomes: evidence from a panel of Italian patients. *Health Economics* **15**, 875–92.

Au, D.W.H., T.F. Crossley and M. Schellhorn (2005) The effect of health changes and long term health on the work activity of older Canadians. *Health Economics* **14**, 999–1018.

Auld, M.C. (2002) Disentangling the effects of morbidity and life expectancy on labor market outcomes. *Health Economics* **11**, 471–83.

Auld, M.C. (2006a) Estimating behavioral response to the AIDS epidemic. *Contributions to Economic Analysis and Policy* **5**, article 12.

Auld, M.C. (2006b) Using observational data to identify the causal effects of health-related behaviour. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Auld, M.C. and P. Grootendorst (2004) An empirical analysis of milk addiction. *Journal of Health Economics* **23**, 1117–33.

Auld, M.C. and N. Sidhu (2005) Schooling, cognitive ability and health. *Health Economics* **14**, 1019–34.

Bago d'Uva, T. (2006) Latent class models for utilisation of health care. *Health Economics* **15**, 329–43.

Baker, M., M. Stabile and C. Deri (2004) What do self-reported, objective, measures of health measure? *Journal of Human Resources* **39**, 1067–93.

Balia, S. and A.M. Jones (2008) Mortality, lifestyle and socioeconomic status. *Journal of Health Economics* **27**(1), 1–26.

Baltagi, B.H., E. Bratberg and T.H. Holmas (2005) A panel data study of physicians' labor supply: the case of Norway. *Health Economics* **14**, 1035–45.

Banks, J., M. Marmot, Z. Oldfield and J.P. Smith (2006) Disease and disadvantage in the United States and in England. *Journal of the American Medical Association* **295**, 2037–45.

Baser, O., J.C. Gardiner, C.J. Bradley, H. Yuce and C. Given (2006) Longitudinal analysis of censored medical cost data. *Health Economics* **15**, 513–25.

Basu, A., B.V. Arondekar and P.J. Rathouz (2006) Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics* **15**, 1091–107.

Basu, A., J. Heckman, S. Navarro and S. Urzua (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* **16**, 1133–57.

Basu, A., W.G. Manning and J. Mullahy (2004) Comparing alternative models: log vs Cox proportional hazard? *Health Economics* **13**, 749–65.
Basu, A. and P.J. Rathouz (2005) Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**, 93–109.
Becker, B., M. Grossman and K. Murphy (1994) An empirical analysis of cigarette addiction. *American Economic Review* **84**, 396–418.
Behrman, J.R. and M.R. Rosenzweig (2004) Returns to birthweight. *Review of Economics and Statistics* **86**, 586–601.
Benitez-Silva, H., M. Buchinsky, H.M. Chan, S. Cheidvasser and J. Rust (2004) How large is the bias in self-reported disability. *Journal of Applied Econometrics* **19**, 649–70.
Bertrand, M., E. Duflo and S. Mullainathan (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 249–75.
Bhattacharya, J., D. Goldman and N. Sood (2003) The link between public and private insurance and HIV-related mortality. *Journal of Health Economics* **22**, 1105–22.
Black, S., P. Devereux and K. Salvanes (2007) From the cradle to the labour market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics* **122**, 409–39.
Bleakley, H. (2007) Disease and development: evidence from hookworm eradication in the American South. *The Quarterly Journal of Economics* **122**, 73–117.
Blough, D.K., C.W. Madden and M.C. Hornbrook (1999) Modeling risk using generalized linear models. *Journal of Health Economics* **18**, 153–71.
Blundell, R. and S. Bond (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115–43.
Bover, O. and M. Arellano (1997) Estimating dynamic limited-dependent variable models from panel data *Investigaciones Economicas* **21**, 141–65.
Bradford, D.W., A.N. Kleit, M.A. Krousel-Wood and R.N. Re (2001) Stochastic frontier estimation of cost models within the hospital. *Review of Economics and Statistics* **83**, 300–8.
Bradley, C.J., D. Neumark, H.L. Bednarek and M. Schenk (2005) Short-term effects of breast cancer on labor market attachment: results from a longitudinal study. *Journal of Health Economics* **24**, 137–60.
Briggs, A. (2006) Statistical methods for cost-effectiveness analysis alongside clinical trials. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Briggs, A., R. Nixon, S. Dixon and S. Thompson (2005) Parametric modelling of cost data: some simulation evidence. *Health Economics* **14**, 421–8.
Brown, T., J. Coffman, B. Quinn, R. Scheffler and D. Schwalm (2005) Do physicians always flee from HMOs? New results using dynamic panel estimation methods. *Health Services Research* **40**, 357–73.
Buntin, M.B. and A.M. Zaslavsky (2004) Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics* **23**, 525–42.
Burgess, J. (2006) Productivity analysis in health care. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Cantoni, E. and E. Ronchetti (2006) A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics* **25**, 198–213.
Carey, K. (2000) Hospital cost containment and length of stay: an econometric analysis. *Southern Economic Journal* **67**, 363–80.
Carpenter, C. (2004) How do zero tolerance drunk driving laws work? *Journal of Health Economics* **23**, 61–83.
Cawley, J., D.C. Grabowski and R.A. Hirth (2006) Factor substitution in nursing homes. *Journal of Health Economics* **25**, 234–47.
Chalkley, M. and C. Tilley (2006) Treatment intensity and provider remuneration: dentists in the British national health service. *Health Economics* **15**, 933–46.
Chamberlain, G. (1980) Analysis of covariance with qualitative data. *Review of Economic Studies* **47**, 225–38.

Chamberlain, G. (1984) Panel data. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics, Volume 1*. Amsterdam: Elsevier.
Chandra, A. and D. Staiger (2007) Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of Political Economy* **115**, 103–40.
Chang, F.-R. and P.K. Trivedi (2003) Economics of self medication: theory and evidence. *Health Economics* **12**, 721–39.
Chay, K.Y. and M. Greenstone (2003) The impact of air polution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* **118**, 1121–67.
Chen, L., W. Yip, M. Chang, H. Lin, S. Lee, Y. Chiu and Y. Lin (2007) The effects of Taiwan's national health insurance on access and health status on the elderly. *Health Economics* **26**, 223–42.
Chen, Y. and L.-A. Zhou (2007) The long term health and economic consequences of the 1959–1961 famine in China. *Journal of Health Economics* **26**, 659–81.
Chou, S.-Y. (2002) Asymmetric information, ownership and quality of care: an empirical analysis of nursing homes. *Journal of Health Economics* **21**, 293–311.
Chou, W.L. (2007) Explaining China's reginal health expenditures using LM-type unit root tests. *Journal of Health Economics* **26**, 682–98.
Clark, A. and F. Etilé (2002) Do health changes affect smoking? Evidence from British panel data. *Journal of Health Economics* **21**, 533–62.
Clark, A.E. and F. Etilé (2006) Don't give up on me baby: spousal correlation in smoking behaviour. *Journal of Health Economics* **25**, 958–78.
Clark, A., F. Etilé, F. Postel-Vinay, C. Senik and K. Van Der Straeten (2005) Heterogeneity in reported well-being: evidence from twelve European countries. *Economic Journal* **115**, C118-32.
Claxton, K., E. Fenwick and M.J. Sculpher (2006) Decision making with uncertainty: the value of information. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Cockx, B. and C. Brasseur (2003) The demand for physician services: evidence from a natural experiment. *Journal of Health Economics* **22**, 881–913.
Contoyannis, P., J. Hurley, P. Grootendorst, S.-H. Jeon and R. Tamblyn (2005) Estimating the price elasticity of expenditure for prescription drugs in the presence of non-linear price schedules: an illustration from Quebec, Canada. *Health Economics* **14**, 909–23.
Contoyannis, P., A.M. Jones and R. Leon-Gonzalez (2004a) Using simulation-based inference with panel data in health economics. *Health Economics* **13**, 101–22.
Contoyannis, P., A.M. Jones and N. Rice (2003) Simulation-based inference in dynamic panel probit models: an application to health. *Empirical Economics* **28**, 1–29.
Contoyannis, P., A.M. Jones and N. Rice (2004b) The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics* **19**, 473–503.
Contoyannis, P. and N. Rice (2001) The impact of health on wages: evidence from the British Household Panel Survey. *Empirical Economics* **26**, 599–622.
Conway, K.S. and P. Deb (2005) Is prenatal care really ineffective? Or, is the "devil" in the distribution? *Journal of Health Economics* **24**, 489–513.
Cooper, N.J., P.C. Lambert, K.R. Abrams and A.J. Sutton (2007) Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis. *Health Economics* **16**, 37–56.
Cowell, A.J. (2006) The relationship between education and health behaviour: some empirical evidence. *Health Economics* **15**, 125–46.
Currie, A., M. Shields, S. Price and Wheatley (2007) The child health/family income gradient: evidence from England. *Journal of Health Economics* **26**, 213–32.
Currie, J. and V.J. Hotz (2004) Accidents will happen? Unintentional childhood injuries and the effects of child care regulations. *Journal of Health Economics* **23**, 25–59.

Currie, J. and B. Madrian (1999) Health, health insurance and the labour market. In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*. Amsterdam: Elsevier.
Currie, J. and M. Neidell (2005) Air pollution and infant health: what can we learn from California's recent experience. *Quarterly Journal of Economics* **120**, 1003–30.
Currie, J. and M. Stabile (2006) Child mental health and human capital accumulation: the case of ADHD. *Journal of Health Economics* **25**, 1094–118.
Dano, A.M. (2005) Road injuries and long run effects on income and employment. *Health Economics* **14**, 955–70.
Das, J. and J. Hammer (2005) Which doctor? Combining vignettes and item response to measure clinical competence. *Journal of Development Economics* **78**, 348–83.
Das, M. and A. van Soest (1999) A panel data model for subjective information on household income growth. *Journal of Economic Behaviour and Organization* **40**, 409–26.
Davidoff, A., L. Blumberg and L. Nichols (2005) State health insurance market reforms and access to insurance for high-risk employees. *Journal of Health Economics* **24**, 725–50.
Dawson, D., H. Gravelle, R. Jacobs, S. Martin and P. Smith (2007) The effects of expanding patient choice of provider on waiting times: evidence from a policy experiment. *Health Economics* **16**, 113–28.
Deb, P. (2001) A discrete random effects probit model with application to the demand for preventive care. *Health Economics* **10**, 371–83.
Deb, P. and A.M. Holmes (2000) Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics* **9**, 475–89.
Deb, P., C. Li, P.K. Trivedi and D.M. Zimmer (2006a) The effect of managed care on use of health care services: results from two contemporaneous household surveys. *Health Economics* **15**, 743–60.
Deb, P., M.K. Munkin and P.K. Trivedi (2006b) Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics* **21**, 1081–99.
Deb, P. and P.K. Trivedi (1997) Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**, 313–36.
Deb, P. and P.K. Trivedi (2002) The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics* **21**, 601–25.
Deb, P. and P.K. Trivedi (2006) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econometrics Journal* **9**, 307–31.
Dee, T.S., D.C. Grabowski and M.A. Morrisey (2005) Graduated driver licensing and teen traffic fatalities. *Journal of Health Economics* **24**, 571–89.
Disney, R., C. Emmerson and M. Wakefield (2006) Ill health and retirement in Britain: a panel data-based analysis. *Journal of Health Economics* **25**, 621–49.
Doyle, J.J. (2005) Health insurance, treatment and outcomes: using auto accidents as health shocks. *Review of Economics and Statistics* **87**, 256–70.
Dranove, D., D. Kessler, M. McClellan and M. Satterthwaite (2003) Is more information better? The effects of "report cards" on health care providers. *Journal of Political Economy* **111**, 555–88.
Dranove, D. and R. Lindrooth (2003) Hospital consolidation and costs: another look at the evidence. *Journal of Health Economics* **22**, 983–97.
Dranove, D. and P. Wehner (1994) Physician-induced demand for childbirths. *Journal of Health Economics* **13**, 61–73.
Duflo, E. (2000) Child health and household resources in South Africa: evidence from the old age pension program. *American Economic Review* **90**, 393–8.
Dusheiko, M., H. Gravelle and R. Jacobs (2004) The effect of practice budgets on patient waiting times: allowing for selection bias. *Health Economics* **13**, 941–58.

Dusheiko, M., H. Gravelle, R. Jacobs and P. Smith (2006) The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment. *Journal of Health Economics* **25**, 449–78.

Dusheiko, M., H. Gravelle, N. Yu and S. Campbell (2007) The impact of budgets for gatekeeping physicians on patient satisfaction: evidence from fundholding. *Journal of Health Economics* **26**, 742–62.

Elliott, R.F., A.H.Y. Ma, A. Scott, D. Bell and E. Roberts (2007) Geographically differentiated pay in the labour market for nurses. *Journal of Health Economics* **26**, 190–212.

Etilé, F. (2006) Who does the hat fit? Teenager heterogeneity and the effectiveness of information policies in preventing cannabis use and heavy drinking. *Health Economics* **15**, 697–718.

Etilé, F. and C. Milcent (2006) Income-related reporting heterogeneity in self assessed health: evidence from France. *Health Economics* **15**, 965–81.

Evans, W.N. and D.S. Lien (2005) The benefits of prenatal care: evidence from the PAT bus strike. *Journal of Econometrics* **125**, 207–39.

Farsi, M. and G. Ridder (2006) Estimating the out-of-hospital mortality rate using patient discharge data. *Health Economics* **15**, 983–95.

Ferrer-i-Carbonell, A. and P. Frijters (2004) How important is methodology for the estimates of the determinants of happiness. *The Economic Journal* **114**, 641–59.

Finkelstein, A. (2004) The interaction of partial public insurance programs and residual private insurance markets: evidence from the US Medicare program. *Journal of Health Economics* **23**, 1–24.

Forster, M. and A.M. Jones (2001) The role of tobacco taxes in starting and quitting smoking: duration analysis of British data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **164**, 517–47.

French, E. (2005) The effects of health, wealth and wages on labour supply and retirement behaviour. *Review of Economic Studies* **72**, 395–427.

Frijters, P., J.P. Haisken-deNew and M.A. Shields (2005) The causal effect of income on health: evidence from German reunification. *Journal of Health Economics* **24**, 997–1017.

Frijters, P., M.A. Shields and S. Wheatley Price (2006) Investigating the quitting decision of nurses: panel data evidence from the British national health service. *Health Economics* **16**, 57–73.

Frolich, M., A. Heshmati and M. Lechner (2004) A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. *Journal of Applied Econometrics* **19**, 375–96.

Galiani, S., P. Gertler and E. Schargrodsky (2005) Water for life: the impact of the privatization of water services on child mortality. *Journal of Political Economy* **113**, 83–120.

Gallet, C.A. and J.A. List (2003) Cigarette demand: a meta-analysis of elasticities. *Health Economics* **12**, 821–53.

Gannon, B. (2005) A dynamic analysis of disability and labour force participation in Ireland 1995–2000. *Health Economics* **14**, 925–38.

García-Ferrer, A., A.D. Juan and P. Poncela (2007) The relationship between road traffic accidents and real economic activity in Spain: common cycles and health issues. *Health Economics* **16**, 603–26.

García-Gómez, P. and A. López-Nicolás (2006) Health shocks, employment and income in the Spanish labour market. *Health Economics* **15**, 997–1009.

Gardner, J. and A.J. Oswald (2007) Money and mental wellbeing: a longitudinal study of medium-sized lottery wins. *Journal of Health Economics* **26**, 49–60.

Geil, P., A. Million, R. Rotte and K.F. Zimmerman (1997) Economic incentives and hospitalization in Germany. *Applied Economics* **12**, 295–311.

Gemmill, M., J. Costa-Font and A. McGuire (2007) In search of a corrected prescription drug elasticity estimate: a meta-regression approach. *Health Economics* **16**, 627–43.

Gertler, P. (2004) Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review* **94**, 336–41.

Geweke, J., G. Gowrisankaran and R.J. Town (2003) Bayesian inference for hospital quality in a selection model. *Econometrica* **71**, 1215–38.
Gilleskie, D.B. and T.A. Mroz (2004) A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**, 391–418.
Girma, S. and D. Paton (2006) Matching estimates of the impact of over the counter emergency birth control on teenage pregnancy. *Health Economics* **15**, 1021–32.
Gravelle, H., M. Sutton, S. Morris, F. Windmeijer, A. Leyland, C. Dibben and M. Muirhead (2003) Modelling supply and demand influences on the use of health care: implications for deriving a needs based capitation formula. *Health Economics* **12**, 985–1004.
Groot, W. (2000) Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics* **19**, 403–20.
Gruber, J. and M. Owings (1996) Physician financial incentives and caesarean section delivery. *RAND Journal of Economics* **57**, 99–123.
Guaraglia, A. and M. Rossi (2004) Private medical insurance and saving: evidence from the British Household Panel Survey. *Journal of Health Economics* **23**, 761–83.
Hadley, J., D. Polsky, J.S. Mandelblatt, J.M. Mitchell, J.C. Weeks, Q. Wang and Y.-T. Hwang (2003) An exploratory instumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Economics* **12**, 171–86.
Hamilton, B.H., V. Ho and D.P. Goldman (2000) Queuing for surgery: is the US or Canada worse off? *Review of Economics and Statistics* **82**, 297–308.
Harris, M.N., P. Ramful and X. Zhao (2006) An ordered generalised extreme value model with application to alchohol consumption in Australia. *Journal of Health Economics* **25**, 782–801.
Harrison, T. (2007) Consolidations and closures: an empirical analysis of exits from the hospital industry *Health Economics* **16**, 457–74.
Hauck, K. and N. Rice (2004) A longitudinal analysis of mental health mobility in Britain. *Health Economics* **13**, 981–1001.
Heckman, J. and B. Singer (1984) A method of minimizing the distributional impact in econometric models for duration data. *Econometrica* **52**, 271–30.
Heckman, J. and E. Vitlacyl (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* **96**, 4730–34.
Heckman, J.J. and E.J. Vitlacyl (2007) Econometric evaluation of social programs, Part I: Causal models, structural models and economic policy evaluation. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics Volume 6*. Amsterdam: Elsevier.
Ho, K. (2006) The welfare effects of restricted hospital choice in the US medical care market. *Journal of Applied Econometrics* **21**, 1039–79.
Ho, V. (2002) Learning and the evolution of medical technologies: the diffusion of coronary angioplasty. *Journal of Health Economics* **21**, 873–85.
Hoch, J.S., A.H. Briggs and A.R. Willan (2002) Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**, 415–30.
Hogelund, J. and A. Holm (2006) Case management interviews and the return to work of disabled employees. *Journal of Health Economics* **25**, 500–19.
Holmas, T.H. (2002) Keeping nurses at work: a duration analysis. *Health Economics* **11**, 493–503.
Holmes, G. (2005) Increasing physician supply in medically underserved areas. *Labour Economics* **12**, 697–725.
Holmlund, H. (2005) Estimating long-term consequences of teenage childbearing. *Journal of Human Resources* **40**, 716–43.
Imbens, G.W. and J. Angrist (1994) Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–75.
Jalan, J. and M. Ravallion (2003) Does piped water reduce diarrhea for children in rural India? *Journal of Econometrics* **112**, 153–73.

Jensen, R.T. and K. Richter (2004) The health implications of social security failure: evidence from the Russian pension crisis. *Journal of Public Economics* **88**, 209–36.
Jewell, R.T. and P. Triunfo (2006) The impact of prenatal care on birthweight: the case of Uruguay. *Health Economics* **15**, 1245–50.
Jiménez-Martin, S., J.M. Labeaga and M. Martínez-Granado (2002) Latent class versus two-part models in the demand for physician services across the European Union. *Health Economics* **11**, 301–21.
Jochmann, M. and R. Leon-Gonzalez (2004) Estimating the demand for health care with panel data: a semiparametric Bayesian approach. *Health Economics* **13**, 1003–14.
Jones, A.M. (2000) Health econometrics. In A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics*. Amsterdam: Elsevier.
Jones, A.M., X. Koolman and E.V. Doorslaer (2007a) The impact of supplementary private health insurance on the use of specialists in European countries. *Annales d'Economie et de Statistiques* **83/84**, 251–75.
Jones, A.M., X. Koolman and N. Rice (2006) Health-related non-response in the British Household Panel Survey and European Community Household Panel: using inverse-probability-weighted estimators in non-linear models. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **169**, 543–69.
Jones, A.M. and J.M. Labeaga (2003) Individual heterogeneity and censoring in panel data estimates of tobacco expenditure. *Journal of Applied Econometrics* **18**, 157–77.
Jones, A.M., N. Rice, T. Bago d'Uva and S. Balia (2007b) *Applied Health Economics*. London: Routledge.
Jurges, H. (2007) True health vs response styles: exploring cross-country diffences in self-reported health. *Health Economics* **16**, 163–78.
Kan, H., D. Goldman, E. Keeler, N. Dhanani and G. Melnick (2003) An analysis of unobserved selcetion in an inpatient diagnostic cost group model. *Health Services and Outcomes Research Methodology* **4**, 71–91.
Kan, K. and W.-D. Tsai (2004) Obesity and risk knowledge. *Journal of Health Economics* **23**, 907–34.
Kerkhofs, M. and M. Lindeboom (1995) Subjective health measures and state dependent reporting errors. *Health Economics* **4**, 221–35.
Kessler, D.P. and M.B. McClellan (2002) How liability law affects medical productivity. *Journal of Health Economics* **21**, 931–55.
Kremer, M. (2003) Randomized evaluations of educational programs in developing countries: some lessons. *American Economic Review* **93**, 102–6.
Kyle, M. (2007) Pharmaceutical price controls and entry strategies. *Review of Economics and Statistics* **89**, 88–99.
Kyriazidou, E. (1997) Estimation of a panel data sample selection model. *Econometrica* **65**, 1335–64.
Lakdawalla, D., N. Sood and D. Goldman (2006) HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics* **121**, 1063–102.
Lalonde, R. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 604–20.
Lee, L.F. (1983) Generalised econometric models with selectivity. *Econometrica* **51**, 507–12.
Lee, M.-C. and A.M. Jones (2004) How did dentists respond to the introduction of global budgets in Taiwan? An evaluation using individual panel data. *International Journal of Health Care Finance and Economics* **4**, 307–26.
Lee, M.-C. and A.M. Jones (2006) Heterogeneity in dentists' activity in Taiwan: an application of quantile regression. *Empirical Economics* **31**, 151–64.
Leigh, A. and C. Jencks (2007) Inequality and mortality: long-run evidence from a panel of countries. *Journal of Health Economics* **26**, 1–24.
Levitt, S.D. and J. Porter (2001) Sample selection in the estimation of air bag and seat belt effectiveness. *Review of Economics and Statistics* **83**, 603–15.

Lindahl, M. (2005) Estimating the effect of income on health and mortality using lottery prizes as exogenous source of variation in income. *Journal of Human Resources* **40**, 144–68.
Lindeboom, M. (2006) Health and work of older workers. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Lindeboom, M., F. Portrait and G.J. Van den Berg (2002) An econometric analysis of the mental health effects of major events in the life of older individuals. *Health Economics* **11**, 505–20.
Lindeboom, M. and E. Van Doorslaer (2004) Cut-point shift and index shift in self-reported health. *Journal of Health Economics* **23**, 1083–99.
Lindrooth, R. and B. Weisbrod (2007) Do religious non-profit and for-profit organisations respond diffrently to financial incentives? The hospice industry. *Journal of Health Economics* **26**, 342–57.
Liu, Z., W.H. Dow and E.C. Norton (2004) Effect of drive-through delivery laws on postpartum length of stay and hospital charges. *Journal of Health Economics* **23**, 129–55.
Lleras-Muney, A. (2005) The relationship between education and adult mortality in the United States. *Review of Economic Studies* **72**, 189–221.
Lourenço, O.D. and P.L. Ferreira (2005) Utilization of public health centres in Portugal: effect of time costs and other determinants. Finite mixture models applied to truncated samples. *Health Economics* **14**, 939–53.
Machado, J.A.F. and J.M.C. Santos Silva (2005) Quantiles for counts. *Journal of the American Medical Association* **100**, 1226–37.
Manning, W.G. (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* **17**, 283–95.
Manning, W.G. (2006) Dealing with skewed data on costs and expenditure. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Manning, W.G., A. Basu and J. Mullahy (2005) Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**, 465–88.
Manning, W.G. and J. Mullahy (2001) Estimating log models: to transform or not to transform? *Journal of Health Economics* **20**, 461–94.
Manning, W., J.P. Newhouse, N. Duan, E. Keeler, A. Leibowitz and M.S. Marquis (1987) Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review* **77**, 251–77.
Marini, G., M. Miraldo, R. Jacobs and M. Goddard (2008) Giving greater financial independence to hospitals – does it make a difference? The case of English NHS trusts. *Health Economics* **17**(6), 751–75.
Martin, S., N. Rice, R. Jacobs and P. Smith (2007) The market for elective surgery: joint estimation of supply and demand. *Journal of Health Economics* **26**, 263–85.
McClellan, M., J.P. Newhouse and B. McNeil (1994) Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association* **272**, 859–66.
McGarry, K. (2004) Health and retirement. *Journal of Human Resources* **39**, 624–48.
Meer, J., D.L. Miller and H.S. Rosen (2003) Exploring the health–wealth nexus. *Journal of Health Economics* **22**, 713–30.
Mello, M.M., S.C. Stearns and E.C. Norton (2002) Do medicare HMOs still reduce health services use after controlling for selection bias? *Health Economics* **11**, 323–40.
Miguel, E. and M. Kremer (2004) Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**, 159–217.
Morris, S. (2006) Body mass index and occupational attainment. *Journal of Health Economics* **25**, 347–64.
Morris, S. (2007) The impact of obesity on employment. *Labour Economics* **14**, 413–33.
Moscone, F., M. Knapp and E. Tosetti (2007) Mental health expenditure in England: a spatial panel approach. *Journal of Health Economics* **26**, 842–64.

Mroz, T.A. (1999) Discrete factor approximations in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome. *Journal of Econometrics* **92**, 233–74.

Mullahy, J. (1998) Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* **17**, 247–81.

Mundlak, Y. (1978) On the pooling of time series and cross section data. *Econometrica* **46**, 69–85.

Munkin, M.K. and P.K. Trivedi (1999) Simulated maximum likelihood estimation of multivariate mixed-poisson regression models, with application. *Econometrics Journal* **2**, 29–48.

Murray, C.J.L., A. Tandon, J. Salomon and C.D. Mathers (2001) Enhancing cross-population comparability of survey results. *GPE Discussion Paper No. 35*. Geneva: World Health Organization.

Nicoletti, C. and F. Peracchi (2005) Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **168**, 763–81.

Nolan, A. (2007) A dynamic analysis of GP visiting in Ireland 1995–2001. *Health Economics* **16**, 129–43.

Or, Z., J. Wang and D. Jamison (2005) International differences in the impact of doctors on health: a multilevel analysis of OECD countries. *Journal of Health Economics* **24**, 531–60.

Paton, D. (2002) The economics of family planning and underage conceptions. *Journal of Health Economics* **21**, 207–25.

Picone, G.A., F.A. Sloan, S.-Y. Chou and D.H.T. Jr (2003b) Does higher hospital cost imply higher quality of care? *Review of Economics and Statistics* **85**, 51–62.

Picone, G., R.M. Wilson and S.-Y. Chou (2003a) Analysis of hospital length of stay and discharge destination using hazard functions with unmeasured heterogeneity. *Health Economics* **12**, 1021–34.

Pop-Eleches, C. (2006) The impact of an abortion ban on socioeconomic outcomes of children: evidence from Romania. *Journal of Political Economy* **114**, 744–73.

Powell, J. (1986) Symmetrically trimmed least squares estimators for Tobit models. *Econometrica* **54**, 1435–60.

Prieger, J.E. (2002) A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics* **17**, 367–92.

Propper, C., S. Burgess and K. Green (2004) Does competition between hospitals improve the quality of care? Hospital death rates and the NHS internal market. *Journal of Public Economics* **88**, 1247–72.

Propper, C., B. Croxson and A. Shearer (2002) Waiting times for hospital admissions: the impact of GP fundholding. *Journal of Health Economics* **21**, 227–52.

Propper, C., J. Eachus, P. Chan, N. Pearson and G.D. Smith (2005) Access to health care resources in the UK: the case of care for arthritis. *Health Economics* **14**, 391–406.

Pudney, S. and M. Shields (2000a) Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics* **15**, 367–99.

Pudney, S. and M.A. Shields (2000b) Gender and racial discrimination in pay and promotion for NHS nurses. *Oxford Bulletin of Economics and Statistics* **62**, 801–35.

Quinn, C. (2005) Generalisable regression methods for cost-effectiveness using copulas. *Health, Econometrics and Data Group Working Paper WP 05/13*, University of York.

Raikou, M. and A. McGuire (2004) Estimating medical care costs under conditions of censoring. *Journal of Health Economics* **23**, 443–70.

Raikou, M. and A. McGuire (2006) Estimating costs for economic evaluation. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Rettenmaier, A.J. and Z. Wang (2006) Persistence in Medicare reimbursements and personal medical accounts. *Journal of Health Economics* **25**, 39–57.

Rice, N., P. Dixon, D. Lloyd and D. Roberts (2000) Derivation of a needs based capitation formula of allocating prescribing budgets to health authorities and primary care groups in England: regression analysis. *British Medical Journal* **320**, 284–88.
Riphahn, R.T., A. Wambach and A. Million (2003) Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics* **18**, 387–405.
Robinson, P. (1998) Root-N consistent semiparametric regression. *Econometrica* **56**, 931–54.
Rosenbaum, P.R. and D.B. Rubin (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Rous, J.J. and D.R. Hotchkiss (2003) Estimation of the determinants of household health care expenditures in Nepal with controls for endogenous illness and provider choice. *Health Economics* **12**, 431–51.
Royalty, A.B. and J.M. Abraham (2006) Health insurance and labor market outcomes: joint decision making within households. *Journal of Public Economics* **90**, 1561–77.
Ruhm, C.J. (2003) Good times make you sick. *Journal of Health Economics* **22**, 637–58.
Ryan, M., K. Gerard and G. Currie (2006) Using discrete choice experiments in health economics. In A.M. Jones (ed.), *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.
Sadana, R., C.D. Mathers, A.D. Lopez, C.J.L. Murray and K. Iburg (2000) Comparative analysis of more than 50 household surveys on health status. *GPE Discussion Paper No. 15.* Geneva: World Health Organization.
Sahn, D.E., S.D. Younger and G. Genicot (2003) The demand for health care services in rural Tanzania. *Oxford Bulletin of Economics and Statistics* **65**, 241–60.
Santos Silva, J.M.C. and F. Windmeijer (2001) Two-part multiple spell models for health care demand. *Journal of Econometrics* **104**, 67–89.
Sarma, S. and W. Simpson (2006) A microeconometric analysis of Canadian health care utilization. *Health Economics* **15**, 219–39.
Sasso, A.T.L. and T.C. Buchmueller (2004) The effect of the State Children's Health Insurance Program on health insurance coverage. *Journal of Health Economics* **23**, 1059–82.
Schellhorn, M. (2001) The effect of variable health insurance deductibles on the demand for physician visits. *Health Economics* **10**, 441–56.
Schmidt, L. (2007) Effects of infertility insurance mandates on fertility. *Journal of Health Economics* **26**, 413–46.
Seshamani, M. and A. Gray (2004) Ageing and health care expenditure: the red herring argument revisited. *Health Economics* **13**, 303–14.
Sloan, F.A., G.A. Picone, D.H. Taylor and S.-Y. Chou (2001) Hospital ownership and cost and quality of care: is there a dime's worth of difference? *Journal of Health Economics* **20**, 1–21.
Smith, J.A. and P.E. Todd (2001) Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* **91**, 112–18.
Smith, M.D. (2003) Modelling sample selection using Archimedean copulas. *Econometrics Journal* **6**, 99–123.
Smith, P.C. and A. Street (2005) Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **168**, 401–17.
Stewart, J.M. (2001) The impact of health status on the duration of unemployment spells and the implications for studies of the impact of unemployment on health status. *Journal of Health Economics* **20**, 781–96.
Tamm, M., H. Tauchmann, J. Wasem and S. Gress (2007) Elasticities of market shares and social health insurance choice in Germany: a dynamic panel data approach. *Health Economics* **16**, 243–56.
Terza, J.V. (2002) Alcohol abuse and employment: a second look. *Journal of Applied Econometrics* **17**, 393–404.
Trivedi, P.K. and D.M. Zimmer (2005) Copula modelling: an introduction for practitioners. *Foundations and Trends in Econometrics* **1**, 1–111.
Van Den Berg, G.J., M. Lindeboom and F. Portrait (2006) Economic conditions early in life and individual mortality. *American Economic Review* **96**, 290–302.

Van Houtven, C.H. and E.C. Norton (2004) Informal care and health care use of older adults. *Journal of Health Economics* **23**, 1159–80.
Van Ours, J.C. (2004) A pint a day raises a man's pay; but smoking blows that gain away. *Journal of Health Economics* **23**, 863–86.
Van Ours, J.C. (2006) Dynamics in the use of drugs. *Health Economics* **15**, 1283–94.
Van Ourti, T. (2004) Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model. *Health Economics* **13**, 705–24.
Vera-Hernandez, M. (2003) Structural estimation of a principal agent model: moral hazard in medical insurance. *RAND Journal of Economics* **34**, 670–93.
Wagstaff, A. (2007) The economic consequences of health shocks: evidence from Vietnam. *Journal of Health Economics* **26**, 82–100.
Wagstaff, A. and S. Yu (2007) Do health sector reforms have their intended impacts? The World Bank's Health VIII project in Gansu province, China. *Journal of Health Economics* **26**, 505–35.
Wang, Z. and A. Rettenmaier (2007) A note on cointegration of health expenditures and income. *Health Economics* **16**, 599–78.
Wildman, J. and A.M. Jones (2007) Disentangling the relationship between health and income. *Journal of Health Economics* **17**(2), 249–65.
Willan, A.R., A.H. Briggs and J.S. Hoch (2004) Regression methods for covariate adjustment and subgroup analysis for non-censored cost effectiveness data. *Health Economics* **13**, 461–75.
Wilson, P.W. and K. Carey (2004) Nonparametric analysis of returns to scale in the US hospital industry. *Journal of Applied Econometrics* **19**, 505–24.
Windmeijer, F., H. Gravelle and P. Hoonhout (2005) Waiting lists, waiting times and admissions: an empirical analysis at hospital and general practice levels. *Health Economics* **14**, 971–85.
Winkelmann, R. (2004a) Health care reform and the number of doctor visits–an economic analysis. *Journal of Applied Econometrics* **19**, 455–72.
Winkelmann, R. (2004b) Co-payments for prescription drugs and the demand for doctor visits–evidence from a natural experiment. *Health Economics* **13**, 1081–9.
Winkelmann, R. (2006) Reforming health care: evidence from quantile regressions for counts. *Journal of Health Economics* **25**, 131–45.
Wolfe, B., T. Kaplan, R. Haveman and Y. Cho (2006) SCHIP expansion and parental coverage: an evaluation of Wisconsin's BadgerCare. *Journal of Health Economics* **25**, 1170–92.
Wooldridge, J. (2005) Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**, 39–54.
Yelowitz, A.S. (2000) Public policy and health insurance choices of the elderly: evidence from the Medicare buy-in program. *Journal of Public Economics* **78**, 301–24.
Yen, S.T., C.-H. Tang and S.-J.B. Su (2001) Demand for traditional medicine in Taiwan: a mixed Gaussian–Poisson model approach. *Health Economics* **10**, 221–32.
Zimmer, D.M. and P.K. Trivedi (2006) Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics* **24**, 63–76.

13

Panel Methods to Test for Unit Roots and Cointegration

Anindya Banerjee and Martin Wagner

Abstract

We provide an up-to-date analytical survey of methods which have been developed to deal with estimation and inference in non-stationary panels. The chapter provides information not only on the tools but also interprets the literature and highlights the important challenges that remain. We discuss the difficulties involved in formulating hypotheses within a panel framework with unit roots and cointegration. These issues include incorporating cross-sectional dependence and structural breaks in the data. Both these features are widely prevalent in the panels and lead to complications in estimation and inference. For example, factor models are a widely used class of methods used to deal with dependence but constitute only one of several ways of formulating the problems involved. We argue that the links between cointegration and factor models in panels need to be considered adequately and the asymptotic theory put on a firmer footing in many respects. The study of cross-sectional dependence, breaks, and multiple cointegrating vectors, all of which are in their relative infancy, mark the way for productive research in the years ahead.

13.1 Introduction

Taking a look at the web page of the Groningen Growth Development Centre (http://www.ggdc.net) is a salutary experience. The Centre is devoted to the "comparative analysis of levels of economic performance and differences in growth rates in the world," both in the short and long durée, and the data resources it brings to this analysis are formidable.

Here in its pages are data for population, employment, annual working hours, gross domestic product (GDP) per capita, GDP per person engaged, and GDP per hour, for nearly 125 countries from 1950 onwards. Many of these series utilize the Maddison (2007) historical series, which can also be used to go back to the nineteenth century and in some cases even further into the past.

The 60-Industry Database provides an equally comprehensive dataset on industrial performance. The coverage is at a detailed industry level for Organization for Economic Cooperation and Development (OECD) countries and Taiwan, and the variables studied include: value added in current and constant prices, value added deflators, persons engaged, hours worked and labor productivity. The data cover industries for the period 1979–2003.

Finally, the Industry Growth Accounting Database provides information on labor skills in three categories and on investment and capital services (three information and communications technologies (ICT) asset categories and three non-ICT categories). The countries covered are Australia, Canada, the United States and four European countries (France, Germany, the Netherlands and the United Kingdom). The information is provided in order to allow for a decomposition of output growth into the contributions of labor and capital and total factor productivity using the growth-accounting methodology.

This chapter deals with particular aspects of the study of such large economic datasets, with a view to using this data to analyze economic hypotheses of interest.

13.1.1 An example: economic convergence in the sense of Evans and Karras (1996)

The available large datasets, containing income and other macroeconomic variables, lead us to choose economic convergence as a motivating example. This allows us to discuss some of the possibilities that macro-panel data offer and some issues that arise.[1]

There is an abundance of definitions of convergence in the literature. Given that we consider panel data, we focus on those put forward in Evans and Karras (1996) (abbreviated henceforth as EK). These essentially coincide with the definitions of Bernard and Durlauf (1995, 1996), formulated within a time series context.[2] Three features of EK are of interest to us, within the context of this chapter. First, that the paper studied convergence within the framework of a dataset in which the cross-section dimension N and time series dimension T were both used (in contrast to using only the T dimension and investigating the hypothesis on a country-by-country basis or only along the cross-sectional dimension). Second, the data could be taken to be integrated of order one (or, loosely speaking, needing first-differencing for stationarity). And third, and most importantly, the hypothesis of interest could be formulated in terms of testing for a unit root, in this case at least, in an autoregressive model. This threefold combination of a so-called macro-panel of data (involving gathering together time series information on a variable or a set of variables with the cross-sectional or cross-country information on these variables) with the problem of testing for a non-stationary root, where the hypothesis is formulated as such, gives our chapter its name and decides its focus.

Thus, denote log per capita GDP by $y_{i,t}$, referred to for simplicity as income, in country i in year t, considered to be valued at constant and common international prices. This variable is observed for a collection of $i = 1, \ldots, N$ countries over the years $t = 1, \ldots, T$. Several underlying theoretical formulations of growth models lead to balanced growth paths for each of the economies to which each of the economies converges in the long run. Under some additional assumptions the balanced growth paths of the economies are parallel to each other. Such a set-up is the starting point of EK's statistical definition of economic convergence, which we give below:

Definition (general): Denote by $y_{i,t+j}$ income in country i at time $t + j$ and assume that there exists a process a_t and finite parameters $\mu_1, \ldots, \mu_N$ for which it holds

that $\lim_{j\to\infty} E\left(y_{i,t+j} - a_t\right) = \mu_i$ for all $i = 1, \ldots, N$. Then the economies are said to converge. Convergence is called conditional if not all $\mu_1, \ldots, \mu_N$ are equal to 0 and is called absolute if all $\mu_1, \ldots, \mu_N$ are equal to 0.

This definition of convergence *potentially* leads to several interesting implications for non-stationary panel analysis since it presumes the existence of one joint trend process a_t (which is typically unobservable and is related to technology), such that the limits of the expected values of the deviations from this trend exist and are constant. Given that income is often found to be a unit root non-stationary process, this allows us to reduce or specialize the general formulation given in the definition above to an $I(1)$ context.

Thus, we consider henceforth the income series in the N countries to be jointly described by a vector $I(1)$ process.[3] In this case the EK definition of convergence implies that the deviations from the trend process a_t are asymptotically stationary. It implies furthermore that the cross-section members will not be independent of each other, given their relation to the single common trend a_t. Clearly, if we assume that the income series are $I(1)$, but the deviations from a_t are stationary, this implies that a_t is also an $I(1)$ process. Thus the panel exhibits within this framework cross-sectional dependence via the stochastic component of the common trend a_t. Cross-sectional dependence is discussed in detail in Appendix B. We can already foresee, given that we formulate the discussion here within an $I(1)$ modeling framework, that the specific formulation of the convergence definition of EK has strong cointegration implications which we detail next.

If we keep, for the moment, the cross-sectional dimension as fixed, the joint vector process $y_t = \left(y_{1,t}, \ldots, y_{N,t}\right)'$ is – under appropriate assumptions – also jointly an $I(1)$ vector process and thus has a Granger-type representation, where for simplicity we abstract from detailing the initial values and their effects, given by:

$$\begin{bmatrix} y_{1,t} \\ \vdots \\ y_{N,t} \end{bmatrix} = \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix} \eta_t + \begin{bmatrix} D_1 \\ \vdots \\ D_N \end{bmatrix} T_t + c^*(L)\varepsilon_t, \tag{13.1}$$

with $\eta_t \in R^r$ the $r \geq 1$ linearly independent common stochastic trends, $T_t \in R^s$ the deterministic components of the data-generation process (DGP) and $c^*(L)\varepsilon_t$ the stationary part, with L denoting the backward shift operator, that is, $L\left(x_t\right)_{t\in Z} = \left(x_{t-1}\right)_{t\in Z}$ and $c^*(L) = \sum_{j=0}^{\infty} c_j^* L^j$ such that $c^*(L)\varepsilon_t$ is a stationary process.

In particular, in an $I(1)$ setting, the EK definition implies that the cointegrating space for the N-dimensional vector of income series is of dimension $N - 1$. This follows immediately from the fact that any pair-wise difference of income between countries in which the unit root process a_t is annihilated is stationary. Note here that this already implies that all non-constant deterministic components are also annihilated by taking pair-wise differences. Consequently, the definition not only specifies the dimension of the cointegrating space but also fully determines the space itself (that is, a basis of the cointegrating space). Note, furthermore, that by the same argument it is also true that the deviations of any individual income

series from the cross-sectional average are stationary, that is, any of the N series $y_{i,t} - \bar{y}_t$ with $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{i,t}$ is also stationary, with mean 0 in the case of absolute convergence. Thus, the definition of convergence in the $I(1)$ context of EK is given by:

Definition (I(1) framework): If each income series $y_{i,t}$ is described by an $I(1)$ process, but all series $y_{i,t} - \bar{y}_t$ are stationary, the set of N economies is said to converge. Convergence is said to be absolute if the means of all the series $y_{i,t} - \bar{y}_t$ are equal to 0 and relative otherwise. The economies are said to diverge if all series $y_{i,t} - \bar{y}_t$ are non-stationary.[4]

Given our Granger representation in (13.1) above, we may ask about the restrictions imposed by this representation, that is, under what conditions are all series $y_{i,t} - \bar{y}_t$ stationary? Computing the cross-sectional average, we obtain:

$$\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} c_{i,1}\eta_{1,t} + \cdots + \frac{1}{N}\sum_{i=1}^{N} c_{i,r}\eta_{r,t} + \frac{1}{N}\sum_{i=1}^{N} d_{i,1}T_{1,t} + \cdots + \frac{1}{N}\sum_{i=1}^{N} d_{i,s}T_{s,t} + \bar{c}^*(L)\varepsilon_t, \tag{13.2}$$

where we use the notation $C_i = \left[c_{i,1}, \ldots, c_{i,r}\right]$ and $D_i = \left[d_{i,1}, \ldots, d_{i,s}\right]$, and $\bar{c}^*(L)\varepsilon_t = \frac{1}{N}\sum_{i=1}^{N} c_i^*(L)\varepsilon_{i,t}$ is stationary.

Now, in order to have convergence in the EK sense, each of the deviations of the income series from the cross-sectional average has to be stationary. This necessitates that all stochastic trends have to be annihilated – as well as all non-constant deterministic components. Consequently, for all $i = 1, \ldots, N$, it has to hold that:

$$c_{i,k} - \frac{1}{N}\sum_{i=1}^{N} c_{i,k} = 0, \tag{13.3}$$

which implies that $c_{i,k} = c_k$ for all $i = 1, \ldots, N$ and for all $k = 1, \ldots, r$. This in turn implies that only one common stochastic $I(1)$ trend, given by $\eta_t = \sum_{k=1}^{r} c_k \eta_{k,t}$, is present and identifiable in the data.[5] This is, of course, not a surprise given that the EK definition implies the presence of $N - 1$ cointegrating relationships (hence of only one stochastic trend), which is furthermore loaded with the same weight into each series. Similar arguments apply to all non-constant elements in the deterministic component: the coefficient to each non-constant deterministic variable has to be equal for each series; thus, for example, if present in the data, all linear trend slopes need to be identical for relative convergence to prevail. For absolute convergence to hold, in addition, all intercepts need to coincide.

Thus, we see that the EK definition of convergence imposes a lot of structure and hence testable restrictions on the panel of time series. Furthermore, as mentioned previously, it also implies that there is long-run cross-sectional dependence in the panel as defined in Appendix B.

It is worth observing at this stage that if one had enough observations to perform a multivariate time series analysis, in relation to the considerations above, one

would not need to resort to panel techniques. However, this is typically not the case and we therefore turn more specifically to the panel (unit root) implications of the convergence definition. In the case of convergence, each of the series $y_{i,t} - \overline{y}_t$ is stationary and divergence has been defined by EK as (unit root) non-stationarity of all the N series $y_{i,t} - \overline{y}_t$. Thus, the null hypothesis of convergence can be tested by a panel stationarity test, as presented, for example, in section 13.2.1.4. More often, however, the null hypothesis of divergence is tested against the alternative of convergence by using panel unit root tests that are specified against the alternative that all series are stationary. Under appropriate assumptions on the DGP, this can be done within panel Dickey–Fuller-type regressions of the form (with individual-specific autoregressive orders p_i):

$$\Delta\left(y_{i,t} - \overline{y}_t\right) = \delta_i + \rho_i\left(y_{i,t-1} - \overline{y}_{t-1}\right) + \sum_{k=1}^{p_i} \varphi_{i,k}\Delta\left(y_{i,t-k} - \overline{y}_{t-k}\right) + u_{i,t}, \qquad (13.4)$$

with $u_{i,t}$ denoting here the residual process of the autoregression. When the cross-sectionally demeaned data are described by an autoregression of the corresponding order, the processes $u_{i,t}$ are white-noise processes. Note here that EK, despite testing the null hypothesis of divergence, proposed such testing in a Dickey–Fuller-type regression including only intercepts and no other deterministic components, which are in fact not restricted under the divergence hypothesis by EK. Thus, in effect, what EK proposed was to test the null hypothesis of divergence with respect to the stochastic trend whilst allowing only for individual specific intercepts and linear trends. This restriction of the deterministic components needs to be investigated statistically, not least to prevent misspecification of the panel Dickey–Fuller-type regression above.

With the specified restriction, the null hypothesis of divergence is given by $H_0 : \rho_i = 0$ for all $i = 1, \ldots, N$ against the alternative hypothesis $H_A : \rho_i < 0$ for all $i = 1, \ldots, N$, where, as is usual in unit root testing, we really consider only stationary alternatives, which imposes restrictions on ρ_i under the alternative that depend upon the unknown coefficients $\varphi_{i,k}$ in case the data are really described by an autoregression and $u_{i,t}$ is white noise. In general there are no simple links between the innovations ε_t from the Granger representation above and the univariate regression errors $u_{i,t}$, with the latter being a function of the former. Generally, unless very strong restrictions are imposed, the errors $u_{i,t}$ will be cross-sectionally dependent. Issues are, however, even more problematic when testing the null hypothesis of divergence. Under divergence, when there are no restrictions on the stochastic trends present in the panel, the deviations from the cross-sectional averages will also in general be linked by stochastic trends, and these deviations can thus be cross-unit cointegrated.[6] Thus, the discussion here highlights strongly that in general we will need to consider panel unit root (and to a certain extent panel stationarity) tests that allow for cross-sectional dependencies, where in particular cross-sectional dependencies may enter via the errors or may arise via common factors. This is a theme to which we shall return in detail during the course of our chapter.

Clearly, if the panel is characterized by several independent stochastic trends, simple cross-sectional demeaning will not be able to remove all these non-stationary components. Allowing for such a situation – bearing in mind that a full systems analysis will typically not be feasible due to data limitations – is one of the major motivations for considering so-called factor models in the non-stationary panel literature. Factor models, by assuming that the data are generated by the sum of two components, common factors and idiosyncratic components, allow for modeling by putting restrictions on the spectrum of the jointly stationary process Δy_t.[7]

To explain further, let us consider the simplest case in relation to our convergence discussion. Ignoring deterministic components for simplicity, assume that each income series can be written as $y_{i,t} = a_t + \nu_{i,t}$, where the processes a_t and $\nu_{i,t}$ are all independent, and both components are either stationary or $I(1)$ processes. Due to the assumption that a_t and $\nu_{i,t}$ are independent, all series $y_{i,t}$ are also either stationary or integrated. The deviations from the cross-sectional averages are given by $\tilde{\nu}_{i,t} = \nu_{i,t} - \frac{1}{N}\sum_{j=1}^{N} \nu_{j,t} = \left(1 - \frac{1}{N}\right)\nu_{i,t} - \frac{1}{N}\sum_{j=1, j\neq i}^{N} \nu_{j,t}$. Now, in the case of convergence, the second term in this equation converges to 0 (under appropriate technical conditions); whereas in the case of divergence, the second term does not converge to 0. This shows that even if we start with cross-sectionally independent processes $\nu_{i,t}$, the processes $\tilde{\nu}_{i,t}$ describing the deviations from the cross-sectional averages are asymptotically independent as $N \to \infty$ only when convergence prevails, but need not be even asymptotically independent in the case of divergence.

The above example shows that, in general, if one is confronted with unobserved factors, cross-sectional averaging will not necessarily lead to asymptotically cross-sectionally independent series by studying the deviations from the cross-sectional averages. In our example above, due to the unobserved factor a_t being *loaded* in all series with the same weight, considering the deviations from the cross-sectional averages in fact allows elimination of this common factor. However, dependencies are introduced via the series $\tilde{\nu}_{i,t}$ that persist when the series are integrated. These remarks all hold similarly in cases where more than one common factor is considered or the factors do not enter all series with the same weights.

Therefore, when focusing only on the set of joint $I(1)$ processes that can be characterized by factor models, the need for appropriate estimation and inference techniques arises. Such techniques should deliver, again formulated for the convergence example considered here, the following: first, they need to construct consistent estimates of the common factor a_t as well as tests for stationarity (respectively unit root) behavior of this series. Second, the procedures need to establish panel unit root tests for the de-factored series, $y_{i,t} - \hat{a}_t$, where the circumflex indicates that only estimates of the unobservable factor are available. With these tools, convergence in the sense of EK can be tested by establishing first that a_t is an integrated process, and all series $\tilde{\nu}_{i,t}$ are stationary. Essentially these tools, allowing for multiple factors and heterogeneous loadings, have been developed by Bai and Ng (2004).

Note, furthermore, that for many empirical applications it may be necessary to also allow for structural change, which is most often modeled in the deterministic component of the processes. In this chapter we put particular emphasis on structural change and present some evidence that allowing for structural change may alter conclusions drastically. This fact is to a certain extent well-established in the time series unit root literature, starting with Perron (1989), but has found less attention in the non-stationary panel literature.

In section 13.2 we start with a general formulation of testing for a unit root in a panel data setting. This is first specialized to the case of panels with cross-sectionally independent members, but is then extended to a discussion of a more general framework, considering both cross-sectional dependence and structural change. Several sub-sections discuss the results of simulation and empirical studies to evaluate the properties of the tests devised for unit roots. As mentioned above, we illustrate the methods by an empirical study of purchasing power parity and the environmental Kuznets curve. Section 13.3 offers an analogous discussion of testing and estimation for cointegration, and as an additional empirical study, contains an analysis of exchange rate pass-through in the euro-area. In this application issues such as structural stability of the import pass-through equations (in the face of policy changes in the euro-area) are studied with reference to the core problem of testing for cointegration. Section 13.4 concludes.

Three appendices follow the main text. In Appendix A we collect some details on the datasets employed in this chapter; Appendix B contains a brief discussion on cross-sectional dependence; and Appendix C mentions a few aspects with respect to limit theory in non-stationary panels.

It is helpful, perhaps, before starting on the main account, to discuss here briefly how our study differs from some of the excellent studies which are already available; see, for example, Breitung and Pesaran (2008) or the special issue of the *Journal of Applied Econometrics* (2007) devoted to the topic of heterogeneity and cross-sectional dependence in panel data models, or more particularly the chapter by Choi (2006b) in the first volume of this handbook. Some of the material which appears in those sources finds some repetition here, since many of the themes which we deal with cannot be presented without context or introduction. Nevertheless, we believe our chapter contributes to the literature in three distinct and important ways. First, it presents a unified and general formulation of testing for unit roots and cointegration in panel data. Second, in doing so, almost uniquely this chapter pays considerable attention to the role of structural change in panel unit root and cointegration analysis. Structural change has been found to be a substantial issue in the time series context and continues to be so in the panel setting. Finally, by presenting a range of evidence on the performance of the methods, based on both simulation and empirical evidence using a variety of data sources, this chapter attempts to demonstrate unit root and cointegration analysis in panels in action. The material collected in Appendices B and C will also be of some interest by adding to the discussion both on modeling and accounting for dependence and on some key concepts of limit theory for integrated panels.

13.2 Unit root analysis in non-stationary panels

In order to highlight the potential advantages that panel data offer, some appropriate assumptions have to be made. Without any restrictions non-stationary panel data, or, to be more precise, integrated panel data, can be written as:

$$y_{i,t} = D_{i,t} + u_{i,t},$$

where $D_{i,t}$ denotes the deterministic part of the process generating the data, $u_{i,t}$ is the stochastic part, $i = 1, \ldots, N$ is the cross-sectional index and $t = 1, \ldots, T$ is the time index. Now, without any restrictions on the joint stochastic behavior of the $u_{i,t}$ series, no gains from pooling the data (that is, the cross-sectional and time series information), for example by constructing a test statistic, can be expected. Essentially, for the approach to be valuable, parsimonious representations of the joint DGP of the processes $u_{i,t}$ are required. The most parsimonious is given by assuming that the stochastic components $u_{i,t}$ are cross-sectionally independent, which is – especially for macroeconometric questions – too strong an assumption. Thus, in order to allow for cross-sectional dependence in a parsimonious way, so-called (approximate) factor models have gained popularity. These model the individual series $u_{i,t}$ as the sum of two components, namely some common factors F_t and an idiosyncratic component $e_{i,t}$, thus arriving at $u_{i,t} = \pi_i' F_t + e_{i,t}$ with π_i denoting the so-called factor loadings. In this set-up the panel unit root testing problem can formulated for the following data-generating process:[8]

$$y_{i,t} = D_{i,t} + \pi_i' F_t + e_{i,t} \tag{13.5}$$

$$(1 - L)F_t = C(L)\eta_t \tag{13.6}$$

$$(1 - \varphi_i L)e_{i,t} = H_i(L)\varepsilon_{i,t}, \tag{13.7}$$

observed for $t = 1, 2, \ldots, T, i = 1, 2, \ldots, N$, with $C(L) = \sum_{j=0}^{\infty} C_j L^j$ and $H_i(L) = \sum_{j=0}^{\infty} H_{i,j} L^j$. In the general case, F_t is an $r \times 1$ vector of "common factors," generated by some multivariate white noise process η_t, which accounts for the dependence that exists across the units of the panel. The unit-specific idiosyncratic terms $e_{i,t}$ are sometimes considered to be cross-sectionally independent, under the assumption that all cross-sectional dependencies are captured by the common factors. In case the idiosyncratic components are not assumed to be cross-sectionally independent, the above model is known as an approximate factor model. For the statistical analysis of the model as outlined above, appropriate assumptions have to be made both for identification as well as to establish the asymptotic behavior of estimators and test statistics. We discuss one set of possible assumptions when describing the method of Bai and Ng (2004) in section 13.2.2.2.

The properties of the functions $C(L)$ and $H_i(L)$ (see the detailed assumptions in section 13.2.2.2) determine the time series properties of the F and e series. For example, if $C(1) = 0$, F_t will be $I(0)$. If, on the other hand, $C(1)$ is of full rank, F_t is an $I(1)$ process composed of r linearly independent stochastic trends. For intermediate

ranks of $C(1)$ the process F_t is a cointegrated $I(1)$ process with the corresponding number of cointegrating relationships and common trends.

One of the key issues to consider is that both the common part of the processes (modeled via the factors) and the idiosyncratic parts are allowed to be integrated or stationary. This implies that determining the time series properties of the $y_{i,t}$ series requires not only consistent estimation but also inference on the properties of the constituent parts. We may, for example, think of these processes as being integrated "unconditionally" if either the factors or the idiosyncratic parts (or both) are integrated of order one, while conditional on taking account of the factors (that is, the dependence across the units of the panel), the $y_{i,t} - \hat{\pi}_i' \hat{F}_t$ series may be tested for integration or stationarity. The circumflexes above the factor loadings and the factors indicate that, for testing, estimates of these unobserved quantities have to be obtained. Note also that when we talk about stationarity of series, we consider stationarity of the stochastic part of the series, whilst also allowing for the presence of deterministic components $D_{i,t}$.

In general, cross-sectional dependence can arise through the unit-specific idiosyncratic terms as well as through the common factors. Both of these channels in general may lead to cointegration between the cross-section members. To make the discussion more precise, we define the concepts of short- and long-run cross-sectional dependence, as well as of cross-unit cointegration, in detail in Appendix B. There we also discuss in some detail the cointegration implications of the (approximate) factor model. For example, the papers of Banerjee, Marcellino and Osbat (2004, 2005) have highlighted that failing to account properly for cross-unit cointegration may severely distort panel cointegration testing. Further simulation evidence in this respect is contained in Wagner and Hlouskova (2007).

Formulating the discussion in terms of (13.5)–(13.7) above allows us to highlight specific important aspects of the theory of testing for unit roots in panels, summarized as follows.

Deterministics and breaks

For example, a typical formulation of the deterministic part of the process in unit root and cointegration analysis is to consider:

$$D_{i,t} = \mu_i + \delta_i t, i = 1, 2, \ldots, N, \tag{13.8}$$

where the i index denotes unit specificity of the deterministic components. This formulation allows for deterministic unit specific intercepts and trend growth rates, but a generalization allows for breaks in intercept and trend. For example, allowing for up to l_i breaks in the intercept and m_i breaks in trend in unit i gives:

$$D_{i,t} = \mu_i + \delta_i t + \sum_{j=1}^{l_i} \theta_{i,j} DU_{i,j,t} + \sum_{k=1}^{m_i} \gamma_{i,k} DT^*_{i,k,t}, i = 1, 2, \ldots, N, \tag{13.9}$$

where the dummy variables are defined as $DU_{i,j,t} = 1$ for $t > T^i_{a,j}$ and 0 elsewhere, and $DT^*_{i,k,t} = (t - T^i_{b,k})$ for $t > T^i_{b,k}$ and 0 elsewhere. In other words, $T^i_{a,j}$ and $T^i_{b,k}$

denote the date of the *j*th break in the intercept and the *k*th break in the trend slope for the *i*th unit. As in the empirical example in section 13.3.1.5, the specification can be simplified by setting $l_i = m_i = 1 \; \forall \; i = 1, 2, \ldots, N$, but the theory is available for the general case (see, for example, Bai and Carrion-i-Silvestre, 2007).

Cross-sectional dependence

The matrix F_t collects the common effects that are present across the cross-section dimension. The non-stationarity (or integration) of F_t will mean that all the units in the panel have common non-stationary (or integrated) components entering into each individual unit $y_{i,t}$ with loadings of magnitude π_i. Within the context of the Evans and Karras (1996) example, the π_i's are all equal to one, and the single common factor is concentrated out of the problem by cross-sectional demeaning – that is, by constructing the variable $y_{i,t} - \frac{1}{N}\sum_{i=1}^{N} y_{i,t}$, so that the focus then lies on testing the null hypothesis of a unit root in the regression model given by (13.4) above. As mentioned above, however, the simple cross-sectional averaging in general introduces correlations in the error terms describing the deviations from the cross-sectional averages.

As alluded to already in the convergence example, the presence of one or more common stochastic factors with heterogeneous loadings necessitates new testing and estimation strategies. One of these approaches, which is given in the work of Bai and Ng (2004) and Pesaran (2006), considers a general alternative to factor extraction by allowing cross-sectional averages (as $N \to \infty$) to approximate the effects of the unobserved or latent common factors. In this way, the need to estimate the factors and their loadings is avoided. The Pesaran approach, in the context of testing for unit roots, is discussed in section 13.2.2.3.

A further very interesting approach has been introduced in Pesaran, Schuerman and Weiner (2004) and Garratt *et al.* (2006) who introduce the concept of global vector autoregressions (GVARs), which consists of specifying VAR models for each cross-sectional member of the panel, including specifically constructed averages of the other cross-section members' variables as exogenous variables. This approach, feasible whenever the data permit VAR modeling for each cross-section member, is a parsimonious way of combining the information of all cross-section members. The construction of the weighted cross-sectional averages of the other countries' variables is a key issue in this modeling approach. One example considered by the authors is to use trade shares as weights. Obtaining further understanding of appropriate weighting schemes, which in general will depend upon the application considered, as well as the performance of these methods, is an important task. Under certain assumptions the joint system can be solved for based on the individual specific VAR models with exogenous variables. Given that a recent book-length discussion of this approach by Garratt *et al.* (2006) is available, we abstain from including this approach here.

Cross-sectional dependence can also emerge through correlations among the $e_{i,t}$ via the $\varepsilon_{i,t}$ variables across the units *i*, as in models of spatial correlation considered in a paper by Baltagi, Bresson and Pirotte (2007) or in purchasing power parity examples such as O'Connell (1998) discussed below. Note that we discuss in detail

in Appendix B the implications of correlation in the variables $e_{i,t}$, which also can lead to long-run cross-sectional dependencies.

In the simplest case of models (13.5)–(13.7), we could think of switching off cross-sectional dependence by setting $\pi_i = 0$ for all the units, specifying that the $\varepsilon_{i,t}$ are independent over i, and allowing specifications for the deterministic process to be given by (13.8) (that is, without breaks).[9] When $\varphi_i = 1$, and hence $\rho_i = 0$, the series $y_{i,t}$ contains a unit root.

13.2.1 Tests without cross-sectional dependence or structural breaks

Several tests, depending on the specification of the null and alternative hypotheses, have been developed to test for unit roots in panels.

13.2.1.1 Levin, Lin and Chu (2002)

As discussed in the many excellent descriptions in the literature (see, for example, Hlouskova and Wagner, 2006), the test is based on running augmented Dickey–Fuller regressions:

$$\Delta y_{i,t} = \mu_i + \delta_i t + \rho_i y_{i,t-1} + \sum_{k=1}^{p_i} \varphi_{i,k} \Delta y_{i,t-k} + v_{i,t}, i = 1, 2, \ldots, N; t = p_i + 2, \ldots, T. \tag{13.10}$$

Here we denote by $v_{i,t}$ the error process corresponding to the autoregressive specification to which the ADF set-up corresponds. Only if the data are really generated by autoregressions of orders $p_i + 1$ will the $v_{i,t}$ be white-noise processes. In practice this implies that lag length selection will be an issue (see below). Three sub-cases concerning the specification of the deterministic component are considered by Levin, Lin and Chu (2002) (henceforth referred to as LLC):

1. no deterministic terms;
2. intercept only;
3. intercept and linear trend.

We index the specification of these three deterministic components with $m = 1, 2, 3$. The null hypothesis of the LLC test is $H_0 : \rho_i = 0 \, \forall i = 1, 2, \ldots, N$ against the homogeneous alternative hypothesis $H_A : \rho_i = \rho < 0 \, \forall i = 1, 2, \ldots, N$.[10] This formulation of the null and alternative hypotheses allows for the construction of pooled tests (once appropriate corrections are made for the cross-sectional heterogeneity arising from other features of the DGP). Pooling may be made in both the within and between dimensions and gives rise to the tests (LLC and IPS, respectively) described below.

As mentioned above, selection of the lag length in (13.10) is an issue. In case the data are not described by finite-order autoregressive processes, the lag lengths have to increase as a function of the T-dimension of the panel (as was first studied by Said and Dickey, 1984, in the time series context) and LLC propose specifically that $p_i(T)$ grows at rate $T^{\kappa}, 0 < \kappa < 0.25$. Careful lag length selection is necessary to

ensure consistency of estimation by choosing the lag lengths in order to eliminate serial correlation in the error terms; that is, to have, at least asymptotically, white-noise processes $v_{i,t}$. In practice this typically means that information criteria are used to choose the lag lengths to ensure that the estimated residuals $\hat{v}_{i,t}$ show no evidence of serial correlation.[11]

In order to construct the LLC test, for chosen lag lengths p_i, two auxiliary regressions are initially estimated – the first consists of regressing $\Delta y_{i,t}$ on its lags $\Delta y_{i,t-k}, k = 1, 2, \ldots, p_i$, and the deterministic terms; denote the residuals from this regression by $\tilde{e}_{i,t}$. The second consists of regressing $y_{i,t-1}$ on the same set of regressors, to yield residuals denoted by $\tilde{f}_{i,t-1}$. Finally, $\tilde{e}_{i,t}$ is regressed on $\tilde{f}_{i,t-1}$ and the regression standard error from this equation, denoted $\hat{\sigma}_{v,i}$, is used to construct the standardized residuals $\hat{e}_{i,t} = \tilde{e}_{i,t}/\hat{\sigma}_{v,i}$ and $\hat{f}_{i,t-1} = \tilde{f}_{i,t-1}/\hat{\sigma}_{v,i}$. This standardization is needed to remove the effects of cross-sectional heterogeneity of the processes $v_{i,t}$ in (13.10) on the limiting distributions.

The next step is to estimate the long-run variance of $\Delta y_{i,t}$. Recall that under the null hypothesis,

$$\Delta y_{i,t} = \mu_i + \delta_i t + \sum_{k=1}^{p_i} \varphi_{i,k} \Delta y_{i,t-k} + v_{i,t}, i = 1, 2, \ldots, N; t = p_i + 2, \ldots, T,$$

so that a direct estimate of the long-run variance of $\Delta y_{i,t}$ is given by: $\hat{\sigma}^2_{v,i}(1 - \sum_{k=1}^{p_i} \hat{\varphi}^2_{i,k})^{-2}$.

In practice the following estimate is preferred in order to improve the size and the power of the test in finite samples. Denoting by $\hat{u}_{i,t} = \Delta y_{i,t} - \hat{\delta}_{mi} d_{mt}$, where d_{mt} denotes the specification of the deterministic terms, the long-run variance is estimated by $\hat{\sigma}^2_{LR} = T^{-1} \sum_{t=1}^{T} \hat{u}^2_{i,t} + \frac{2}{T} \sum_{j=1}^{L} w(j, L) \sum_{t=j+1}^{T} \hat{u}_{i,t} \hat{u}_{i,t-j}$, with the lag truncation parameter chosen by criteria given in Andrews (1991) or Newey and West (1994). The weights $w(j, L)$ are, in most applications, given by $w(j, L) = 1 - \frac{j}{L+1}$, referred to as the Bartlett kernel. For future reference, define $\hat{s}^2_i = \hat{\sigma}^2_{LR}/\hat{\sigma}^2_{v,i}$ and $\hat{S}_{N,T} = N^{-1} \sum_{i=1}^{N} \hat{s}_i$.

The essential component of the LLC test statistic (under all three deterministic specifications) is given by:

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \sum_{t=p_i+2}^{T} \hat{e}_{i,t} \hat{f}_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=p_i+2}^{T} \hat{f}^2_{i,t-1}},$$

computed from the pooled regression of $\hat{e}_{i,t}$ on $\hat{f}_{i,t-1}$. The t-form of this statistic, to test the null hypothesis $H_0 : \rho = 0$, is given by standardizing $\hat{\rho}$ by the standard deviation of $\hat{\rho}$, denoted $STD(\hat{\rho})$, from the pooled regression.

For the case with no deterministic terms, $t_{\rho=0} \Rightarrow N(0, 1)$ as $T \to \infty$ followed by $N \to \sigma$.[12] However, when either a constant or trend (or both) is present in the model, the t-statistic diverges (due to the presence of the so-called Nickell bias; see Nickell, 1978) to minus infinity (even under the null) and consequently needs to

be re-centered and re-normalized for convergence to a well-behaved distribution. Therefore,

$$t_{\rho *} = \frac{t_\rho - N\tilde{T}\hat{S}_{N,T}STD(\hat{\rho})\mu_{mT}}{\sigma_{mT}} \Rightarrow N(0,1).$$

Here $\tilde{T}$ denotes the effective average sample size, to take into account individual specific numbers of lagged terms across the individual units, while μ_{mT} and σ_{mT} denote the mean and variance corrections for the three different specifications of the deterministic terms. The latter are tabulated by LLC for various dimensions of N and T.

These are interesting results, deriving from the application of sequential (with first $T \to \infty$ followed by $N \to \infty$) central limit theorems, applicable under the assumption of *independent* cross-sectional units whenever the individual building blocks, which are identically distributed once T has passed to infinity, have finite second moments. As we show below, violations of this assumption can lead to severe difficulties, which may be soluble for a number of cases using a range of additional techniques.

Various modifications of the LLC procedure have been proposed, *inter alia*, by Harris and Tzavalis (1999), who derive asymptotic results for fixed T, allowing only the N dimension to tend to infinity, with closed form expressions for the correction factors for serially uncorrelated errors. As for LLC above, appropriately scaled and re-centered t-statistics tend to $N(0,1)$ densities as N tends to infinity, leading to tests which have better properties when the T dimension is relatively small and little or no serial correlation is permitted in the $v_{i,t}$ processes. Breitung (2000) develops, by means of appropriate variable transformations, a modified LLC test which, while coinciding with the LLC test when no deterministic terms are present, does not require bias correction factors for the cases where a constant or trend is present.

Several features of the testing framework above lend themselves to extensions and we shall try to deal with each of these in turn in the sections which follow:

1. relaxing homogeneity;
2. relaxing cross-sectional independence;
3. relaxing structural stability of the deterministic component.

13.2.1.2 Relaxing homogeneity – Im, Pesaran and Shin (2003)

Im, Pesaran and Shin (2003) (henceforth IPS), propose a class of group-mean panel unit root tests to allow for a heterogeneous alternative specified as:

$$H_A^1 : \rho_i < 0 \text{ for } i = 1, 2, \ldots, N_1 \text{ and } \rho_i = 0 \text{ for}$$
$$i = N_1 + 1, \ldots, N, \text{ where } \lim_{N\to\infty} \frac{N_1}{N} = k > 0.$$

IPS present two tests for the case of serially uncorrelated and correlated errors, and for specifications of the deterministic terms allowing for a constant and a constant and linear trend. The two tests are (i) a t-test based on ADF regressions, denoted IPS_t, and (ii) a Lagrange multiplier test, denoted IPS_{LM}. We concentrate here on (i),

where individual specific serial correlation structures are allowed in (13.10), with the assumption that the $y_{i,t}$ follow $AR(p_i + 1)$ processes.

We start by establishing some notation, assuming for simplicity that all required lagged observations are available. From (13.10), let:

$$\begin{aligned}
\tilde{\varphi}_i &= (\varphi_{i,1}, \ldots, \varphi_{i,p_i})' \\
y_{i,-1} &= (y_{i,0}, y_{i,1}, \ldots, y_{i,T-1})' \\
\Delta y_{i,-s} &= (\Delta y_{i,1-s}, \Delta y_{i,2-s}, \ldots, \Delta y_{i,T-s})', s = 0, 1, \ldots, p_i \\
\Delta y_i &= \Delta y_{i,-0} \\
d_{2T} &= (1, 1, \ldots, 1)' \\
t_{VEC} &= (1, 2, \ldots, T)' \\
d_{3T} &= (d'_{2T}, t'_{VEC}) \\
Q_{i,m} &= (d_{mT}, \Delta y_{i,-1}, \ldots, \Delta y_{i,-p_i}), m = 2, 3 \\
M_{Q_{i,m}} &= I_T - Q_{i,m}(Q'_{i,m}Q_{i,m})^{-1}Q'_{i,m}, m = 2, 3 \\
X_{i,m} &= (y_{i,-1}, Q_{i,m}), m = 2, 3 \\
M_{X_i,m} &= I_T - X_{i,m}(X'_{i,m}X_{i,m})^{-1}X'_{i,m}, m = 2, 3.
\end{aligned}$$

Then:

$$t_{iT,m}(p_i, \tilde{\varphi}_i) = \frac{\sqrt{T - p_i - m}(y'_{i,-1}M_{Q_i,m}\Delta y_i)}{(y'_{i,-1}M_{Q_i,m}y_{i,-1})^{1/2}(\Delta y'_i M_{X_i,m}\Delta y_i)^{1/2}}, m = 2, 3.$$

As before, $m = 2$ refers to the specification with a constant but without a linear trend, while $m = 3$ includes both a constant and a linear trend.

IPS_t is based on the cross-sectional average of the corrected t-statistics, that is,

$$IPS_{t,m} = \frac{\sqrt{N}\{\bar{t}_m - \frac{1}{N}\sum_{i=1}^{N} E(t_{iT,m}(p_i, 0)|\rho_i = 0)\}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{var}(t_{iT,m}(p_i, 0)|\rho_i = 0)\}}} \Rightarrow N(0, 1), \quad m = 2, 3,$$

where:

$$\bar{t}_m = \frac{1}{N}\sum_{i=1}^{N} t_{iT,m}(p_i, \tilde{\varphi}_i),$$

and $E(t_{iT,m}(p_i, 0)|\rho_i = 0)$ and $\mathrm{var}(t_{iT,m}(p_i, 0)|\rho_i = 0)$ are the mean and variance of the Dickey–Fuller statistic respectively for finite T and depend on the nuisance parameters $\tilde{\varphi}_i$. As $T \to \infty$, this dependence disappears and $E(t_{iT,m}(p_i, 0)|\rho_i = 0)$ and $\mathrm{var}(t_{iT,m}(p_i, 0)|\rho_i = 0)$ converge to the mean and variance of the Dickey–Fuller density corresponding to the model estimated (with intercept or with intercept and linear trend).

IPS tabulate these so-called correction terms for a set of values for T and $p_i = p$ for both specifications of the deterministic terms. Use of these correction terms (for

non-asymptotic values of T) is therefore restricted to balanced panels and equal lag lengths in all units of the panel. For large values of T, the use of simulated critical values (specific to the nuisance parameters) can be avoided and the asymptotic critical values can be used instead.

Similar principles apply to the use of the IPS_{LM} test under the additional restriction that $\lim N/T = c > 0$, as first T and then $N \to \infty$.

13.2.1.3 Fisher tests – Maddala and Wu (1999) and Choi (2001)

Continuing to focus on cross-sectional independence, Maddala and Wu (1999) and Choi (2001) developed tests based on combining p-values, an idea due to Fisher (1932). Fisher started his considerations by observing that the p-values of a continuous test statistic are uniformly distributed over the unit interval. Combining this with the fact that minus two times the log of a uniform distribution over the unit interval is distributed as a $\chi^2_{(2)}$ random variable, panel unit root tests can easily be constructed for cross-sectionally independent panels. If the N units are independent, the sum of the p-values, $-2\sum_{i=1}^{N} \log p_i \sim \chi^2_{(2N)}$, under the null hypothesis of a unit root in each unit of the panel. As long as p-values are computable for a test for a unit root on the individual units, either asymptotic, or via means of simulations or response surfaces (see, for example, MacKinnon, 1994, and MacKinnon, Haug and Michaelis, 1995, for augmented Dickey–Fuller tests), the Fisher test can be used to test for unit roots in panels. The key assumption remains that of cross-sectional independence, although its relaxation is possible in some cases (to allow for certain special forms of short-run dependence as described in section 13.2.2.1) by means of bootstrap techniques.

13.2.1.4 Tests with stationarity as the null hypothesis

Tests have also been developed which take as the null that of (heterogeneous) stationary roots against the alternative of a unit root in all cross-section members. Among this class of tests are those due to Hadri (2000) and Hadri and Larsson (2005), which apply the idea developed in Kwiatkowski *et al.* (1992) to the panel framework.

Looking at the specification where a linear trend is present under the null hypothesis, the LM-tests are based on looking at the partial sums of the residuals of the regressions (estimated for each unit):

$$y_{i,t} = \mu_i + \gamma_i t + u_{i,t}, \tag{13.11}$$

where, under the null hypothesis, and only for the sake of illustration, $u_{i,t}$ may be taken to be a serially uncorrelated stationary process.

Denoting by $\hat{u}_{i,t}$ the estimated residuals in (13.11), and their partial sums by $S_{i,t} = \sum_{j=1}^{t} \hat{u}_{i,j}$, the Hadri statistic, denoted by H_{LM}, is given by:

$$H_{LM} = \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{S_{i,t}^2}{\hat{\sigma}_{u,i}^2},$$

where:

$$\hat{\sigma}_{u,i}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{i,t}^2.$$

Under the null, using sequential convergence (that is, as before, $T \to \infty$ followed by $N \to \infty$), the re-centered and re-scaled Hadri statistic is given by:

$$Z_{LM} = \frac{\sqrt{N}(H_{LM} - \xi)}{\zeta} \Rightarrow N(0, 1).$$

The correction terms depend on the specification of the deterministic process and are given by Hadri (2000). The extension to the case of serially correlated (but stationary under the null) errors is easily dealt with by using an estimator of the long-run variance of $u_{i,t}$.

Hadri and Larsson (2005) allow for fixed T by deriving the finite sample mean and variance of $\kappa_{i,T} = \frac{1}{T^2}\sum_{t=1}^{T}\frac{S_{i,t}^2}{\hat{\sigma}_{u,i}^2}$, so that, by a simple application of central limit theory, it follows that:

$$H_T = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\kappa_{i,T} - \mathrm{E}\kappa_{i,T}}{\mathrm{var}(\kappa_{i,T})}\right) \Rightarrow N(0, 1) \text{ as } N \to \infty.$$

As before, the correction terms depend on the specification of the deterministic component and are given by Hadri and Larsson (2005). Note also that the closed form expressions for the correction factors derived in Hadri and Larsson are only correct for the case of serially uncorrelated errors, which limits the usefulness of the test to some extent.

13.2.1.5 A summary of simulation evidence (Hlouskova and Wagner, 2006)

We describe briefly here the results of a large-scale simulation study undertaken by Hlouskova and Wagner (2006) on many of the tests described above. For further details, we refer the reader to their paper, which provides a meticulous account of the behavior of the tests for unit roots in panels that are cross-sectionally independent as a function of numerous features of the DGP and estimation methods, including the dimensions of T and N, lag selection algorithms in the augmentation of the Dickey–Fuller tests, the presence of moving average terms in the error processes, and certain forms of cross-sectional correlation, given by either constant correlation or geometrically declining correlations by assuming the correlation matrix to be of Toeplitz form.

Hlouskova and Wagner conclude that the best performance in terms of power, where the evidence is based on simulations, for the case where the model has an intercept, is displayed by the LLC test or its modification proposed by Breitung (2000). For short panels, the Harris and Tzavalis (1999) modification of LLC offers some gains in panels where the errors are not serially correlated or the amount of serial correlation is small. They also observe that, when T is small

relative to N, the size and power properties of panel tests are affected adversely. The presence of moving average terms distorts the size of tests especially as the moving average coefficient θ in the process $v_{i,t} = \varepsilon_{i,t} + \theta\varepsilon_{i,t-1}$ tends to -1. This is a feature of tests for unit roots that is also evident in the time series case (see Molinas, 1986, or Schwert, 1989). Appropriate selection of the lag length in the ADF regressions is an issue here, since one way to account for the moving average dynamics is by sufficiently expanding the order of the autoregressions. As noted in a previous section, the impact of lag length selection, by various criteria such as AIC or BIC, is found to be ambiguous – beneficial in some cases and harmful in others. Evidence overall tends to suggest that, for θ close to zero, smaller lag lengths than those selected by the BIC tend to lead to better performance of the unit root tests, while the converse is true for values of θ close to -1.

13.2.2 Allowing for cross-sectional dependence

An important assumption underlying the so-called first-generation tests for unit roots in panel data, discussed in the previous section, is that of cross-sectional independence of the units of the panel. It is important, therefore, to assess the consequences of making this assumption when applying these techniques to datasets where it is not sustainable. As in our discussion of EK, in our view this appears to be relevant for many of the datasets for which one would have occasion to use unit root or cointegration tests in panels.

The motivation for the research presented next is to begin by analyzing, within the framework of some simulation studies, the consequences of departures from cross-sectional independence for the size and power properties of the first generation of commonly-used tests for unit roots in panels, and then to consider in detail the second generation of unit root tests that have emerged to allow for dependence. It is possible to generalize this framework to allow for structural breaks and we shall return to this issue in a later section.

Some particular examples of cross-sectional dependence are presented here, which we classify as short-run and long-run dependence according to the discussion and definition given in Appendix B. We distinguish between these two forms of dependence according to whether cross-unit cointegrating relationships, also defined in Appendix B, are present in the panel or not. While there are many ways of formulating dependence in panels, what is needed is the development of a general framework to incorporate the different possibilities and analyze their consequences. Appendix B contains a few simple observations.

13.2.2.1 Exemplifying the effects of cross-sectional dependence (O'Connell, 1998)

The study of O'Connell (1998) is a powerful demonstration of the oversizing of the LLC tests in the presence of short-run dependence. His simulation results have since been replicated and generalized, for example, by Hlouskova and Wagner (2006) and Baltagi, Bresson and Pirotte (2007). In the latter paper, cross-sectional dependence is derived from so-called spatial correlations. Describing some of this

research is the main motivation for this section, with a view to describing some of the general principles involved.

Our discussion is in several parts. We start by discussing O'Connell's contribution to the debate on testing for purchasing power parity in panels. By setting out the data-generation processes in some detail, we describe how O'Connell introduces short-run dependence into the panel. Second, we present some of O'Connell's results to describe the size distortions arising from this dependence. We discuss some solutions proposed to deal with that dependence. Based on this discussion, we then proceed in the following sub-sections to a discussion of the second generation of unit root tests aimed at modeling cross-sectional dependence, focusing on the work of Bai and Ng (2004) and other authors who have made use of factor models.

Consider, for illustration, the DGP given by:

$$\Delta y_{i,t} = \varepsilon_{i,t}, \quad i = 1, 2, \ldots, N; \quad t = 2, 3, \ldots, T,$$

where, for the purposes of the illustration, we may take $\varepsilon_{i,t} \sim$ i.i.d. $N(0, 1)$ without much loss of generality and where we also abstract from deterministic components. $y_{i,t}$ is therefore a simple random walk process. The model to be estimated is given by a simplified version of equation (13.10):

$$\Delta y_{i,t} = \mu_i + \rho_i y_{i,t-1} + \varepsilon_{i,t}, \quad i = 1, 2, \ldots, N; \quad t = 2, 3, \ldots, T.$$

The model is used to test:

$$H_0 : \mu_i = \rho_i = 0,$$

against the alternative:

$$H_A : \mu_i \neq 0; \quad \rho_i < 0,$$

although the results described below (taken from O'Connell, 1998) are based on looking only at the t-statistic for the estimate of the autoregressive parameter. It should be noted that, as usual within the LLC framework, ρ_i is restricted to be the same across all N units under both the null and alternative hypotheses. The testing framework can be augmented by polynomials of time, in particular by a linear trend t.

Now, under the normality and independent and identically distributed (i.i.d.) assumptions, cross-sectional dependence between the series is fully characterized by the *dynamic* covariance structure of the joint process $\varepsilon_t = \left(\varepsilon_{1,t}, \ldots, \varepsilon_{N,t}\right)'$. To make the discussion as simple as possible, consider a situation, like O'Connell (1998) does implicitly, where the joint random vector is also normally distributed and where the only correlations occur contemporaneously, in which case the dependence structure is fully characterized by $\Omega = E(\varepsilon_t \varepsilon_t')$. The joint vector of all disturbances $\varepsilon = \left(\varepsilon_1, \ldots, \varepsilon_T\right)'$ has its block-diagonal covariance, in our set-up

equal to the correlation, matrix given by:

$$\Sigma = \begin{pmatrix} \Omega & 0 & \dots & 0 \\ 0 & \Omega & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \Omega \end{pmatrix}.$$

In order to analyze the effect of cross-sectional dependence, O'Connell (1998) studied the effects of relaxing the assumption of diagonality of the matrix Ω on the performance of the LLC tests, using the original LLC correction factors and critical values derived under the assumption of cross-sectional independence. In particular, O'Connell considers the following simple shape of the correlation matrix, also used by Hlouskova and Wagner (2006) as one of their designs which they label, for obvious reasons, constant correlation:

$$\Omega = \begin{pmatrix} 1 & \omega & \dots & \omega \\ \omega & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \omega \\ \omega & \dots & \omega & 1 \end{pmatrix}.$$

This formulation imples that the "distance" between the cross-sectional disturbances is not considered relevant economically. Various justifications for such a covariance structure can be offered from examples dealing with spatial structures or for purchasing power parity based applications.[13] However, here it serves only for the purposes of illustration within the context of a Monte Carlo study. The simple O'Connell design also ensures that there is no long-run dependence for $-1 < \omega < 1$.

The simulations were conducted with the following choices of sample sizes and parameters in the DGP:[14]

$$N \in \{10, 50, 90\}$$
$$T \in \{20, 60, 100\}$$
$$\omega \in \{0, 0.3, 0.5, 0.7, 0.9\},$$

with the disturbances distributed as described above. It is worth pointing out that the $\{N, T\}$ combinations considered do have relevance for assessing the results of the simulations, since a number of the asymptotic results on the properties of unit root tests in panels impose joint conditions on N and T tending to infinity in order to prove the theorems. This is also a crucial issue that arises later when discussing factor estimation.

For given values of N, T and ω, the t-statistic $\hat{\rho}/\widehat{\sigma}_{\hat{\rho}}$ (denoted t_{OLS}) was computed and compared (for each of the 5,000 replications/panels) with critical values (at 1%, 5% and 10%) tabulated by LLC, designed for the case without correlation, that is, for $\omega = 0$.

The results, as reported in O'Connell (1998, p. 6, Table 1) for non-zero values of ω, were quite dramatic. He showed that, for example for T= 20, N= 10 and ω = 0.3, the true size of the test at 5% (respectively 10%) nominal size, if LLC critical values are used, was 9% (respectively 15%). These distortions increased as ω increased – for example, when $\omega = 0.9$, for the same configuration of T and N, the rejections of the true null hypothesis at 5% (respectively 10%) increased to 37% (respectively 43%). The size distortions were not affected significantly by increasing T for fixed N. Thus the configurations $T = 60$, $N = 10$, $\omega = 0.3$ and $T = 100$, $N = 10$, $\omega = 0.3$ give rejections of 9% (respectively 15%) at 5% (respectively 10%) nominal size, which are unchanged from their $T = 20$ values. The distortions, however, do increase with N for fixed T and ω. For $T = 20$, $N = 50$, $\omega = 0.3$, the rejection at 5% (respectively 10%) increases to 21% (respectively 28%) and to 31% (respectively 37%) when N is raised to 90, and these numbers are even higher when ω is also increased.

Adjusting the critical values to control for the size of the test leads to a severe reduction in power. For example, the power of the t-test to reject $H_0 : \rho_i = 0$ against the alternative $H_0 : \rho_i = -0.04$ (reported in O'Connell, 1998, p. 6), where this non-zero value of ρ_i gives a half-life of deviation from purchasing power parity of between four and five years and is coherent with empirical estimates, is 8% when $T = 20$, N = 10, 14% when T =20, N= 50, and 13% when $T = 20$, $N = 90$. ω here is 0.3 and the confidence level of the tests is 5%. For T = 60 and N = 50, power drops from 92% (when ω = 0) to 30% (when ω = 0.5) and to 9% (when ω = 0.9). Thus the necessity to counteract distortions neutralizes any beneficial effects of increasing N, the longitudinal dimension of the panel, and the loss of power involved – as a result of the necessary adjustment – is clearly a serious one. Indeed, the trade-off between the N and T dimensions, which is evident from both the asymptotic theory and the empirical implementation, is a topic which is relatively ill-studied in the literature.

In order to account for the non-zero off-diagonal terms in Ω, O'Connell proposed the following generalized least squares (GLS) estimator:

$$\hat{\rho}_{GLS} = \frac{tr(X'\Delta Y\Omega^{-1})}{tr(X'X\Omega^{-1})},$$

where ΔY is a $T \times N$ matrix of the first-differenced ys and X is a matrix of lagged ys. When $\omega = 0$, we recover the usual ordinary least squares (OLS) estimator. Thus:

$$\Delta Y_{T\times N} = \begin{pmatrix} \Delta y_{1,1} & \Delta y_{2,1} & \cdot & \cdot & \cdot\cdot & \Delta y_{N,1} \\ \Delta y_{1,2} & \Delta y_{2,2} & \cdot & \cdot & \cdot & \Delta y_{N,2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ \Delta y_{1,T} & \Delta y_{2,T} & \cdot & & & \Delta y_{N,T} \end{pmatrix}.$$

Computation of the feasible GLS estimator, denoted by $\hat{\rho}_{FGLS}$, requires a consistent estimator of Ω. Allowing only for contemporaneous correlation, a consistent

estimator of Ω is given by the (estimated) covariance matrix of the first differences of the vector y_t, that is,

$$\hat{\Omega} = T^{-1} \sum_{t=1}^{T} \Delta y_t \Delta y_t'$$

$$\Delta y_t = (\Delta y_{1t}, \Delta y_{2,t}, \ldots, \Delta y_{N,t})'$$

and, therefore,

$$\hat{\rho}_{FGLS} = \frac{tr(X' \Delta Y \hat{\Omega}^{-1})}{tr(X' X \hat{\Omega}^{-1})}.$$

Clearly, with this estimator not all correlations between the cross-section members need to be identical, as assumed by O'Connell to illustrate the issues in the simulation part of his paper. The results reported in O'Connell (1998, p. 12, Table 3) are encouraging because the rejection frequencies under both the null hypothesis (size) and the alternative hypothesis (power) are now invariant to the value of ω in the DGP, and depend on T and N alone. Provided T is much larger than N, the power of the FGLS test comes close to the OLS panel unit root test when the disturbances are independently and identically distributed. Efficiency losses ensue as N comes closer and closer to T, which Hlouskova and Wagner (2006) refer to as size divergence when N is "too large" compared to T.

The discussion above serves to introduce and illustrate some of the main issues involved, but is nevertheless too specialized, applying as it does to only one highly restrictive specification of cross-sectional dependence. More general methods for dealing with short-run cross-sectional dependence will be discussed below (including dealing with nonparametric methods for estimating the variance covariance matrix of the disturbances when there is dependence across the units).

13.2.2.2 Cross-sectional dependence via approximate factor models – Bai and Ng (2004)

An important class of tests developed to allow for long-run cross-sectional dependence is due to Bai and Ng (2004). Returning to the formulation given by (13.5)–(13.7), the basic idea is to think of the series comprising the panel as consisting of the sum of a set of deterministic terms, common factors and idiosyncratic components. The detailed assumptions made by Bai and Ng are given by:

(i) For non-random π_i, $||\pi_i|| \leq A$; for random π_i,

$E||\pi_i||^4 \leq A$, $\frac{1}{N} \sum_{i=1}^{N} \pi_i \pi_i' \to \Sigma_\Pi$, an $r \times r$ positive definite matrix, where $\to$ denotes convergence in probability.

(ii) $\eta_t \sim i.i.d.(0, \Sigma_\eta), E||\eta_t||^4 \leq A, \text{var}(\Delta F_t) = \sum_{j=0}^{\infty} C_j \Sigma_\eta C_j' > 0, \sum_{j=0}^{\infty} j||C_j|| < A$, $C(1)$ has rank r_1, $0 \leq r_1 \leq r$.

(iii) For each i,

$$\varepsilon_{i,t} \sim i.i.d.(0, \sigma_{\varepsilon_i}^2), E|\varepsilon_{i,t}|^8 \leq A, \sum_{j=0}^{\infty} j||H_{i,j}|| < A, \omega_i^2 = H_i(1)^2\sigma_{\varepsilon_i}^2 > 0;$$

$\varepsilon_{i,t}$ are independent over i.[15]

(iv) The errors $\varepsilon_{j,t}$, $\eta_{s,i}$ and the loadings π_i form three mutually independent groups for all (j, t, s, i).[16]

(v) $E||F_0|| \leq A$, and for all i =1,2,...,N, $E|e_{i,0}| \leq A$.

In the assumptions A is taken to be a positive number not depending on either T or N. The notation $||B|| = \text{trace}(B'B)^{1/2}$.

Both the factors and the idiosyncratic components can be integrated or stationary, so that short- and long-run dependence can be modeled both via the common factors and also the idiosyncratic components. However, in our discussion we assume that in (13.7) the idiosyncratic terms are taken to be independent across i, which is slightly stronger than the assumptions necessary for applicability of the Bai and Ng (2004) methods.

The heart of the unit root analysis consists of making the decomposition (between common factors and idiosyncratic terms) and then testing each of these components for a unit root. Thus, returning to the set-up described by (13.5), let:

$$y_{i,t} = \mu_i + \pi_i' F_t + e_{i,t},$$

where F_t is an $r \times 1$ vector of "common factors." The model can be rewritten in first differences:

$$\Delta y_{i,t} = \pi_i' f_t + z_{i,t}, \quad i = 1, 2, \ldots, N; \; t = 2, 3, \ldots, T,$$

where:

$$f_t = \Delta F_t, \; t = 2, 3, \ldots, T$$
$$z_{i,t} = \Delta e_{i,t}, \; i = 1, 2, \ldots, N; \; t = 2, 3, \ldots, T.$$

Next, define:

$$Y = (y_1, y_2, \ldots, y_N),$$

as the $T \times N$ matrix of all observations, where:

$$y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})',$$

and:

$$\tilde{y} = \Delta Y = (\Delta y_1, \ldots, \Delta y_N),$$

is the corresponding $(T-1)\times N$ matrix of first differences. The principal component estimator $\hat{f}$ of $f = (f_2, f_3, \ldots, f_T)'$ is $\sqrt{T-1}$ times the r eigenvectors corresponding to the r largest eigenvalues of the $(T-1) \times (T-1)$ matrix $\tilde{y}\tilde{y}'$ and the estimated factor loading matrix given by $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_N)'$ is obtained from the relationship

$\hat{\pi} = \tilde{y}'\hat{f}/(T-1)$ under the normalization $\hat{f}'\hat{f}/(T-1) = I_r$, where I_r is the $r \times r$ identity matrix.

The estimated factors can now be recovered by summation:

$$\hat{F}_t = \sum_{s=2}^{t} \Delta\hat{f}_s, t = 2, \ldots, T.$$

If r= 1, this single factor can be tested for a unit root using an augmented Dickey–Fuller test with an intercept.

If (13.5) contains a linear trend in addition to an intercept, the method of principal components is applied to the *demeaned* and *differenced* data and the ADF test for the factor contains an intercept and a trend. The critical values for both these sets of tests are as provided by Dickey and Fuller (1979).

When the number of common factors r is greater than 1, Bai and Ng (2004) also develop tests for the number of linearly independent $I(1)$ common trends (equivalently the cointegrating rank) contained in the common factors. After an appropriate basis change the number of stationary factors is given by r_0, and there are r_1 linearly independent integrated factors or common trends, such that, as always in the $I(1)$ framework, $r_0 + r_1 = r$.

The two test statistics ($MQ_c(m)$ and $MQ_f(m)$) follow, up to a transformation to ensure real valued test statistics, Stock and Watson (1988). The test statistics for testing the null hypothesis of m common trends in the common factors are computed recursively with the first test statistic based on $r = m$ common factors. The statistic $MQ_c(m)$ is based on estimating a VAR(1) process for $\hat{Y}_t$, where $\hat{Y}_t = \hat{\beta}_{ORTH}\tilde{F}_t$, and where, in the first step, $\tilde{F}_t$ is the m-dimensional vector of factors computed from the demeaned, respectively demeaned and detrended observations, while $\hat{\beta}_{ORTH}$ is the matrix of m eigenvectors associated with the m eigenvalues of the matrix given by $\frac{1}{T^2}\sum_{t=2}^{T}\tilde{F}_t\tilde{F}_t'$. The statistic $MQ_f(m)$ is constructed similarly, except that a pth order VAR is fitted first to get $\hat{y}_t = \hat{\Pi}(L)\hat{Y}_t$ and the test is based on the filtered $\hat{y}_t$ series.

The testing procedure consists of constructing a sequence of these MQ statistics, starting with testing the null hypothesis $m = r$ (that is, r stochastic trends) against the alternative hypothesis $m = r - 1$, and testing down until the first non-rejection of the null hypothesis occurs; for example, in the second step, the test statistic is based on only $r-1$ eigenvectors $\hat{\beta}_{ORTH}$ corresponding to the $r-1$ largest eigenvalues (for a detailed description see Bai and Ng, 2004, pp. 1133–4).

Two versions of these statistics are considered by the authors to allow for demeaning and/or detrending of the observations, depending upon the model considered. Critical values are provided by them for up to six stochastic trends.

Similarly, the estimated idiosyncratic components can also be tested for unit roots. These are obtained from $\hat{e}_{i,t} = \sum_{s=2}^{t}\Delta\hat{z}_{i,s}$, $i = 1, 2, \ldots, N$; $t = 2, \ldots, T$, with $\Delta\hat{z}_{i,s} = \Delta\hat{y}_{i,s} - \hat{\pi}_i'\hat{f}_s$. The additional complication here is that the estimates that are available are not, in general, cross-sectionally independent in finite samples due to their dependence upon the estimated factors and loadings. We believe that

this issue, which complicates the use of first-generation panel unit root tests on estimated de-factored data, is not considered carefully enough in the literature.

Bai and Ng derive the asymptotic behavior of tests paralleling those of Choi (2001) and Maddala and Wu (1999), henceforth referred to as BN_N and BN_{χ^2}, under the assumption of cross-sectional independence of the idiosyncratic components.

The Choi (2001)-type test is constructed as follows. Denote the p-value associated with the ADF test on the residual series from the ith unit, $\hat{e}_{i,t}$, $i = 1, 2, \ldots, N$, by $p_{\hat{e}}(i)$. Then the following result:

$$BN_N = \frac{-2\sum_{i=1}^{N} \log p_{\hat{e}}(i) - 2N}{\sqrt{4N}} \Rightarrow N(0, 1),$$

can be established by careful analysis, paralleling the type of panel unit root test proposed by Choi (2001) for cross-sectionally independent panels. The test statistic BN_{χ^2} is, of course, given by $-2\sum_{i=1}^{N} \log p_{\hat{e}}(i) \sim \chi^2_{(2N)}$.

Note that simple sequential limit theory with first $T \to \infty$ followed by $N \to \infty$, as used in the panel unit root tests for cross-sectionally independent data, does not apply here. This stems from the fact that consistent estimation of factors and loadings is based on joint limit theory with minimum rate restrictions for both dimensions of the panel, which implies that the observed $\hat{e}_{i,t}$ are, in general, not cross-sectionally independent for finite samples even under the assumption that the $e_{i,t}$ are independent. The panel unit root test result is established both when the factors are computed after differencing (intercept only case) and after differencing and demeaning (intercept and trend), and requires that the idiosyncratic terms be independent across i. It should be noted, however, that the main results of the Bai and Ng (2004) paper (consistent estimation of the factors and loadings and testing for common trends in the common factors) allow for weak cross-sectional correlation of the idiosyncratic errors.

Two general observations should be made of the Bai and Ng method. First, it is sufficient that at least one integrated factor be present for all the $y_{i,t}$ series to have unit roots, if this factor is loaded into all series. Bai and Ng (2004) call this integration or non-stationarity due to "a pervasive source." The integration properties of the idiosyncratic components, however, feed, under the assumption of their cross-sectional independence, uniquely into each series. That is, if all common factors are stationary the series $y_{i,t}$ has a unit root if and only if $e_{i,t}$ has a unit root. This is referred to as a "series specific source" and may be thought to be a measure of the existence or otherwise of a unit root in a given unit of a panel once account has been taken of all the common trends driving the data (and describing the dependence). The judgment on which is the more important source depends upon the phenomenon being studied – that is, whether each series has its "own root" or whether it comes from the same factor(s).

Second, the importance of the Bai and Ng (2004) analysis is in showing that, by applying the method of principal components to first-differenced data, it is possible to obtain consistent estimators of the factors and the idiosyncratic terms regardless

of the dynamic properties of these series (under the assumptions posited above). That is, the tests for the number of common stochastic trends do not depend on whether the idiosyncratic components are stationary, while the tests for whether the errors are stationary do not depend on the presence or absence of common stochastic trends.

To be set off against these advantages are the restrictions (for example, the restrictions with respect to cointegration discussed Appendix B) implied by modeling dependence via common factors – and whether alternative approaches such as GVAR models should be considered. In addition the finite sample properties of factor-based methods, which have not always been shown to be encouraging, need to be considered in more detail.

13.2.2.3 Specializations of the Bai and Ng (2004) framework

Three specializations of the Bai and Ng procedure may be considered briefly, since they serve to illustrate some principles for dealing with cross-sectional dependence, which have been explored in more general contexts (see, for example, Pesaran, 2006).

Pesaran (2007)

The first specialization, due to Pesaran (2007), allows for the dependence among the cross-sectional units of the panel to derive only from one *stationary* common factor in the disturbances of each unit, with this common factor entering into the units with heterogeneous loadings.

His DGP takes the following form:

$$\begin{aligned} y_{i,t} &= (1-\varphi_i)\mu_i + \varphi_i y_{i,t-1} + u_{i,t} \\ u_{i,t} &= \pi_i f_t + \varepsilon_{i,t} \\ & i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T. \end{aligned}$$

The following assumptions are put in place:

(i) The idiosyncratic errors $\varepsilon_{i,t}$ are independently distributed across both i and t, have mean zero, variance σ_i^2 and finite fourth-order moment.
(ii) The common factor f_t is serially uncorrelated with mean zero, constant variance σ_f^2 and finite fourth-order moment.
(iii) $\varepsilon_{i,t}, f_t$ and π_i are mutually independent groups.

The assumption of serially uncorrelated error $u_{i,t}$ can be relaxed (see Pesaran, 2006, for details).

The null and alternative hypotheses are as for the IPS test (remember that $\rho_i = \varphi_i - 1$). Under the null hypothesis ($\varphi_i = 1, \rho_i = 0$) there is no trend in the data. With this rather restrictive specification, Pesaran proposes the use of a cross-sectionally augmented version of the IPS_t test described above. The procedure consists of including cross-sectional averages of the level and of lagged differences

in the IPS-type regressions, where the former are taken to act as a proxy for the single common factor for N sufficiently large. Lags of the cross-sectional averages may be utilized if necessary to take account of serial correlation in the $u_{i,t}$ process.

Thus, using (for the case without linear trend),

$$\Delta y_{i,t} = \mu_i + \rho_i y_{i,t-1} + c_i \bar{y}_t + d_i \Delta \bar{y}_t + \sum_{k=1}^{p_i} \varphi_{ik} \Delta y_{i,t-k} + v_{i,t}, \; i = 1, 2, \ldots, N; \; t = p_i + 2, \ldots, T, \tag{13.12}$$

where the cross-sectional average of the $y_{i,t}$ terms is:

$$\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{i,t}, \text{ and}$$
$$\Delta \bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} \Delta y_{i,t},$$

the $CIPS_t$ test is given by:

$$CIPS_t = \frac{1}{N} \sum_{i=1}^{N} CADF_i,$$

with $CADF_i$ denoting the Dickey–Fuller t-statistic for testing $H_0 : \rho_i = 0 \; \forall \; i$ in (13.12). This test harks back both to the idea of using group mean tests to allow for heterogeneity of the autoregressive root under the alternative hypothesis and of using cross-sectional averaging to allow for cross-sectional dependence across the units. The latter may not be effective, depending on the nature of the dependence being modeled.

Pesaran (2007) investigates the asymptotic null distribution of the individual $CADF_i$ statistics as well as of the associated $CIPS_t$ test statistic. The former allows for the construction of p-values for the individual (that is, unit by unit) $CADF_i$ test statistics so that tests in the spirit of Maddala and Wu (1999) or Choi (2001) can be constructed. The asymptotic distributions are derived both for sequential asymptotics as well as joint asymptotics (N and T tending to infinity such that $N/T \to k$, where k is a fixed finite non-zero positive constant). Pesaran shows that the $CADF_i$ statistics do not depend on the factor loadings but are asymptotically correlated through their dependence on the common factor. This has the consequence that standard central limit theorems cannot be used to derive the asymptotic distribution of either $CIPS_t$ or of the pooled p-value tests. A truncated $CIPS_t$ test statistic, denoted $CIPS_t^*$, is also proposed with better properties. Critical values for both $CIPS_t$ and $CIPS_t^*$ are presented for the three main specifications of the deterministic components, and not just for the specification outlined here for illustration.

Moon and Perron (2004)

A second specialization, somewhat less restrictive than Pesaran, is due to Moon and Perron (2004), who consider a DGP of the following form (which may also be

rewritten in the form given by (13.5)–(13.7) above):

$$
\begin{aligned}
y_{it} &= \mu_i + y^0_{i,t} \\
y^0_{i,t} &= \varphi_i y^0_{i,t-1} + u_{i,t} \\
u_{i,t} &= \pi_i' f_t + e_{i,t} \\
i &= 1, 2, \ldots, N;\ t = 1, 2, \ldots, T.
\end{aligned}
$$

Note that f_t is an r-dimensional vector of common factors, taken here to be stationary, see assumptions (ii) and (v) below, where r may be taken to be known.[17]

Key assumptions within this framework include the following:

(i) $e_{i,t} = \sum_{j=0}^{\infty} d_{i,j}\varepsilon_{i,t-j}$, where $\varepsilon_{i,t}$ are i.i.d. (0,1) across i and over t, have finite eighth moment, $\inf_i \sum_{j=0}^{\infty} d_{i,j} > 0$ and $\overline{d}_j = \sup_i |d_{i,j}|$, $\sum_{j=0}^{\infty} j^m \overline{d}_j < M$ for some $m > 1$;

(ii) $f_t = \sum_{j=0}^{\infty} c_j \eta_{t-j}$, where c_j are $r \times r$ matrices of real numbers and the $r \times 1$ vectors $\eta_t = (\eta_{1,t}, \ldots, \eta_{j,t}, \ldots, \eta_{r,t})'$ are i.i.d. $(0, I_r)$, so that $\eta_{j,t}$ is i.i.d across j and over t. It is also assumed that $\sum_{j=0}^{\infty} j^m ||c_j|| < M$, for some $m > 1$;

(iii) $\varepsilon_{i,t}$ and $\eta_{i,s}$ are independent;

(iv) as $N \to \infty$, $\frac{1}{N}\sum_{i=1}^{N} \pi_i \pi_i' \to \Sigma_\Pi > 0$;

(v) as $T \to \infty$, $\frac{1}{T}\sum_{i=1}^{N} f_t f_t' \to \Sigma_f > 0$.

The unit root in the $y_{i,t}$ process comes solely from φ_i being equal to one. This is a key restriction from the overall Bai and Ng (2004) framework (which allows for integrated factors and integrated idiosyncratic components).

Following from above, the unit root null hypothesis of $H_0 : \varphi_i = 1\ \forall i$ is therefore tested against the heterogeneous alternative of $H_A : \varphi_i < 1$ for some of the units, as long as the number of these units remains a positive fraction of the total number of units as $N \to \infty$. The vector π_i determines the loadings of the factors into the ith unit and, if $r = 1$, the system simplifies to the DGP considered by Pesaran (2007).[18]

The procedure consists of first computing the pooled OLS estimator $\hat{\varphi}_{Pooled}$ (of φ), given by setting $\varphi_i = \varphi\ \forall i$ and calculating the residuals from the pooled regression as $\hat{u}_{i,t} = y_{i,t} - \hat{\varphi}_{Pooled} y_{i,t-1}$. Extracting the factors from the $\hat{u}_{i,t}$ series is the second step.

Next, we establish some notation:

$$
\begin{aligned}
\hat{\Lambda} &= (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N)', \text{ where } \hat{\pi}_i, i = 1, 2, \ldots, N, \text{ are the estimated factor loadings} \\
\hat{Q}_{\hat{\Lambda}} &= I_N - \hat{\Lambda}(\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}' \\
y_{i,-1} &= (y_{i,0}, y_{i,1}, \ldots, y_{i,T-1})' \\
y_i &= (y_{i,1}, \ldots, y_{i,T})' \\
Y &= (y_1, y_2, \ldots, y_N)
\end{aligned}
$$

$$
\begin{aligned}
Y_{-1} &= (y_{1,-1}, y_{2,-1}, \ldots, y_{N,-1}) \\
\hat{u}_i &= (\hat{u}_{i,1}, \ldots, \hat{u}_{i,T})' \\
\hat{u} &= (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N) \\
\hat{e} &= \hat{u}\hat{Q}_{\hat{\Lambda}} \\
\omega_{e,i}^2 &= \left(\sum_{j=0}^{\infty} d_{i,j} \right)^2 \\
\lambda_{e,i} &= \sum_{l=1}^{\infty} \sum_{j=0}^{\infty} d_{i,j} d_{i,j+l} \\
\omega_e^2 &= \frac{1}{N} \sum_{i=1}^{N} \omega_{e,i}^2 \\
\phi_e^4 &= \frac{1}{N} \sum_{i=1}^{N} \omega_{e,i}^4 \\
\lambda_e^N &= \frac{1}{N} \sum_{i=1}^{N} \lambda_{e,i}.
\end{aligned}
$$

Consistent estimators of ω_e^2, ϕ_e^4 and λ_e^N are provided by constructing kernel estimators based on sample covariances provided by $\frac{1}{T}\sum_t \hat{e}_{i,t}\hat{e}_{i,t+j}$, where $\hat{e}_{i,t}$ is the (i,t)th element of $\hat{e}$ defined above and $1 \leq t, t+j \leq T$. These are denoted by $\hat{\omega}_e^2$, $\hat{\phi}_e^4$ and $\hat{\lambda}_e^N$, respectively.

Two test statistics, denoted MP_a and MP_b, follow, defined as:

$$
\begin{aligned}
MP_a &= \frac{\sqrt{N}T(\varphi^*_{Pooled} - 1)}{\sqrt{\frac{2\hat{\phi}_e^4}{\hat{\omega}_e^4}}} \\
MP_b &= \frac{\sqrt{N}T(\varphi^*_{Pooled} - 1)}{\hat{\phi}_e^2} \sqrt{\frac{1}{NT^2} tr(Y_{-1}\hat{Q}_{\hat{\Lambda}} Y'_{-1})}\hat{\omega}_e, \\
\text{where } \varphi^*_{Pooled} &= \frac{tr(Y_{-1}\hat{Q}_{\hat{\Lambda}} Y') - NT\hat{\lambda}_e^N}{tr(Y_{-1}\hat{Q}_{\hat{\Lambda}} Y'_{-1})}.
\end{aligned}
$$

Under the null hypothesis $H_0 : \varphi_i = 1 \; \forall i$, Moon and Perron (2004) show that:

$$
MP_a, MP_b \Rightarrow N(0,1), \text{ as } N, T \to \infty \text{ with } N/T \to 0.^{19}
$$

No inference concerning the unit root behavior is conducted on the estimated factors, since these are taken to be stationary *a priori*. However, under the null hypothesis it is the accumulated factors that enter the data y.

Choi (2006a)

The final specialization discussed here, due to Choi (2006a), is based on a two-way error component model. Consider the model given by:

$$\begin{aligned} y_{i,t} &= \beta_0 + x_{i,t} \\ x_{i,t} &= \mu_i + \lambda_t + u_{i,t} \\ u_{i,t} &= \sum_{j=1}^{p_i} \alpha_{ij} u_{i,t-j} + \varepsilon_{i,t}. \end{aligned}$$

The μ_i, λ_t and $\varepsilon_{i,t}$ processes (which are uncorrelated with each other) are assumed to have the following structures:

$$E(\mu_i) = 0 \ \forall i, E(\mu_i^2) = \sigma_\mu^2 < \infty \ \forall i, E(\mu_i \mu_j) = 0 \ \forall i \neq j$$

$$E(\lambda_t) = 0 \ \forall t, E(\lambda_t \lambda_s) = \sigma_\lambda(|t-s|) < \infty$$

$$\varepsilon_{i,t} \sim \text{i.i.d. } (0, \sigma_\varepsilon^2), \quad \varepsilon_{i,t} \text{ is independent of } \varepsilon_{j,s} \ \forall i \neq j, s \neq t.$$

The test therefore consists of purging the $x_{i,t}$ process of its "common" time effects, given by λ_t, its individual specific effects, given by μ_i, and testing for a unit root in $\hat{u}_{i,t}$. The null and alternative hypotheses have the form:

$$H_0 : \sum_{j=1}^{p_i} \alpha_{ij} = 1 \ \forall i; \quad H_A : \sum_{j=1}^{p_i} \alpha_{ij} < 1 \text{ for } 0 < N_1 < N \text{ units.}$$

As earlier, the fraction of the units with no unit roots is required to fulfill the following property under the alternative:

$$\lim_{N \to \infty} \frac{N_1}{N} = k > 0.$$

Consider testing the demeaned and detrended residuals $\hat{u}_{i,t}$, for each unit i, using ADF tests (and critical values) and suppose the corresponding p-value of the test for each unit is given by p_i.

Choi (2006a) proposes the use of three different group mean tests based on the Fisher principle (generalizing his earlier work on Fisher tests for cross-sectionally independent panels):

$$\begin{aligned} C_P &= -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\log(p_i) + 1) \\ C_Z &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\Phi^{-1}(p_i)) \\ C_{L*} &= \frac{1}{\sqrt{\pi^2 \frac{N}{3}}} \log\left(\frac{p_i}{1-p_i}\right). \end{aligned}$$

Here Φ denotes the distribution function of the standard normal distribution. Under the null, all three statistics tend to $N(0,1)$, $T, N \to \infty$. Under the alternative, $C_Z \to \infty$ while the remaining two statistics, $C_P, C_Z \to -\infty$, $T, N \to \infty$. This framework can, of course, be generalized to incorporate a trend in the model.

13.2.2.4 Nonlinear instrumental variables unit root test – Chang (2002)

Chang (2002) allows for cross-sectional dependence through correlation in the noise processes. Thus, the starting point is correlation between the series $u_{i,t}$ in $y_{i,t} = \rho_i y_{i,t-1} + u_{i,t}$, abstaining here from deterministic components for simplicity. Chang assumes that all $u_{i,t}$ series are stationary autoregressive processes of some orders p_i generated by some innovations $\varepsilon_{i,t}$. In particular, Chang (2002, Assumption 2.2, p. 264) assumes that $\varepsilon_t = \left(\varepsilon_{1,t}, \ldots, \varepsilon_{N,t}\right)'$ is an i.i.d. sequence with (non-diagonal) positive definite covariance matrix, which precludes cross-unit cointegration. The unit root test itself is based on instrumental variable estimation with the instrument given by integrable functions of the lagged levels of $y_{i,t}$. The test statistic, labeled NL, for the null hypothesis $H_0 : \rho_i = 1$ for $i = 1, \ldots, N$ against the heterogeneous alternative, is given by an appropriately weighted sum of the individual t-statistics. Chang (2002, p. 277) proposes the following instrument generating function $F(y_{i,t-1}) = y_{i,t-1} e^{-c_i |y_{i,t-1}|}$, where c_i is related to the sample standard error of $\Delta y_{i,t}$. Im and Pesaran (2003) show that the asymptotic behavior established in Chang (2002) holds only when $N \ln T / \sqrt{T} \to 0$, which suggests that N has to be quite small compared to T in practice.

13.2.2.5 Summary of section 13.2.2

We have described above a broad class of methods developed to deal with cross-sectional dependence. These are largely dependent on using factor models of varying degrees of generality to model dependence across the units. It is thus of interest to investigate the size and power of these methods within the context of a simulation study, and to apply these methods to real-world datasets (although a direct comparison of the results cannot be made since the tests operate under different assumptions).

Some general principles can be identified. First, cross-sectional dependence is mainly modeled by resorting to factor models, which allows us to model both short- and long-run dependence, albeit with some restrictions. Second, for certain special configurations – as, for example, the one considered by O'Connell (1998) – short-run correlation can be relatively easily accounted for in simple testing approaches by resorting to some corrections; for example, feasible GLS. Third, under appropriate assumptions, many of the test statistics are, under the null hypothesis, asymptotically standard normally distributed (after appropriate centering and re-scaling). This is most easily established for cross-sectionally independent panels and sequential limits with $T \to \infty$ followed by $N \to \infty$, but can also be established (under appropriate assumptions) in panels with cross-sectional dependencies.

We turn next to the most general formulation (13.5)–(13.7), which allows not only for cross-sectional dependence but also for the presence of structural breaks in the deterministic components of the series comprising the panel. These structural breaks are typically assumed to take the form of changes in the intercept or linear trend of the process (in a given unit or a set of units) at a potentially unknown date in-sample. As in time series unit root problems, inference about unit roots in

integrated panels needs to take account of shifts in the deterministic parts of the processes, since breaks in series may lead (as in Perron, 1989) to spurious findings of a unit root. The analysis can, in principle, also be extended to allow for breaks in the factors or in the loading coefficients on the factors, but this takes us beyond the scope of the current discussion.

13.2.3 Relaxing structural stability – Bai and Carrion-i-Silvestre (2007)

The restrictions underlying (13.5)–(13.7) above, with specification of the breaks given in (13.9), are relatively mild. This model formulation allows for a very general specification of the testing framework, which allows not only for cross-sectional dependence via common factors but also for structural instability. As in the assumptions underlying the Bai and Ng model (given in section 13.2.2.2) the factor loadings need to be identifiable; there are conditions on the short- and long-run variance of ΔF_t and the $\varepsilon_{i,t}$ processes are assumed to be weakly serially correlated but cross-sectionally independent (as given by the set of conditions (iii) for $\varepsilon_{i,t}$ in section 13.2.2.2).[20] Under certain assumptions, the machinery of feasible GLS-type corrections could be employed for cases where the errors are thought to be correlated in the short-run, but within the broader problem we wish to address in this section, this issue remains a somewhat mild technicality. Finally, some restrictions on the initial conditions are also needed.

The key part of the testing strategy of Bai and Carrion-i-Silvestre (2007) is based on constructing so-called modified Sargan–Bhargava statistics (henceforth MSB), due to Sargan and Bhargava (1983) and Stock (1999). The unit-specific MSB statistics are based on computing:

$$MSB_i = \frac{T^{-2} \sum_{t=1}^{T} \hat{e}_{i,t}^2}{\hat{\omega}_i^2}, \tag{13.13}$$

where the denominator is a measure of the long-run variance of $e_{i.t}$.

The most important aspect of the analysis is therefore to extract consistent estimates of $e_{i,t}$ (denoted $\hat{e}_{i,t}$) in the presence of common factors and structural breaks, and to test these series for the presence of unit roots. In addition to the different specifications for the breaks which may be considered (to allow for breaks only in the intercept or also to allow for changes in the trend slope), a further crucial distinction among the methods is to deal with the cases where the break date(s) are known versus cases where they need to be estimated consistently (for example, by means of the algorithms considered in Bai and Perron, 1998).[21] We touch upon all these issues in turn.

Under the assumption of cross-sectional independence of the idiosyncratic components, the final form of the test statistic proposed by Bai and Carrion-i-Silvestre takes the form of pooling the individual MSB_i statistics, using correction terms for the mean and variance so that the limiting distribution is standard normal. An alternative strategy involves pooling p-values in the spirit of Maddala and Wu

(1999). For pooling to lead to tests with power requires all the idiosyncratic processes to be integrated under the null hypothesis, while under the alternative a strictly positive fraction of these processes is stationary even as $N \to \infty$.

13.2.3.1 Break dates known

Two models are considered, given by the specification of the deterministic processes.

$$\textit{Model 1}: \quad D_{i,t} = \mu_i + \sum_{j=1}^{l_i} \theta_{i,j} DU_{i,j,t}, i = 1, 2, \ldots, N.$$

$$\textit{Model 2}: \quad D_{i,t} = \mu_i + \delta_i t + \sum_{j=1}^{l_i} \theta_{i,j} DU_{i,j,t} + \sum_{k=1}^{m_i} \gamma_{i,k} DT^*_{i,k,t}, i = 1, 2, \ldots, N.$$

Thus in Model 1 only the intercept of the process is broken, with the ith unit being characterized by l_i breaks at fractions (of the time-span of the sample) given by $\frac{T^i_{a,j}}{T} = \gamma_{i,j}$, which are fractions that remain constant as $T \to \infty$. In Model 2, by contrast, both the intercept and trend are broken (the latter at fractions given by $\frac{T^i_{b,k}}{T} = \lambda_{i,k}, k = 1, 2, \ldots, m_i$). The break dates can be positioned heterogeneously across i, the breaks can be of different magnitudes, each unit may have different numbers of breaks, and the breaks in the intercept can be located at different time periods from the breaks in trend.

To presage the results somewhat, it is simple to note that in Model 1, since the factors are extracted from a model for the differenced observations, the breaks in intercept reduce to impulse dummies which do not, in fact, have any impact on the asymptotic distribution of the MSB statistics. Thus whether or not the breaks are known or unknown does not make any difference, as long as these breaks are restricted to the intercept. By contrast, for Model 2, where the breaks in trend reduce upon differencing to changes in the intercept of the differenced process, the statistics do depend upon the nuisance parameters $\lambda_{i,k}$ and critical values need to be computed to take this into account. In addition, if the break dates are not known *a priori*, consistent estimates of the break fractions are needed.

Analysis of Model 1

Differencing Model 1 yields:

$$\Delta y_{i,t} = \pi_i' \Delta F_t + \Delta e^*_{i,t}, \tag{13.14}$$

where:

$$\Delta e^*_{i,t} = \Delta e_{i,t} + \sum_{j=1}^{l_i} \theta_{i,j} D(T^i_{a,j})_t,$$

and $D(T^i_{a,j})_t$ are the impulse dummies with $D(T^i_{a,j})_t = 1$ if $t = T^i_{a,j} + 1$ and 0 elsewhere.

As in section 13.2.2.2, let:

$$\Delta y_i = (\Delta y_{i,2}, \ldots, \Delta y_{i,T})'$$
$$\Delta e_i^* = (\Delta e_{i,2}^*, \ldots, \Delta e_{i,T}^*)'$$
$$\Delta F = (\Delta F_2, \ldots, \Delta F_T)' \quad ((T-1) \times r \text{ matrix of differenced factors for all the units})$$
$$\pi_i = (\pi_{i,1}, \ldots, \pi_{i,r})' \quad (r \times 1 \text{ vector of loadings of factors for } i\text{th unit}).$$

Then we may write the model (in vector notation) as:

$$\tilde{y}_i = f\pi_i + z_i,$$

where $\tilde{y}_i = \Delta y_i$, $f = \Delta F$ and $z_i = \Delta e_i^*$.

The estimated factors $\hat{f}$ and their loadings $\hat{\pi}_i$ are calculated as in section 13.2.2.2, and:

$$\hat{z}_i = \tilde{y}_i - \hat{f}\hat{\pi}_i; \text{ and, finally, } \hat{e}_{i,t} = \sum_{s=2}^{t} \hat{z}_{i,s}.$$

Since the MSB statistic is not affected by the impulse dummies, and hence by the break fractions γ_i, Bai and Carrion-i-Silvestre prove that under the null hypothesis of a unit root and the assumption that $\frac{T_{a,j}^i}{T} = \gamma_i$ remain constant as $T \to \infty$:

$$MSB(i) \Rightarrow \int_0^1 W_i^2(r)dr,$$

where $W_i(r)$ is a standard Brownian motion independent across i.

The pooled statistic has the form:

$$Z = \sqrt{N}\frac{\overline{MSB} - 0.5}{1\Big/\sqrt{3}},$$

and has a limiting normal distribution, as $N \to \infty$, where $\overline{MSB} = \frac{1}{N}\sum_{i=1}^{N} MSB_i$.

The mean and variance correction terms are those appropriate for the individual $MSB(i)$ statistics.

A pooled Fisher-type test can also be constructed. Denoting by p_i the p-value of the $MSB(i)$ test for the ith unit, then:

$$BC_{\chi^2} = -2\sum_{i=1}^{N} \log p_i \sim \chi_{2N}^2.$$

This is a result applicable to cases where the cross-section dimension N is finite. When N is large, we may also use the asymptotic approximation to the chi-squared statistic proposed by Choi (2001), that is,

$$BC_N = \frac{-2\sum_{i=1}^{N} \log p_i - 2N}{\sqrt{4N}} \Rightarrow N(0,1).$$

These results apply even when the break dates are unknown.[22]

Analysis of Model 2

Here things are slightly more complicated, since differencing the processes does not eliminate the dependence of the relevant test statistics on the break fractions. For Model 2 we have, upon differencing:

$$\Delta y_{i,t} = \Delta F_t' \pi_i + \delta_i + \sum_{k=1}^{m_i} \gamma_{i,k} DU_{i,k,t} + \Delta e^*_{i,t}, \tag{13.15}$$

where $DU_{i,k,t} = 1$ when $T > T^i_{b,k}$ and zero elsewhere are the step dummies which arise from differencing the trend breaks.

To follow Bai and Carrion-i-Silvestre's notation, let:

$$d_i = (\delta_i, \gamma_{i,1}, \ldots, \gamma_{i,m_i})'$$
$$a_{i,t} = (1, D_{i,1,t}, \ldots, D_{i,m_i,t})'$$
$$a_i = (a_{i,2}, \ldots, a_{i,T})'.$$

Then the first-differenced model can be rewritten (using the notation established previously) as:

$$\tilde{y}_i = f\pi_i + a_i d_i + z_i.$$

Since the break dates are assumed known the matrix a_i is completely specified. Conditional on δ_i being known, the variable $w_i = \tilde{y}_i - a_i d_i$ has a "complete" factor structure, in the sense of assumption (i), which follows (13.5)–(13.7) above. f and Π can therefore be estimated based on $w = (w_1, w_2, \ldots, w_N)$. If, on the other hand, $f\pi_i$ is known, regressing $\tilde{y}_i - f\pi_i$ on a_i leads to consistent estimates of d_i. Bai and Carrion-i-Silvestre therefore propose (conditional upon the break dates being known) an iterative procedure for estimating the model as follows:

Step 1: Estimate d_i by least squares – that is, $\tilde{d}_i = (a_i' a_i)^{-1} a_i' \tilde{y}_i$ – ignoring the presence of the factors, which are assumed to have zero mean and will thus be included in the regression errors.

Step 2: Given $\tilde{d}_i$, construct the series $\tilde{w}_i = \tilde{y}_i - a_i \tilde{d}_i$ and estimate the factors and factor loadings to give $\tilde{f}\tilde{\pi}_i$.

Step 3: Regress $\tilde{w}_i$ on a_i to obtain updated estimates of d_i and iterate until convergence.

Step 4: Denoting the final estimates by $\hat{f}$, $\hat{\pi}$ and $\hat{d}_i$, compute $\hat{z}_i = \tilde{y}_i - \hat{f}\hat{\pi}_i - \hat{a}_i \hat{d}_i$ and cumulate to obtain $\hat{e}_{i,t} = \sum_{s=2}^{t} \hat{z}_{i,s}$. Compute the *MSB* statistic for each unit.

Then, as $T \to \infty$,

$$MSB(i, \lambda_i) \Rightarrow \sum_{k=1}^{m_i+1} (\lambda_{i,k} - \lambda_{i,k-1})^2 \int_0^1 V^2_{i,k}(b) db,$$

where:

$\lambda_{i,k} = \frac{T^i_{b,k}}{T}, k = 1, \ldots, m_{i+1},$ which are fractions that remain constant as $T \to \infty$

$\lambda_i = (\lambda_{i,0}, .\lambda_{i,k}, \ldots, \lambda_{i,m_i+1})', \text{ with } \lambda_{i,0} = 0, \lambda_{i.m_i+1} = 1, i = 1, 2, \ldots, N,$

and:

$$V^b_{i,k} = W_{i,k}(b) - bW_{i,k}(1),$$

are Brownian bridges independent across i and k.

Note the dependence of the statistics on the break fractions. Thus, denoting $\lambda = (\lambda_1, \ldots, \lambda_N)'$, we have:

$$Z = \sqrt{N}\frac{\overline{MSB} - \overline{\xi}}{\overline{\zeta}} \Rightarrow N(0,1)$$

$$\overline{MSB} = \frac{1}{N}\sum_{i=1}^{N} MSB(i, \lambda_i)$$

$$\xi_i = \frac{1}{6}\sum_{k=1}^{m_i+1} (\lambda_{i,k} - \lambda_{i,k-1})^2$$

$$\zeta_i^2 = \frac{1}{45}\sum_{k=1}^{m_i+1} (\lambda_{i,k} - \lambda_{i,k-1})^4$$

$$\overline{\xi} = \frac{1}{N}\sum_{i=1}^{N} \xi_i$$

$$\overline{\zeta}^2 = \frac{1}{N}\sum_{i=1}^{N} \zeta_i^2.$$

The Fisher or p-value versions of these tests *à la* Choi (2006a), referred to as BC_N, and *à la* Maddala and Wu (1999), referred to as BC_{χ^2}, can similarly be constructed, as long as the break fractions are known. This is because the p-values for the individual tests will depend on the break fractions in the presence of breaks in trend.

13.2.3.2 Break dates unknown

The results and methods discussed above go through as before for Model 2 (note that for Model 1 knowledge of the break fractions is not a relevant consideration) provided consistent estimates of the break fractions can be obtained. From (13.15), defining the composite error $n_{i,t} = \Delta F_t' \pi_i + \Delta e^*_{i,t}$, and assuming without loss of generality that ΔF has zero mean, we can write the model as:

$$\Delta y_{i,t} = \delta_i + \sum_{k=1}^{m_i} \gamma_{i,k} DU_{i,k,t} + n_{i,t}.$$

Then (13.15) can be thought of as a model with intercept breaks, where the number and timing of the breaks can be calculated (unit by unit) using the Bai–Perron dynamic programming algorithm. Thus, due to the fast convergence of the estimated break points to their "true" values, the a_i matrix can be assumed to be known upon substituting a_i by $\hat{a}_i$. Asymptotically, given consistency, the theoretical results are unaffected by replacing the true break dates with their estimated values. The accuracy of the estimation of the break dates and its impact on the properties of the testing for unit roots using the MSB statistic may well be a pragmatic issue in cases where T is relatively small. This can again be addressed within the framework of simulation studies, an example of which is contained in Bai and Carrion-i-Silvestre's paper but for which further work is clearly justifiable.

13.2.4 Two empirical examples

In this section, we illustrate the arguments and tests developed above by means of two examples, analyzing in particular the effect of incorporating cross-sectional dependence (via factors) and structural breaks into the testing procedures. The two examples are based on Wagner (2008a) and Wagner (2008b) respectively.

13.2.4.1 Purchasing power parity

The empirical analysis of (weak) purchasing power parity (PPP), in an $I(1)$ modeling framework, typically formulated as stationarity of the real exchange rate (RER), is a prime application of both time series and panel unit root (respectively cointegration) testing. We have already referred to the influential paper by O'Connell (1998) when discussing the consequences of overlooking the impact of cross-sectional dependence.

In logarithms, the RER for country i is given by:

$$q_{i,t} = e_{i,t} + p_{i,t} - p_t^*, \tag{13.16}$$

where $q_{i,t}$ is the RER, $e_{i,t}$ is the nominal exchange rate, $p_{i,t}$ is the price level in country i and p_t^* is the price level of the base country (all in logarithms). In our application below the base country is the United States, the nominal exchange rates are thus vis-à-vis the US dollar (per unit of local currency) and the prices are given by the consumer price indices, which implies that, like almost all of the literature, we do not study real exchange rates but real exchange rate indices.

Early panel studies of PPP, including Coakley and Fuertes (1997), Frankel and Rose (1996), Lothian (1997) and Wu (1996), have used first-generation panel unit root tests to test stationarity (respectively unit root behavior) of RER panels. Panel methods have been used to overcome the deficiencies of time series unit root tests, as highlighted in the PPP context by Engel (2000) and as discussed in the introductory sections to this chapter.

Such studies have often "found support for PPP," that is, rejections of the unit root null hypothesis, although it is presumably well understood that rejection of the null hypothesis (that is, a unit root in RER) is not acceptance of the alternative. Much more seriously, the use of first-generation tests designed for cross-sectionally independent panels appears to be troublesome in the PPP context.

Looking at (13.16), it becomes clear that there are several potential sources of cross-sectional dependence in RER panels. First, all nominal exchange rates are coupled to each other by no arbitrage restrictions which appear to hold quite well in liquid foreign exchange markets. Given that nominal exchange rates, if not fixed to the base country currency, typically fluctuate more than prices, this dependence will not be wiped out by price level movements across countries. Second, all RERs contain the same (non-stationary) base country price index p_t^*. Since the goods contained in the consumer baskets generally differ across countries, it is realistic to assume that the permanent components in $(e_{i,t} + p_{i,t})$ and p_t^* do not exactly coincide. If they do not perfectly coincide, then in general the $q_{i,t}$ series contain a common permanent component related to p_t^*. Third, the no arbitrage (in the goods markets) arguments underlying the law of one price and PPP rely upon economic interaction of one form or the other. The world economy becomes ever more integrated and thus shocks can also be expected to be transmitted more strongly across countries. The type of short-run dependence considered by O'Connell (1998) and discussed above may not be sufficient since, as we have just argued, RER panels may be prone to the presence of common non-stationary components.

Lyhagen (2000) studies a special case of this situation with one common stochastic trend and shows that several first-generation tests, including those of Levin, Lin and Chu (2002) and Im, Pesaran and Shin (2003), are severely affected in this case. In particular, Lyhagen shows that, in the presence of one common stochastic trend, the size of the tests tends to one with increasing cross-sectional dimension.

To assess the impact of cross-sectional dependence when testing for PPP, Wagner (2008a) applies a battery of first- and second-generation panel unit root tests to four monthly RER panels. The sample periods, as well as the cross-section and time dimensions of the panels, are contained in Table 13.1. Details are given in Appendix A.

The euro-area dataset (the data are taken from the IMF IFS, OECD MEI and ECB databases) consists of 11 of the 12 countries that were starting members of the euro-area in January 1999, with Ireland missing due to constraints on data availability. The Central and Eastern European countries (CEEC) dataset consists of 11 transition

Table 13.1 Time periods and panel dimensions of the four monthly datasets considered

	Start	*End*	*T*	*N*
Euro-area	1980/1	1998/12	228	11
CEEC	1993/1	2004/6	138	11
Industrial	1980/1	1998/12	228	29
Worldwide	1981/1	2004/4	280	57

Note: The number of observations over time is denoted by T while N denotes the cross-sectional dimension.

Table 13.2 Results of first-generation panel unit root and stationarity tests

	LLC	*Breitung*	IPS_t	*MW*	*Hadri*
Euro-area	*−1.31*	*−1.37*	*−1.34*	24.65	*12.79*
CEEC	**–8.87**	−0.58	**–4.17**	**98.69**	**14.52**
Industrial	−1.04	**–2.49**	**–1.73**	66.10	**22.68**
Worldwide	**–9.64**	6.48	**–2.60**	**207.30**	**51.70**

Notes: Bold entries indicate rejection of the unit root null hypothesis at the 5% critical level and italic entries indicate rejection at the 10% level.
LLC: Levin, Lin and Chu test in section 13.2.1.1.
Breitung: Breitung modification of LLC test in section 13.2.1.1.
IPS_t: t-test proposed by Im, Pesaran and Shin in section 13.2.1.2.
MW: Maddala and Wu's p-value test in section 13.2.1.3.
Hadri: Test with stationarity as null in section 13.2.1.4.

economies with the sample ranging from January 1993 to June 2004. The start date is chosen to exclude the high inflation period of the early 1990s. Two other datasets containing a larger number of countries are also considered. One of these is an industrial countries dataset consisting of 29 countries, including the countries of the euro-area dataset, for which the sample period coincides with the sample period for the euro-area. The other, labeled Worldwide, contains 57 non-euro-area countries for which monthly data are available back to January 1981.

The detailed discussion in Wagner (2008a) shows clearly that all four panel datasets exhibit cross-sectional dependence, investigated by computing the long-run covariance matrix of $\Delta q_t = \left[\Delta q_{1,t}, \ldots, \Delta q_{N,t}\right]'$ (respectively sub-vectors thereof), by inspecting the cross-correlation functions and by cointegration analysis. In particular there appears to be evidence for the presence of common non-stationary components, in line with the discussion above.

The results obtained when applying a battery of first-generation panel unit root tests are reported in Table 13.2.

Depending upon the panel considered, two or three of the four reported panel unit root tests lead to a rejection of the unit root null hypothesis. For the euro-area dataset the three rejections only occur at the 10% level. The quite substantial number of rejections, and hence the strong evidence "in favor" of PPP, is a typical finding in this literature and is in line with the results discussed in Lyhagen (2000).

When the null is taken to be stationarity of the $q_{i,t}$ series, however, the reported results for the Hadri test show rejections of the null hypothesis of stationarity for all the panels. This is consistent with the evidence concerning non-stationarity in the RER panels, as documented in detail in Wagner (2008a), and runs counter to the evidence from the unit root tests. However, as illustrated by Hlouskova and Wagner (2006), the poor performance of the Hadri test, with rejections occurring far too often whenever the data exhibit sizeable serial correlation, severely limits the usefulness of this test, even if the panels were cross-sectionally independent.

Table 13.3 Results of Bai and Ng (2004) analysis

	Factors	BN_N	$BN_{\chi 2}$	$MQ_c(m)$	$MQ_f(m)$
Euro-area	6	0.17	23.14	6	6
CEEC	5	1.24	30.25	5	5
Industrial	4	**1.93**	**78.76**	4	4
Worldwide	4	0.25	117.78	3	3

Notes: Bold entries indicate rejection of the unit root null hypothesis at the 5% critical level; BN_N and $BN_{\chi 2}$ denote the Bai and Ng tests on the estimated idiosyncratic components described in section 13.2.2.2; and $MQ_c(m)$ and $MQ_f(m)$ are the Bai and Ng tests for common trends in section 13.2.2.2.

Further evidence concerning the cross-sectional dependence structure in the RER panels is collected by applying the Bai and Ng (2004) methodology for computing common factors and the results are reported in Table 13.3. The second column contains the estimated number of common factors chosen according to the information criterion BIC_3 of Bai and Ng (2002), while the third and fourth columns provide the results of tests on the idiosyncratic components based on using the pooled inverted normal test and the Maddala and Wu test respectively. The fifth and sixth columns give the number of common trends amongst the common factors according to the two different tests described previously.

These results reflect a well-known weakness of the information criteria to determine the number of common factors (see the second column), which tend to favor large numbers of estimated common factors (given that the upper bound for the number of common factors is six). For the euro-area six factors are selected, five factors are selected for the CEEC panel and four for the other two datasets. Onatski (2006) reports simulations showing that correlation between the idiosyncratic components of the individual units (for which evidence is presented in Wagner, 2008a, for the RER panels at hand) leads to overestimation of the number of common factors when using the information criteria of Bai and Ng (2002). Thus, the results concerning the *number* of common factors should be interpreted with caution, given that the cross-sectional dimensions are rather small in the panels.

The second striking feature is that the number of common trends is selected to be equal to the number of common factors, with the exception of the worldwide dataset where the number of common trends in the common factors is selected to be three (that is, a lot of evidence for unit roots in the common factors). Simulations performed by the authors indicate a tendency of the tests to lead to too large a number of common trends unless, for example, the time series dimension T is very large. Hence, the results concerning the number of common trends also have to be interpreted with some caution.

The Bai and Ng (2004) panel unit root tests are designed for cross-sectionally independent data. In order to verify whether the de-factored data are cross-sectionally independent, one can again look at the cross-correlation functions or the long-run

covariance matrices of the first differences of the estimated idiosyncratic components. Doing so (for details, see again Wagner, 2008a) indicates that even the de-factored data do not appear to be uncorrelated. In this respect, note that the correlation structure between the estimated idiosyncratic components depends upon the number of factors chosen, which is intuitively clear because extracting too many factors may actually induce cross-sectional dependence. Having this potential caveat in mind, for the idiosyncratic components the unit root hypothesis is only rejected for the industrial country dataset. Thus, from the perspective of the Bai and Ng (2004) approach, there is little support for stationarity in the RER panels.

When the computations are performed with only one common factor as an additional robustness check, the unit root null hypothesis is not rejected for this common factor for the euro-area, CEEC and industrial countries datasets but is rejected for the common factor in the worldwide dataset. With one common factor, the panel unit root tests on the idiosyncratic components lead to no rejection of the unit root null hypothesis.

The second-generation test results in Table 13.4 show that, much like for the first-generation tests and for the more restrictive second-generation tests, the unit root null hypothesis is by and large rejected for our datasets, with the exception of the Chang (2002) test.

It is important to remember that these tests are designed for more restricted DGPs than the Bai and Ng (2004) framework allows for. The Pesaran (2007) and Choi (2006a) tests allow for only one common stationary serially uncorrelated factor, with identical loadings for the latter test. The Moon and Perron (2004) approach restricts the factors to be stationary under the alternative. Thus for all these approaches key assumptions necessary for the panel unit root tests are most likely violated, which is consistent with the rejections of the null hypothesis (compare also Breitung and Das, 2008, and Gengenbach, Palm and Urbain, 2006). Finally, it is unclear what drives the non-rejections of the unit root null hypothesis obtained by applying the Chang (2002) test, since this is designed for panels

Table 13.4 Results of other second-generation panel unit root tests

	MP_a	MP_b	*CIPS*	C_p	C_Z	C_{L*}	*NL*
Euro-area	**–9.67**	**–4.88**	−1.96	**5.08**	**–4.56**	**–4.43**	1.52
CEEC	**–12.44**	**–7.86**	**–2.91**	**7.34**	**–4.57**	**–5.31**	1.95
Industrial	**–15.81**	**–6.62**	−1.85	**9.87**	**–8.01**	**–8.04**	1.50
Worldwide	**–20.77**	**–9.37**	**–2.43**	**18.05**	**–12.27**	**–13.67**	−0.05

Notes: The abbreviations are as in the discussion of the tests in section 13.2.2. Bold entries indicate rejection of the unit root null hypothesis at the 5% critical level.
MP_a and MP_b: Moon and Perron tests in section 13.2.2.3.
CIPS: Pesaran test with cross-sectional demeaning in section 13.2.2.3.
C_p, C_Z and C_{L*}: Choi tests in section 13.2.2.3.
NL: Chang test in section 13.2.2.4.

Table 13.5 Results of the tests of Bai and Carrion-i-Silvestre (2007) for unit roots in non-stationary panels with common factors and structural breaks

	Euro-area	*CEEC*	*Industrial*	*Worldwide*
	Z and p-value tests			
Z	6.36	−1.61	2.36	−1.55
BC_N	−2.33	1.44	−0.24	−0.04
$BC_{\chi 2}$	6.54	31.57	55.46	113.37
	Simplified Z and p-value tests			
Z	17.51	3.03	15.96	0.51
BC_N	−2.61	0.49	−1.08	0.05
$BC_{\chi 2}$	4.69	25.25	46.40	114.81
Number of breaks	9	5	11	8

Notes: The number of factors is chosen according to the information criterion BIC_3. The test results reported allow for at most one break in both the intercept and linear trend, except for the euro-area dataset where up to two breaks are allowed for. The critical value (at the 5% nominal level) is given by –1.645 for the Z test, by 1.645 for the BC_N test and by 32.92 (N = 11), 76.78 (N = 29) and 139.92 (N = 57) for the $BC_{\chi 2}$ test. Number of breaks indicates the number of countries in which breaks have been detected. The tests are as described in section 13.2.3.

without non-stationary common components. All in all, the findings show that a careful modeling of cross-sectional dependence that leads to an appropriate choice of panel unit root test is of key importance.

Let us finally turn to the issue of structural change, neglected up to now. Table 13.5 contains the results obtained when applying the Bai and Carrion-i-Silvestre (2007) tests. Bai and Carrion-i-Silvestre also propose simplified approximate tests whose limiting distributions are independent of the break fractions. The results obtained with these tests are also included in Table 13.5.

The major observation that emerges is that the unit root null hypothesis is not rejected for any of the datasets by any of the tests. Thus, even when allowing for structural breaks and common factors, no evidence for stationarity of the RER panels emerges. In particular, contrasting the findings obtained by allowing for multiple factors (and breaks), with no evidence for stationarity of the RER, with the findings obtained with first-generation tests or more restrictive second-generation tests, where seemingly more evidence for PPP prevails, indicates that the resurrection of PPP due to the usage of panel methods may not yet have been accomplished.

13.2.4.2 The environmental Kuznets curve

The empirical example, based on Wagner (2008b), considered in this sub-section is based partly upon using the Groningen dataset mentioned in the introduction. Since the seminal work of Grossman and Krueger (1995) many econometric studies of the relationship between measures of economic development (typically per

capita GDP) and pollution (often proxied by emissions) have been conducted. Most of the papers focus on a specific conjecture, the so-called environmental Kuznets curve (EKC) hypothesis, which postulates an inverted U-shaped relationship between the level of economic development and pollution. This hypothesis therefore states that pollution first rises with increasing income up to a certain point to then fall for income increasing further. The term EKC refers by analogy to the inverted U-shaped relationship between the level of economic development and the degree of income inequality, postulated by Kuznets (1955) in his 1954 presidential address to the American Economic Association.

The theoretical underpinnings for EKCs are discussed, for example, in Andreoni and Levinson (2001), Brock and Taylor (2004, 2005), Jones and Manuelli (2001) and Stokey (1998). Brock and Taylor (2005) identify three different mechanisms that link economic activity with pollution (respectively emissions). These are the scale, composition and technique effects. For unchanging composition of output and unchanging technology, emissions rise alongside the scale of economic activity. For given scale and technique, emissions can rise or fall when the composition of output changes towards a more or less emissions-intensive composition. Finally, emissions (per unit of output, that is, emissions intensity), can decrease with improvements in technology, that is, via improved abatement technology. Depending upon the relative magnitudes of these three effects, a monotonic, a U-shaped, or an inverted U-shaped relation (among the different patterns possible) between per capita GDP and per capita emissions may emerge.

Disentangling the relative importance of the three effects requires detailed structural modeling. The empirical EKC literature is typically less ambitious and focuses on reduced form modeling to address the issue of whether the three distinct mechanisms are jointly sufficiently strong to allow for an inverted U-shaped relationship.

The basic parametric panel formulation of a homogeneous EKC is given by:

$$e_{i,t} = \alpha_i + \gamma_i t + \beta_1 y_{i,t} + \beta_2 y_{i,t}^2 + u_{i,t}, \tag{13.17}$$

where $e_{i,t}$ here denotes the logarithm of per capita emissions, $y_{i,t}$ the logarithm of per capita GDP, $u_{i,t}$ denotes the stochastic error term, $i = 1, \ldots, N$ is the country index and $t = 1, \ldots, T$ is the time index as before. The formulation also includes country specific fixed effects and country specific linear trends, to allow for income independent level and slope effects that could *inter alia* reflect country specific preferences with respect to emissions.

An EKC, that is, an inverted U-shaped relationship, occurs if $\beta_1 > 0$ and $\beta_2 < 0$, in which case the turning point with respect to income (in the homogeneous formulation identical across countries) is given by $y_{TP} = \exp\left(-\beta_1/2\beta_2\right)$. The primary task in an empirical EKC analysis is to estimate the relationship (13.17) by appropriate methods and to test the relevant hypothesis concerning the coefficients.

As discussed in the introduction and also in the convergence example at the beginning of this chapter, $y_{i,t}$ is often considered to be a unit root process. Often, due to the relatively short time series available for many countries, for both per

capita GDP and emissions, the use of a panel perspective may appear fruitful to enhance the performance of unit root and cointegration tests. In any empirical analysis of an EKC like (13.17), using panel unit root and cointegration techniques, all the complications discussed above, such as cross-sectional dependence or structural changes in the deterministic component, may arise and have consequently to be addressed. We turn to each of these problems below. Furthermore, the homogeneity restriction (that is, $\beta_{1i} = \beta_1$ and $\beta_{2i} = \beta_2$ for $i = 1, \ldots, N$) needs to be investigated, which we address in a later section in this chapter (having discussed testing for cointegration and estimation of cointegrating vectors).

There is one additional problem that arises in an equation like (13.17): log per capita GDP and its square are both present. However, at most one of the two processes can be a unit root process, with the other process necessarily being a nonlinear transformation thereof. We return to this issue and its implications later when considering estimation of (13.17).

In our empirical application we consider sulphur dioxide (SO_2) emissions using a balanced panel of 97 countries over the period 1950–2000 (described in detail in a table in Appendix A).

Let us start the discussion of the empirical results by looking at the results obtained with the Bai and Ng (2004) methodology, summarized in Table 13.6. For SO_2 two common factors are found and for GDP one common factor is found. Depending upon the test chosen, $MQ_c(m)$ or $MQ_f(m)$, one or both common factors in SO_2 are non-stationary and the ADF test on the single common factor in GDP also does not lead to a rejection of the unit root null hypothesis. The GDP common factor reflects the evolution of average worldwide GDP and also tracks the observed slowdown of the mid 1970s quite well. The SO_2 common factors are less clearly interpretable and are, furthermore, not cointegrated with the GDP common factor. This implies that there is some long-run disconnect between per capita GDP and per capita SO_2 emissions.

The idiosyncratic components for GDP (as tested by BN_N and BN_{χ^2}) appear to be stationary, since the unit root null hypothesis is rejected by both tests. This implies that the deviations of the individual countries' GDP from the common global factor are stationary and that none of the countries' GDP deviates permanently from the single global stochastic trend.

For SO_2 emissions the null hypothesis of a unit root in the idiosyncratic components is not rejected. This is consistent with large inter-country systemic and technological differences in industry and energy production as well as differences in environmental legislation.

The other, more restrictive, second-generation panel unit root tests, collected in Table 13.7, provide very mixed results. The most relevant results from this table are the Moon and Perron (2004) results, since the other results collected in this table are for tests designed for panels without long-run cross-sectional dependencies.

In the specification with only fixed effects included, the unit root null hypothesis is rejected for the idiosyncratic component for both GDP and SO_2 emissions, whereas it is not rejected when both fixed effects and linear trends are included.

Table 13.6 Results of Bai and Ng (2004) analysis

	Factors	BN_N	BN_{χ^2}	$MQ_c(m)$	$MQ_f(m)$	CF_{ADF}
SO_2	2	−0.70	180.22	1	2	–
GDP	1	**2.89**	**250.96**	–	–	−2.30

Notes: Bold entries indicate rejection of the unit root null hypothesis at the 5% critical level.
CF_{ADF}: ADF test on single common factor as in section 13.2.2.2.
Also see notes to Table 13.3.

Table 13.7 Results of second-generation tests

	MP_a	MP_b	*CIPS*	*Cp*	C_Z	C_{L*}	*NL*
Fixed effects							
SO_2	**–14.70**	**–7.56**	–1.78	**9.45**	**–5.83**	**–6.22**	4.05
GDP	**–17.72**	**–9.05**	–1.64	**1.75**	–0.56	–0.79	8.29
Fixed effects and linear trends							
SO_2	–0.81	–0.49	–2.44	**1.81**	3.89	4.91	6.69
GDP	1.38	1.41	–2.46	–3.01	6.30	7.26	5.82

Notes: Bold entries indicate rejection of the unit root null hypothesis at the 5% critical level. Also see notes to Table 13.4.

The other tests are included to show again that by a strategic choice of panel unit root test any desired conclusion can be supported.

Neglected up to now and potentially important for the case of emissions series is the issue of structural change (for example, due to technological or legal changes). In this respect the short time dimension of the panels represents a major limitation, since the panel unit root tests that allow for structural change in the deterministic component, as described in section 13.2.3, require, when the break dates are assumed to be unknown, the individual specific estimation of these break dates. Nevertheless, applying the Bai and Carrion-i-Silvestre (2007) tests, see Table 13.8, provides evidence for structural change in about a third to a half of the countries included in the panel. All the unit root test results lead to non-rejection of the unit root null hypothesis even when allowing for structural breaks. The results in the table refer to the case with at most one structural break per unit, since even when allowing for more structural breaks only one break is detected.[23]

This example is continued in section 13.3.5 where, in particular, the estimation of (13.17) is discussed.

13.2.5 Some concluding remarks

We have continued to bring together the key strands of the testing strategies within the panel integration framework. Important considerations have included:

Table 13.8 Results of the tests of Bai and Carrion-i-Silvestre (2007) for unit roots in non-stationary panels with common factors and structural breaks

	GDP	SO_2
Z and p-value tests		
Z	−1.38	5.59
BC_N	0.51	−1.01
$BC_{\chi 2}$	210.14	174.02
Simplified Z and p-value tests		
Z	0.63	9.10
BC_N	−1.12	−2.10
$BC_{\chi 2}$	177.70	152.61
Number of breaks	44	39

Notes: The number of factors is chosen according to the information criterion BIC_3. The test results reported allow for at most one break in both the intercept and linear trend. The critical value (at the 5% nominal level) is given by –1.645 for the *Z* test, by 1.645 for the BC_N test and by 227.50 for the $BC_{\chi 2}$ test. Number of breaks indicates the number of countries in which a break has been detected. Also see notes to Table 13.5.

(a) Homogeneity or heterogeneity of the roots under the null and alternative hypotheses.
(b) Deciding upon testing strategy – pooled (LLC) versus group mean methods (IPS), for example, governed largely by the form of the alternative hypothesis adopted.
(c) The need to allow for dependence; based upon the definition presented in Appendix B, we distinguish short-run and long-run dependence, with the latter being related to cross-unit cointegrating relationships and, hence, in a sense made precise in Appendix B, the prevalence of joint common trends across cross-section members. Up to now the most popular model to allow for cross-sectional dependencies is the approximate factor model of Bai and Ng (2004), whose cross-unit cointegration implications are also discussed in Appendix B. For very special cases, short-run cross-sectional dependence can be handled by resorting to relatively simple corrections such as the feasible GLS procedure of O'Connell (1998), which, however, can only be used if the time dimension of the panel exceeds the cross-section dimension.
(d) The construction of the statistics themselves, involving mean and variance corrections to center and standardize the densities of the statistics derived from the individual units.
(e) The startling features of Gaussianity in the limit for many of these statistics, based on constructing statistics that are a weighted aggregate of the individual unit-by-unit statistics. If these units are taken to be independent, under appropriate assumptions (that, for example, ensure the existence of the required

moments) relatively simple sequential central limit theory provides the asymptotic normality of these densities. As soon as the assumption of cross-sectional dependence is lifted, however, issues become much more intricate, as outlined in the highly stylized discussion in Appendix C. Note that cross-sectional dependence of the data often requires the use of joint limit theory, as, for example, in the work of Bai and Ng (2004). More generally, the joint limit theory applied in factor models, where consistent estimation of the factors requires the cross-sectional dimension to tend to infinity, has implications for all unit root and cointegration test procedures that use de-factored observations. This is an issue that we believe deserves more attention in the literature.

(f) The simple ways in which factor structures are utilized and the intuitive manner in which the tests are constructed.

(g) The natural ways in which structural breaks are incorporated.

All of these features are important to the development and use of this methodology, and further work, theoretical, simulation-based and empirical, to investigate the efficacy of the various testing strategies proposed is still largely necessary.

Many of (a)–(g) carry over when it is not a unit root hypothesis that we are interested in testing, but a hypothesis of cointegration among variables of interest, and it is to a consideration of these methods to which we now turn. We could think, for example, of PPP, where we might be interested in seeing co-movement between foreign and domestic prices expressed in the same currency; or exchange rate pass-through, where one could look at how changes in exchange rates are transmitted to the price of imported goods; or the Feldstein–Horioka puzzle where the apparent co-movement of savings and investments (which runs contrary to commonly held beliefs on the consequences of quasi-perfect capital markets); or, of course, the growth literature and issues of convergence discussed earlier.

13.3 Cointegration analysis in non-stationary panels

The majority of cointegration analysis in multivariate time series panels is conducted within the single equation set-up, in which the m-dimensional time series $Y_{i,t}$ are separated into $Y_{i,t} = \left[y_{i,t}, x'_{i,t}\right]'$, where subsequently $y_{i,t}$ is the dependent variable and $x_{i,t}$ are the regressors. This approach is subject to the same limitations as Engle and Granger (1987) single-equation cointegration analysis in the time series case. The most important restriction is that the analysis is limited to situations in which there is either no cointegration in $Y_{i,t}$ (under the null hypothesis of the "no cointegration" tests) or only one cointegrating relationship (under the alternative of the "no cointegration tests").[24] This assumption is in general, with $x_{i,t}$ a multivariate vector of regressor variables, not easily sustainable, except perhaps in special cases. Clearly, methods that allow for higher-dimensional cointegrating spaces are therefore also relevant in the panel cointegration context. Such methods have until now been based on panel extensions of VAR cointegration analysis and are discussed in section 13.3.2.

Single-equation methods, however, offer some advantages, since they allow us to consider – paralleling much of the developments in panel unit root analysis – both cross-sectional dependence via factor models and structural changes in the deterministic components. None of these two aspects has yet been studied in system methods for panel cointegration analysis.

This section starts with a general formulation of the single equation panel cointegration set-up and then continues with discussing tests for cointegration that abstract from cross-sectional dependence and structural change. Structural change is considered next in tests for cointegration, following which allowing for cross-sectional dependence is also added to the testing structure. In section 13.3.1.6 we discuss single equation estimators of the cointegrating vector, for the situation without cross-sectional dependence or structural change. Section 13.3.2 then discusses testing for cointegration as well as estimation of the cointegrating spaces in panel VAR models, under the assumptions of cross-sectional independence and no structural change. Empirical examples are again used throughout to illustrate the techniques.

13.3.1 Single equation analysis of cointegration

Paralleling the set-up of the DGP for studying the unit-root problem given by (13.5)–(13.7) above, we may consider describing the general set-up of testing for cointegration in panels:

$$y_{i,t} = D_{i,t} + x_{i,t}'\beta_{i,t} + u_{i,t} \tag{13.18}$$

$$u_{i,t} = \pi_i' F_t + e_{i,t} \tag{13.19}$$

$$(1-L)F_t = C(L)\eta_t \tag{13.20}$$

$$\left(1-\varphi_i L\right) e_{i,t} = H_i(L)\varepsilon_{i,t} \tag{13.21}$$

$$(1-L)x_{i,t} = v_{i,t}, \tag{13.22}$$

where the major difference to the set-up considered in (13.5)–(13.7) is the presence of additional regressors $x_{i,t}$ in (13.22) that are potentially related to the dependent variable $y_{i,t}$ via a cointegrating relationship. Note that in general one can consider, as in (13.18), the cointegrating relationship to be both individual specific and time varying.

It is important to note that cointegration in the usual sense occurs if the error process $u_{i,t}$ in (13.18) is stationary. However, this is not the only possible form of relationship in which one could be interested. One could, for example, also consider cointegration after taking out the common factors,[25] and this is the additional aspect that the cross-sectional dimension and the modeling of cross-sectional dependence brings into play.

Short-run dependence across the units may, under appropriate circumstances and assumptions, be dealt with as before by considering the variance–covariance structure of the error processes.

As in the unit root discussion, before dealing with the general problem, let us consider a set of simplifications that help to illustrate many of the issues involved. Suppose we start, as in section 13.2, by switching off the factor dependence structure and assuming that both the deterministic processes and the cointegrating vector are unbroken. Then a very simple version of the system reduces to:

$$y_{i,t} = D_{i,t,m} + x'_{i,t}\beta_i + u_{i,t}, \quad m = 1, 2, 3 \tag{13.18′}$$

$$(1 - L)x_{i,t} = v_{i,t}, \tag{13.22′}$$

where the index m again describes the usual specifications of the deterministic component. $D_{i,t,m}$ can either be empty (that is, contain no deterministic terms), $m = 1$, or have a constant, $m = 2$, or a constant and a linear trend, $m = 3$.

Let us further assume that the vectors $x_{i,t}$, and therefore β_i and $v_{i,t}$, are l-dimensional and that, if $y_{i,t}$ and $x_{i,t}$ are cointegrated, the *unique* cointegrating vector is given by $(1, -\beta'_i)'$. Assume further that, when cointegration prevails, the processes $e_{i,t} = (u_{i,t}, v'_{i,t})'$ are cross-sectionally *independent* stationary autoregressive moving average (ARMA) processes. The ARMA assumption is stronger than we need but serves well for the purposes of illustration. In particular we are able to appeal, under this ARMA assumption, to the existence of a finite long-run covariance matrix for $e_{i,t}$, given by:

$$\Omega_i = \begin{bmatrix} \omega^2_{u,i} & \Omega_{uv,i} \\ \Omega'_{uv,i} & \Omega_{v,i} \end{bmatrix}.$$

$\Omega_{v,i}$ is taken to be of full rank, which excludes cointegration amongst the variables $x_{i,t}$. We also need to define the conditional long-run variance as:

$$\omega^2_{u.v,i} = \omega^2_{u,i} - \Omega_{uv,i}\Omega^{-1}_{v,i}\Omega'_{uv,i},$$

and the matrix:

$$\Lambda_i = \sum_{j=0}^{\infty} E(e_{i,t}e'_{i,t-j}),$$

partitioned conformably with Ω_i. Finally, let us take $\beta_i = \beta \ \forall i$ if cointegration exists. This is an assumption that can be generalized easily, as discussed by Pedroni (2004).

If there is no cointegration, equation (13.18′) is a spurious relationship with $y_{i,t}$ being an $I(1)$ process not cointegrated with $x_{i,t}$. In this case we can assume, analogously to the ARMA assumption above, that $\Delta y_{i,t} = u^{sp}_{i,t}$ and that $e^{sp}_{i,t} = \left(u^{sp}_{i,t}, v'_{i,t}\right)'$ are cross-sectionally independent stationary ARMA processes with full rank long-run covariance matrices. Of course, this includes the case of independence of $y_{i,t}$ and $x_{i,t}$. The superscript "sp" here is chosen to indicate the spurious regression nature of the relationship in the case of no cointegration.

13.3.1.1 Testing for the null hypothesis of no cointegration – Pedroni (1999, 2004)

Write:

$$\Delta u_{i,t} = \rho_i u_{i,t-1} + \upsilon_{i,t}, \tag{13.23}$$

where $\upsilon_{i,t}$ is a stationary ARMA process for all the units. As in the tests for unit roots, two combinations of the null and alternative hypotheses may be considered – the first applying to pooled tests, the second to group mean tests. Under the former, the null hypothesis is given by $H_0 : \rho_i = 0 \; \forall \; i = 1, 2, \ldots, N$ against the homogeneous alternative hypothesis $H_A : \rho_i = \rho < 0 \; \forall \; i = 1, 2, \ldots, N$, where we again restrict attention to stationarity under the alternative. The group mean tests are based on $H_A^1 : \rho_i < 0$ for $i = 1, 2, \ldots, N_1$ and $\rho_i = 0$ for $i = N_1 + 1, \ldots, N$, where $\lim_{N\to\infty} \frac{N_1}{N} = k > 0$. The analogy with the unit-root testing framework is obvious and many of the same estimation and testing principles apply. If $u_{i,t}$ were known the procedures would be exactly the same – in practice, since we must estimate $u_{i,t}$, the analogy is not exact and the tests must be based on estimating equations of the form (13.18′) instead.

Let us denote by $\hat{u}_{i,t}$ the regression residuals from (13.18′) estimated by OLS. Two adjustments to the OLS coefficient are needed, first to account for the endogeneity of the regressors and second to account for the ARMA structure in $\upsilon_{i,t}$.

The first adjustment requires an estimate of $\omega^2_{u.\upsilon,i}$, denoted $\hat{\omega}^2_{u.\upsilon,i}$. This may be done by first estimating the OLS regression of $\Delta y_{i,t}$ on the deterministic components and $\Delta x_{i,t}$, extracting the residuals $\hat{\upsilon}_{i,t}$ and fitting an ARMA or AR model to this derived process (and computing its long-run variance.) Alternatively, a nonparametric estimator of the form prescribed by Newey and West (1987) may also be used.

The correction for serial correlation can also be dealt with either parametrically, by means of ADF regressions on the residuals $\hat{u}_{i,t}$, or nonparametrically from the nonaugmented regressions given by (13.23), with $\hat{u}_{i,t}$ replacing $u_{i,t}$.

In order to compute the nonparametric correction for serial correlation, denote the estimated variance of the residuals in (13.23) by $\hat{\sigma}^2_{\upsilon,i}$ and the corresponding long-run variances by $\hat{\omega}^2_{\upsilon,i}$. The serial correlation correction factors are $\psi_i = \frac{1}{2}\left(\hat{\omega}^2_{\upsilon,i} - \hat{\sigma}^2_{\upsilon,i}\right)$ and let $\hat{\omega}^2_{N,T} = \frac{1}{N}\sum_{i=1}^{N} \frac{\hat{\omega}^2_{\upsilon,i}}{\hat{\omega}^2_{u.\upsilon,i}}$. The following test statistics for "no cointegration" can now be defined, laid out in four different categories. Each of the statistics requires re-scaling and centering in order for the asymptotic distributions to hold.

(i) Pooled tests (nonparametric corrections)

$$\textit{variance ratio}: \qquad N^{1/2}\left(N^{-1}\sum_{i=1}^{N}\hat{\omega}^{-2}_{u.\upsilon,i}\left(T^{-2}\sum_{t=2}^{T}\hat{u}^2_{i,t-1}\right)\right)^{-1}$$

$$\textit{pooled } \rho\textit{-test}: \quad N^{1/2}\frac{N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-1}\sum_{t=2}^{T}\hat{u}_{i,t-1}\Delta\hat{u}_{i,t}-\psi_i\right)}{N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-2}\sum_{t=2}^{T}\hat{u}_{i,t-1}^{2}\right)}$$

$$\textit{pooled } t\textit{-test}: \quad N^{1/2}\frac{N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-1}\sum_{t=2}^{T}\hat{u}_{i,t-1}\Delta\hat{u}_{i,t}-\psi_i\right)}{\hat{\omega}_{N,T}\left(N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-2}\sum_{t=2}^{T}\hat{u}_{i,t-1}^{2}\right)\right)^{1/2}}.$$

(ii) Pooled tests (parametric corrections)
In order to compute the parametrically corrected versions of the t-test, a similar device to that discussed previously in the construction of the LLC test is used. Two auxiliary regressions are estimated:

$$\Delta\hat{u}_{i,t}=\sum_{k=1}^{K_i}\gamma_{1,ik}\Delta\hat{u}_{i,t-k}+\zeta_{1,i,t}$$

$$\hat{u}_{i,t-1}=\sum_{k=1}^{K_i}\gamma_{2,ik}\Delta\hat{u}_{i,t-k}+\zeta_{2,i,t},$$

where the lag-length selection (of K_i) may be undertaken using automatic selection criteria such as AIC. Next, $\hat{\zeta}_{1,i,t}$ is regressed on $\hat{\zeta}_{2,i,t}$:

$$\hat{\zeta}_{1,i,t}=\rho_i\hat{\zeta}_{2,i,t}+\theta_{i,t},$$

and $\hat{\sigma}_{NT}^{2}=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=K_i+2}^{T}\hat{\theta}_{i,t}^{2}$ is computed. The variance of the estimated residuals $\hat{\theta}_{i,t}$, denoted $\hat{\sigma}_{\theta_i}^{2}$, needed for the computation of the group mean tests is also computed. The parametrically corrected pooled t-test is then:

pooled t-test – parametrically corrected:

$$N^{1/2}\frac{N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-1}\sum_{t=K_i+2}^{T}\hat{\zeta}_{1,i,t}\hat{\zeta}_{2,i,t}\right)}{\hat{\sigma}_{N,T}\left(N^{-1}\sum_{i=1}^{N}\hat{\omega}_{u.v,i}^{-2}\left(T^{-2}\sum_{t=K_i+2}^{T}\hat{\zeta}_{2,i,t}^{2}\right)\right)^{1/2}}.$$

The group mean tests are defined as follows:

(iii) Group mean tests (nonparametric corrections)

$$\textit{group mean } \rho\textit{-test}: N^{-1/2} \sum_{i=1}^{N} \frac{\left(T^{-1} \sum_{t=2}^{T} \hat{u}_{i,t-1} \Delta \hat{u}_{i,t} - \psi_i \right)}{\left(T^{-2} \sum_{t=2}^{T} \hat{u}_{i,t-1}^2 \right)}$$

$$\textit{group mean } t\textit{-test}: N^{-1/2} \sum_{i=1}^{N} \frac{\left(T^{-1} \sum_{t=2}^{T} \hat{u}_{i,t-1} \Delta \hat{u}_{i,t} - \psi_i \right)}{\hat{\omega}_{\upsilon,i} \left(T^{-2} \sum_{t=2}^{T} \hat{u}_{i,t-1}^2 \right)^{1/2}}$$

(iv) Group mean tests (parametric corrections)

$$\textit{group mean } t\textit{-test – parametrically corrected}: N^{-1/2} \sum_{i=1}^{N} \frac{\left(T^{-1} \sum_{t=K_i+2}^{T} \hat{\zeta}_{1,i,t} \hat{\zeta}_{2,i,t} \right)}{\hat{\sigma}_{\theta,i} \left(T^{-2} \sum_{t=K_i+2}^{T} \hat{\zeta}_{2,i,t}^2 \right)^{1/2}}.$$

The mean and variance correction terms (for large N and T) for each of these seven tests depend upon the deterministic specification considered, as well as upon the dimension of the $x_{i,t}$ vector. Once recentered and scaled by the appropriate mean and variance corrections, the standardized statistics tend in the limit to the $N(0, 1)$ density (under sequential convergence, $T \to \infty, N \to \infty$). Hlouskova and Wagner (2008) contains a large set of finite sample and asymptotic correction factors for up to 12 regressors.

13.3.1.2 Some general remarks

It is worth re-emphasizing that the testing principles derived earlier for unit root tests in panels carry over in direct ways to the testing for cointegration, with the obvious (but by no means trivial) embellishments of endogeneity correction, and the need for methods that estimate the cointegrating vector efficiently and consistently when required. A class of tests due to Kao (1999) is constructed under the same general framework as described above for the Pedroni tests, while Westerlund (2005) develops two simple nonparametric tests, one against a homogeneous alternative whilst the other is a group mean test against a heterogeneous stationary alternative.

Wagner and Hlouskova (2007), in a companion study to Hlouskova and Wagner (2006), report the results of an extensive simulation exercise, one part of which is devoted to looking at the behavior of the Pedroni and Westerlund tests. The study explores the size and power properties of the tests under the baseline case where cross-sectional independence is imposed in the DGP, but also reports results when cross-sectional independence is violated both via a correlation structure, as described in section 13.2.2.1 (or a slight modification, where the correlation matrix

has geometrically declining weights – that is, the (i, j)th element of the matrix Ω is given by $\omega^{|i-j|}$) and where there is also cross-unit cointegration.

They report that, among the seven Pedroni single-equation tests for the null hypothesis of no cointegration, the two parametric tests (based on estimating ADF regressions) are the ones that have the best size properties. The remaining five are severely undersized and also have low power, especially for small values of T. The authors conjecture that this may be due to the use of asymptotic correction terms (used to standardize the test statistics) when finite sample correction terms or bootstrapped values may be more beneficial.

However, the conservative properties of the parametric tests also imply that these are the ones least affected by the presence of short-run cross-sectional correlation or cross-unit cointegration of the form considered by Wagner and Hlouskova (2007). Systems tests for cointegration, discussed in section 13.3.2, are in fact outperformed by Pedroni-type tests (although the latter are not applicable in the presence of multiple cointegrating vectors). There is a tendency to overestimate the rank of the cointegrating space (that is, the number of cointegrating vectors), and size and power distortions occur not only when T is relatively small but also when the N-dimension is large (because of the difficulty with dealing with high-dimensional systems alluded to earlier.) The use of finite sample corrections or bootstrapped values is found to be efficacious here, as also reported in earlier work by Banerjee, Marcellino and Osbat (2004).

Banerjee and Carrion-i-Silvestre (2007) consider extensions of the Pedroni tests to allow for structural breaks, discussed in the following sub-section, and for structural breaks and cross-sectional dependence, discussed in section 13.3.1.4 thereafter. Structural breaks are allowed to occur in both the deterministic components and the cointegrating vector.

13.3.1.3 Allowing for structural breaks in the Pedroni tests

As mentioned, the considered structural changes may take the form of breaks in the deterministic processes and/or in the slopes of the cointegrating coefficients β. In the next sub-section, we also allow for cross-sectional dependence via a factor structure.

In this sub-section, we focus on the consequences of relaxing the assumption that the deterministic processes are not broken. Referring back to (13.18), under the null hypothesis $y_{i,t}$ and $x_{i,t}$ are not cointegrated, that is, $u_{i,t}$ is an $I(1)$ process. Under the alternative hypothesis, $u_{i,t}$ is stationary but either the deterministic terms or the cointegrating vectors are time-variant (defined more precisely below), so that, while cointegration exists under the alternative, it does so in the presence of instabilities.

That such a situation is not of pure academic interest becomes evident both from considering the simulation results in Banerjee and Carrion-i-Silvestre (2007), and from the empirical example presented below, where the presence of structural breaks is crucial for establishing the presence of cointegration. The presence of structural breaks severely undermines the size and power properties of the tests for cointegration, especially when the break occurs in the deterministic trend of the processes.

We focus here on the parametric version of the pooled t-test corrected parametrically (given the good properties of this test identified by Wagner and Hlouskova, 2007), but the theoretical and simulation analysis can be repeated for any of the tests developed by Pedroni discussed above or in the related literature.

Let us begin, as in the previous section, with the following flexible representation of the model:

$$y_{i,t} = D_{i,t} + x'_{i,t}\beta_{i,t} + u_{i,t} \tag{13.18''}$$

$$(1-L)x_{i,t} = v_{i,t}, \tag{13.22''}$$

where the further generalization at this stage is to allow for the $D_{i,t}$ and $\beta_{i,t}$ terms to be broken, although for the moment the factor structure remains switched off. For simplicity, the restriction (in relation to the Bai and Carrion-i-Silvestre, 2007, paper) imposed here is that there is only one break allowed for, and that the breaks in intercept and/or trend and/or cointegrating coefficients (under the alternative) all occur at the same, possibly unknown, time period. It is important to note, however, that the timing of these breaks is allowed to vary across the units. These breaks are specified in six ways, constituting six different sub-models nested within (13.18$''$).

Start with the general functional form for the deterministic term $D_{i,t}$:

$$D_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + \gamma_i DT^*_{i,t},$$

where:

$$\begin{aligned} DU_{i,t} &= 0 \ \forall t \leq T_{b,i} \\ &= 1 \ \forall t > T_{b,i}, \end{aligned}$$

and:

$$\begin{aligned} DT^*_{i,t} &= 0 \ \forall t \leq T_{b,i} \\ &= (t - T_{b,i}) \ \forall t > T_{b,i}. \end{aligned}$$

Note that $T_{b,i}$ denotes the time of the break for the ith unit and $\lambda_i = \frac{T_{b,i}}{T}$, which is the fraction of the sample at which the break occurs in the ith unit, remains constant as $T \to \infty$ and belongs to a closed sub-set of (0,1). The time-varying cointegrating vector is specified as a function of time so that:

$$\beta_{i,t} = \beta_i + b_i \cdot DU_{i,t},$$

where $DU_{i,t}$ is as defined above.

With the allowed breaks, at least six different model specifications may be considered and we list these below. The list is not exhaustive, but probably includes the most relevant specifications, especially for the empirical applications that we would have in mind.

Model 1: Constant term with a change in intercept but stable cointegrating vector:

$$y_{i,t} = \mu_i + \theta_i DU_{i,t} + x'_{i,t}\beta_i + u_{i,t}. \tag{13.24}$$

Model 2: Time trend with a change in intercept but stable cointegrating vector:

$$y_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + x'_{i,t}\beta_i + u_{i,t}. \tag{13.25}$$

Model 3: Time trend with change in both intercept and trend but stable cointegrating vector:

$$y_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + \gamma_i DT^*_{i,t} + x'_{i,t}\beta_i + u_{i,t}. \tag{13.26}$$

Model 4: Constant term with change in both intercept and changing cointegrating vector:

$$y_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + x'_{i,t}\beta_{i,t} + u_{i,t}. \tag{13.27}$$

Model 5: Time trend with change in intercept and changing cointegrating vector (the slope of the trend does not change):

$$y_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + x'_{i,t}\beta_i + u_{i,t}. \tag{13.28}$$

Model 6: The intercept, the trend slope and the cointegrating vector all change:

$$y_{i,t} = \mu_i + \delta_i t + \theta_i DU_{i,t} + \gamma_i DT^*_{i,t} + x'_{i,t}\beta_{i,t} + u_{i,t}. \tag{13.29}$$

Under any of these specifications, Banerjee and Carrion-i-Silvestre (2007) propose methods for testing the null hypothesis of no cointegration against the alternative hypothesis of cointegration, with corresponding breaks.

The Banerjee and Carrion-i-Silvestre (2007) proposal is conceptually extremely simple. The analysis may be specialized to the case where the λ_i (break fractions) are assumed to be known but, since this is perhaps not a likely scenario, we present the more general analysis where the break dates are assumed to be unknown. The tests for cointegration in panels with structural breaks are based on time series cointegration tests developed by Gregory and Hansen (1996) and the panel cointegration tests of Pedroni (1999, 2004) discussed above. Testing entails several steps. First, one of the models (depending upon the precise empirical motivation) given in (13.24)–(13.29) above is estimated by OLS for each unit *i* of the panel. Since this requires us to specify the break fraction λ_i, and we assume here that that break dates are unknown, the models must be estimated for each unit for every possible choice of break fraction within a bounded interval, taken here to be the interval $\Lambda = [0.15,\ 0.85]$. This gives rise to a set of ADF regressions, derived from the particular choice of λ_i.

Next, for a given choice of λ_i, and for a particular choice of model, residuals $\hat{u}_{i,t}(\lambda_i)$ are extracted for each unit *i*. These residuals are then used to estimate the augmented Dickey–Fuller-type regression given by:

$$\Delta\hat{u}_{i,t}(\lambda_i) = \rho_i \hat{u}_{i,t-1}(\lambda_i) + \sum_{j=1}^{k} \phi_{i,j} \Delta\hat{u}_{i,t-j}(\lambda_i) + \varepsilon_{i,t}.$$

Since under the null hypothesis $\rho_i = 0\ \forall\ i$, the next step is to compute $t_{\rho_i=0}(\lambda_i)$ for each of these ADF regressions and to take as the estimate of the break-point/fraction

for the ith unit the fraction which minimizes the sequence of the individual t-statistics. That is, for unit i,

$$\hat{T}_{b,i} = \underset{\lambda_i \in [0.15,0.85]}{\arg\min}\ t_{\hat{\rho}_i}(\lambda_i).$$

The limiting distribution of $\inf_{\lambda_i \in [0.15,0.85]} t_{\hat{\rho}_i}(\lambda_i)$ is shown by Gregory and Hansen (1996) not to depend on the break fraction parameters and it is for this reason that the minimization is undertaken over the sequence of break fractions in unit i. The procedure is repeated for all units i, leading to a sequence of "estimated" break fractions for the N units, denoted in vector notation as $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_N)'$.

Finally, the pooled test (in the spirit of Pedroni, 1999, 2004), which allows for breaks under the alternative hypothesis, is given by:

$$N^{-1/2} Z_{\hat{t}_{N,T}}(\hat{\lambda}) = N^{-1/2} \sum_{i=1}^{N} t_{\hat{\rho}_i}(\hat{\lambda}_i).$$

The asymptotic distribution of $N^{-1/2} Z_{\hat{t}_{N,T}}(\hat{\lambda})$ is given by the theorem below. It is important to note that in this framework (apart from the restriction to cross-sectional independence, which is lifted in the following sub-section) a high degree of heterogeneity is allowed across the units, since the cointegrating vector, the short-run dynamics and the break date are all allowed to differ among units. In the spirit of much of this literature, and in particular of Pedroni (1999, 2004), the panel test statistics are shown to converge to standard normal distributions once they have been properly standardized. The correction terms, of course, differ from those tabulated by Pedroni (2004) since the models are considerably more complicated due to the presence of breaks at unknown points of time across the units, but the principles involved remain the same. The following result may now be shown:

Theorem (Banerjee and Carrion-i-Silvestre, 2007, Theorem 1): Let Θ *and* Ψ *denote the mean and variance for the vector Brownian motion functional*:

$$\Upsilon' = \left(\inf_{\lambda_i \in \Lambda} \int_0^1 Q(\lambda_i, s) dQ_i(\lambda_i, s) \left[\int_0^1 Q(\lambda_i, s)^2 ds \right]^{-1}, \inf_{\lambda_i \in \Lambda} \int_0^1 Q(\lambda_i, s) dQ_i(\lambda_i, s) \times \right.$$
$$\left. \left[\int_0^1 Q(\lambda_i, s)^2 ds \cdot (1 + \rho(\lambda_i)' D(\lambda_i) \rho(\lambda_i)) \right]^{-1/2} \right).$$

$Q(\lambda_i, s)$ *and* $\rho(\lambda_i)$ *are functions of vector Brownian motions and the deterministic components and* $D(\lambda_i)$ *depends on the model chosen (see Gregory and Hansen, 1996, for details). Then, as* $T \to \infty, N \to \infty$ *in sequence, under the null hypothesis of no cointegration the asymptotic distribution of the statistic* $Z_{\hat{t}_{N,T}}(\hat{\lambda})$ *is given by* $N^{-1/2} Z_{\hat{t}_{N,T}}(\hat{\lambda}) - \Theta_2 \sqrt{N} \Rightarrow N(0, \Psi_2)$.

Several remarks are appropriate in connection with this theorem. First, we may derive similar theorems for the remaining statistics proposed by Pedroni (1999,

2004) and would reach very similar characterizations of the limiting densities (with different correction terms in each case).

Second, the asymptotic moments of the form Θ_2 and Ψ_2 can be approximated by Monte Carlo simulation for all the different models (for all the different tests.) Banerjee and Carrion-i-Silvestre compute these moments for up to seven stochastic regressors and $T = 1,000$.

Third, since the large sample correction terms (or moments) may perform poorly in finite samples, the moments of the test statistics for different values of T, specifically $T = \{30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 400, 500\}$ are calculated. To generalize the applicability of these techniques yet further, some response surfaces are also computed to approximate the critical values for different values of T.

13.3.1.4 Allowing for cross-sectional dependence

In some sense it is worth thinking of (13.18)–(13.22) above as the most general formulation (within the framework adopted by this chapter) of the elements necessary for testing unit roots and cointegration in a panel, and we propose this structure as one that encompasses almost all of the issues involved. There are difficulties which are not addressed in sufficient detail, for example to do with multiple cointegrating vectors, but in terms of the three key elements of (i) testing for unit roots or cointegration, (ii) cross-sectional dependence, and (iii) instability in the deterministic or stochastic processes, (13.18)–(13.22) offers all the generality that is needed.

One can regard the problem as a series of switches which, when on, introduce the technology relevant to the investigator. For example, the switch from a cointegration to a unit root framework is more or less immediate by abstracting from the $x_{i,t}$ vector, formally setting its dimension equal to zero. The switch for structural breaks in the process is given by the vectors (or scalars in the case of a single break of parameters) $(\theta_i', \gamma_i', b_i')'$, which if set to zero (together or in part) removes structural instability from the relevant parts of the process.[26] Finally, the π_i vectors multiplying the common factors F_t introduce cross-sectional dependence. Cross-sectional dependence can also be introduced via correlation in the idiosyncratic components.

This sub-section deals with testing for cointegration when all the remaining switches (governing dependence and breaks) are, in principle, also on. Thus, in addition to structural breaks, we reintroduce cross-sectional dependence through the factors. We still work within the frameworks of Models 1–6 above, but with the additional element of cross-sectional dependence.

The assumptions governing the $u_{i,t}$ processes in (13.18′) and (13.22′) are as given in section 13.2.2.2 (as they apply to (13.5)–(13.7)). The factor structure (that is, equation (13.19)) is now switched on. An important restriction (applying also to the simpler Pedroni tests described in section 13.3.1.2) should be considered explicitly concerning the relationship between the $e_{i,t}$ process in (13.19) and $v_{i,t}$ in (13.22″).[27] If $E(e_{i,t}|v_{i,t}) = 0$, the regressors are said to be (strictly) exogenous and

the limiting distributions of the statistics do not depend on the stochastic regressors as given in Theorem 2 of Banerjee and Carrion-i-Silvestre (2007). However, if, for example, $E(e_{i,t}|v_{i,t}) = \Delta x_{i,t}' A_i(L) + \xi_{i,t}$, with $A_i(L)$ being a vector of lag and lead polynomials and $\xi_{i,t} \sim$ i.i.d. $(0, \Sigma_\xi)$, the regressors are no longer strictly exogenous and modifications must be introduced to account for the endogenous regressors. The two theorems given below are for the case of exogenous regressors but a simple modification of the testing procedure is suggested to allow for endogeneity.

Consider the case where the break dates are known, so that for the deterministic components only the corresponding parameters need to be estimated.

Compute the first difference of (13.18) to give:

$$\Delta y_{i,t} = \Delta D_{i,t} + \Delta x_{i,t}' \delta_{i,t} + \Delta F_t' \pi_i + \Delta e_{i,t}.$$

Note that, depending on the specification of the deterministic breaks, the differenced deterministic component $\Delta D_{i,t}$ is a mixture of step functions (in the case of a break in trend) and impulse dummies (when there is a break in intercept).

$\Delta y_i = (\Delta y_{i,2}, \ldots, \Delta y_{i,T})'$

$\Delta e_i = (\Delta e_{i,2}, \ldots, \Delta e_{i,T})'$

$\Delta F = (\Delta F_2, \ldots, \Delta F_T)'((T-1) \times r)$ matrix of differenced factors for all the units)

$\pi_i = (\pi_{i,1}, \ldots, \pi_{i,r})'$ ($r \times 1$ vector of loadings of factors for ith unit)

$\Delta x_i = (\Delta x_{i,2}, \ldots, \Delta x_{i,T})'$.

Defining the projection matrix $M_i = (I_{T-1} - \Delta x_i^D (\Delta x_i^{D\prime} \Delta x_i^D)^{-1} \Delta x_i^{D\prime})$, with $\Delta x_{i,t}^D = (1, \Delta D_{i,t}, \Delta x_{i,t}')'$ and $\Delta x_i^D = (\Delta x_{i,2}^D, \ldots \Delta x_{i,T}^D)'$, we have:

$$\begin{aligned} M_i \Delta y_i &= M_i \Delta F \pi_i + M_i \Delta e_i \\ &= f \pi_i + z_i, \end{aligned}$$

where $f = M_i \Delta F$ and $z_i = M_i \Delta e_i$. The projection matrix M_i is sensitive to the specification of the deterministic terms, as indicated by the index D in the definition of this matrix; yet $M_i \Delta F$ cannot be sensitive to i if the common factor representation is to be valid. This is a restriction of the framework used to model dependence, and can be satisfied in two different ways – (a) either the deterministic components do not matter, as in the case where differencing leads to impulse dummies, or (b) where the timing of the breaks in the deterministic terms is the same across the units. This is relevant in the case of trend breaks, where differencing leads to shifts in intercept which are relevant for the derivation of the asymptotic distributions of the statistics, that is, for Models 3 and 6.

Pre-multiplication by this projection matrix serves to isolate the factor components, so that we can now proceed to extract the factors. The estimated factors, denoted $\tilde{f} = (\tilde{f}_2, \ldots, \tilde{f}_T)$ are given by $\sqrt{T-1}$ times the r eigenvectors corresponding to the r largest eigenvalues of the matrix $y^* y^{*\prime}$, where $y^* = (y_1^*, \ldots, y_N^*)'$ and

$y_i^* = M_i \Delta y_i$. Normalizing, $\frac{\tilde{f}'\tilde{f}}{T-1} = I$, we have the estimate of the loading matrix:

$$\tilde{\Pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_N)' = \frac{y^{*\prime}\tilde{f}}{T-1},$$

and:

$$\tilde{z}_{i,t} = y_{i,t}^* - \tilde{f}_t' \tilde{\pi}_i.$$

Finally:

$$\tilde{e}_{i,t} = \sum_{s=2}^{t} \tilde{z}_{i,s},$$

which can now be tested for a unit root, via ADF regressions. This is, in effect, the test for the null of no cointegration, once the common factor structure and the structurally unstable deterministic processes have been accounted for.

Thus, for each unit i, we estimate the regressions:

$$\Delta \tilde{e}_{i,t}^m(\lambda_i) = \alpha_{i,0} \tilde{e}_{i,t-1}^m(\lambda_i) + \sum_{j=1}^{k} \alpha_{i,j} \Delta \tilde{e}_{i,t-j}^m(\lambda_i) + \varepsilon_{i,t},\ m = c, \tau, \gamma,$$

and test the null hypothesis $H_0 : \alpha_{i,0} = 0$ either unit by unit or by constructing a pooled panel test similar in spirit to LLC.

Three sub-cases, $m = c, \tau, \gamma$, are considered by Banerjee and Carrion-i-Silvestre (2007), which index the t-tests $t_{\alpha_{i,0}}^m(\lambda_i)$. $m = c$ denotes models which do not include a time trend and the structural change affects either the intercept and/or the cointegrating vector, $m = \tau$ stands for models with a trend where, as before, the break affects the intercept and/or the cointegrating vector, and $m = \gamma$ stands for models where a change in trend is allowed.

The dependence on the break dates (which had been suppressed earlier) is made clear here. No added estimations are needed at this stage because the breaks are assumed known. In the more general case, however, the tests will have to be based on estimates of the break dates for each of these units, as discussed for the generalization of the Pedroni tests above, but with the further complication of accounting for cross-sectional dependence. For the models considered, the following result has been derived.

Theorem (Banerjee and Carrion-i-Silvestre, 2007, Theorem 2): Under the null hypothesis that $\rho_i = \varphi_i - 1 = 0$:

(a) $t_{\alpha_{i,0}}^c(\lambda_i) \Rightarrow \frac{\frac{1}{2}(W_i(1)^2 - 1)}{\left(\int_0^1 W_i(s)^2 ds\right)^{1/2}}$

(b) $t_{\alpha_{i,0}}^\tau(\lambda_i) \Rightarrow -\frac{1}{2}\left(\int_0^1 V_i(s)^2 ds\right)^{-1/2}$
where $V_i(s) = W_i(s) - sW_i(1)$

(c) $t_{\alpha_{i,0}}^\gamma(\lambda_i) \Rightarrow -\frac{1}{2}\left(\lambda^2 \int_0^1 V_i(b_1)^2 db_1 + \left(1 - \lambda^2\right)\right) \int_0^1 V_i(b_2)^2 db_2$
where $V_i(b_j) = W_i(b_j) - b_j W_i(1)$, $j = 1, 2$ are two independent detrended Brownian motion processes.

Moreover, if we assume for the moment that there is only one factor, then the estimated factor $\tilde{F}_t$ can also be tested for a unit root using an ADF regression of the form:

$$\Delta\tilde{F}_t^m(\lambda_i) = \delta_0\tilde{F}_{t-1}^m(\lambda_i) + \sum_{j=1}^{k}\delta_j\Delta\tilde{F}_{t-j}^m(\lambda_i) + u_t.$$

Then, for cases where there is no break in trend:

(d) $$t_{\delta_0} \Rightarrow \frac{\int_0^1 W^d(s)dW^d(s)}{\left(\int_0^1 W^d(s)^2ds\right)^{1/2}}$$

where $W^d(s)$ denotes a detrended Brownian motion. However, allowing for a change in trend leads to dependence on the break fraction, such that:

(e) $$t_{\delta_0}(\lambda) \Rightarrow \frac{\int_0^1 W^d(s,\lambda)dW^d(s,\lambda)}{\left(\int_0^1 W^d(s,\lambda)^2ds\right)^{1/2}}.$$

It is key to note the following features of the results:

(i) The densities derived above show no dependence on the stochastic regressors (that is, the $x_{i,t}$ processes). This follows from assuming orthogonality of the stochastic regressors to the factors and exogeneity with respect to the idiosyncratic errors. It may be shown that this implies that the above results also hold in cases where breaks in the cointegrating vector occur.

(ii) The limiting distributions, as long as a change in trend is not involved, do not depend on the break dates. This implies that, in principle, multiple changes in constants and slopes of the cointegrating vectors are allowed to occur. Heterogeneous break dates are allowed, and the issue of break dates being known or unknown is no longer relevant.

(iii) The situation changes substantially when a break in trend is allowed. The break fraction is important and thus must be estimated if not known. Moreover, recalling the discussion above, for the factor structure to be preserved, we need the breaks to occur at the same time across the units, so that the projection matrix M is not indexed by i. Thus trend breaks, if they occur, must have the same λ in all the units

(iv) Finally, for the case of several common factors, that is, $r > 1$, variations of the MQ tests proposed by Bai and Ng (2004) can be used and are subject to the same remarks as above. Models that do not allow for changes in trend show no dependence either on the stochastic regressors or, more importantly, on the timing of the breaks. Models that allow for a change in trend lead to statistics that depend upon the common break date.

The pooled test statistics are given by:

$$N^{-1/2}Z_{t_{NT}}^c, N^{-1/2}Z_{t_{NT}}^\tau \text{ and } N^{-1/2}Z_{t_{NT}}^\gamma,$$

where:

$$Z^{m}_{t_{NT}} = \sum_{i=1}^{N} t^{m}_{\alpha_{i,0}}; \quad m = c, \tau$$
$$Z^{\gamma}_{t_{NT}}(\lambda) = \sum_{i=1}^{N} t^{\gamma}_{\alpha_{i,0}}(\lambda),$$

and, as $T \to \infty, N \to \infty$ in sequence,

$$N^{-1/2} Z^{m}_{t_{NT}} - \sqrt{N}\Theta^{m}_{2} \Rightarrow N(0, \Psi^{m}_{2}), m = c, \tau$$
$$N^{-1/2} Z^{\gamma}_{t_{NT}}(\lambda) - \sqrt{N}\Theta^{m}_{2}(\lambda) \Rightarrow N\left(0, \Psi^{m}_{2}(\lambda)\right).$$

The moments for the pooled tests when $m = c, \tau$ are the same as those derived by Bai and Ng (2004), while for $m = \gamma$, the moments (which depend on the break fractions) are presented by Banerjee and Carrion-i-Silvestre (2007).

The analysis described above generalizes both the results of Bai and Carrion-i-Silvestre (2007) and Bai and Ng (2004) and also of Pedroni (1999, 2004). We have proposed here an encompassing framework that allows investigators to study dependence and breaks within the context of testing for cointegration and unit roots in macro-panels.

Two further extensions may be considered to allow exogenous regressors and for unknown break dates. The former is relatively straightforward and uses dynamic OLS (D-OLS) regressions to extract the residuals to be tested for integration. The methods are as rehearsed in section 13.3.2, albeit for the case of systems estimators without breaks, but no conceptual novelties are involved. It is shown by Banerjee and Carrion-i-Silvestre (2007) that using D-OLS leads to the asymptotic distributions being the same as those given above, even in the presence of endogeneity of regressors.

The case of unknown break dates is somewhat more complicated but relies on consistent estimation of the break fractions, so that wherever knowledge of the break date is necessary (as in the case where there are changes in trend) the true break date can be replaced by a consistent estimate. Since the estimation algorithms must allow for the breaks to occur at every point in time (within the closed interval), the densities depend on computing the infimum of the standardized t-statistics and are therefore non-standard. The critical values for this case are also given by Banerjee and Carrion-i-Silvestre (2007).

13.3.1.5 Empirical illustration with exchange rate pass-through in the euro-area

This sub-section is based on material from Banerjee and Carrion-i-Silvestre (2007) and de Bandt, Banerjee and Kozluk (2008) and provides an illustration of single-equation panel cointegration techniques when both cross-sectional dependence and structural breaks are allowed to be present.

Prominent in the exchange rate pass-through literature are the papers by Campa and González-Minguez (2006) (henceforth CM), Campa, Goldberg and González-Minguez (2005) (henceforth CGM), and Frankel, Parsley and Wei (2005), *inter alia*, who have investigated the issue of exchange rate pass-through (ERPT) of foreign to domestic prices – that is, how changes in prices of imported goods are transmitted

to the domestic market in the face of exchange rate fluctuations (say, relative to the dollar if foreign prices are denominated in dollars) and pricing strategies of the exporters in the foreign countries. Studies of ERPT have been conducted both for the United States and for countries of the euro-area to study the importance of institutional arrangements (such as the inauguration of the euro-area) in generating responses to exchange rate institutions and changes.

An important feature missing from the discussion is a connection between the theoretical arguments surrounding the key determinants of pass-through, and the inappropriate techniques used to estimate equations measuring import or export exchange rate pass-through. For example, while almost all the theories contain a long-run or steady-state relationship in the levels of a measure of import unit values (in domestic currency), the exchange rate (relating the domestic to the numeraire currency) and a measure of foreign prices (unit values in the numeraire currency, typically US dollars), this long-run relationship is routinely disregarded in most of the empirical implementations.

Since there is substantial consensus in the literature that the time series being studied are integrated variables, one way of defining the long run is in the sense of Engle and Granger (1987) (henceforth EG), where it is given by the cointegrating relationship. A reason often given for ignoring this long-run relationship, and substituting it by an *ad hoc* measure, is a failure to find evidence for cointegration in the data. We argue that a more satisfactory approach is to look for the long-run relationship using more appropriate and powerful methods, such as those which allow for changes in the long-run or use more powerful panel data methods. In doing so, it is also important to allow for breaks in the long-run theoretical relationship so as to take due account of potential changes in pass-through rates in response to changes in financial regimes, such as the introduction of the euro in January 1999.

Exchange rate pass-through into import prices

By definition, import prices for any country i and type of good j, $MP_t^{i,j}$, are a transformation of the export prices of a country's trading partners, $XP_t^{i,j}$, using the bilateral exchange rate ER_t (say with respect to the dollar, if prices are denominated in dollars). Thus dropping superscripts i, j for clarity:

$$MP_t = ER_t \times XP_t.$$

In logarithms (depicted in lower case):

$$mp_t = er_t + xp_t.$$

If the export price consists of the exporters marginal cost and a mark-up:

$$XP_t = FMC_t \times FMKUP_t,$$

we have, in logarithms, by substituting for xp_t:

$$mp_t = er_t + fmc_t + fmkup_t.$$

The industrial organization literature offers explanations for why the effect of the change in er_t on mp_t may differ from one (with typically less than full pass-through, where the latter is defined as pass-through equal to one), using determinants of the mark-up such as competitive conditions among exporters in the destination markets. Mark-up responsiveness depends on the market share of domestic producers relative to foreign producers, the form of competition that takes place in the market for the industry, and the extent of price discrimination. Other factors affecting pass-through are the currency denomination of exports and the structure and importance of intermediate goods markets.

For example, the empirical set-up of CGM is based on assuming unity translation of exchange rate movements. If pass-through is complete (for example, in the case of producer currency pricing), and the mark-ups of producers do not fluctuate in response to fluctuations of the exchange rates, this leads to a pure currency translation. At the other extreme, the exporter can decide not to vary the prices in the destination country currency (local currency pricing) and absorb the fluctuations within the mark-up. Thus, mark-ups in an industry are assumed to consist of a component specific to the type of good, independent of the exchange rate, and a reaction to exchange rate movements:

$$fmkup_t = a + \Phi er_t.$$

It is also important to consider effects working through the marginal cost. These are a function of demand conditions in the importing country, denoted y_t; marginal costs of production in the exporting country (labor wages in domestic currency), denoted fw_t; and commodity prices denominated in foreign currency, fcp_t:

$$fmc_t = c_o y_t + c_1 fw_t + c_2 er_t + c_3 fcp_t.$$

We therefore have:

$$mp_t = a + (1 + \Phi + c_2) er_t + c_0 y_t + c_1 fw_t + c_3 fcp_t + \varepsilon_t$$

where the coefficient $b \equiv (1 + \Phi + c_2)$ on the exchange rate er_t is the pass-through elasticity and ε_t is a stochastic error term. In the CGM "integrated world market" specification, $c_0 y_t + c_1 fw_t + c_3 fcp_t$ is independent of the exchange rate. It is called the opportunity cost of allocating those same goods to other customers and is reflected in the world price of the product fp_t in the world currency (here taken to be the US dollar). Thus the final pass-through equation can be rewritten as follows:

$$mp_t = a + ber_t + cfp_t + \varepsilon_t,$$

which gives the long-run relation between the import price, exchange rate and a measure of foreign price.

Testing for ERPT

Both economic theory and relevant tests lead us to think of each of the series (import price, exchange rate and world price) as being characterized by a unit root.

However, despite the underlying levels equation described as the pass-through equation above, CM are not able to reject the null hypothesis of the non-existence of a cointegrating relationship among the three series. This implies that no evidence can be found in favor of an Engle–Granger long-run relationship among the three series, and thus of an estimate of ERPT in this sense. Hence, they proceed by estimating the long-run equation above *in first differences* (with some dynamic augmentation):

$$\Delta mp_t = a + \sum_{k=0}^{4} b_k \Delta er_{t-k} + \sum_{k=0}^{4} c_k \Delta fp_{t-k} + \varepsilon_t,$$

for industrial sector i in country j, where superscripts have been omitted for clarity. Since CM do not find evidence of a long-run relationship in the EG sense, they propose their own working definition of the long run. They define the coefficient b_k and the sum of coefficients $\sum_{k=0}^{4} b_k$ as the short-run and long-run ERPT respectively.

An alternative route, based on retaining use of the original EG formulation whilst not losing power to look at the long run, is to use the panel cointegration technology developed in this chapter, where for each (i, j) pair there are roughly 110–20 observations. Given that we have ten countries and nine industrial sectors, a panel-based test could use up to approximately $9 \times 10 \times 110$ observations).

The number of observations in the panel is dependent on our need to use a balanced panel. In order to obtain the longest *time* dimension (that is, from 1995), three countries, Austria, Finland and Portugal, need to be deleted from the whole sample since we do not have observations before 1996 for these countries. In order to maximize the *cross-section dimension*, however, so that no country is dropped, our sample needs to start in 1996:1 and end in 2004:12. The estimation results reported below are for this choice of the sample since fewer observations are lost under this configuration and, by allowing for heterogeneity, we should in principle obtain a far clearer idea of the common trends underlying the series and hence of the long run. In the spirit of the discussion above, any such estimation procedure in panels would of course need to allow for structural change. We look at these issues in turn after a brief consideration of the data.

Data

An unbalanced sample of 1995–2005 from Eurostat is available. The construction of the variables follows CM, and is described in Appendix A. The indicator we use for import prices, the index of import unit values (IUV), has a series of caveats associated with their use but we are constrained in our investigations by the quality of the publicly available data.

It is also important to support our claim that there are a number of reasons why we expect there may be a change in the long-run ERPT within our sample. First, on January 1, 1999, 11 European countries fixed their exchange rates by adopting the euro. Greece failed to fulfill the Maastricht Treaty criteria, and therefore joined two years later, effective January 1, 2001. This constituted a change in monetary

policy, especially for countries that previously had less credible policy regimes. Especially in countries with previously rather less successful monetary policy, the perceived stabilization of monetary policy may well have induced the producers to change their pricing strategies and would thus be expected to have an influence on the long-run ERPT. Moreover, the adoption of a common currency has changed competitive conditions by increasing the share of goods denominated in the (new) domestic currency. Finally, virtually all the currencies were depreciating against the US dollar in the period 1995–2000, and especially since 1996. Thereafter, following a short period of a stable euro–dollar exchange rate, the euro started appreciating till the end of our sample. This asymmetry of exchange rate developments may have different implications for ERPT.

Panel cointegration tests

There would essentially be three ways of proceeding in order to construct panels from the datasets – (1) creating country panels of industry cross-sections, (2) industry panels with country cross-sections, and (3) a pooled panel in which every country and industry combination constitutes a separate unit. In search of the existence of a cointegrating relationship in the series we try to maximize the dimensions of our panel, and thus will focus on (3). Results for (1) and (2) are available from us upon request.

In Table 13.9 we present the results of the modified Pedroni tests due to Banerjee and Carrion-i-Silvestre (2007), allowing for structural breaks (as described in section 13.3.1.3) and allowing for both structural change and cross-sectional dependence (as described in section 13.3.1.4). As noted earlier, results are presented for the longest available panel which includes all countries. Throughout the analysis it is assumed that in each cross-section member at most one break occurs, with, depending upon the model, this break occurring at the same time in all breaking components (intercept, trend, cointegrating vector).

The panel headed "Cross-sectionally independent" reports the results from the modified Pedroni (pooled panel t-) tests on the idiosyncratic components under the assumption that these are cross-sectionally independent. The results for all six model specifications concerning the deterministic components as outlined in section 13.3.1.3 – for all of which heterogeneous break dates are permitted – are reported in this panel.

The panel headed "Cross-sectionally dependent" provides the results for the tests of Banerjee and Carrion-i-Silvestre (2007), where cross-sectional dependence is also allowed. The test results are again reported for the idiosyncratic components and we display the results for both cross-sectionally homogeneous and heterogeneous break dates. For Models 3 and 6, we are restricted to imposing a cross-sectionally homogeneous (but unknown) break point. The remaining models allow for both heterogeneous and homogeneous breaks. The maximum number of factors allowed is six, the column labeled $\hat{r}$ provides estimates of the number of common factors, and under the heading $\hat{r}_1$ the number of integrated common trends detected from the MQ statistic is also reported. The break dates detected by the cross-sectionally dependent test, when homogeneous breaks are imposed, are also given.

Table 13.9 Banerjee and Carrion-i-Silvestre (2007) cointegration test results

Model	*Cross-sectionally independent*	*Cross-sectionally dependent*						
		Homogeneous break dates				*Heterogeneous break dates*		
	Panel test t-ratio	*Idiosyn.*	*Break date*	$\hat{r}$	$\hat{r}_1$	*Idiosyn.*	$\hat{r}$	$\hat{r}_1$
1	−22.596	−5.009	1999.02	6	1	−2.744	6	2
2	−26.500	−6.409	2002.11	6	1	−2.933	6	1
3	−25.018	−5.816	2002.03	6	1			
4	−22.523	−5.201	2000.03	6	1	−3.060	6	3
5	−24.677	−6.189	2002.08	6	1	−2.239	6	2
6	−23.498	−6.353	2002.10	6	1			

It is clear that in each case the tests overwhelmingly reject the null hypothesis of no cointegration. The reported breaks all occur in the neighborhood of the introduction of the euro in 1999 or the beginning of its strong appreciation in 2002. A refinement of our tests to allow for multiple breaks would perhaps allow us to detect both these "regime" changes, although it may be argued that the time dimension of the panel will not permit such detailed discrimination. Finally, if one were to think of Model 4 as the most plausible choice, there is very clear evidence, under every configuration, for a long-run relationship with a structural break in 2000.

There are a number of further issues that may be considered, including the magnitude of the pass-through coefficient and its change in response to the new monetary arrangements (or exchange rate movements). On the whole, allowing for a structural break in the relationship, we find that ERPT generally increases in the vicinity of the introduction of the euro. This may be the effect of stabilization in the monetary regime, leading to less noisy exchange rate behavior. Thus actual changes in the exchange rate may be perceived as more permanent and based on macro-fundamentals and exporters may therefore be more willing to pass these on to prices. An alternative explanation relates to the effect of the appreciation of the euro. In a world with a depreciating euro, exporters to the euro-area would be expected to hold back from passing through exchange rate changes to the price (since this would lead to their becoming more uncompetitive relative to the local producers in the euro-area). An appreciating currency means, however, that dollar prices become cheaper in the intra-euro market, leading producers to shift away from local currency pricing. Passing through more of their dollar price, to maintain their revenue in dollars, would still not erode their ability to keep an edge on the market and compete with local products. This asymmetrical response to currency appreciation versus its depreciation could explain the higher pass-through following the likely changes in regime identified above. All these results are tabulated in great detail in de Bandt, Banerjee and Kozluk (2008) and are also available from us on request.

Having studied testing for cointegration in some detail, we move next to a consideration of methods of estimating cointegrating vectors, both in single equations and in systems.

13.3.1.6 Single equation estimation of the cointegrating vector

In this sub-section we consider estimation of the cointegrating vector β for data given by (13.18′) and (13.22′), that is, when abstracting from the presence of cross-sectional dependence via factors and without structural breaks in the deterministic component. This is a topic of further research, as is the study of systems estimators of cointegrating vectors in a similar set-up (see section 13.3.2).

Note that in this case (that is, without cross-sectional dependence or breaks) we have already imposed the assumption that the cointegrating vectors are cross-sectionally identical, that is, $\beta_i = \beta$ for all $i = 1, \ldots, N$. This restriction appears reasonable (although it may be generalized) since in order to gain by resorting to panel methods some of the coefficients should be considered identical for all cross-section members. Given that cointegration is the prime focus, it appears natural to assume identical cointegrating relationships and to allow for heterogeneity in the other characteristics of the DGP.

It has to be noted, however, that, as distinct from the pure time series case, the pooled OLS estimator of β in (13.18′) also converges to a well-defined limit when the cointegrating vectors are not cross-sectionally identical and, more interestingly, converges to a well-defined limit even in the spurious regression case. This limit is given by the so called *average long-run regression coefficient* (for a detailed discussion and the precise assumptions, see Theorems 4 and 5 of Phillips and Moon, 1999). As in the time series case, the limiting distribution of the OLS estimator depends upon so called second-order bias terms, which necessitates the use of modified estimation methods to result in mean zero mixed normal limiting distributions, which are required, for example, to perform valid inference. The literature has, similar to the time series case, proposed two modifications of the OLS estimator to take account of second-order bias. These are given by fully modified OLS (FM-OLS) estimation, as proposed by Phillips and Hansen (1990), and D-OLS estimation, introduced in Saikkonen (1991) (to which we have already referred in section 13.3.1.4). Furthermore, similar to the cointegration tests, estimation can be performed in a pooled or group mean fashion.

For simplicity, we focus in the description of the estimation procedures on the case $m = 2$, that is, the case including only fixed effects and note that the other two cases for m – no deterministic components or both fixed effects and individual specific linear trends – can be handled analogously. If $m = 1$ the original observations are used as inputs in the procedure and if $m = 3$ the variables are individually demeaned and detrended first. Let $\bar{y}_i = \frac{1}{N}\sum_{t=1}^{T} y_{i,t}, \bar{x}_i = \frac{1}{N}\sum_{t=1}^{T} x_{i,t}$ and denote the cross-sectionally demeaned variables by $\tilde{y}_{i,t} = y_{i,t} - \bar{y}_i$ and $\tilde{x}_{i,t} = x_{i,t} - \bar{x}_i$.[28]

Fully Modified OLS

Estimation of the cointegrating vector in the panel context by using FM-OLS estimation is discussed in Phillips and Moon (1999), Kao and Chiang (2000) and Pedroni (2000). In a first step obtain estimators $\hat{\Omega}_{uv,i}, \hat{\Omega}_{v,i}$ and $\hat{\Lambda}_{v,i}$ from the residuals $(\hat{u}_{i,t}, v'_{i,t})'$. The residuals can be obtained by either individual specific OLS estimation or by using the least squares dummy variable (LSDV) estimator

in (13.18′), which puts the homogeneity restriction on the cointegrating relationship in place. Next, defining the endogeneity corrected variable $\tilde{y}^{+}_{i,t} = \tilde{y}_{i,t} - \hat{\Omega}_{uv,i}\hat{\Omega}^{-1}_{v,i}\Delta\tilde{x}_{i,t}$ leads to the following pooled FM-OLS estimator:

$$\hat{\beta}_{FM} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{x}'_{i,t}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{x}_{i,t}\tilde{y}^{+}_{i,t} - \left(\hat{\Lambda}^{+}_{uv,i}\right)'\right)\right),$$

where $\Lambda^{+}_{uv,i} = \hat{\Lambda}_{uv,i} - \hat{\Omega}_{uv,i}\hat{\Omega}^{-1}_{v,i}\hat{\Lambda}_{v,i}$. Phillips and Moon (1999) use in their formulation of the FM-OLS estimator averaged correction factors, for example, $\hat{\Omega}_{v} = \frac{1}{N}\sum_{i=1}^{N}\hat{\Omega}_{v,i}$ and similarly for the other required matrices. The limiting distribution of the FM-OLS estimator (see, for example, Theorem 9 of Phillips and Moon, 1999) is given by:

$$N^{1/2}T\left(\hat{\beta}_{FM} - \beta\right) \Rightarrow N\left(0, 6\omega^{2}_{u.v}\Omega^{-1}_{v}\right),$$

with $\omega^{2}_{u.v} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\omega^{2}_{u.v,i}$ and $\Omega_{v} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\Omega_{v,i}$, with these limits (in most papers implicitly) assumed to exist.[29] For the case $m = 0$ (without deterministic components), the multiplicative factor 6 in the variance term of the limiting distribution has to be replaced by 2. Also note that the limiting covariance matrix is composed of cross-sectional averages.

Standard, up to the factor 2 or 6 depending upon the deterministic specification considered, normally distributed pooled FM-OLS estimators are also easily constructed. These are popular due to their implementation in freely available software. Define the following:

$$\begin{aligned}
\tilde{y}^{0}_{i,t} &= \hat{\omega}^{-1}_{u.v,i}\tilde{y}^{+}_{i,t} - \left[\left(\hat{\omega}^{-1}_{u.v,i}I_{\dim(x)} - \hat{\Omega}^{-1/2}_{v,i}\right)\tilde{x}_{i,t}\right]'\hat{\beta}\\
\tilde{x}^{0}_{i,t} &= \hat{\Omega}^{-1/2}_{v,i}\tilde{x}_{i,t}\\
\hat{\Lambda}^{0}_{uv,i} &= \hat{\omega}^{-1}_{u.v,i}\hat{\Lambda}^{+}_{uv,i}\hat{\Omega}^{-1/2}_{v,i},
\end{aligned}$$

where $I_{\dim(x)}$ denotes the identity matrix with dimensional equal to the number of regressors $x_{i,t}$. and $\hat{\beta}$ denotes the LSDV estimator. Then, the normalized FM-OLS estimator is given by:

$$\hat{\beta}^{0}_{FM} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}^{0}_{i,t}\tilde{x}^{0\prime}_{i,t}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{x}^{0}_{i,t}\tilde{y}^{0}_{i,t} - \left(\hat{\Lambda}^{0}_{uv,i}\right)'\right)\right),$$

for which it holds that $N^{1/2}T\left(\hat{\beta}^{0}_{FM} - \beta\right) \Rightarrow N\left(0, 6I_{\dim(x)}\right)$, where again the factor 6 has to be replaced by the factor 2 in the case of no deterministic components.

Furthermore, group mean FM-OLS estimation is considered in Pedroni (2000), with the estimator in its un-normalized form simply being given by the cross-sectional average of the individual FM-OLS estimators:

$$\hat{\beta}_{FM}^{G} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{x'}_{i,t}\right)^{-1}\left(\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{y}_{i,t} - \left(\hat{\Lambda}_{uv,i}^{+}\right)'\right)\right).$$

Dynamic OLS

We now turn to dynamic OLS estimation of the cointegrating relationship as discussed in Kao and Chiang (2000) and Mark and Sul (2003). The idea of D-OLS estimation is to correct for the correlation between $v_{i,t}$ and $u_{i,t}$ (see equations (13.18′)–(13.22′)) by including leads and lags of $\Delta x_{i,t}$ as additional regressors in the cointegrating regression. As in the time series case (in general), the number of leads and lags has to be increased with the (time dimension of the) sample size at a suitable rate to induce the noise process in the lead and lag augmented regression and $v_{i,t}$ to be uncorrelated asymptotically. Thus, considering again the case $m = 2$, let the augmented cointegrating regression be given by:

$$\begin{aligned}\tilde{y}_{i,t} &= \tilde{x}_{i,t}^{'}\beta + \sum_{j=-p_i}^{p_i}\Delta\tilde{x}_{i,t-j}^{'}\gamma_{i,j} + u_{it}^{*}\\ &= \tilde{x}_{i,t}^{'}\beta + \tilde{Z}_{i,t}^{'}\gamma_i + u_{it}^{*},\end{aligned}$$

where the last equation defines $\tilde{Z}_{i,t}$ and γ_i. The pooled D-OLS estimator for β is then obtained from OLS estimation of the above equations. Let $\tilde{Q}_{i,t} = \left[\tilde{x}_{i,t}^{'}, 0^{'}, \ldots, 0^{'}, \tilde{Z}_{i,t}^{'}, 0^{'}, \ldots, 0^{'}\right] \in R^{2\dim(x)\left(1+\sum_{i=1}^{N}p_i\right)}$, where the variables $\tilde{Z}_{i,t}$ are at the ith position in the second block of the regressors. Using this notation we arrive at:

$$\begin{bmatrix}\hat{\beta}_D\\ \hat{\gamma}_1\\ \vdots\\ \hat{\gamma}_N\end{bmatrix} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Q}_{i,t}\tilde{Q}_{i,t}^{'}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Q}_{i,t}\tilde{y}_{i,t}\right).$$

Mark and Sul (2003) obtain the asymptotic distribution of $\hat{\beta}_D$, which has a sandwich type limit covariance matrix. Denote this as $\overline{V} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\omega_{u.v,i}^{2}\Omega_{v,i}$ (again assumed to exist), then it holds that:

$$N^{1/2}T\left(\hat{\beta}_D - \beta\right) \Rightarrow N\left(0, 6\Omega_v^{-1}\bar{V}\Omega_v^{-1}\right).$$

Kao and Chiang (2000) discuss a normalized version of the D-OLS estimator that corresponds to $\hat{\beta}_{FM}^{0}$. This estimator, $\hat{\beta}_D^0$ say, is obtained when, in the above discussion of the D-OLS estimator, $\tilde{y}_{i,t}$ and $\tilde{x}_{i,t}$ are replaced by $\tilde{y}_{i,t}^{0}$ and $\tilde{x}_{i,t}^{0}$. These changes

lead to an estimator with a limiting covariance matrix proportional to the identity matrix.

Pedroni (2001) considers a group mean D-OLS estimator. Denote $\tilde{R}_{i,t} = \left[\tilde{x}_{i,t}^{'}, \tilde{Z}_{i,t}^{'}\right]^{'}$ and estimate, for each $i = 1, \ldots, N$,

$$\begin{bmatrix} \hat{\beta}_{D,i} \\ \hat{\gamma}_i \end{bmatrix} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{R}_{i,t} \tilde{R}_{i,t}^{'} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{R}_{i,t} \tilde{y}_{i,t} \right).$$

Then the group mean D-OLS estimator is given by $\hat{\beta}_D^G = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_{D,i}$. The group mean D-OLS estimator can also be computed in a normalized fashion.

13.3.2 Testing for cointegration and estimation of the cointegrating vectors in systems

The analysis of multiple cointegrating vectors in panels is still very limited. The analysis is plagued by problems of arriving at a proper, yet statistically tractable, formulation of multiple cointegration (since one needs not only to consider the cointegration possibilities within the units of a panel, but also to identify those that link across the units of the panel). Estimation and inference are thus hampered by theoretical difficulties, as well as dimensionality issues, even when the panel dimensions are large.

Breitung (2005), Groen and Kleibergen (2003) and Larsson, Lyhagen and Löthgren (2001) have attempted analysis of this framework under restrictive assumptions such as a homogeneous cointegrating space (for the units of the panel) and, more unrealistically, cross-sectional independence of the units of the panel. Some of these methods are described in detail below.

If cross-unit cointegration is allowed then, as shown by Banerjee, Marcellino and Osbat (2004), tests for cointegration in panels (in the presence of multiple cointegrating vectors) suffer from size distortions and loss of power. Inference on the full cointegration structure in such situations remains an extremely difficult exercise and much further work is needed to understand the theoretical properties of feasible estimators and their performance in finite and large samples. For a detailed simulation study of some of the issues involved, see Wagner and Hlouskova (2007).

The established system methods are panel extensions of vector autoregressive (VAR) cointegration analysis (see Johansen, 1995). Compared to the single-equation methods several differences are worth mentioning. First, the systems approach allows us to test for and model multiple cointegrating relationships. Second, the cointegrated VAR approach allows for the incorporation of a richer specification concerning (restricted) deterministic components which are considered relevant in the applied cointegration literature. Third, specifying a parametric model incorporates the modeling of the short-run dynamics of the data, which are treated as nuisance parameters in the single-equation methods.

Given that the system methods are based on VAR estimates, substantial biases arise with short time series. Thus, for practical applications the time series dimension has to be sufficiently large, which is also required since the specification of a

dynamic model necessitates the estimation of a substantial number of parameters. Clearly, it has to be mentioned that accurate estimation of the long-run variances for the nonparametric methods requires sufficiently long time series.

Open issues are allowing for a factor structure in the error processes of the VAR models to allow for cross-sectional dependencies, and considering structural change in the deterministic components.

Without imposing any cross-sectional homogeneity assumption, the cross-sectionally independent panel VAR DGP considered is given in *error correction form* by:

$$\Delta Y_{i,t} = C_{1i} + C_{2i}t + \alpha_i \beta_i^{'} Y_{i,t-1} + \sum_{j=1}^{p_i} \Gamma_{ij} \Delta Y_{i,t-j} + w_{i,t}, \tag{13.30}$$

with $Y_{i,t} = (y_{i,t}, x_{i,t})^{'} \in R^m, m = l + 1, C_{1i}, C_{2i} \in R^m, \alpha_i, \beta_i \in R^{m \times k_i}$ with full rank $k_i, \Gamma_{ij} \in R^{m \times m}$ and w_{it} are cross-sectionally independent m-dimensional white-noise processes with covariance matrices $\Sigma_i > 0$.[30] To ensure that all the processes described by (13.30) are (up to the deterministic components) $I(1)$ processes, the matrices $\alpha_{i\perp}^{'} \Gamma_i \beta_{i\perp}$ have to be invertible, where $\alpha_{i\perp} \in R^{m \times (m-k_i)}, \beta_{i\perp} \in R^{m \times (m-k_i)}$ are full rank matrices such that $\alpha_{i\perp}^{'} \alpha_i = 0, \beta_{i\perp}^{'} \beta_i = 0$ and $\Gamma_i = I_m - \sum_{j=i}^{p_i - 1} \Gamma_{ij}$. One possible choice is given by $\alpha_{i\perp} = I_m - \alpha_i (\alpha_i^{'} \alpha_i)^{-1} \alpha_i^{'}$ and similarly for β_i. Under these assumptions the column space of the matrix β_i is the k_i-dimensional cointegrating space of unit i.

In the VAR cointegration literature the following five specifications of the deterministic components are usually discussed. Case 1 is without any deterministic components. In case 2 restricted intercepts of the form $C_{1i} = \alpha_i \tau_i$ are contained in the cointegrating space whereas in case 3 unrestricted intercepts C_{1i} that induce linear time trends in $Y_{i,t}$ are included. In case 4 unrestricted intercepts and restricted trend coefficients $C_{2i} = \alpha_i \kappa_i$ are included, which allows for linear trends in both the data and the cointegrating relationships. Finally, in case 5 unrestricted intercepts and trend coefficients are included, which leads to quadratic deterministic trends in $Y_{i,t}$. For a detailed discussion of these specifications concerning the deterministic variables see Johansen (1995, sec. 5.7). This monograph also includes a very detailed discussion of the statistical analysis, that is, parameter estimation via reduced rank regression, testing for the cointegrating rank as well as hypothesis testing. Therefore we do not include a description of this widely-used method here.

In the following two sub-sections we discuss two approaches, due to Larsson, Lyhagen and Löthgren (2001) and Breitung (2005), to test for cointegration in panel VAR models. In common with the single-equation approaches described above, both these methods put *at least* the restriction of a cross-sectionally homogeneous cointegrating space, that is, $\beta_i = \beta$, in place. Again the argument in favor of such a restriction is that in order to apply panel methods fruitfully, some of the coefficients have to be considered cross-sectionally identical. Since in cointegration analysis the main focus is on the cointegrating relationships, it is natural to assume identical cointegrating spaces and allow for individual specific short-run dynamics. This is

very similar to the analysis relating to single-equation estimators. In panel VAR cointegration analysis, however, the testing for cointegration and the estimation steps are much more intertwined than in the single-equation tests discussed above, for most of which the null hypothesis is that of no cointegration. The two tests to be discussed differ in this respect, since the test of Larsson, Lyhagen and Löthgren (2001) does not impose cross-sectional homogeneity in the construction of the test statistic (but needs this assumption for the derivation of the asymptotic distribution of the test statistic) whereas the test of Breitung (2005) incorporates this restriction.

13.3.3 Larsson, Lyhagen and Löthgren (2001)

Larsson, Lyhagen and Löthgren (2001) (henceforth LLL) consider testing for cointegration in the above framework under the assumption that $\Pi_i = \alpha_i \beta_i^{'} = \alpha\beta^{'} = \Pi$ for all $i = 1, \ldots, N$. For some further technical assumptions we refer the reader to their paper. Note here that their test is based simply on the cross-sectional average of the Johansen (1995) trace statistic, where the cross-sectional homogeneity assumption just mentioned is not put in place anywhere in the construction of their test. However, the asymptotic distribution (under the null hypothesis) of their test statistic is established only under this assumption (see Assumption 3 and Theorem 1 of LLL). The null hypothesis of their test is $H_0 : rk(\Pi) = k$ and the test is consistent against the alternative hypothesis $H_1 : rk(\Pi_i) > k$ for a non-vanishing fraction of the cross-section members.

The construction of the test statistic is similar to that of IPS and it is given by a suitably centered and scaled cross-sectional average of the individual trace statistics. Thus, denote by $LR_i^s(k, m)$ the trace statistic for the null hypothesis of a k-dimensional cointegrating space for cross-section unit i, where the superscript $s = 1, \ldots, 5$ indicates the specification of the deterministic components. Further, denote by $\mu_{LR}^s(k, m)$ and $\sigma_{LR}^{2,s}(k, m)$ the expected value and variance of $LR_i^s(k, m)$. In this respect LLL is a most notable but unfortunately underrated exception in the panel unit root and cointegration literature, insofar as the authors derive, admittedly only for the model without deterministic components and with normally distributed innovations, the existence of the necessary moments as well as uniform integrability and Lindeberg-type conditions. These are needed to derive formally the asymptotic normality of the test statistic.[31] Using a sequential limit with first $T \to \infty$ followed by $N \to \infty$, it holds that:

$$LLL^s(k, m) = N^{-1/2} \sum_{i=1}^{N} \frac{LR_i^s(k, m) - \mu_{LR}^s(k, m)}{(\sigma_{LR}^{2,s}(k, m))^{1/2}} \Rightarrow N(0, 1). \tag{13.31}$$

Finite-sample correction factors for the LLL test statistic are given in Hlouskova and Wagner (2008) for all five mentioned specifications of the deterministic components, where we note again that the theoretical result is only derived for the case without deterministic components and with normally distributed innovations.

LLL do not explicitly consider estimation of the cointegrating space. Given that their test is based on the assumption of a cross-sectionally identical cointegrating space, one possibility to obtain an estimate of the cointegrating space is given by the cross-sectional average of identically normalized individual specific cointegrating

spaces, for example, normalized as $\hat{\beta}_i = \left[I_k, \hat{\beta}_{i,2}^{'}\right]^{'}$, which are in any case computed in the derivation of the test statistic.

13.3.4 Breitung (2005)

Breitung (2005) proposes a two-step estimation and test procedure that extends the Ahn and Reinsel (1990) and Engle and Yoo (1991) approach from the time series to the panel case.

Breitung considers the homogeneous cointegration case where, however, only the cointegrating spaces are assumed to be identical for all cross-section members. In the first step of his procedure the parameters are estimated individual specifically (applying the method outlined in Johansen, 1995). This includes the first-step estimates of β_i. In the second step the common cointegrating space β is estimated in a pooled fashion.

For simplicity we describe the method here for the VAR(1) model excluding any deterministic component. In the general case lagged differences as well as potentially restricted deterministic components are treated in the usual way by being concentrated out at the beginning of the procedure. Thus, consider the following:

$$\Delta Y_{i,t} = \alpha_i \beta^{'} Y_{i,t-1} + w_{i,t}. \tag{13.32}$$

Next, define $T_i = \left(\alpha_i^{'} \Sigma_i^{-1} \alpha_i\right)^{-1} \alpha_i^{'} \Sigma_i^{-1}$ and pre-multiply (13.32) with this quantity to obtain:

$$\begin{aligned}
\left(\alpha_i^{'} \Sigma_i^{-1} \alpha_i\right)^{-1} \alpha_i^{'} \Sigma_i^{-1} \Delta Y_{i,t} &= \beta^{'} Y_{i,t-1} + \left(\alpha_i^{'} \Sigma_i^{-1} \alpha_i\right)^{-1} \alpha_i^{'} \Sigma_i^{-1} w_{i,t} \\
T_i \Delta Y_{i,t} &= \beta^{'} Y_{i,t-1} + T_i w_{i,t} \\
\Delta Y_{i,t}^{+} &= \beta^{'} Y_{i,t-1} + w_{i,t}^{+},
\end{aligned} \tag{13.33}$$

where the last equation defines the variables with superscript +. Note also that $E[w_{i,t}^{+}(w_{i,t}^{+})^{'}] = \left(\alpha_i^{'} \Sigma_i^{-1} \alpha_i\right)^{-1}$. Next use the normalization $\beta = \left[I_k, \beta_2^{'}\right]^{'}$ and partition $Y_{it} = \left[(Y_{i,t}^1)^{'}, (Y_{i,t}^2)^{'}\right]^{'}$ with $Y_{i,t}^1 \in R^k$ and $Y_{i,t}^2 \in R^{m-k}$. Using this notation the above equation (13.33) can be rewritten as:

$$\Delta Y_{i,t}^{+} - Y_{i,t-1}^1 = \beta_2^{'} Y_{i,t-1}^2 + w_{i,t}^{+}. \tag{13.34}$$

Breitung suggests estimating (13.34) by pooled OLS using the estimates $\hat{T}_i = \left(\hat{\alpha}_i^{'} \hat{\Sigma}_i^{-1} \hat{\alpha}_i\right)^{-1} \hat{\alpha}_i^{'} \hat{\Sigma}_i^{-1}$ based on the individual specific Johansen estimates. Note that, given that the covariance structure of the errors in (13.34) is known and an estimate is readily available, pooled feasible GLS estimation of (13.34) can also be performed.

Breitung's estimation procedure stops here. However, an iterative estimator based on the above procedure is easily conceived. With the estimated $\hat{\beta}_2$, all individual-specific parameters in (13.32) can be re-estimated. Since we have chosen the VAR(1)

set-up without deterministic components for illustration, these parameters are contained in the matrices α_i and Σ_i. Based on the new estimates of α_i and Σ_i, equation (13.34) can be re-estimated. This can be repeated until convergence, according to some numerical criterion, occurs.

Such an iterative procedure corresponds by and large to the iterative estimator proposed in Larsson and Lyhagen (1999), with the only difference being that Larsson and Lyhagen propose using $\hat{\beta}_2 = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{i,2}$ as the initial estimator. $\hat{\beta}_{i,2}$ denotes as before the (not-normalized coordinates of the) individual specific Johansen (1995) estimates. Let us note as a side remark that the set-up of Larsson and Lyhagen (1999) is more general, since these authors consider, in the VAR(1) case without deterministic components, the specification:

$$\begin{bmatrix} \Delta Y_{1,t} \\ \vdots \\ \Delta Y_{N,t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1N} \\ \vdots & \ddots & \vdots \\ \alpha_{N1} & \cdots & \alpha_{NN} \end{bmatrix} \begin{bmatrix} \beta' & & \\ & \ddots & \\ & & \beta' \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ \vdots \\ Y_{N,t-1} \end{bmatrix} + \begin{bmatrix} w_{1,t} \\ \vdots \\ w_{N.t} \end{bmatrix},$$

with a full covariance matrix of the stacked noise process and all necessary assumptions such that the joint process is an $I(1)$ process. Thus, Larsson and Lyhagen (1999) consider a VAR model for the stacked vector of variables, where only the cointegrating space is restricted to be cross-sectionally identical and where no cross-unit cointegration occurs (meaning, in the notation of Appendix B, that the B matrix that collects all cointegrating relationships is block-diagonal).

It is not clear whether such an almost unrestricted VAR model for the stacked process should really be interpreted as a panel model or as a time series model for a high-dimensional process with some cross-equation restrictions. Pragmatically, the loose parameterization implies that the time dimension of the panel has to be very large compared to the cross-sectional dimension. Similar comments apply to the work of Groen and Kleibergen (2003), who also consider a very general (panel) VAR model and consequently present simulation evidence for a bivariate example for $N = 1, 3$ and 5 and $T = 1{,}000$.

Breitung (2005) shows that his two-step estimator, $\tilde{\beta}_2$ say, is asymptotically normally distributed in the sequential limit with first $T \to \infty$ followed by $N \to \infty$:

$$N^{1/2}Tvec(\tilde{\beta}_2 - \beta_2) \Rightarrow N(0, \Omega_2^{-1} \otimes \Sigma_\alpha),$$

where $\otimes$ denotes the Kronecker product,

$$\Omega_2 = \lim_{N\to\infty}\lim_{T\to\infty} E[\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{i,t-1}^2 (Y_{i,t-1}^2)'],$$

and $\Sigma_\alpha = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i'\Sigma_i^{-1}\alpha_i\right)^{-1}$, with these limits, in case of Ω_2 non-singular, implicitly assumed to exist by Breitung (2005).

Let us now turn to the test for cointegration that Breitung considers, which is based on Saikkonen (1999). The main difference to the test of LLL discussed above is

that Breitung incorporates the homogeneity restriction on the cointegrating spaces in the construction of the test statistics. We continue the discussion for the VAR(1) model without deterministic components. Denote with $\gamma_i \in R^{m\times(m-k)}$ full column rank matrices and consider:

$$\Delta Y_{i,t} = \alpha_i \beta^{'} Y_{i,t-1} + \gamma_i \beta_{\perp}^{'} Y_{i,t-1} + w_{i,t}. \tag{13.35}$$

Under the null hypothesis of a k-dimensional cointegrating space it holds that $\gamma_i = 0$ for $i = 1, \ldots, N$ and, under the alternative of an m-dimensional cointegrating space, γ_i is unrestricted to allow for $\Pi_i = \alpha_i \beta^{'} + \gamma_i \beta_{\perp}^{'}$ of full rank. For test consistency the alternative has to comprise a non-vanishing fraction of cross-section members. Pre-multiply (13.35) with $\alpha_{i,\perp}^{'}$ to obtain:

$$\begin{aligned} \alpha_{i,\perp}^{'} \Delta Y_{i,t} &= \alpha_{i,\perp}^{'} \gamma_i \beta_{\perp}^{'} Y_{i,t-1} + \alpha_{i,\perp}^{'} w_{i,t} \\ \alpha_{i,\perp}^{'} \Delta Y_{i,t} &= \phi_i (\beta_{\perp}^{'} Y_{i,t-1}) + \tilde{w}_{i,t}, \end{aligned} \tag{13.36}$$

where the last equation defines the coefficient matrices and variables. Replacing $\alpha_{i,\perp}^{'}$ and $\beta_{\perp}$ by estimates allows us to estimate equations (13.36) separately for each cross-section unit by OLS and to construct test statistics for the hypothesis $H_0 : \phi_i = 0, \ i = 1, \ldots, N$.

Any of the classical testing principles, that is, the likelihood ratio, Wald or Lagrange multiplier principle, can be used. Breitung discusses the Lagrange multiplier test statistic, which has the advantage that it only requires estimation under the null hypothesis. Denoting by $\hat{f}_{i,t} = \hat{\alpha}_{i,\perp}^{'} \Delta Y_{i,t}$ and with $\hat{g}_{i,t} = \hat{\beta}_{\perp}^{'} Y_{i,t}$, the Lagrange multiplier test statistic for unit i is given by:

$$LM_i(k,m) = Ttr\left[\sum_{t=2}^{T} \hat{f}_{i,t}\hat{g}_{i,t-1}^{'} \left(\sum_{t=2}^{T} \hat{g}_{i,t-1}\hat{g}_{i,t-1}^{'}\right)^{-1} \sum_{t=2}^{T} \hat{g}_{i,t-1}\hat{f}_{i,t}^{'} \left(\sum_{t=2}^{T} \hat{f}_{i,t}\hat{f}_{i,t}^{'}\right)^{-1}\right],$$

which is sequentially computed for the different values of $k = 0, \ldots, m$.

The panel test statistic is then, as usual, given by the corresponding centered and scaled cross-sectional average, where we now use again the superscript s to indicate the dependence upon deterministic components. Under the null hypothesis we hence arrive at:

$$B^s(k,m) = N^{-1/2} \sum_{i=1}^{N} \frac{LM_i^s(k,m) - \mu_{LM}^s(k,m)}{(\sigma_{LM}^{2,s}(k,m))^{1/2}} \Rightarrow N(0,1). \tag{13.37}$$

The correction factors $\mu_{LM}^s(k,m)$ and $\sigma_{LM}^{2,s}(k,m)$ coincide exactly with those of LLL above. The method also shares with LLL the limitations concerning the availability of proper asymptotic theory. Clearly, instead of basing the panel tests upon the trace statistic, the max statistic of Johansen (1995) could be used as the underlying time series test statistic in both the LLL and Breitung tests.

The final substantive section of our chapter is to apply some of the estimation methods described in this section to the EKC analysis, building on our findings in section 13.2.4.2.

13.3.5 The environmental Kuznets curve analysis continued

As indicated above, estimation of an equation of the form (13.17) using usual cointegration methods is troublesome given the presence of nonlinear transformations of integrated regressors (GDP and its square). Regressions involving nonlinear transformations of integrated regressors behave differently from the regressors usually considered and have been studied for the time series case in Chang, Park and Phillips (2001) and Park and Phillips (1999, 2001). Hong and Wagner (2008a) develop an FM-OLS estimator for nonlinear cointegrating relationships including integer powers of integrated regressors as well as specification and cointegration tests for this set-up. Hong and Wagner (2008b) derive first results for a simple panel setting by considering a seemingly unrelated nonlinear cointegrating regressions framework.

Ignoring the problems caused by nonlinear transformations, cross-sectional dependencies and structural change, and applying the seven tests of Pedroni (1999, 2004) leads to seemingly overwhelming support for cointegration (with the detailed results available upon request). Again these findings are in line with the Lyhagen result that in the presence of common non-stationary components the unit root null hypothesis is over-rejected, which leads to seemingly strong support for cointegration. Note that these findings are obtained both when using the quadratic formulation (13.17) as well as the specification when only GDP is included in the regression, with the latter specification being subject only to the cross-sectional dependence and structural change problems.

Given the seemingly "strong evidence for cointegration," the final step in panel cointegration analysis is the estimation of relationship (13.17). In Table 13.10 we present the results for sulphur dioxide emissions when applying the fully modified OLS estimator discussed in detail in Phillips and Moon (1999), the dynamic OLS estimator of Mark and Sul (2003) and the two-step estimator of Breitung (2005). With the exception of the D-OLS estimator, seemingly strong evidence for the prevalence of an EKC for sulphur dioxide appears, even with plausible turning points with respect to income. Especially when applying fully modified estimation the turning points are well within the sample. However, these findings should be treated with some caution, given that the properties of the estimation methods in the presence of cross-sectional dependence and, in particular, of nonlinear transformations of integrated regressors are far from fully worked out.

The finding of stationary de-factored GDP allows us to use standard econometric methods to estimate equation (13.17) with de-factored observations. Table 13.11 shows the results from a variety of specifications, for example, one-way and two-way fixed effects estimation and equations including the extracted factors as additional regressors. For none of these specifications is there evidence for an inverted U-shaped relationship, since the coefficients to income and squared income are significantly positive.

Table 13.10 Panel cointegration estimation results for SO_2

	FM-OLS	*D-OLS*	*Two-step*
	Fixed effects		
$\ln y_{it}$	**5.26**	**1.35**	**4.80**
	(35.65)	(6.93)	(13.34)
$(\ln y_{it})^2$	**–0.26**	−0.01	−0.22
	(-28.22)	(−0.60)	(−10.18)
TP (in $)	24,720	–	54,671
	Fixed effects and linear trends		
$\ln y_{it}$	**4.82**	**−12.63**	**4.12**
	(34.45)	(-68.80)	(11.74)
$(\ln y_{it})^2$	**−0.23**	**0.86**	**-0.18**
	(−26.35)	(16.12)	(−8.34)
TP (in $)	35,534	–	93,381

Notes: Estimation results obtained with the panel cointegration estimation methods described in sections 13.3.1–4. The upper panel contains the results with only fixed effects included in the regression and the lower panel reports the results with both fixed effects and linear trends included. Bold indicates significance of the estimated coefficients at the 5% level and the numbers in brackets are the *t*-values. TP denotes the implied turning point.

Table 13.11 Estimation results with de-factored observations

	DF-1			*DF-2*		
	One-way	*Two-way*	*Incl.* F_t	*One-way*	*Two-way*	*Incl.* F_t
$\ln y_{it}$	**0.86**	**0.79**	**0.78**	**0.92**	**0.66**	**0.69**
	(14.75)	(13.90)	(13.47)	(14.86)	(10.80)	(11.30)
$(\ln y_{it})^2$	**0.48**	**0.44**	**0.38**	**0.55**	0.10	**0.22**
	(12.24)	(11.30)	(9.60)	(6.34)	(1.09)	(2.52)

Notes: One-way refers to equations including fixed individual effects, Two-way refers to equations including fixed individual and time effects and Incl. F_t refers to equations including fixed individual effects and the estimated common factors of both SO_2 emissions and GDP. *DF*-1 indicates that the common factors are obtained from the model without linear trends and *DF*-2 indicates that the common factors are obtained from the model with linear trends. Bold indicates significance of the estimated coefficients at the 5% level and the numbers in brackets are the *t*-values.

The estimation results just mentioned are robust in a number of ways. First, they are qualitatively unchanged if the de-factorization method allows for structural breaks, as found when using the Bai and Carrion-i-Silvestre (2007) methodology. The results are also robust to using smaller sets of countries or shorter time periods to rule out the potential structural changes. Also, when allowing for parameter

heterogeneity across countries, only very limited evidence for an EKC arises, with the main observation being that for several countries very implausible parameter estimates occur.

The scarcity of evidence for an EKC for sulphur dioxide in the panel at hand is in line with the observation that the common factors driving GDP and SO_2 appear not to be cointegrated. This long-run disconnect between these two variables potentially shows the limitation of reduced form EKC modeling and it may be necessary to resort to more structural modeling approaches to shed further light on the EKC hypothesis, from either a time series or a panel perspective.

13.4 Conclusions

We have sought in this chapter to provide an up-to-date analysis of the methods involved for estimation and inference in non-stationary panels. Our overriding objective has been not only to provide information on the tools but also to interpret the literature and to highlight the considerable challenges that remain. Starting with the motivating example and concluding with the appendices, this chapter has sought to emphasize the difficulties involved in formulating hypotheses within a panel framework, estimating them and conducting inference coherently.

Features of the data that the methods need to incorporate include (in addition to non-stationarity) cross-sectional dependence and structural instability. For example, we have argued that assuming cross-sectional independence of the units leads to relatively simple sequential central limit theory that provides the asymptotic normality of many of the test statistics. As soon as this assumption is relaxed, matters become much more complicated and in ways which are still not fully understood in the literature.

Modeling dependence via factors is a popular device but is only one of the many ways of formulating the problem. Account must be taken of short-run versus long-run dependence, which must be dealt with appropriately. The link between cointegration and factor models in panels needs to be adequately explored (see Appendix B).

More generally, the asymptotic theory must be put on a surer footing to deal not only with many of the joint limiting arguments that arise in the consideration of any form of dependence in the panel but also to deal with cases where there is potential structural instability in the data. In this chapter we have demonstrated some preliminary (and fairly crude) ways of modeling structural instability but a closer look is clearly warranted. The extension of systems methods to panels – to allow for multiple cointegrating vectors – in the possible presence of cross-sectional dependence and structural breaks is an important task but one of considerable complexity. Dealing with these problems should constitute fruitful areas of research in the years ahead.

13.5 Appendix A: Datasets employed

Data for exchange rate pass-through example

Sources: Eurostat, COMEXT.

Table A.13.1 Country lists for the four monthly RER panels

Euro-area (1980/1–1998/12)			
Austria	France	Italy	Portugal
Belgium	Germany	Luxembourg	Spain
Finland	Greece	Netherlands	
CEEC (1993/1–2004/6)			
Albania	**Estonia**	**Lithuania**	Slovak Republic
Bulgaria	Hungary	Poland	Slovenia
Czech Rep.	**Latvia**	**Romania**	
Industrial (1980/1–1998/12)			
Argentina	Germany	Malaysia	South Africa
Austria	Greece	*Mexico*	Spain
Belgium	Indonesia	Netherlands	Sweden
Brazil	Italy	Norway	Switzerland
Canada	Japan	Philippines	Thailand
Denmark	Korea	Portugal	Turkey
Finland	Luxembourg	Singapore	United Kingdom
France			
Worldwide (1981/1–1998/12)			
Algeria	*Dominican Republic*	**Kenya**	*Samoa*
Argentina	Ecuador	Korea	**Saudi Arabia**
Bahamas	Egypt	Madagascar	Senegal
Bolivia	El Salvador	Malaysia	Seychelles
Botswana	Fiji	Malta	Singapore
Brazil	Ghana	**Mauritius**	South Africa
Burkina Faso	**Guatemala**	**Mexico**	**Swaziland**
Burundi	Haiti	**Morocco**	Sweden
Canada	Honduras	Niger	Switzerland
Chile	Hong Kong	Norway	Thailand
Colombia	India	Pakistan	**Turkey**
Costa Rica	Indonesia	Paraguay	United Kingdom
Cote d'Ivoire	Israel	Peru	Uruguay
Cyprus	Japan	Philippines	Venezuela
Denmark			

Note: Italic entries indicate rejection of the unit root null hypothesis at the 10% level and bold entries indicate rejection at the 5% level when applying the ADF test to the time series individually.
Source: IMF IFS, OECD MEI and ECB.

Import prices – monthly indexes of import unit values (calculated based on local currency) for imports originating outside the euro-area.

Foreign prices – monthly indexes of import unit values (calculated based on US dollars) from imports originating outside the euro-area into the euro-zone.

Exchange rates – index of monthly average exchange rate of local currency against the US dollar.

Table A.13.2 List of countries for EKC computations for SO_2

Afghanistan	Equatorial Guinea	Libya	South Korea
Albania	Finland	Madagascar	Spain
Argentina	France	Mauritius	Sri Lanka
Australia	Germany	Mexico	Sudan
Austria	Ghana	Mongolia	Swaziland
Bahrain	Greece	Morocco	Sweden
Belgium	Guatemala	Mozambique	Switzerland
Bolivia	Guinea Bissau	Nepal	Syria
Brazil	Haiti	Netherlands	Taiwan
Bulgaria	Honduras	New Zealand	Thailand
Burma	Hong Kong	Nicaragua	Togo
Canada	Hungary	Nigeria	Trinidad Tobago
Cape Verde	India	North Korea	Tunisia
Chile	Indonesia	Norway	Turkey
China	Iraq	Panama	Uganda
Colombia	Iran	Paraguay	United Kingdom
Costa Rica	Ireland	Peru	Uruguay
CSSR	Israel	Philippines	USA
Cuba	Italy	Poland	USSR
Denmark	Jamaica	Portugal	Venezuela
Djibouti	Japan	Qatar	Yugoslavia
Dominican Republic	Jordan	Reunion	Zaire
Ecuador	Kenya	Romania	
Egypt	Lebanon	Sierra Leone	
El Salvador	Liberia	South Africa	

Notes: The data for CSSR, USSR and Yugoslavia are constructed by taking the values for all the countries that previously constituted these countries before the separations in the early 1990s. The GDP series are taken from the Groningen Growth and Development Center dataset already mentioned in the introduction, and the SO_2 emissions series are from Stern (2006).

All variables are used in logarithms in the econometric analysis

SITC code – Industry
0 – Food and live animals chiefly for food
1 – Beverages and tobacco
2 – Crude materials, inedible, except fuels
3 – Mineral fuels, lubricants and related materials
4 – Animal and vegetable oils, fats and waxes
5 – Chemicals and related products
6 – Manufactured goods classified chiefly by materials
7 – Machines, transport equipment
8 – Manufactured goods

CM dataset 1989–2001 – series for 1989:1–2001:3 for Belgium + Luxembourg, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal and Spain.

Series for 1996:1–2001:3 for Austria and Finland.

Our "new" dataset 1995–2005 – 1995:1–2005:3 for ten out of eleven countries of the CM dataset (Belgium + Luxembourg excluded, Austria and Finland start 1996:1, Portugal and Austria stop 2004:12)

Full panel – reduced version of 1995–2005 dataset: trimmed in order to obtain a balanced panel. Covers 1996:1–2004:12 for all ten countries.

13.6 Appendix B: Cross-sectional dependence

Allowing for cross-sectional dependence complicates matters substantially compared to the case of cross-sectionally independent panels. In the early literature on non-stationary panels, cross-sectional dependence had been modeled by allowing for common fixed (respectively random) effects. It was, however, quickly realized that in a non-stationary time series panel context the modeling of cross-sectional dependence needed to allow for richer types of dependencies that allow us, in particular, to model both transitory (short-run) and permanent (long-run) cross-sectional dynamic dependencies.

During the course of the chapter, in modeling cross-sectional dependence, we have typically followed the route prescribed by Bai and Ng (2004), *inter alia*, in using approximate factor models. Dependence can also be introduced, as in Banerjee *et al*. (2004), by considering cointegrating relationships across the variables in the cross-sectional units. In this brief appendix, we attempt to provide an exploration of these issues, including a discussion of the restrictions on the cointegrating space (of the full panel) implied by factor models and of the links between these alternative formulations of cross-sectional dependence.

In order to clarify the concepts, consider the case of a panel of *univariate* time series, neglecting for simplicity deterministic components, and denote the stacked joint vector (for given N) as $y_t = (y_{1,t}, \ldots, y_{N,t})'$.

The first assumption that is usually, unfortunately typically only implicitly, made is that the joint vector process is a vector $I(1)$ process, or, in the case of all series being stationary, a multivariate stationary process. Clearly, this is an assumption that has to be put in place over and above the assumption that all the *individual* series are $I(1)$ or stationary. This stems from the fact that stacking stationary processes does not in general lead to a jointly stationary vector process. Stationarity of the stacked process occurs if and only if all the processes are *stationary correlated*, that is, if all cross-correlation functions between the individual processes are stationary. For cross-sectionally independent processes this latter condition is fulfilled by construction. Since stacked stationary processes are not necessarily stationary it is clear also that stacked $I(1)$ processes are not necessarily $I(1)$ processes.

Given that we consider panels of non-stationary time series, dependence concepts well-established in the time series literature are also of prime importance in this context. In particular we will distinguish between short-run and long-run dependence. To be precise, in these definitions we will focus on the dependence

structure of the second moments, that is, the covariance, respectively correlation, structure. Obviously, this focus stems from the fact that when using the $I(1)$ framework we are dealing with cumulated weakly stationary processes which are characterized by their second moment structure. Focusing on the second moment structure is enough for studying the structure of dependencies although, in general, it will not be enough for the statistical analysis, since, for example, the cross-section members might be uncorrelated but not independent. However, for our discussion here we use dependence synonymously with correlation.

Considering the non-stationary panel (for any given cross-sectional dimension N) as a high-dimensional time series, we will first define cross-unit cointegration (following Wagner and Hlouskova, 2007), which will then be used to define short-run and long-run dependence.

Let us start our discussion with panels of univariate time series, noting here that the concept of cross-unit cointegration becomes more interesting for panels of multivariate time series. Take without loss of generality in the panel of time series the first N_1 series to be (jointly) stationary and the remaining $N - N_1$ series to be $I(1)$. With the first N_1 series stationary, a set of *trivial* cointegrating relationships emerges for y_t, with a basis given by $[I_{N_1}, 0]'$. Cross-unit cointegration is present if, over and above these, cointegrating relationships that are not contained in the above trivial cointegrating space are present.

This concept can be formalized by denoting as B (a basis of) the cointegrating space of the vector y_t. Then, we define the *cross-unit cointegrating space* as the projection of B on the orthogonal-complement of the trivial cointegrating space (that is, $[I_{N_1}, 0]'$,) given by $B^{CU} = \begin{bmatrix} 0 & 0 \\ 0 & I_{N-N_1} \end{bmatrix} B$, which in our currently simple framework of panels of univariate series amounts to nothing but the cointegrating space of the $N - N_1$ integrated series contained in the panel, assumed to be ordered last in the discussion. The dimension of this space is called the cross-unit cointegrating rank.

Let us now turn to panels of multivariate time series and consider a panel comprised of Nm-dimensional vectors of time series that are assumed to be jointly $I(1)$, respectively stationary, that is, $Y_t = (Y_{1,t}', \ldots, Y_{N,t}')' \in R^{Nm}$.[32] Next denote with $\beta_i \in R^{m \times k_i}$ the k_i-dimensional cointegrating space of $Y_{i,t}$. As before we denote the cointegrating space of the stacked process as $B \in R^{Nm \times \sum_i k_i}$. We stack the individual specific cointegrating spaces in:

$$\beta\!\!\!/ = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \beta_N \end{bmatrix} \in R^{Nm \times \sum_i k_i}.$$

By definition it holds that $sp\{B\} \supseteq sp\{\beta\!\!\!/\}$. The cross-unit cointegrating space is defined as the projection of B on the orthogonal-complement of $\beta\!\!\!/$, that is, it is

given by:

$$B^{CU} = (I_{Nm} - \not{\beta}(\not{\beta}'\not{\beta})^{-1}\not{\beta}')B. \tag{B.1}$$

The dimension of the space (spanned by) B^{CU} is defined as the cross-unit cointegrating rank. The definition merely formalizes the notion that, considering here the simplest example, cointegrating relationships of the form $[\beta_1^{'}, \beta_2^{'}, 0, \ldots, 0]^{'}$ that involve variables from different cross-sections but lead to a stationary transformed process $\beta_1^{'}Y_{1,t} + \beta_2^{'}Y_{2,t}$ merely via combining already stationary processes from different cross-section units (in the example $\beta_1^{'}Y_{1,t}$ and $\beta_2^{'}Y_{2,t}$) should not be considered as *genuine* cross-unit cointegrating relationships. The above definition of the space B^{CU} as the projection of B on the ortho-complement of $\not{\beta}$ delivers all cointegrating relationships that are not given by linear combinations of individual specific cointegrating relationships, that is, $B = B^{CU} \oplus \not{\beta}$, with $\oplus$ denoting the direct sum. We say that the panel exhibits long-run cross-sectional dependence if the cross-unit cointegrating rank is larger than 0. In case the cross-sectional units are correlated, but there is no cross-unit cointegration, we speak of (*pure*) short-run cross-sectional dependence.

Note for completeness that the previous sentence implies that panels comprised of independent random walks are short-run dependent – since independent random walks are asymptotically correlated, this is the famous spurious correlation (for example, Phillips, 1986). Therefore, if one were to be keen on excluding spurious asymptotic correlations, the above definition of short-run dependence can simply be reformulated and stated in terms of cross-sectional correlation once all variables have been transformed to stationarity, that is, by first-order differencing all integrated variables in the panel. In other words, the definition might be based on the innovations of the individual series to avoid the necessity to explicitly discuss spurious correlations.

Given that we consider processes that have a representation as a solution to autonomous stochastic difference equations, both forms of dependence, short-run and long-run, originate in the dependence structure of the error processes driving the individual series in the respective difference equations. Consider, for example, the DGP considered for testing for cointegration by single equation methods in (13.18)–(13.22), abstracting here from deterministic components. In this system the stochastic behavior is governed by the three random processes $\eta_t, \varepsilon_{i,t}$ and $\nu_{i,t}$. The *raison d'être* of the common factors is to induce cross-sectional dependencies via the common factors, but short- and long-run dependence can also arise via the other two components as soon as the assumption that they are cross-sectionally independent is relaxed. To illustrate the issue consider (13.21), that is, $(1-\varphi_i L)e_{i,t} = H_i(L)\varepsilon_{i,t}$, with the $\varepsilon_{i,t}$ being white-noise processes. Noting that the fact that the $\varepsilon_{i,t}$ are individually white-noise processes does not imply that the stacked vector $\varepsilon_t = \left(\varepsilon_{1,t}, \ldots, \varepsilon_{N,t}\right)^{'}$ is a vector white-noise process, it becomes immediately clear

that even via $\varepsilon_{i,t}$ both short- and long-run dependence can be introduced into the panel. The same applies for the vector $u_t = \left(u_{1,t}, \ldots, u_{N,t}\right)'$.

In a way, short-run dependence can, in several respects be potentially less problematic than long-run dependence. If we think of, for example, unit root testing, then short-run cross-sectional correlation will lead to distorted inference, of which account can be taken. Compare the discussion of O'Connell (1998) or the set-up of Chang (2002). Long-run dependence typically is found to have more detrimental effects, see Lyhagen (2000), that cannot be remedied easily, for example, by simple feasible GLS-type corrections.

There is one further issue with respect to long-run dependence or cross-unit cointegration. Our discussion up to here has been for the case of fixed cross-sectional dimension. In general, the dimension of the cross-unit cointegrating space is itself dependent upon N since, due to the inclusion of additional panel members, additional cross-unit cointegrating relationships may emerge. Consequently, a thorough analysis of cross-unit dependence and its effects needs to consider the dependence behavior when $N \to \infty$. Part of the literature takes short-cuts in this respect; for example, Evans and Karras (1996) and Lyhagen (2000) consider a panel set-up with exactly one common trend and with all pair-wise differences of the series being stationary for all values of N. In this way they avoid a proper consideration of the dependence of the cointegration structure on the cross-sectional dimension.

Considering all series together as a high-, or in the limit infinitely-, dimensional time series is useful for understanding the algebraic structure of cointegration, cross-unit cointegration and short-run cross-sectional dependence. However, from a panel modeling perspective, restrictions on the joint DGP have to be put in place in order to materialize gains from pooling in one way or another.

Looking at factor models, consider the joint process $u_t = \left(u_{1,t}, \ldots, u_{N,t}\right)'$, where for simplicity we ignore deterministic components:

$$u_t = \Pi' F_t + e_t,$$

with $\Pi = \left[\pi_1', \ldots, \pi_N'\right]' \in R^{r \times N}$, the r common factors $F_t \in R^r$, and the idiosyncratic components $e_t = \left(e_{1,t}, \ldots, e_{N,t}\right)'$.

Since we focus here on the cointegration implications of the factor model, we assume for simplicity that the factor loadings matrix Π is non-stochastic and that the idiosyncratic components are cross-sectionally independent. As noted previously, the fact that Bai and Ng (2004) also allow the factor loadings Π to be stochastic is mainly a mathematical achievement but does not really change any of the properties of the time series panel since all observations are generated from *one single realization* of the factor loadings. Bai and Ng (2004) allow for a certain form of correlation between the components of e_t (described in section 13.2.2.2) which is why they dub their model an approximate factor model. The corresponding assumption (see Bai and Ng, 2004, p. 1130, Assumption C) is, however, of mainly a technical character and simply allows for bounded correlation between the

innovation sequences. Under some further simplifying assumptions, the long-run covariance matrix of Δu_t, Ω_Δ say, is given by:

$$\Omega_\Delta = \Pi' \Omega_{\Delta F} \Pi + diag\left(\omega_{\Delta i}\right),$$

with $\Omega_{\Delta F} \in R^{r \times r}$ denoting the long-run covariance matrix of the first difference of the factors and $diag(\omega_{\Delta i})$ collecting the long-run variances of Δe_{it} for $i = 1, \ldots, N$. As is well known, cointegration prevails when the long-run covariance matrix has reduced rank. Starting with the idiosyncratic components, we know that the long-run variance of $\Delta e_{i,t}$ is equal to 0 when $e_{i,t}$ is itself already stationary. The first part of Ω_Δ has at most rank r. A cointegrating relationship has to be contained in the left kernel of both terms above. These observations allow us to immediately study the cointegration properties of u_t.

Assume, without loss of generality, that the first $0 \leq r_1 \leq r$ factors are integrated but not cointegrated and the remaining ones are stationary. Also assume that the units are ordered in such a way that the first $0 \leq N_1 \leq N$ coordinates of e_t are stationary and the remaining ones integrated. With this set-up, cointegration prevails if, in $\beta' u_t$, the first r_1 factors and the lower $N - N_1$ coordinates of e_t are annihilated. Partition the factors as $F_t = \left[(F_t^1)', (F_t^2)'\right]'$, with $F_t^1 \in R^{r_1}, F_t^2 \in R^{r-r_1}$ and $\Pi = \left[\Pi_1', \Pi_2'\right]'$ accordingly. Then the cointegrating space is given by the intersection of the orthogonal-complement of Π_1' and $\left[I_{N_1}, 0'_{(N-N_1)\times N_1}\right]'$ to wipe out the two non-stationary contributions to the vector u_t. Without further assumptions on the various spaces the cointegrating rank cannot be determined.

Note in addition that the identification of the factors for $N \to \infty$ rests upon the assumption that $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \pi_i \pi_i' = \Sigma_\Pi > 0$, which does not, determine the ranks of the loading matrices for finite cross-sectional dimension and hence does not, in particular, determine the number of linearly independent integrated common factors for finite N.

However, some simple observations can be made. First, cointegration can only occur if some of the idiosyncratic components are stationary, since it can only occur between series with stationary idiosyncratic components.

Second, if all common factors are stationary, the dimension of the cross-unit cointegrating space is zero.

Third, to allow for a bit more detailed analysis of the cointegrating space assume now that, for the given cross-sectional dimension N, it holds that $rk\left(\Pi_1\right) = r_1$. We already know from the first observation that we only need to consider the first N_1 series $\left(u_{1,t}, \ldots, u_{N_1,t}\right)'$ corresponding to the stationary idiosyncratic components to analyze cointegration. Amongst these series the cointegrating space is given by the left kernel of:

$$\begin{bmatrix} \left(\pi_1^1\right)' \\ \vdots \\ \left(\pi_{N_1}^1\right)' \end{bmatrix} \in R^{N_1 \times r_1}. \tag{B.2}$$

The assumption of full rank of Π_1 does not imply that the above sub-matrix composed of the first N_1 columns of Π_1 also has rank equal to r_1. Clearly, it necessarily has rank smaller than r_1 if $N_1 < r_1$, but can also have reduced rank otherwise. A reduced rank, s_1 say, of the matrix in (B.2) implies that only this smaller number s_1 of common trends is distinguishable within the first N_1 coordinates of u_t. The cointegrating space is given by the orthogonal-complement in R^{N_1} of the space spanned by the matrix (B.2), which is of dimension $N_1 - s_1$ in our discussion. The cross-unit cointegrating space can be determined as in the definition above, again by projecting on the orthogonal-complement of the stacked individual specific cointegrating spaces (β in the notation introduced above). In our univariate context individual specific cointegration means that a series $u_{i,t}$, for some $i = 1, \ldots, N_1$, is stationary, which happens when the corresponding $\lambda_i^1 = 0$. In this case the corresponding entry in β is set equal to 1. Note that, due to possible rank reduction in (B.2), for this to happen it is not necessary for the condition $\pi_i^1 = 0$ to hold, although it is of course sufficient. The above analysis also applies when all idiosyncratic components are stationary, in which case $N_1 = N$.

Altogether the discussion highlights the price that has to be paid, in terms of very restricted structures of the cointegrating spaces, when reducing model complexity (to be precise the modeling of cross-sectional dependence) by resorting to (approximate) factor models.

Clearly, the limitations discussed above are also present when using factor models in a multivariate panel context, either by using factor DGPs as error processes in single equation cointegration analysis (as in section 13.3.1.4) or when considering systems inference procedures with a factor structure in the stochastic component.

13.7 Appendix C: Limiting concepts for integrated panels

One major difference to classical microeconometric panels is that for time series panel applications the assumption of cross-sectional independence is often untenable. For many applications from the realms of international macroeconomics or finance, dependence of the variables across countries appears to be the norm rather than the exception.

The first generation of the literature, by which we label all methods (that is, tests and estimation procedures) that are based on the assumption of cross-sectional independence, has made this strong assumption of cross-sectional independence not because of its empirical validity but because of methodological simplicity. The assumption of cross-sectional independence facilitates many parts of the theoretical analysis considerably.

To illustrate the relative simplicity that the cross-sectional independence assumption – in conjunction with sequential limit theory where $N \to \infty$ after $T \to \infty$ – consider the following simple example. Assume that for each cross-section member the data are generated according to $\Delta y_{it} = \rho_i y_{it-1} + \varepsilon_{it}$, that is, by an autoregression of order 1 without any deterministic components and with ε_{it} a white-noise

process with variance σ_i^2. The OLS estimates and corresponding t-statistics of ρ_i are given by:

$$\hat{\rho}_i = \frac{\sum y_{it-1} \Delta y_{it}}{\sum y_{it-1}^2} \tag{C.1}$$

$$t_{\rho_i} = \frac{\sum y_{it-1} \Delta y_{it}}{\hat{\sigma}_i \left(\sum y_{it-1}^2 \right)^{1/2}}, \tag{C.2}$$

with $\hat{\sigma}_i^2 = \frac{1}{T} \sum \left(\Delta y_{it} - \hat{\rho}_i y_{it-1} \right)^2$. Under the null hypothesis that $\rho_i = 0$, it holds under rather general assumptions on ε_{it} for $T \to \infty$ that:

$$T\hat{\rho}_i \to \frac{\int W(r) dW(r)}{\int W^2(r) dr} \tag{C.3}$$

$$t_{\rho_i} \to \frac{\int W(r) dW(r)}{\left(\int W^2(r) dr \right)^{1/2}}. \tag{C.4}$$

For example, Nabeya (1999) shows the existence of moments up to order six for the above two limiting quantities (C.3) and (C.4) under the assumption that the innovation ε_{it} is an i.i.d. sequence. Denote the expected value, respectively the variance, of the limit in (C.4) by μ_{t_ρ} and $\sigma_{t_\rho}^2$. Now, if we assume that the individual series are cross-sectionally independent and that it holds for all cross-section members that $\rho_i = 0$, then it immediately follows for $N \to \infty$ after $T \to \infty$ that:

$$\bar{t}_\rho = \frac{1}{N} \sum_{i=1}^{N} \frac{t_{\rho_i} - \mu_{t_\rho}}{\sigma_{t_\rho}} \Rightarrow N(0, 1).$$

This result also holds true in the sequential limit if one allows for heterogeneity in the individual series, for example, by allowing for different autoregressive orders and performing corresponding ADF regressions since, as is well known, the same time series limit (C.4) prevails also in this case.

If one wants to consider a joint limit where N and T tend to infinity together, matters become more complicated even if we stick to the assumption of cross-sectional independence. Two properties need to be established to allow for the applicability of joint central limit theorems, which we illustrate again for the t-test. First, the existence of the necessary moments of the finite T quantities (C.2) has to be established. Such results have been derived only under relatively strong assumptions, mainly relying upon normality of the innovations ε_{it} and often only for simple DGPs (see, for example, Evans and Savin, 1981, or Larsson 1997). In this respect it is also important to note that the finite sample distributions of, for example, the t-values depend, in the case of higher-order autoregressive processes, upon all autoregressive coefficients, since the dependence upon these nuisance parameters vanishes only in the limit for $T \to \infty$. This dependence upon characteristics of the DGPs of the cross-section members is often the reason for joint limits being

only established with certain *rate restrictions*, for example, of the form $\frac{\sqrt{N}}{T} \to 0$ or $\frac{N}{T} \to 0$. Phillips and Moon (1999) contains an insightful discussion in this respect.

To return to our example, as long as T is finite, the individual specific t-statistics will in general not be identically distributed. One way of overcoming this problem is to assume cross-sectionally identically distributed series $y_{i,t}$, which is, however, far too strong an assumption to be of any practical relevance. Therefore, since identical distributions for finite T samples are usually out of the question, other conditions that allow the use of joint central limit theorems for independent but not identically distributed random variables have to be established. A prime candidate in this respect is to establish a Lindeberg-type condition, or some special cases formulated in terms of uniform integrability conditions, in the joint limit. On a detailed discussion and an example of this approach see Phillips and Moon (1999).

The major problem for the panel unit root and cointegration literature is to establish the required conditions for finite values of T, which are partly not even derived for the time series limits of the building blocks of the panel statistics. One underrated but good exception in this respect is the work of Larsson, Lyhagen and Löthgren (2001), discussed in section 13.3.2.1, who work out the proofs in detail for a panel version of the Johansen (1995) trace test for VAR models without deterministic components and with normally distributed errors. However, substantial parts of the current literature need theoretical strengthening. It is clearly a major task for the literature to work out the correctness of necessary intermediate results for the numerous test statistics and estimators. Until this task has been accomplished, the panel unit root and cointegration literature urgently needs to make further significant progress in terms of establishing mathematical rigor.

Clearly, relaxing the assumption of cross-sectional independence complicates matters even further, since now limit theory for dependent doubly indexed random sequences has to be invoked to establish results. It is clear that without modeling the extent and form of cross-sectional dependence carefully it will not, in general, be possible to establish any well-defined limiting behavior. Up to now no general modeling strategies for cross-sectionally dependent unit root non-stationary panels have been devised and only certain special modeling approaches are in use to date. Two approaches appear to be prominent. One is actually to treat the N dimension as finite and fixed, which means that essentially high-dimensional time series problems are considered (see, for example, Pedroni *et al.*, 2008). The other approach is given by modeling cross-sectional dependence by resorting to factor models, as we have discussed in the main body of the chapter.

Note as a final remark that the usage of panel techniques with both the time series and the cross-sectional dimension tending to infinity allows us to solve some problems that cannot be addressed in a pure time series setting. These include, for example, the consistent estimation of the non-centrality parameter in a local-to-unity framework or of distant initial conditions (see Moon and Phillips, 2000).

Acknowledgments

We are very grateful to our students, colleagues and co-authors, whose insights and results we share in the results reported in this chapter. Particular thanks go to Josep Carrion-i-Silvestre, Jaroslava Hlouskova, Massimiliano Marcellino and Chiara Osbat for participating in research projects on panel unit roots and cointegration with us over several years. AB also wishes to thank Victor Bystrov for valuable research assistance and the Department of Economics of the European University Institute for funding his research projects over the past eight years. MW thanks the Jubiläumsfonds of the Oesterreichische Nationalbank for partially funding his research over the last years. The help of Kerry Patterson and an anonymous referee is also gratefully acknowledged. Responsibility for any errors that remain rests with us.

Notes

1. Later in our chapter we illustrate the methods described by some other examples. These include the frequently studied topic of purchasing power parity, but we also look at less common examples such as the analysis of environmental Kuznets curves and exchange rate pass-through. Further applications for which non-stationary panel methods have been used include business-cycle synchronization, house price convergence, regional migration and household income dynamics.
2. The discussion in this subsection draws on Wagner (2008c) who considers the precise econometric implications of several economic convergence definitions.
3. Throughout our chapter, when referring to $I(1)$ processes we refer to processes whose stochastic component is integrated of order 1 and allow also for deterministic components.
4. Clearly, this definition leaves lots of possibilities, for example, multiple convergence clubs, entirely unexplored. See Wagner (2008c) for a discussion of these issues.
5. Remember that in the Granger representation r indicates the number of linearly independent common stochastic trends, which in our case is 1 in the case of convergence. Clearly, algebraically this single stochastic trend can be split into r components in infinitely many ways, that is, the constituent parts are not identified. An alternative way of formulating the same thing is to simply remember the fact that for $I(1)$ processes the sum of the dimension of the cointegrating space and the number of common trends equals the dimension of the process.
6. The discussion in EK at pp. 254–5 shows that the authors focus in their considerations on potential correlation in the processes $u_{i,t}$ but ignore the potential of both the presence of deterministic components as well as stochastic trends in the panel of series $y_{i,t} - \overline{y}_t$. The authors make the assumption that as the cross-sectional dimension tends to infinity the series $u_{i,t}$ become uncorrelated. This assumption, coupled with assuming that the Dickey–Fuller regression also describes the DGP, with the series $u_{i,t}$ being indeed white-noise processes, then implies that there is no cointegration between the series $y_{i,t} - \overline{y}_t$ under the null hypothesis of divergence.
7. See Appendix B for further details.
8. This formulation is similar to Bai and Carrion-i-Silvestre (2007), see also Bai and Ng (2004).
9. Breaks may of course coexist with cross-sectional independence. We merely look here at the simplest cases first – that is, without dependence and without breaks – and introduce some of the complications in later sub-sections.
10. For both the unit root and cointegration tests consistency of the tests is established for the case that the process is stationary under the alternative. This implies restrictions on the coefficients φ_{ik} and ρ_i to ensure $I(1)$ behavior under the null hypothesis and stationarity under the alternative, which is the common framework in unit root and cointegration analysis.

11. The efficacy of information criteria such as AIC and BIC in dealing with serial correlation is an issue of some debate. Simulation evidence suggests that, due to the choice of very low lag lengths, especially with BIC, serial correlation remains.
12. Phillips and Moon (1999) contains an excellent discussion concerning sequential (that is, first T to infinity followed by N to infinity) versus joint (that is, T and N tend to infinity simultaneously, potentially with some restrictions on the divergence rates), as well as the relationships between the different asymptotic concepts. Appendix C of our chapter discusses some of these issues.
13. Hlouskova and Wagner (2006) discuss the consequences of assuming, instead of constant covariance, a Toeplitz structure for the covariance matrix which corresponds to a geometrically declining correlation with distance. This same idea is pursued in more detail in Baltagi, Bresson and Pirotte (2007). For simplicity, since it serves to make the substantive point, in our illustration this geometrical decline is absent.
14. Hlouskova and Wagner (2006) present a larger experimental design and also consider more tests.
15. This is a simplifying assumption adopted here, and is stronger than is needed for the Bai and Ng (2004) results to hold. Bai and Ng allow for weak cross-correlation in the errors, "weak in the sense that the column sum of the error covariance matrix remains bounded" (Bai and Ng, 2004, p. 1131). This leads to what is termed an approximate factor model.
16. Considering the factor loadings to be stochastic appears to be more of mathematical rather than practical value, given that for each loading it is only one realization that generates the data.
17. If the number of factors r is unknown, it must be estimated consistently using, for example, the Bai and Ng information criteria. The factors themselves are estimated via principal components, as described above in the discussion of the Bai and Ng (2004) method.
18. The model discussed here does not allow for a linear trend. Section 3 of Moon and Perron (2007) extends the analysis to allow for linear trends.
19. Moon and Perron (2004) derive results for the distribution of the estimator under local-to-unity alternatives for the root. They also need a further restriction on the rate of convergence of N and T to infinity in order to obtain consistent estimates $\hat{\omega}_e^2$,$\hat{\phi}_e^4$ and $\hat{\lambda}_e^N$. Assumption 10 (p. 91) of their paper gives this as $a = \liminf_{(N,T\to\infty)} \log T / \log N > 1$. The parameter a is related to the speed of N/T tending to zero.
20. As noted in note 15, cross-sectional independence of the $e_{i,t}$ processes, via the cross-sectional independence of the $\varepsilon_{i,t}$ processes, is not needed for estimation and inference concerning the factors. We assume it here (and earlier) for simplicity – since this allows for the construction of pooled panel tests for the idiosyncratic terms.
21. The problem is formulated in terms of obtaining consistent estimates of the *break fractions* (as discussed below.) This allows the derivations to deal with $T \to \infty$ in order to provide large-sample results. If instead the break date were fixed, the problem would become degenerate in cases where T was large.
22. The Bai and Carrion-i-Silvestre result presented above, as well as those discussed below, rely upon limiting arguments not fully discussed either by these authors or in the seminal Bai and Ng (2004) paper. A major problem is posed by the fact that, for finite N, the de-factored observations and, consequently, the test statistics based upon these, are not cross-sectionally independent, since only the product of the estimated loadings and the estimated factors are subtracted. Cross-sectional independence requires both N and T to go to infinity to have consistent factors (and loadings). Therefore, one cannot simply appeal to sequential limit theory to derive the asymptotic distribution for $MSB(i)$ above by letting T tend to infinity first (to derive the p-values) and then derive the asymptotic distribution of BC_N by letting N "tend to infinity again." A similar issue arises with the derivation of the Z statistic above. In fact, similar problems plague all of the rapidly growing panel unit root literature based on de-factored observations that fails to take into account that joint limit theory for (in finite samples) cross-sectionally dependent

quantities has to be employed. Such limit theory is, to the best of the authors' knowledge, very sparse at best.

23. For completeness we have also performed the computations including only those countries where no structural breaks are detected or with a shorter panel that excludes the structural breaks. The latter experiment, however, suffers from the poor performance of panel unit root and cointegration tests for short panels, compare again Hlouskova and Wagner (2006) and Wagner and Hlouskova (2007). The potentially poor performance notwithstanding, the additionally obtained results, available upon request, are highly similar in terms of findings and conclusions to the results for the full country set and full period panel discussed in the chapter.
24. As in testing for unit roots in panels, the role of the null and alternative hypotheses can be reversed in the construction of the tests – that is, the null hypothesis can be taken to be that a cointegration vector exists while the alternative hypothesis is that of no cointegration. The limitations nevertheless remain.
25. The analogous discussion with respect to unit roots is contained in section 13.2.2.2.
26. There are other ways in which one can think of introducing instability, in particular through instability in the factors themselves or in the factor loadings π_i. We choose not to address this issue here, since the general problem is already over-detailed.
27. η_t in (13.20) may be taken to be independent of $v_{i,t}$.
28. Remember that in sections 13.2.2.2 and 13.2.3.1 we used the notation $\tilde{y}_{i,t}$ to denote first differences whereas we use this notation here to denote the deviations from the cross-sectional averages. Some overlap in notation appears to be unavoidable but we hope that this does not lead to any confusion.
29. Note that Phillips and Moon (1999) is the only paper that formulates explicitly stochastic assumptions on the underlying DGP that ensure the existence of the required cross-sectional limits by introducing so-called stochastic linear processes, in which the coefficients describing the DGP are cross-sectionally independent random variables with certain properties. This implies that one can, under appropriate assumptions, determine the required limits by laws of large numbers. Clearly the case of deterministic coefficients with corresponding assumptions is nested in this framework.
30. Again some overlap in notation occurs since the index m has been used before to indicate the three main specifications of the deterministic components in unit root and single equation and cointegrating testing whereas here and in Appendix B it denotes the dimension of the multivariate time series in the panel data.
31. The authors consider also a specific joint limit where the rate condition $\frac{N^{1/2}}{T} \to 0$ is put in place. For this specific joint limit, a Lindeberg-type condition is required since the cross-section specific building blocks of the panel test statistic are not i.i.d. for finite values of T, whereas once T has passed to infinity, the time series building blocks of the test statistics are identically distributed in the cross-section dimension.
32. Note that the restriction to the same dimension m for each panel member is not required for the discussion of the cointegration properties.

References

Ahn, S.K. and G.C. Reinsel (1990) Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* **85**, 813–23.

Andreoni, J. and A. Levinson (2001) The simple analytics of the environmental Kuznets curve. *Journal of Public Economics* **80**, 269–86.

Andrews, D.W.K. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–58.

Bai, J. and J.L. Carrion-i-Silvestre (2007) Structural changes, common stochastic trends, and unit roots in panel data. Mimeo.

Bai, J. and S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.
Bai, J. and S. Ng (2004) A PANIC attack on unit roots and cointegration. *Econometrica* **72**, 1127–77.
Bai, J. and P. Perron (1998) Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78.
Baltagi, B.H., G. Bresson and A. Pirotte (2007) Panel unit root tests and spatial dependence. *Journal of Applied Econometrics* **22**, 339–60.
Banerjee, A. and J.L. Carrion-i-Silvestre (2007) Cointegration in panel data with breaks and cross-section dependence. Mimeo.
Banerjee, A., M. Marcellino and C. Osbat (2004) Some cautions on the use of panel methods for integrated series of macroeconomic data. *Econometrics Journal* **7**, 322–40.
Banerjee, A., M. Marcellino and C. Osbat (2005) Testing for PPP: should we use panel methods? *Empirical Economics* **30**, 77–91.
Bernard, A.B. and S.N. Durlauf (1995) Convergence in international output. *Journal of Applied Econometrics* **10**, 97–108.
Bernard, A.B. and S.N. Durlauf (1996) Interpreting tests of the convergence hypothesis. *Journal of Econometrics* **71**, 161–73.
Breitung, J. (2000) The local power of some unit root tests in panel data. In B.H. Baltagi (ed.), *Nonstationary Panels, Panel–Cointegration, and Dynamic Panels*, pp. 161–77. Amsterdam: Elsevier.
Breitung, J. (2005) A parametric approach to the estimation of cointegration vectors in panel data. *Econometric Reviews* **24**, 151–73.
Breitung, J. and S. Das (2008) Testing for unit roots in panels with a factor structure. *Econometric Theory* **24**, 88–108.
Breitung, J. and M.H. Pesaran (2008) Unit roots and cointegration in panels. In L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 279–322. Boston: Kluwer Academic Publishers.
Brock, W.A. and M.S. Taylor (2004) The green Solow model. NBER Working Paper No. 10557.
Brock, W.A. and M.S. Taylor (2005) Economic growth and the environment: a review of theory and empirics. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth, Vol. 1B*, pp. 1749–821. Amsterdam: North-Holland.
Campa, J.M., L. Goldberg and J. González-Minguez (2005) Exchange-rate pass-through to import prices in the euro-area. NBER Working Paper No. 11632.
Campa, J.M. and J. González-Minguez (2006) Differences in exchange rate pass-through in the euro-area. *European Economic Review* **50**, 121–45.
Chang, Y. (2002) Nonlinear IV unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* **110**, 261–92.
Choi, I. (2001) Unit root tests for panel data. *Journal of International Money and Finance* **20**, 249–72.
Choi, I. (2006a) Combination unit root tests for cross-sectionally correlated panels. In D. Corbae, S. Durlauf and B. Hansen (eds.), *Econometric Theory and Practice: Frontiers of Analysis and Applied Research. Essays in Honor of Peter C.B. Phillips*, pp. 311–33. New York: Cambridge University Press.
Choi, I. (2006b) Nonstationary panels. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory*, pp. 511–39. New York: Palgrave Macmillan.
Coakley, J. and A.M. Fuertes (1997) New panel unit root tests of PPP. *Economics Letters* **57**, 17–22.
de Bandt, O., A. Banerjee and T. Kozluk (2008) Measuring long-run exchange rate pass-through. *Economics E-Journal* **2**, 2008–6.
Dickey, D.A. and W.A. Fuller (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**, 427–31.

Engel, C. (2000) Long-run PPP may not hold after all. *Journal of International Economics* **51**, 243–73.

Engle, R.F. and C.W.J Granger (1987) Co-integration and error correction: representation, estimation and testing. *Econometrica* **55**, 251–76.

Engle, R.F. and B.S. Yoo (1991) Cointegrated economic time series: an overview with new results. In R.F. Engle and C.W.J. Granger (eds.), *Long Run Economic Relationships: Readings in Cointegration*, pp. 237–66. Oxford: Oxford University Press.

Evans, G.B.A. and N.E. Savin (1981) Testing for unit roots: 1. *Econometrica* **49**, 753–97.

Evans, P. and G. Karras (1996) Convergence revisited. *Journal of Monetary Economics* **37**, 249–65.

Fisher, R.A. (1932) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Frankel, J.A., D. Parsley and S. Wei (2005) Slow passthrough around the world: a new import for developing countries? NBER Working Paper No. 11199.

Frankel, J.A. and A.K. Rose (1996) A panel project on purchasing power parity: mean reversion within and between countries. *Journal of International Economics* **40**, 209–24.

Garratt, A., K. Lee, M.H. Pesaran and Y. Shin (2006) *Global and National Macroeconometric Modelling: A Long-run Structural Approach*. Oxford: Oxford University Press.

Gengenbach, C., F.C. Palm and J.-P. Urbain (2006) Panel unit root tests in the presence of cross-sectional dependencies: comparison and implications for modelling. *Oxford Bulletin of Economics and Statistics* **68**, 683–719.

Groen, J.J.J. and F. Kleibergen (2003) Likelihood-based cointegration analysis in panels of vector error correction models. *Journal of Business and Economic Statistics* **21**, 295–318.

Gregory, A.W. and B.E. Hansen (1996) Residual-based tests for cointegration in models with regime shifts. *Journal of Econometrics* **70**, 99–126.

Grossman, G.M. and A.B. Krueger (1995) Economic growth and the environment. *Quarterly Journal of Economics* **110**, 353–77.

Hadri, K. (2000) Testing for stationarity in heterogeneous panel data. *Econometrics Journal* **3**, 148–61.

Hadri, K. and R. Larsson (2005) Testing for stationarity in heterogeneous panel data where the time dimension is fixed. *Econometrics Journal* **8**, 55–69.

Harris, R.D.F. and E. Tzavalis (1999) Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics* **90**, 1–44.

Hlouskova, J. and M. Wagner (2006) The performance of panel unit root and stationarity tests: results from a large scale simulation study. *Econometric Reviews* **25**, 85–116.

Hlouskova, J. and M. Wagner (2008) Finite sample correction factors for panel cointegration tests. Forthcoming in *Oxford Bulletin of Economics and Statistics*.

Hong, S.H. and M. Wagner (2008a) Nonlinear cointegration analysis and the environmental Kuznets curve. Mimeo.

Hong, S.H. and M. Wagner (2008b) Seemingly unrelated nonlinear cointegrating regressions with an application to the environmental Kuznets curve. Mimeo.

Im, K.S. and M.H. Pesaran (2003) On the panel unit root test using nonlinear instrumental variables. Mimeo.

Im, K.S., M.H. Pesaran and Y. Shin (2003) Testing for unit roots in heterogeneous panels. *Journal of Econometrics* **115**, 53–74.

Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Jones, L.E. and R.E. Manuelli (2001) Endogenous policy choice: the case of pollution and growth. *Review of Economics Dynamics* **4**, 369–405.

Kao, C. (1999) Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics* **90**, 1–44.

Kao, C. and M.-H. Chiang (2000) On the estimation and inference of a cointegrated regession in panel data. In B.H. Baltagi (ed.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, pp. 179–222. Amsterdam: Elsevier.

Kuznets, S. (1955) Economic growth and income inequality. *American Economic Review* **45**, 1–28.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* **54**, 159–78.

Larsson, R. (1997) On the asymptotic expectations of some unit root tests in a first order autoregressive process in the presence of trend. *Annals of the Institute of Statistical Mathematics* **49**, 585–99.

Larsson, R. and J. Lyhagen (1999) Likelihood-based inference in multivariate panel cointegration models. Working Paper Series in Economics and Finance 331, Stockholm School of Economics.

Larsson, R., J. Lyhagen and M. Löthgren (2001) Likelihood-based cointegration tests in heterogeneous panels. *Econometrics Journal* **4**, 109–42.

Levin, A., C.F. Lin and C-S. J. Chu (2002) Unit roots in panel data: asymptotic and finite sample properties. *Journal of Econometrics* **108**, 1–22.

Lothian, J.R. (1997) Multi-country evidence on the behavior of purchasing power parity under the current float. *Journal of International Money and Finance* **16**, 19–35.

Lyhagen, J. (2000) Why not use standard panel unit root tests for testing PPP. Stockholm School of Economics, mimeo.

MacKinnon, J.G. (1994) Approximate asymptotic distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics* **11**, 609–18.

MacKinnon, J.G., A. Haug and L. Michelis (1999) Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics* **14**, 563–77.

Maddala, G.S. and S. Wu (1999) A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* **61**, 631–52.

Maddison, A. (2007) *Contours of the World Economy: The Pace and Pattern of Change 1–2030 A.D.* Cambridge: Cambridge University Press.

Mark, N.C. and D. Sul (2003) Cointegration vector estimation by panel dynamic OLS and long-run money demand. *Oxford Bulletin of Economics and Statistics* **65**, 655–80.

Molinas, C. (1986) A note on spurious regressions with integrated moving average errors. *Oxford Bulletin of Economics and Statistics* **48**, 279–82.

Moon, H.R. and B. Perron (2004) Testing for a unit root in panels with dynamic factors. *Journal of Econometrics* **122**, 81–126.

Moon, H.R. and P.C.B. Phillips (2000) Estimation of autoregressive roots near unity using panel data. *Econometric Theory* **16**, 927–97.

Nabeya, S. (1999) Asymptotic moments of some unit root test statistics in the null case. *Econometric Theory* **15**, 139–49.

Newey, W.K. and K.D. West (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–8.

Newey, W.K. and K.D. West (1994) Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* **61**, 631–53.

Nickell, S.J. (1978) Biases in dynamic models with fixed effects. *Econometrica* **49**, 1417–26.

O' Connell, P.J. (1998) The overvaluation of purchasing power parity. *Journal of International Economics* **44**, 1–19.

Onatski, A. (2006) Determining the number of factors from empirical distributions of eigenvalues. Mimeo.

Park, J.Y. and P.C.B. Phillips (1999) Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory* **15**, 269–98.

Park, J.Y. and P.C.B. Phillips (2001) Nonlinear regressions with integrated time series. *Econometrica* **69**, 117–61.

Pedroni, P. (1999) Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics and Statistics* **61**, 653–70.

Pedroni, P. (2000) Fully modified OLS for heterogeneous cointegrated panels. In B.H. Baltagi (ed.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, pp. 93–130. Amsterdam: Elsevier.

Pedroni, P. (2001) Purchasing power parity tests in cointegrated panels. *Review of Economics and Statistics* **83**, 1371–5.
Pedroni, P. (2004) Panel cointegration, asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric Theory* **20**, 597–625.
Pedroni, P.L., T.J. Vogelsang, M. Wagner and J. Westerlund (2008) Nonparametric unit root and cointegration rank tests for time series panels. Mimeo.
Perron, P. (1989) The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* **57**, 1361–401.
Pesaran, M.H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74**, 967–1012.
Pesaran, M.H. (2007) A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics* **22**, 265–312.
Pesaran, M.H., T. Schuerman and S. Weiner (2004) Modelling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business and Economics Statistics* **22**, 129–62.
Phillips, P.C.B. (1986) Understanding spurious regressions in econometrics. *Journal of Econometrics* **33**, 311–40.
Phillips, P.C.B. and B. Hansen (1990) Statistical inference in instrumental variables regressions with $I(1)$ processes. *Review of Economic Studies* **57**, 99–125.
Phillips, P.C.B. and H.R. Moon (1999) Linear regression limit theory for nonstationary panel data. *Econometrica* **67**, 1057–111.
Said, S.E. and D.A. Dickey (1984) Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **71**, 599–607.
Saikkonen, P. (1991) Asymptotically efficient estimation of cointegrating regressions. *Econometric Theory* **8**, 1–27.
Saikkonen, P. (1999) Testing the normalization and overidentification of cointegrating vectors in vector autoregressive processes. *Econometric Reviews* **18**, 235–57.
Sargan, D. and A. Bhargava (1983) Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle. *Econometrica* **51**, 799–820.
Schwert, G.W. (1989) Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics* **7**, 147–59.
Stern, D.I. (2006) Reversal of the trend in global anthropogenic sulfur emissions. *Global Environmental Change* **16**, 207–20.
Stock, J.H. (1999) A class of tests for integration and cointegration. In R.F. Engle and H. White (eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W.F. Granger*, pp. 135–67. Oxford: Oxford University Press.
Stock, J.H. and M.W. Watson (1988) Testing for common trends. *Journal of the American Statistical Association* **83**, 1097–107.
Stokey, N. (1998) Are there limits to growth? *International Economic Review* **39**, 1–31.
Wagner, M. (2008a) On PPP, unit roots and panels. *Empirical Economics* **35**, 229–49.
Wagner, M. (2008b) The carbon Kuznets curve: a cloudy picture emitted by bad econometrics? *Resource and Energy Economics* **30**, 388–408.
Wagner, M. (2008c) On definitions of economic convergence. Mimeo.
Wagner, M. and J. Hlouskova (2007) The performance of panel cointegration methods: results from a large scale simulation study. Forthcoming in *Econometric Reviews*.
Westerlund, J. (2005) New simple tests for panel cointegration. *Econometric Reviews* **24**, 297–316.
Wu, X. (1996) Are real exchange rates nonstationary? Evidence from a panel-data test. *Journal of Money, Credit and Banking* **28**, 54–63.

Part V
Microeconometrics

This page intentionally left blank

14

Microeconometrics: Current Methods and Some Recent Developments

A. Colin Cameron

Abstract

This chapter surveys microeconometrics methods with the emphasis being on recent developments in these methods. The survey presumes the basic theory for the standard estimation methods (LS, ML and IV). Estimation methods surveyed include GMM, empirical likelihood, simulation-based estimation, Bayesian methods, quantile regression and semiparametric estimation. Inference methods include robust inference and bootstrap methods. The chapter addresses the recent literature on estimation of marginal effects that can be given a causative interpretation, notably treatment effects. The common data complications of nonrandom sampling, missing data and mismeasured data are also discussed.

14.1 Introduction

Applied microeconometrics primarily applies regression methods to cross-section and longitudinal economics-related data. Most often the goal is to obtain estimates of one or more marginal effects. A stereotypical example is estimation of the effect on earnings of a one-year increase in education. A simple approach is ordinary least squares (OLS) estimation of a linear cross-section regression of log-earnings on years of schooling and other control variables. Potential complications include nonlinearity (with implications for estimation and statistical inference); endogeneity of the regressor schooling (that is chosen by the individual); unobserved individual heterogeneity (the marginal effect even after controlling for regressors may differ across individuals); and missing or mismeasured data.

In this chapter I survey various methods to deal with these complications. Some of these methods have already become well established and command little current theoretical research. Other methods, especially those that are currently active areas of research, may or may not ultimately become part of the toolkit. An impetus for many of these methods is increased computing power and data availability, discussed in the next chapter in this volume, by Jacho-Chávez and Trivedi.

The survey presumes the basic theory for least squares (LS), maximum likelihood (ML) and instrumental variables (IV) estimation of nonlinear cross-section models and linear panel data models, methods well established by the late 1970s. Section 14.2 presents a summary of identification that includes more recent semiparametric identification and partial identification. Section 14.3 presents estimation methods that enable the use of richer models, notably generalized methods of moments (GMM), empirical likelihood, simulation-based methods (classical and Bayesian), quantile regression, and semiparametric estimation. Even when more basic LS, ML and IV estimators are used, there have been considerable developments in statistical inference, most notably the use of robust standard errors and bootstrap methods. These are presented in section 14.4. Section 14.5 presents a wide range of methods that have been developed to obtain marginal effects that can be given a causative interpretation, even when observational data are used. A fundamental change in thinking is the use of the potential outcomes framework and quasi-experimental approaches to tease out causation. Section 14.6 discusses methods to control for unobserved heterogeneity. Section 14.7 presents adjustments to standard methods that incorporate the practical data complications of survey sampling schemes, missing data, and measurement error.

The following notation is used. The typical observation is the ith, with scalar dependent variable y_i, $k \times 1$ regressor vector $\mathbf{x}_i$, and, where relevant, $m \times 1$ instrument vector $\mathbf{z}_i$. Unless otherwise noted independence over i is assumed. At times

it is convenient to denote the ith observation by $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ or $\mathbf{w}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i)$. The parameter vector in general is a $q \times 1$ vector $\boldsymbol{\theta}$. In some cases this is specialized to a $k \times 1$ parameter vector $\boldsymbol{\beta}$. Combining all N observations, $\mathbf{y}$ is the $N \times 1$ vector of dependent variables, and $\mathbf{X}$ is the $N \times K$ regressor matrix. The linear regression model is written as $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

The reader should be aware that this is a methods survey, rather than a literature survey. It is not possible to cite more than a few relevant references for each topic, leading to omission of the important contributions of many authors. More complete references are given in the relevant texts by T. Amemiya (1985), Greene (2003, first edition 1990), Davidson and MacKinnon (1993), Wooldridge (2008, first edition 2002) and Cameron and Trivedi (2005), and in the lectures by Imbens and Wooldridge (2007). The most recent references given in this chapter should provide a useful start to the current literature.

14.2 Identification

Most applied econometrics studies use methods and models for which identification is not an issue, with a notable exception being the need to have sufficient instruments in linear regression with endogenous regressors. Identification does come to the forefront when more complex models are estimated, or when models are incompletely parameterized.

14.2.1 Point identification

Introductory treatments of econometrics focus on specifying a parametric model for the conditional distribution $f(y|\mathbf{x}, \boldsymbol{\theta})$, or for the conditional mean, $\mathrm{E}[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta})$. Given a specification of $f(\cdot)$ or $g(\cdot)$ and a sampling process, such as random sampling or an exogenous stratified sampling that provides no additional complication, the emphasis is on estimation of the parameters $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$, and on statistical inference based on these parameter estimates. Identification, meaning unique determination of $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$, is discussed briefly in the context of rank conditions to ensure identification in linear simultaneous equations models.

For nonlinear parametric models, identification can be more challenging. A standard result is that in, for example, Newey and McFadden (1994, p. 2134), who state that "the identification condition for consistency of an extremum estimator is that the limit of the objective function has a unique maximum at the truth."

For semiparametric modeling, the identification question is whether a model, or key features of that model, can be estimated assuming an infinitely large sample is available and given the relevant sampling scheme. Only after identification is secured can one move on to estimation and inference given a finite sample. An example is a censored regression model, with observed data $y_i = y_i^*$ if $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i \geq 0$ and $y_i = 0$ otherwise. The goal is to (uniquely) identify $\boldsymbol{\beta}$ given assumptions on the distribution of u_i that fall short of complete parameterization of the distribution of u_i (such as assuming normality). In general there is no unified theory and identification conditions vary with the model being considered and the sampling

process. Also, not all parameters may be identified. For example, regression coefficients may be identified only up to scale or an intercept may not be identified while slope parameters are; Pagan and Ullah (1999) provide many examples.

A nonparametric nonlinear simultaneous equations model is $\mathbf{r}(\mathbf{y}_i, \mathbf{x}_i) = \mathbf{u}_i$, where $\mathbf{y}$ and $\mathbf{u}$ are $G \times 1$ vectors and $\mathbf{x}$ is $K \times 1$. The model is nonparametric identified if it is possible to recover the unknown function $\mathbf{r}(\cdot)$ and the distribution of $\mathbf{u}$ from the joint distribution of $(\mathbf{y}, \mathbf{x})$. Matzkin (2008) provides identification conditions when $\mathbf{u}$ is independent of $\mathbf{x}$.

14.2.2 Partial identification

Manski (1995, 2008) and related papers emphasize partial identification or set identification that merely provides bounds, rather than stronger point identification or complete identification. Partial identification can be possible under weaker assumptions about the data-generating process (DGP) than those needed for point identification.

For example, suppose data on y are missing for 20% of the sample, potentially due to self-selection. Then, without any further assumptions, the median is necessarily bounded by the 0.375 and 0.625 quantiles of the nonmissing data. For example, with 80 observed values of y suppose that the 20 missing values are all less than the smallest observed value. Then the median of all 100 observations is the 30th or 31st of the 80 observed values, or the $30/80 = 0.375$ quantile. Bounding $\mathrm{E}[y]$, rather than the median, is more challenging as it requires additional assumptions on the minimum and maximum value of the mean for the missing data. Qualitatively similar results exist if a fraction of the data are mismeasured rather than missing. In practice the bounds obtained can be wide, but additional information or assumptions can tighten bounds considerably.

Manski and Pepper (2000) provide an upper bound for returns to schooling, controlling for schooling level being endogenously chosen. Haile and Tamer (2003) provide bounds on the quantiles of the distribution of bidders' valuations using auction outcomes. Blundell *et al.* (2007) provide bounds on the interquartile range of wages, controlling for changing composition of the employed and unemployed. Statistical inference, using a framework of estimation based on moment inequalities, is presented in Chernozhukov, Hong and Tamer (2007).

Finally, it should be noted that while much of the literature focuses on identification of parameters, this may not be necessary. In particular, many studies in microeconometrics seek to calculate the marginal effect on the conditional mean of, say, the jth regressor, $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j\big|_{\mathbf{x}=\mathbf{x}^*}$, and this can be achieved by nonparametric or semiparametric regression. Even where a model for $\mathrm{E}[y|\mathbf{x}]$ is posited, complete identification of $\mathrm{E}[y|\mathbf{x}]$ may not be necessary. For example, consider a linear panel fixed effects model where $\mathrm{E}[y_{it}|\mathbf{x}_{it}] = \mathbf{x}_{it}'\boldsymbol{\beta}$ and $\mathbf{x}_{it}$ includes a time-invariant variable, the kth say, with $x_{ik} = x_k$. Then even if x_k is unobserved, fixed effects estimation provides consistent estimates of the components of $\boldsymbol{\beta}$ corresponding to time-varying regressors, and hence the marginal effect.

14.3 Estimation

Most applied microeconometric studies include estimation of parametric models or the conditional mean, $\mathrm{E}[y_i|\mathbf{x}_i]$, or the conditional density $f(y_i|\mathbf{x}_i)$. The specific models used vary with the type of outcome y that is being modeled. The commonly used estimation methods are ML and quasi-ML (where appropriate) for fully parameterized models, and LS and IV for linear conditional mean models.

In the simplest case y is continuous on $(-\infty, \infty)$ and the linear model $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ is used. But often the outcome is restricted in some way, leading to various nonlinear models. In the most extreme case y can take only one of two values, such as whether or not employed. Then the distribution is necessarily a Bernoulli (the binomial with one trial) and different models for the probability parameter correspond to logit and probit models. When there are only a few possible categorical outcomes a wide range of multinomial models exist, including multinomial and ordered logit and probit. For count outcome y that takes only non-negative integer values, such as number of doctor visits, the standard parametric models are Poisson and negative binomial with $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. For duration outcome y, such as length of employment spell, the standard parametric models are exponential and Weibull.

Econometrics packages for cross-section data provide estimators for these models, as well as for the standard corrections for the common complications of truncation and censoring. We do not detail these models and their standard estimators, though the appropriate statistical inference is detailed in section 14.4.

This section instead reviews more advanced estimation methods that permit estimation of more flexible parametric models, when models are fully parameterized, as well as methods that permit estimation when models are not fully parameterized. The most important of these methods is GMM, which provides a very general framework for estimation of nonlinear models that nests OLS, IV and ML estimation. Empirical likelihood is an adaptation of GMM that has different finite sample properties. Simulation methods permit classical and Bayesian methods to be applied to a much wider range of parametric models. Quantile regression and semiparametric methods place less structure on the data generating process.

Among these methods only GMM, quantile regression and nonparametric regression (with single regressor) appear in one or more standard econometrics software packages, so the methods are currently not as widely used as they might be.

14.3.1 Generalized method of moments

The starting point for GMM is the moment condition:

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \tag{14.1}$$

where $\mathbf{h}(\cdot)$ is an $r \times 1$ vector.

The analogy principle, emphasized by Manski (1988) who attributes it to Goldberger, proposes estimation using the sample analog of the population condition (14.1). In the just-identified case this leads to a method of moments (MM) estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MM}}$ that solves $N^{-1}\sum \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}$. A simple example is that

$\mathrm{E}[y_i - \mu] = 0$ leads to the estimator $\widehat{\mu} = \bar{y}$. OLS, ML and just-identified IV estimators can be interpreted as examples of MM estimators.

In the overidentified case in which the dimension r of h_i is greater than q, there are more moment conditions than parameters. Hansen (1982) proposed the GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ that minimizes the quadratic form:

$$Q(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum\nolimits_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right]' \mathbf{W}_N \left[\frac{1}{N}\sum\nolimits_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right], \tag{14.2}$$

where $\mathbf{W}_N$ is an $r \times r$ symmetric full rank weighting matrix that is usually data-dependent. The resulting estimator sets a $q \times r$ linear combination of $\frac{1}{N}\sum_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})$ equal to $\mathbf{0}$. Under appropriate assumptions, including that (14.1) holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance matrix of "sandwich form":

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}] = \frac{1}{N}\left(\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{G}}\right)^{-1}\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{S}}\mathbf{W}_N\widehat{\mathbf{G}}\left(\widehat{\mathbf{G}}'\mathbf{W}_N\widehat{\mathbf{G}}\right)^{-1}, \tag{14.3}$$

where $\widehat{\mathbf{G}} = N^{-1}\sum_i \partial \mathbf{h}_i/\partial \boldsymbol{\theta}'\big|_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{S}}$ is a consistent estimate of $\mathbf{S}_0 = \mathrm{plim}\frac{1}{N}\sum_i\sum_j \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)\mathbf{h}(\mathbf{w}_j, \boldsymbol{\theta}_0)'$. Given independence over i, $\widehat{\mathbf{S}}$ simplifies to $\widehat{\mathbf{S}} = \frac{1}{N}\sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})'$, while for clustered observations adaptations similar to those given in section 14.4.1 are used.

A leading overidentified example is IV estimation. The condition that instruments $\mathbf{z}_i$ are uncorrelated with the error term $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ in a linear regression model implies that $\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. In the just-identified case the MM estimator solves $\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$, which yields the IV estimator. In the overidentified case the GMM estimator minimizes $\left[\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right]' \mathbf{W}_N \left[\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = \mathbf{0}$. The two-stage least squares (2SLS) estimator is the special case $\mathbf{W}_N = \left[N^{-1}\sum_i \mathbf{z}_i\mathbf{z}_i'\right]^{-1}$. Estimators for dynamic panel data models, such as that of Arellano and Bond (1991), are also overidentified GMM estimators that are in common use.

The GMM estimator reduces to the MM estimator, regardless of the choice of $\mathbf{W}_N$, for just-identified models. For overidentified models the most efficient GMM estimator based on the moment conditions (14.1), called the optimum GMM (OGMM) or two-step GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{OGMM}}$, sets $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ is a consistent estimate of $\mathbf{S}_0$. Given independence over i, $\widehat{\mathbf{S}} = \frac{1}{N}\sum_i \mathbf{h}(\mathbf{w}_i, \widetilde{\boldsymbol{\theta}})\mathbf{h}(\mathbf{w}_i, \widetilde{\boldsymbol{\theta}})'$, where $\widetilde{\boldsymbol{\theta}}$ is a first-step GMM estimator based on an initial choice of $\mathbf{W}_N$. Then the OGMM estimator has estimated asymptotic variance

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}_{\mathrm{OGMM}}] = N^{-1}(\widehat{\mathbf{G}}'\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{G}})^{-1}.$$

Chamberlain (1987) showed that the OGMM estimator is the fully efficient estimator based on condition (14.1). In practice, however, it is found that the optimal GMM estimator suffers from small sample bias (see Altonji and Segal, 1996), and other simpler choices of $\mathbf{W}_N$ may be better. This has spawned an active literature, including Windmeijer (2005) and the empirical likelihood methods given in section 14.3.2.

There are several attractions to GMM. First, it provides a unifying framework to estimation as it nests many estimation procedures, including LS, with $\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, and ML, with $\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$, as special cases. Second, it provides a natural extension of instrumental variables methods in overidentified models from linear to nonlinear models, and can be viewed as a generalization of nonlinear 2SLS. Third, it views estimation as a sample analog to population moment conditions, the analogy principle emphasized by Manski (1988). Fourth, taking this view leads naturally to conditional moment tests (see section 14.4.2) that lead to model moment specification tests based on model moment conditions that are not exploited in estimation. Finally, it is relatively simple computationally as standard iterative methods such as Newton–Raphson can be employed, though not all econometric packages provide a general GMM command for nonlinear models.

A method closely related to GMM, though one less used, is minimum distance estimation. Suppose that the relationship between q structural parameters and $r > q$ reduced form parameters is that $\boldsymbol{\pi} = \mathbf{g}(\boldsymbol{\theta})$. Given a consistent estimate $\widehat{\boldsymbol{\pi}}$ of the reduced form parameters, an obvious estimator is $\widehat{\boldsymbol{\theta}}$ such that $\widehat{\boldsymbol{\pi}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$. But this is infeasible since $q < r$. Instead, the minimum distance (MD) estimator $\widehat{\boldsymbol{\theta}}_{\text{MD}}$ minimizes, with respect to $\boldsymbol{\theta}$, the objective function:

$$Q_N(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta}))'\mathbf{W}_N(\widehat{\boldsymbol{\pi}} - \mathbf{g}(\boldsymbol{\theta})), \tag{14.4}$$

where $\mathbf{W}_N$ is an $r \times r$ weighting matrix. The optimal MD estimator uses the weighting matrix $\mathbf{W}_N = \widehat{\mathbb{V}}[\widehat{\boldsymbol{\pi}}]^{-1}$ in (14.4). This estimator is used mainly in panel data analysis (see Chamberlain, 1982, 1984), especially in estimation of covariance structures (see Abowd and Card, 1987).

The statistics literature rarely uses the GMM framework. This may be because GMM is particularly useful for overidentified models, notably IV with surplus instruments, that are much more often used in econometrics. Instead, for nonlinear models the statistics literature emphasizes the more restrictive generalized linear models and generalized estimating equations frameworks (see McCullagh and Nelder, 1983).

14.3.2 Empirical likelihood

Empirical likelihood is based on the same moment conditions as GMM, but is a different estimation method with better second-order asymptotic properties. Empirical likelihood may be a better estimator in settings where optimum GMM is known to perform poorly in finite samples, but it is not widely used, in part due to computational difficulty.

Let $\pi_i = f(y_i|\mathbf{x}_i)$ denote the probability that the ith observation on y takes the realized value y_i. The empirical likelihood (EL) approach, introduced by Owen (1988), maximizes the empirical log-likelihood function:

$$Q_N(\pi_1, \ldots, \pi_N) = N^{-1}\sum_{i=1}^{N} \ln \pi_i, \tag{14.5}$$

subject to any model constraints.

With no model the only constraint is that probabilities sum to one. This leads to maximum EL estimates $\widehat{\pi}_i = 1/N$, so the estimated density function $\widehat{f}(y|\mathbf{x})$ has mass $1/N$ at each of the realized values y_i, $i = 1, \ldots, N$, and the resulting distribution function estimate is just the usual empirical distribution function.

With a model introduced, attention focuses on the estimates for parameters of that model. In the simplest case of estimation of a common population mean μ, the maximum EL estimate can be shown to be the sample mean. A more general example is to specify a model that imposes r moment conditions:

$$\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \tag{14.6}$$

the same condition as in (14.1) for MM or GMM estimation. The EL approach maximizes the empirical likelihood function $N^{-1} \sum_i \ln \pi_i$ subject to the constraint $\sum_i \pi_i = 1$, since probabilities sum to one, and the additional sample constraint based on the population moment condition (14.6) that:

$$\sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{14.7}$$

Thus maximize with respect to $\boldsymbol{\pi} = [\pi_1 \ldots \pi_N]'$, η, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ the Lagrangian:

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ln \pi_i - \eta \left(\sum_{i=1}^{N} \pi_i - 1 \right) - \boldsymbol{\lambda}' \sum_{i=1}^{N} \pi_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}), \tag{14.8}$$

where the Lagrangian multipliers are a scalar η and an $r \times 1$ column vector $\boldsymbol{\lambda}$.

This maximization is not straightforward. First concentrate out the N parameters $\pi_1, \ldots, \pi_N$. Differentiating $\mathcal{L}(\boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \boldsymbol{\theta})$ with respect to π_i yields $1/(N\pi_i) - \eta - \boldsymbol{\lambda}'\mathbf{h}_i = 0$. Then find $\eta = 1$ by multiplying by π_i and summing over i and using $\sum_i \pi_i \mathbf{h}_i = \mathbf{0}$. It follows that the Lagrangian multipliers $\boldsymbol{\lambda}$ solve $\pi_i(\boldsymbol{\theta}, \boldsymbol{\lambda}) = 1/[N(1 + \boldsymbol{\lambda}'\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))]$. The problem is now reduced to a maximization problem with respect to $(r + q)$ variables $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the Lagrangian multipliers associated with the r moment conditions (14.7) and the q parameters $\boldsymbol{\theta}$. Solution at this stage requires numerical methods, even for just-identified models with $r = q$. After some algebra, the log-likelihood function evaluated at $\boldsymbol{\theta}$ is:

$$\mathcal{L}_{\mathrm{EL}}(\boldsymbol{\theta}) = -N^{-1} \sum_{i=1}^{N} \ln[N(1 + \boldsymbol{\lambda}(\boldsymbol{\theta})'\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}))]. \tag{14.9}$$

The maximum empirical likelihood (MEL) estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MEL}}$ maximizes this function with respect to $\boldsymbol{\theta}$.

Qin and Lawless (1994) show that the MEL estimator has the same limit distribution as the optimal GMM estimator. In finite samples, however, $\widehat{\boldsymbol{\theta}}_{\mathrm{MEL}}$ differs from $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$. Furthermore, inference can be based on sample estimates $\widehat{\mathbf{G}} = \sum_i \widehat{\pi}_i \partial \mathbf{h}_i / \partial \boldsymbol{\theta}' \big|_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{S}} = \sum_i \widehat{\pi}_i \mathbf{h}_i(\widehat{\boldsymbol{\theta}}) \mathbf{h}_i(\widehat{\boldsymbol{\theta}})'$ that weight by the estimated probabilities $\widehat{\pi}_i$ rather than the proportions $1/N$. Newey and Smith (2004) show that MEL

has better second-order asymptotic properties than GMM, and it appears that using the weights $\widehat{\pi}_i$ in forming $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{S}}$ leads to improved finite sample performance.

Generalized empirical likelihood estimators (GEL) use objective functions other than $N^{-1}\sum_i \ln \pi_i$. Exponential tilting uses $N^{-1}\sum_i \pi_i \ln \pi_i$, and the continuous updating GMM estimator of Hansen, Heaton and Yaron (1996) is shown by Newey and Smith (2004) to fall in the class of GEL estimators.

Computational methods for these estimators are presented in Mittelhammer, Judge and Schoenberg (2005), and in the surveys of empirical likelihood by Imbens (2002) and Kitamura (2006). The continuous updating GMM estimator has the attraction of not requiring estimation of Lagrange multipliers, but it does not always converge. More generally, the MEL and GEL estimators have an objective function that is less well-behaved than the quadratic form for regular GMM.

14.3.3 Simulation-based ML and MM estimation

Simulation-based estimation methods enable ML estimation in cases where the conditional density of y given $\mathbf{x}$ includes an integral for which a closed-form solution does not exist, so that conventional ML is not possible. The integral is approximated by Monte Carlo integration, making many draws from an appropriate distribution.

Specifically, let the conditional density of y given regressors $\mathbf{x}$ and parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ be an integral:

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \int f(y|\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}_1) g(\mathbf{u}|\boldsymbol{\theta}_2) d\mathbf{u}, \tag{14.10}$$

where $f(y|\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}_1)$, which depends in part on unobservables $\mathbf{u}$, is of closed form, but there is no closed form for the desired density $f(y|\mathbf{x}, \boldsymbol{\theta})$.

A leading example is unobserved heterogeneity. Then $\boldsymbol{\theta}_1$ denotes parameters of intrinsic interest, $\mathbf{u}$ denotes unobserved heterogeneity that may depend on unknown parameters $\boldsymbol{\theta}_2$, and the integral will not have a closed-form solution except in some special cases. A second example is the multinomial probit model. Then $\boldsymbol{\theta}_1$ denotes regression parameters, $\mathbf{u}$ denotes an error term in a latent model that may have unknown error variances and covariances $\boldsymbol{\theta}_2$, and, given m alternatives, the probability that a specific alternative is chosen is given by an $(m-1)$-dimensional integral that has no closed form solution.

If the integral is of low dimension, then numerical integration by Gaussian quadrature may provide a reasonable approximation to $f(y|\mathbf{x}, \boldsymbol{\theta})$. But these methods can work poorly in the higher dimensions often encountered in practice. For example, for multinomial probit Gaussian quadrature is felt to work poorly if there are more than four alternatives.

Instead, the maximum simulated likelihood (MSL) method makes many draws of the unobservables $\mathbf{u}$ from density $g(\mathbf{u}|\boldsymbol{\theta}_2)$. The MSL estimator maximizes the simulated log-likelihood function:

$$\widehat{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta}), \tag{14.11}$$

where $\widehat{f}(\cdot)$ is the Monte Carlo estimate or simulator:

$$\widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\widetilde{f}(y_i|\mathbf{x}_i, \mathbf{u}_i^s, \boldsymbol{\theta}), \tag{14.12}$$

where $\mathbf{u}_i^{(S)} = (\mathbf{u}_i^1, \ldots, \mathbf{u}_i^S)$ denotes S draws with marginal density $g(\mathbf{u}_i|\boldsymbol{\theta}_2)$, and $\widetilde{f}(\cdot)$ is a subsimulator such as $f(y|\mathbf{x}, \mathbf{u}^s, \boldsymbol{\theta}_1)$. Many possible simulators may be used – the essential requirement is that $\widehat{f}_i \xrightarrow{p} f_i$ as $S \to \infty$. The MSL estimator is consistent and asymptotically equivalent to the ML estimator, provided that $S \to \infty$, in addition to the usual assumption that $N \to \infty$, with $\sqrt{N}/S \to \infty$ so that S grows at a rate slower than N.

The MSL estimator opens up the possibility of using a much wider range of parametric models, such as richer models for unobserved heterogeneity that may be more robust to model misspecification. At the same time the method can be computationally demanding. An early application of MSL was by Lerman and Manski (1981) for the multinomial probit model. Then $I \times N \times S$ draws of $\mathbf{u}_i^s$ are made if analytical derivatives are used, where I is the number of iterations, and even more draws are needed if numerical derivatives are used. A currently popular application is to the random parameters logit or mixed logit model.

For models with unobserved heterogeneity, an alternative is to treat heterogeneity as being discretely distributed. Such finite mixture or latent class models are especially popular in the duration and count (number of health services) literatures (see Meyer, 1990; Deb and Trivedi, 2002). These models do not require simulation methods, and can be more easily estimated using quasi-Newton methods or the expectation maximization algorithm. Often a heterogeneity distribution with just two or three points of support is sufficient.

The MSL can be extended to MM and GMM estimation. In that case, theory leads to a moment condition $\mathrm{E}[m(y_i|\mathbf{x}_i, \boldsymbol{\theta})] = 0$, where $m(\cdot)$ is a scalar for simplicity, but there is no closed form expression for $m(y, \mathbf{x}, \boldsymbol{\theta})$. Instead $m(y, \mathbf{x}, \boldsymbol{\theta})$ is an integral:

$$m(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int h(y_i|\mathbf{x}_i, \mathbf{u}_i, \boldsymbol{\theta}_1)g(\mathbf{u}_i|\boldsymbol{\theta}_2)d\mathbf{u}_i, \tag{14.13}$$

for some functions $h(\cdot)$ and $g(\cdot)$, where $m(\cdot)$ has no closed form. Let $\widehat{m}_i = \widehat{m}(y_i|\mathbf{x}_i, \mathbf{u}_i^{(S)}, \boldsymbol{\theta})$ be a simulator for $m(y_i, \mathbf{x}_i, \boldsymbol{\theta})$. Then the method of simulated moments (MSM) estimator uses $\widehat{m}_i$ in place of m_i in GMM estimation. A key result, due to McFadden (1989) and Pakes and Pollard (1989), is that the MSM estimator is consistent for $\boldsymbol{\theta}$ as $N \to \infty$ even if S is very small, provided that an unbiased simulator is used, meaning $\mathrm{E}[\widehat{m}_i] = m_i$. Furthermore, small S may lead to little loss of precision. In the special case that $\widehat{m}(\cdot)$ is the frequency simulator, the MSM estimator has variance $(1 + (1/S))$ times that of the MM estimator.

There are several subtleties in the use of MSL and related estimators. Book references are Gouriéroux and Monfort (1996), who also discuss indirect inference, and Train (2003), who focuses on applications to multinomial choice. First, because the simulated likelihood is usually maximized by iterative gradient methods, the simulator $\widehat{f}_i$ should be differentiable (or smooth) in $\boldsymbol{\theta}$. For example, for limited

dependent variables models with normal errors the Geweke–Hajivassiliou–Keane (GHK) simulator is often used. Second, to enable convergence and avoid "chatter" the same underlying random numbers used to obtain $\mathbf{u}_i^S$ should be used at each iteration. Third, the draws from $g(\mathbf{u}_i|\boldsymbol{\theta}_2)$ need not be independent. For example, better approximations for given S may be obtained by using dependent quasi-random numbers, such as Halton sequences, rather than independent pseudo-random numbers, and by the use of antithetic sampling. Fourth, it may be difficult to make draws from $\mathbf{u}_i^S$ using standard methods such as inverse transformation and accept–reject methods. Newer Markov chain Monte Carlo methods, widely used in Bayesian analysis, may be then used.

14.3.4 Simulation-based Bayesian analysis

Bayesian analysis can serve two purposes. First, it can provide a quite different method of inference as it views parameters as random variables, with the goal being to combine the prior distribution on parameters and the sample distribution of the data to recover the posterior distribution of these parameters. By contrast, classical frequentist inference views parameters as taking fixed values that are unknown, with data used to make inference on those unknown values. Second, if priors are chosen to be sufficiently uninformative so that they have little effect on the posterior distribution, then it is possible to directly use the posterior distribution of parameters to perform classical frequentist inference.

Econometric applications most often use Bayesian methods for the second purpose. Recent advances in computational methods, outlined below, can make Bayesian methods especially useful in analytically intractable models with many parameters that may be very difficult to estimate using conventional ML or even simulated ML methods. Even so, Bayesian methods are used sparingly in econometrics when compared to the statistics literature. One reason is a hesitation to use fully parametric methods, though Bayesian methods do allow quite flexible parametric models to be specified.

Let $L\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}\right) = f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$ denote the sample joint density or likelihood, and $\pi(\boldsymbol{\theta})$ denote the prior distribution. Then by Bayes' rule the posterior density for $\boldsymbol{\theta}$ is:

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{L(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{X})}, \tag{14.14}$$

where $f(\mathbf{y}|\mathbf{X}) = \int_{R(\boldsymbol{\theta})} L(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $R(\boldsymbol{\theta})$ denotes the support of $\pi(\boldsymbol{\theta})$. Because the denominator $f(\mathbf{y}|\mathbf{X})$ is free of $\boldsymbol{\theta}$, it is standard simply to write:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{14.15}$$

where the regressors $\mathbf{X}$ are suppressed for notational simplicity. The posterior density is proportional to the product of the likelihood and prior.

The heart of Bayesian analysis is the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. In the simplest cases a closed form expression for this exists. For example, if $\mathbf{y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and known variance and the prior for $\boldsymbol{\beta}$ is the normal with specified mean and variance, then the posterior is normal.

But for most models, especially standard nonlinear regression models, the posterior is unknown. One approach is to then obtain key moments, such as the posterior mean $\mathrm{E}[\boldsymbol{\theta}] = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$, using Monte Carlo integration methods that do not require draws from $p(\boldsymbol{\theta}|\mathbf{y})$. In particular, importance sampling methods can be used (see Kloek and van Dijk, 1978; Geweke, 1989).

The more modern approach is instead to obtain many draws, say $\widehat{\boldsymbol{\theta}}^1, \ldots, \widehat{\boldsymbol{\theta}}^S$, from $p(\boldsymbol{\theta}|\mathbf{y})$. The posterior mean can then be estimated by $S^{-1} \sum_{s=1}^{S} \widehat{\boldsymbol{\theta}}^s$. Furthermore, the distribution of any quantity of interest, such as the distribution of marginal effects in a nonlinear model, can be similarly computed given these draws from the posterior.

The key ingredient is the recent development of methods to obtain draws of $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ even when $p(\boldsymbol{\theta}|\mathbf{y})$ is unknown (see Gelfland and Smith, 1990). The starting point is the Gibbs sampler. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ and suppose that it is possible to draw from the conditional posteriors $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$, even though it is not possible to draw from $p(\boldsymbol{\theta}|\mathbf{y})$. The Gibbs sampler obtains draws from $p(\boldsymbol{\theta}|\mathbf{y})$ by making alternating draws from each conditional distribution. Thus, given an initial value $\boldsymbol{\theta}_2^{(0)}$, obtain $\boldsymbol{\theta}_1^{(1)}$ by drawing from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(0)}, \mathbf{y})$, then $\boldsymbol{\theta}_2^{(1)}$ by drawing from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(1)}, \mathbf{y})$, then $\boldsymbol{\theta}_1^{(2)}$ by drawing from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(1)}, \mathbf{y})$, and so on. When repeated many times it can be shown that this process ultimately leads to draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y})$, even though in general $p(\boldsymbol{\theta}|\mathbf{y}) \neq p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}) \times p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$. The sampler is an example of a Markov chain Monte Carlo (MCMC) method. The term "Markov chain" is used because the procedure sets up a Markov chain for $\boldsymbol{\theta}$ whose stationary distribution can be shown to be the desired posterior $p(\boldsymbol{\theta}|\mathbf{y})$. The method extends immediately to more partitions for $\boldsymbol{\theta}$. For example, if $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2' \ \boldsymbol{\theta}_3']'$ then draws need to be made from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \mathbf{y})$, and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \mathbf{y})$ and $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{y})$.

In many applications some of the conditional posteriors are unknown, in which case MCMC methods other than the Gibbs sampler need to be used. A standard method is the Metropolis–Hastings (MH) algorithm, which uses a trial or jumping distribution. The Gibbs sampler can be shown to be an example of an MH algorithm, one with relatively fast convergence.

The MCMC methods in principle permit Bayesian analysis to be applied to a very wide range of models. In practice, there is an art to ensuring that the chain converges in a reasonable amount of computational time. The first B draws of $\boldsymbol{\theta}$ are discarded, where B is chosen to be large enough that the Markov chain has converged. The remaining S draws of $\boldsymbol{\theta}$ are then used. Various diagnostic methods exist to indicate convergence, although these do not guarantee convergence. MCMC methods yield correlated draws from $p(\boldsymbol{\theta}|\mathbf{y})$, rather than independent draws, but this correlation only effects the precision of posterior analysis and often the correlation is low. Many Bayesian models include both components with closed form solutions for the posterior and components that require the use of MCMC methods – the Gibbs sampler, if possible, and, failing that, the MH algorithm with hopefully a good choice of jumping distribution.

Bayesian methods are particularly attractive in models entailing latent variables, such as tobit models (see Chib, 1992, 2001), and multinomial probit models

(see McCulloch, Polson and Rossi, 2000). The data augmentation method then generates draws of unobserved values of the latent variables that can then be treated as observed values, greatly simplifying analysis. A recent application is that by Geweke, Gowrisankaran and Town (2003), who model mortality at 114 Los Angeles County hospitals allowing for the complication that better hospitals may attract more difficult cases, with difficulty depending in part on unobservables correlated with hospital mortality rates. Recent econometrics books are Koop (2003), Lancaster (2004), and Koop, Poirier and Tobias (2007).

14.3.5 Quantile regression

Quantiles, such as deciles and quartiles, are often used to summarize the distribution of variables such as income, earnings and wealth. Quantile regression is an extension to the regression case where, for example, interest may lie in the different response of earnings to education at different points of the conditional earnings distribution.

A leading example is the least absolute deviations (LAD) estimator that minimizes the sum of absolute residuals $\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$. This is a generalization of the median in the independent and identically distributed (i.i.d.) case, since with $\mathbf{x}_i'\boldsymbol{\beta} = \beta$ the resulting estimate of β is the sample median.

More generally, conditional quantiles other than the median may be estimated. The qth quantile regression estimator $\widehat{\boldsymbol{\beta}}_q$ minimizes over $\boldsymbol{\beta}_q$:

$$Q_N(\boldsymbol{\beta}_q) = \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}}^{N} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}}^{N} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|,$$

where the subscript q in $\boldsymbol{\beta}_q$ is needed as different choices of q estimate different values of $\boldsymbol{\beta}$. The special case $q = 0.5$ is the LAD estimator. The objective function is not differentiable, so linear programming methods are used rather than more familiar gradient methods. These enable relatively fast computation of $\widehat{\boldsymbol{\beta}}_q$. The quantile regression estimator is consistent and asymptotically normal. Estimation of the analytical asymptotic variance of $\widehat{\boldsymbol{\beta}}_q$ requires estimation of $f_{u_q}(0|\mathbf{x})$, the conditional density of the error term $u_q = y - \mathbf{x}'\boldsymbol{\beta}_q$ evaluated at $u_q = 0$. An easier method is to instead obtain bootstrap standard errors for $\widehat{\boldsymbol{\beta}}_q$ using a paired bootstrap.

Quantile regression was proposed by Koenker and Bassett (1978). Powell (1984, 1986) adapted the method to permit consistent estimation in censored linear regression models. With censoring the conditional median can be recovered without the strong distributional assumptions, such as normality, needed to recover the conditional mean. Buchinsky (1994) provided a much-cited application that documented recent US changes in the quantiles of the conditional wage distribution. Such analysis is now easily implemented as Stata includes a quantile regression command. Koenker and Hallock (2001) provide an early summary of applications.

Existing results specify all quantile regression functions to be linear. Angrist, Chernozhukov and Ferandez-Val (2006) provide interpretation of the quantile

regression if instead the unknown true quantile function is nonlinear, and provide the asymptotic distribution in that case. The situation is analogous to that for OLS and ML under model misspecification.

Quantile regression has recently become a very active area of research with extensions including instrumental variables estimation (see Chernozhukov and Hansen, 2005), and a richer range of models for censored data (see Honore and Hu, 2004). Koenker (2005) provides many results on quantile regression.

14.3.6 Nonparametric and semiparametric methods

Consider the regression model:

$$\mathrm{E}[y_i|\mathbf{x}_i] = m(\mathbf{x}_i), \tag{14.16}$$

where the function $m(\mathbf{x})$ is unspecified. Nonparametric regression provides a consistent estimate of $m(\mathbf{x})$. At the specific point $\mathbf{x} = \mathbf{x}_0$, $m(\mathbf{x}_0)$ can be estimated by taking a local weighted average of y_i over those observations with $\mathbf{x}_i$ in a neighborhood of $\mathbf{x}_0$. There are many variations on this approach, including kernel regression, nearest neighbors regression, local linear, local polynomial, Lowess, smoothing spline and series estimators. Less than N observations are effectively used at any point $\mathbf{x}_0$, because a local average is taken, so $\widehat{m}(\mathbf{x}_0) \xrightarrow{p} m(\mathbf{x}_0)$ at a rate less than the usual $N^{-1/2}$, although asymptotic normality still holds.

Fully nonparametric regression works best in practice when there is just a single regressor. Even then, empirical results vary greatly with the choice of bandwidth or window width that defines the size of the neighborhood. "Plug-in" estimates of the bandwidth that work well for density estimation often work poorly for regression. The standard method is to use leave-one-out cross-validation to select the bandwidth, but this method is by no means perfect.

There is no theoretical obstacle to using nonparametric regression when there are many regressors. But in practice nonparametric methods usually work poorly with more than very few regressors, due to a curse of dimensionality that arises because the local averages will be made over fewer observations. For example, if averaging is over 10 bins with one regressor then averaging may need to be over $10^2 = 100$ bins when there are two regressors. More formally, the optimal convergence rate using mean squared error as a criterion is $N^{-2/(\dim[\mathbf{x}]+4)}$, so the convergence rate decreases as $\dim[\mathbf{x}]$ increases. This problem is less severe when some regressors take only a few values, such as binary indicator variables. Racine and Li (2004) present results for kernel regression when some regressors are discrete and some are continuous.

The microeconometrics literature focuses on semiparametric methods that overcome the curse of dimensionality by partially parameterizing a model, so that there is a mix of parametric and nonparametric components. A very early example is the maximum score estimator for the binary choice model of Manski (1975).

Theoretically, a first step is to determine whether parameters are identified given only partial specification of the model. Ideally $\sqrt{N}$-consistent and asymptotically normal estimates of the parameters can be obtained. Furthermore, it is preferred

that they be fully efficient in that they attain semiparametric efficiency bounds (see Chamberlain, 1987; Newey, 1990; Severini and Tripathi, 2001) that are extensions of Cramer–Rao lower bounds or the Gauss–Markov theorem to semiparametric models.

There are many semiparametric models, and for each model there can be several different ways to obtain estimators. Partial linear and single index models are two leading examples that are also the building blocks for more general models.

The partial linear model specifies the conditional mean to be the usual linear regression function plus an unspecified nonlinear component, so:

$$\mathrm{E}[y_i|\mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \lambda(\mathbf{z}_i), \tag{14.17}$$

where the scalar function $\lambda(\cdot)$ is unspecified. An example is the estimation of a demand function for electricity, where $\mathbf{z}$ reflects time of day or weather indicators such as temperature. A second example is a sample selection model where $\lambda(\mathbf{z})$ is the expected value of a model error, conditional on the sample selection rule. In applications interest may lie in $\boldsymbol{\beta}$, $\lambda(\mathbf{z})$ or both. Various estimators for the partial linear model have been proposed. The differencing method proposed by Robinson (1988) estimates $\boldsymbol{\beta}$ by OLS regression of $(y_i - \widehat{m}_{yi})$ on $(\mathbf{x}_i - \widehat{\mathbf{m}}_{\mathbf{x}i})$, where $\widehat{m}_{yi}$ and $\widehat{\mathbf{m}}_{\mathbf{x}i}$ are predictions from nonparametric regression of, respectively, y and $\mathbf{x}$ on $\mathbf{z}$. Robinson used kernel estimates that may need to be oversmoothed. Other methods that additionally estimate $\lambda(z)$, at least for scalar z, include a generalization of the cubic smoothing spline estimator and using a series approximation for $\lambda(z)$.

The single index model specifies the conditional mean to be an unknown scalar function of a linear combination of the regressors, with:

$$\mathrm{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i'\boldsymbol{\beta}), \tag{14.18}$$

where the scalar function $g(\cdot)$ is unspecified and the parameters $\boldsymbol{\beta}$ are then only identified up to location and scale. An example is a binary choice model with $\Pr[y = 1|\mathbf{x}] = g(\mathbf{x}'\boldsymbol{\beta})$ where $g(\cdot)$ is unknown. The single index formulation is attractive as the marginal effect of a change in the jth regressor is $g'(\mathbf{x}_i'\boldsymbol{\beta})\beta_j$, so that the ratio of parameter estimates equals the ratio of marginal effects. Estimators for the single index model include an average derivative estimator, a density weighted average derivative estimator (see Powell, Stock and Stoker, 1989), and semiparametric least squares.

Microeconometricians have focused on semiparametric estimation for limited dependent variable models – binary choice with an unspecified function for the probabilities, censored regression and sample selection. Nonparametric and semiparametric methods are also used in the treatment effects literature detailed in section 14.5.1. The literature is vast. References include the applied study by Bellemare, Melenberg and van Soest (2002), and books by Pagan and Ullah (1999) and Li and Racine (2007). The latter book is accompanied by many routines in R for nonparametric and semiparametric regression.

14.4 Statistical inference

There have been considerable advances in statistical inference methods, even for the standard LS, IV and ML estimators (MLEs).

The most notable change is the use of robust statistical inference, notably robust standard error computation, that relies on distributional assumptions that are as weak as possible. Various model specification tests have been developed, such as the Hausman test and White's information matrix test. The bootstrap provides an alternative way to perform statistical inference that can be simpler than conventional asymptotic methods. Furthermore, the bootstrap can produce more accurate asymptotic approximations for test statistics that lead to tests with actual size close to nominal size in small samples.

14.4.1 Robust inference for Wald tests

Analysis begins with the cross-section case of independent observations, before moving to clustered observations, which includes short panels.

Consider an m-estimator $\widehat{\boldsymbol{\theta}}$ that maximizes with respect to $\boldsymbol{\theta}$ the objective function $Q_N(\boldsymbol{\theta}) = N^{-1}\sum_i q(y_i, \mathbf{x}_i, \boldsymbol{\theta})$. For ML estimation $q(\cdot)$ is the log-density, and for least squares estimation $q(\cdot)$ is minus the squared error (or a rescaling of this). The m-estimator solves the first-order conditions:

$$N^{-1}\sum\nolimits_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{14.19}$$

where $\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}) = \partial q(y_i, \mathbf{x}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Under suitable assumptions, notably that $\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ in the population, it can be shown that $\widehat{\boldsymbol{\theta}}$ is $\sqrt{N}-$consistent, with limit distribution:

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0'^{-1}], \tag{14.20}$$

where $\mathbf{A}_0 = \operatorname{plim} N^{-1}\sum_i \partial\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}'\Big|_{\boldsymbol{\theta}_0}$ and $\mathbf{B}_0 = \operatorname{plim} N^{-1}\sum_i \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)'$, and $\boldsymbol{\theta}_0$ is the value of $\boldsymbol{\theta}$ in the DGP.

Inference is based on $\widehat{\boldsymbol{\theta}}$ being asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and estimated asymptotic variance matrix of sandwich form:

$$\widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}] = N^{-1}\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}'^{-1}, \tag{14.21}$$

where $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are consistent estimates of $\mathbf{A}_0$ and $\mathbf{B}_0$. The Wald test statistic for $H_0 : \theta_j = r$ is then $W = (\widehat{\theta}_j - r)/s_j$, where s_j is the jth diagonal entry of $\widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}]$, and $W \overset{a}{\sim} \mathcal{N}[0, 1]$ under H_0. The more general hypothesis $H_0 : \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$ is tested using $W = \mathbf{c}(\widehat{\boldsymbol{\theta}})'(\widehat{\mathbf{R}}\widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}]\widehat{\mathbf{R}}')^{-1}\mathbf{c}(\widehat{\boldsymbol{\theta}})$, where $\widehat{\mathbf{R}} = \partial\mathbf{c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\Big|_{\widehat{\boldsymbol{\theta}}}$ and $W \overset{a}{\sim} \chi^2(rank[\mathbf{R}])$ under H_0.

There are several possible ways to form $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, depending in part on the strength of the distributional assumptions made. Robust variance estimates are those that rely on minimal distributional assumptions, provided $N \to \infty$.

Given data independent over i, the robust variance matrix estimate uses:

$$\widehat{\mathbf{A}} = N^{-1} \sum_i \left. \frac{\partial \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right|_{\widehat{\boldsymbol{\theta}}},$$
$$\widehat{\mathbf{B}} = N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})'. \tag{14.22}$$

The resulting standard errors are called robust standard errors. In some cases the Hessian $\widehat{\mathbf{A}}$ in (14.22) may be replaced by the expected Hessian, and $\widehat{\mathbf{B}}$ may use a degrees-of-freedom correction such as $(N-q)^{-1}$ rather than N^{-1}.

A leading example is the heteroskedastic-consistent estimate of the variance-covariance matrix of the OLS estimator. Then $q_i(\boldsymbol{\beta}) = -\frac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$, where the multiple $\frac{1}{2}$ is added for convenience, so that $\mathbf{h}_i(\boldsymbol{\beta}) = \partial q_i / \partial \boldsymbol{\beta} = (y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i$, and $\partial \mathbf{h}_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}' = -\mathbf{x}_i \mathbf{x}_i'$. It follows that:

$$\widehat{\mathbb{V}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] = \left[\sum_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_i \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'\right] \left[\sum_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1}, \tag{14.23}$$

where $\widehat{u}_i = (y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})$.

For ML estimation use of (14.22) relaxes the traditional information matrix equality assumption that $\mathbf{A}_0 = -\mathbf{B}_0$, which gives the simplification $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} = -\mathbf{A}_0^{-1}$. Failure of the information matrix equality generally implies inconsistency of the MLE. Then $\boldsymbol{\theta}_0$ needs to be reinterpreted as a "pseudo-true value," which is the value of $\boldsymbol{\theta}$ that maximizes the probability limit of $1/N$ times the log-likelihood function. However, the MLE does retain consistency in many standard models with specified density in the linear exponential family, notably linear, Poisson, logit and probit, provided the conditional mean function is correctly specified. Robust standard errors are then especially applicable.

For independent errors the key early reference is White (1980), who proposed the special case (14.23). Robust standard errors have been applied to many estimators, including instrumental variables and generalized method of moments (see (14.3)). T. Amemiya (1985) and Newey and McFadden (1994) provide quite general treatments of inference and estimation (see also White, 1984).

The estimates in (14.22) can be extended to clustered data. In that case observations are grouped into clusters, with correlation permitted within a cluster but independence assumed across clusters. An example is panel data where the cluster unit is the individual: observations for a given individual over time are correlated, but observations across individuals are independent. Let $c = 1, \ldots, C$ denote clusters and let $j = 1, \ldots, N_c$ denote the N_c observations in cluster c. Then the cluster-robust variance matrix estimate is (14.21), where $\widehat{\mathbf{A}}$ is again given in (14.22) but now:

$$\widehat{\mathbf{B}} = N^{-1} \sum_{c=1}^{C} \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} \mathbf{h}(\mathbf{w}_{jc}, \widehat{\boldsymbol{\theta}}) \mathbf{h}(\mathbf{w}_{kc}, \widehat{\boldsymbol{\theta}})'. \tag{14.24}$$

This estimator, proposed by Liang and Zeger (1986), permits both error heteroskedasticity and quite flexible error correlation within cluster. It has largely supplanted the use of a more restrictive random effects or error components model,

although it does require $C \to \infty$. It is essential to control for clustering, as failure to do so can lead to greatly under estimated standard errors (see Moulton, 1990). With few clusters the asymptotic normal distribution can perform poorly. Small sample corrections include using the T_{C-1} distribution, using a cluster bootstrap with asymptotic refinement (Cameron, Gelbach and Miller, 2008), and using an alternative estimator that, under some assumptions, is exactly T_{C-2} distributed as $N_c \to \infty$ (Donald and Lang, 2007). Wooldridge (2003) provides a survey that is updated and newer results given in Wooldridge (2006). Cameron, Gelbach and Miller (2006) propose an extension of (14.24) to multi-way clustering.

Cluster-robust variance matrix estimators have also been proposed for models where $\mathbf{h}_i(\boldsymbol{\theta})$ is spatially correlated. Driscoll and Kraay (1998) do so for panel data where the time dimension is large. Conley (1999) presents a quite general estimator.

Models for clustered or spatial data may allow the conditional mean to be affected, in addition to conditional correlations, and then standard estimators become inconsistent. For clustered data one can use a model with cluster-specific fixed effects, analogous to fixed effects for panel data. For spatial data recent references include Anselin (2001) and Lee (2004).

The theory for robust inference is well-established and, in the independent observations case at least, is well incorporated into microeconometrics practice. In particular, for LS problems it is standard to estimate by OLS and then use robust standard errors, even though there may be efficiency loss compared to doing feasible generalized least squares (GLS). Note, however, that one can still employ feasible GLS but then compute robust standard errors that guard against misspecification of the model for the error variance matrix.

14.4.2 Hypothesis tests

For hypotheses on parameters of the form:

$$H_0 : \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$$
$$H_a : \mathbf{c}(\boldsymbol{\theta}) \neq \mathbf{0},$$

the classical tests in the likelihood framework are the Wald, Lagrange multiplier (LM), and likelihood ratio tests. For a correctly specified likelihood function these tests are first-order asymptotically equivalent under the null hypothesis and under local alternatives, so choice between them is one of convenience.

More recent work has focused on generalization to the non-likelihood framework, and on finite-sample properties of the tests.

The Wald test has become the most popular of these three tests, as it generalizes easily to non-likelihood models and is most easily robustified as detailed in section 14.4.1. But it does have the limitation of lack of invariance to parameterization. For example, a test of $H_0 : \theta_1/\theta_2 = 1$ will lead in finite samples to a Wald test statistic that differs from that for the equivalent hypothesis $H_0 : \theta_1 - \theta_2 = 0$. A bootstrap with asymptotic refinement (see section 14.4.4), should reduce this invariance.

The LM or score test is less commonly used, in part because it is usually implemented by a convenient auxiliary regression that has poor finite sample properties.

Specifically, Monte Carlo studies find considerable over-rejection due to the finite sample test size being considerably larger than the asymptotic size. A bootstrap with asymptotic refinement, however, can correct this problem. The LM test can be extended to non-likelihood settings, and can be robustified.

The likelihood ratio test generally does not extend to non-likelihood settings, though it does for optimal GMM estimation. Newey and West (1987) generalize the three classical tests from the likelihood framework to the GMM framework.

14.4.3 Model specification tests

Various model specification tests have been developed that do not rely on hypothesis tests of the form $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$.

The Hausman (1978) test contrasts two estimators that may be the same under a null hypothesis and differ under an alternative hypothesis. For example, one can compare OLS to the 2SLS estimator and conclude that there is endogeneity if the two estimators differ. Denote the two estimators by $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, and test H_0 : plim$(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) = \mathbf{0}$ using the statistic:

$$H = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'(\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}])^{-1}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}],$$

which is chi-squared distributed under H_0. Implementation requires estimating the variance matrix of the difference in the estimators. The original approach was to assume that one estimator, say $\widehat{\boldsymbol{\theta}}$, is efficient under the null, in which case $\mathrm{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}] = \mathrm{V}[\widetilde{\boldsymbol{\theta}}] - \mathrm{V}[\widehat{\boldsymbol{\theta}}]$. This is the standard method used today, even though it is generally incorrect since, from section 14.4.1, most applied studies use heteroskedastic-robust or cluster-robust standard errors that presume the estimator is, in fact, inefficient. One should instead use alternative methods to estimate $\mathrm{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}]$, such as the bootstrap.

Moment tests are tests of whether or not a population moment condition is supported by the data. Specifically, they test:

$$H_0 : \mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$$
$$H_a : \mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] \neq \mathbf{0}.$$

An obvious test is based on whether the corresponding sample moment $\widehat{\mathbf{m}} = N^{-1} \sum_i \mathbf{m}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$ is close to zero. The test statistic is:

$$M = \widehat{\mathbf{m}}'(\widehat{\mathrm{V}}[\widehat{\mathbf{m}}])^{-1}\widehat{\mathbf{m}},$$

where M is chi-squared distributed under H_0 and the challenge is to estimate $\widehat{\mathrm{V}}[\widehat{\mathbf{m}}]$.

One leading example is an overidentifying restrictions (OIR) test. Then GMM estimation based on $\mathrm{E}[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ cannot exactly impose $\widehat{\mathbf{m}} = \mathbf{0}$ if the model is overidentified. If GMM with an optimal weighting matrix is used then Hansen (1982) showed that M is chi-squared distributed under H_0 with degrees of freedom equal to the number of overidentifying restrictions.

A second class of examples are conditional moment tests, where some model restrictions are used in estimation while other restrictions, not imposed in

estimation, are used for specification testing. For example, in linear regression of y on $\mathbf{x}_1$, the hypothesis that $\mathbf{x}_2$ can be excluded as a regressor implies $\mathrm{E}[(y - \mathbf{x}_1'\boldsymbol{\beta}_1)|\mathbf{x}_2] = 0$. This can be specified as a test of $H_0 : \mathrm{E}[(y - \mathbf{x}_1'\boldsymbol{\beta}_1)\mathbf{x}_2] = \mathbf{0}$. Here it can be difficult to obtain $\widehat{\mathrm{V}}[\widehat{\mathbf{m}}]$, though auxiliary regressions are available to compute an asymptotically equivalent version of M in the special case that $\widehat{\boldsymbol{\theta}}$ is the MLE. Examples of conditional moment tests, proposed by Newey (1985) and Tauchen (1985), include the information matrix test of White (1982) and chi-squared goodness-of-fit tests.

The Hausman test and OIR tests are routinely used in GMM applications. Conditional moment tests are less commonly used, even though they are easy to implement in likelihood settings, where they would seem especially useful due to concerns of reliance on distributional assumptions. One reason is that the convenient auxiliary regressions used to compute them can have poor finite-sample size properties, but this can be rectified by a bootstrap with asymptotic refinement (see, for example, Horowitz, 1994). A second reason is the more practical one that, especially with large samples, any model is quite likely to be rejected at conventional 5% significance levels.

For model selection when models are nested the standard hypothesis testing methods can be used. For model selection with non-nested models there is an extensive literature that is not addressed here. A recent survey is provided by Pesaran and Weeks (2001).

14.4.4 Bootstrap

Inference in microeconometrics is based on asymptotic results that provide only an approximation given typical sample sizes. The bootstrap, introduced by Efron (1979), provides an alternative approximation by Monte Carlo simulation.

The motivation of the bootstrap is to view the data in hand, or the fitted DGP, as the population. Then draw B resamples from this population, and for each resample compute a relevant statistic. The empirical distribution of the resulting B statistics is used to approximate the distribution of the original statistic.

The most common use of the bootstrap is as a way to calculate standard errors. The data $\mathbf{w}_1, \ldots, \mathbf{w}_N$ are assumed to be i.i.d. The bootstrap standard error procedure is:

1. Do the following B times:
   - Draw a bootstrap resample $\mathbf{w}_1^*, \ldots, \mathbf{w}_N^*$ by sampling with replacement from the original data (called a paired bootstrap).
   - Obtain estimate $\widehat{\theta}^*$ of θ, where for simplicity θ is scalar.
2. Use the B estimates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ to approximate the distribution of $\widehat{\theta}$. In particular, the bootstrap estimate of the standard error of $\widehat{\theta}$ is:

$$s_{\widehat{\theta},\mathrm{Boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \bar{\widehat{\theta}}^*)^2}, \tag{14.25}$$

where $\bar{\widehat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_b^*$. This is simply the standard deviation of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$.

This method is convenient whenever standard errors are difficult to obtain by conventional methods. Leading examples are (i) two-step estimators when estimation at the first step complicates inference at the second step; (ii) Hausman tests that require computation of the variance of the difference between two estimators when neither estimator is efficient under the null hypothesis; and (iii) estimation with clustered errors when a package does not compute cluster-robust standard errors (in this case a cluster bootstrap that resamples over clusters is used). Given bootstrap standard errors, a standard Wald test of $H_0 : \theta = \theta_0$ uses $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta},\text{Boot}}$ and asymptotic normal critical values.

The preceding bootstrap is theoretically no better than usual first-order asymptotic theory. The attraction is the practical one of convenience.

Some bootstraps, however, provide a better asymptotic approximation, called an asymptotic refinement. The econometrics literature focuses on asymptotic refinement for test statistics. Consider a test of $H_0 : \theta = \theta_0$ with nominal significance level or nominal size α. An asymptotic approximation yields an actual rejection rate or true size $\alpha + O(N^{-j})$, where $O(N^{-j})$ means is of order N^{-j} and $j > 0$, with often $j = 1/2$ or $j = 1$. Then the true size goes to α as $N \to \infty$. Larger j is preferred, however, as then convergence to α is faster. A method with asymptotic refinement (or higher-order asymptotics) is one that yields j larger than that obtained using conventional asymptotics. The hope is that such asymptotic refinement will lead to tests with true size closer to α for moderate sample sizes, though this is not guaranteed. Asymptotic refinement may be possible if the bootstrap is applied to an asymptotically pivotal statistic, meaning one with asymptotic distribution that does not depend on unknown parameters.

The bootstrap standard error procedure does not lead to asymptotic refinement for the Wald test. Nor does the percentile method which rejects $H_0 : \theta = \theta_0$ if θ_0 falls outside the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$. The problem is that $\widehat{\theta}$ is bootstrapped and $\widehat{\theta}$ is not asymptotically pivotal, since even under H_0 its asymptotic normal distribution depends on an unknown parameter (the variance).

Instead, the Wald statistic itself should be bootstrapped, as $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$ is asymptotically pivotal, since it is asymptotically $\mathcal{N}[0, 1]$ under H_0, an asymptotic distribution with no unknown parameters. The bootstrap-t or percentile-t procedure for a two-sided test of $H_0 : \theta = \theta_0$ at level α is:

1. Do the following B times:

   - Draw a bootstrap resample $\mathbf{w}_1^*, \ldots, \mathbf{w}_N^*$ by sampling with replacement from the original data (called a paired bootstrap).
   - Obtain an estimate $\widehat{\theta}^*$, standard error $s_{\widehat{\theta}^*}$ and t-statistic $t^* = (\widehat{\theta}^* - \widehat{\theta})/s_{\widehat{\theta}^*}$.

2. Use the B statistics $t_1^*, \ldots, t_B^*$ to approximate the distribution of $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$. For an equal-tailed (or nonsymmetric) test reject H_0 if the original sample t-statistic falls outside the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates

$t_1^*, \ldots, t_B^*$. For a symmetric test reject H_0 if the original sample t-statistic falls outside the α quantile of $|t_1^*|, \ldots, |t_B^*|$.

Note that t^* in step 1 is centered on $\widehat{\theta}$ as the bootstrap views the original sample, with $\theta = \widehat{\theta}$, as the DGP. For equal-tailed two-sided tests (or for one-sided tests) this procedure leads to asymptotic refinement with true size $\alpha + O(N^{-1})$, rather than $\alpha + O(N^{-0.5})$, using bootstrap standard errors (or standard errors obtained using equation (14.21)). For a two-sided symmetrical test (or a chi-squared test) the corresponding rates are instead, respectively, $\alpha + O(N^{-1.5})$ and $\alpha + O(N^{-1})$.

There are as many ways to bootstrap as there are different ways to obtain resamples, and there are many ways to use these resamples.

The resampling method used above is called a paired bootstrap, as often $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ and here both y_i and $\mathbf{x}_i$ are being resampled. By contrast, a residual bootstrap, for a model with additive error, holds $\mathbf{x}_i$ fixed and resamples over residuals $\widehat{u}_1, \ldots, \widehat{u}_N$ to yield resampled values $\mathbf{w}_i^* = (y_i^*, \mathbf{x}_i)$, where $y_i^* = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{u}_i^*$. A parametric bootstrap uses distributional knowledge, such as a specified distribution for $y_i|\mathbf{x}_i$, to resample y_i given $\mathbf{x}_i$. For clustered data any resampling is over clusters. For hypothesis tests it is best, if possible, to impose H_0 in drawing the bootstrap sample. More bootstrap replications are needed when the goal of the bootstrap is asymptotic refinement for a test statistic.

Much development of the bootstrap has been done in the statistics literature. The econometrics literature is surveyed in Horowitz (2001), and MacKinnon (2002) provides much useful practical advice. Econometric studies have focused on bootstraps for estimation methods used mainly by econometricians. For overidentified GMM models one should recenter so that the population moment condition is imposed in the sample; see Hall and Horowitz (1996).

The bootstrap needs to be used with caution, as standard bootstraps can provide inconsistent standard error estimates for nonsmooth estimators and for less than $\sqrt{N}$-consistent estimators. This has led to a currently active literature. Abrevaya and Huang (2005) consider the maximum score estimator, Abadie and Imbens (2008) consider matching treatment effects estimators, and Moreira, Porter and Suarez (2004) consider IV with weak instruments. Sub-sampling, due to Politis and Romano (1994), works in a wider range of settings than the bootstrap.

In applied microeconometrics the main use of the bootstrap is to obtain standard errors. Bootstraps with asymptotic refinement are rarely done, as sample sizes are felt to be fairly large. But a bootstrap with asymptotic refinement can correct for many well-documented problems associated with standard tests, including the lack of invariance to parameterization for the Wald test and the poor finite-sample performance of auxiliary regressions used in computing LM tests and conditional moment tests.

14.5 Causation

The preceding sections presented estimation and inference methods for quite general regression models. Econometrics is distinguished by a desire to go beyond

correlative data summary to obtaining estimates of a causative effect, meaning measures of how an outcome changes in response to exogenous changes in a regressor. The current microeconometrics toolkit contains many methods to do so.

The "treatment effects" or "natural experiment" approach seeks to measure causation by extending randomized experiment methods to observational data. This major innovation in microeconometrics research uses a potential outcomes notation that differs from the simultaneous equations framework developed at the Cowles Commission. Developments have also been made in other more traditional methods to tease out causation, notably instrumental variables estimation, use of panel data, and estimation of structural models.

14.5.1 Treatment effects

The treatment effects literature focuses on the simplest case of estimating the causal effect of a binary regressor. A stereotypical example is to consider the impact on earnings of participation in a training program. The terminology of a medical trial is used. Enrollment in a training program is viewed as treatment, having no training is viewed as control, and the objective is to estimate the causative effect of the treatment on the outcome variable, earnings.

The ideal way to calculate this effect is to observe earnings for a person with the training, observe earnings for the same person without training, and subtract. But this is impossible. Instead the outcome is observed in only one state, while the other state is a hypothetical unobserved value, called a potential outcome or counterfactual.

The randomized experiment approach solves the inability to observe the counterfactual by comparing average outcomes, rather than individual outcomes, for two groups that are randomly assigned to either treatment or control. This approach is used at times in the social sciences, in social experiments. But most economics studies must instead rely on observational data.

The treatment effects literature seeks to extend the experimental approach to nonrandomized settings. Again averages across groups are compared, but now individuals select their treatment. Different assumptions about the nature of the self-selection of treatment and data availability lead to different methods to compute average effects of treatment. A key consideration is whether or not it is reasonable to assume that self-selection can be controlled for using observed variables, or whether self-selection additionally depends on unobservables. The latter case requires much stronger assumptions to make progress.

The following framework is used. The binary treatment variable d takes value 1 if treatment is assigned and value 0 if untreated (a control). The observed outcome of interest y is a continuous variable that then takes values:

$$y_i = \begin{cases} y_{1i} & \text{if treated } (d_i = 1) \\ y_{0i} & \text{if control } (d_i = 0). \end{cases} \tag{14.26}$$

The individual treatment effect is defined to be:

$$\alpha_i = (y_{1i} - y_{0i}). \tag{14.27}$$

Note that (14.26) and (14.27) imply:

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i} = y_{0i} + \alpha_i d_i. \tag{14.28}$$

Since only one of y_{1i} and y_{0i} are observed, α_i is not observable. Instead, the goal is to estimate population averages of α_i, notably the average treatment effect (ATE):

$$\alpha_{\text{ATE}} = \text{E}[\alpha_i], \tag{14.29}$$

and the average treatment effect on the treated (ATET):

$$\alpha_{\text{ATET}} = \text{E}[\alpha_i | d_i = 1]. \tag{14.30}$$

These are conceptually quite different quantities. ATET gives the average gain in earnings for a person who actually receives training. ATE gives the earnings gain averaged across those who did and those who did not receive the training.

The evaluation problem can be illustrated by decomposing ATET into two terms as:

$$\alpha_{\text{ATET}} = \{\text{E}[y_{1i} | d_i = 1] - \text{E}[y_{0i} | d_i = 0]\} - \{\text{E}[y_{0i} | d_i = 1] - \text{E}[y_{0i} | d_i = 0]\}. \tag{14.31}$$

A naive estimate of α_{ATET} uses just the first term. But this ignores the second term, a selection term that arises if the treated and untreated are different in that, on average, they would have different untreated outcome. Methods differ according to whether this selection term can be solely controlled for by regressors, or whether it additionally depends on unobservables.

Given regressors $\mathbf{x}$, similar average effects can be defined, now varying with regressors. The ATE is:

$$\alpha_{\text{ATE}}(\mathbf{x}) = \text{E}[\alpha_i | \mathbf{X}_i = \mathbf{x}]. \tag{14.32}$$

and the ATET is:

$$\alpha_{\text{ATET}}(\mathbf{x}) = \text{E}[\alpha_i | \mathbf{X}_i = \mathbf{x}, d_i = 1]. \tag{14.33}$$

Treatment effects are called heterogeneous if these quantities vary with the evaluation point $\mathbf{x}$, and are called homogeneous if $\alpha_{\text{ATE}}(\mathbf{x}) = \alpha_{\text{ATET}}(\mathbf{x}) = \alpha$. In practice, researchers usually report estimates of the population measures $\alpha_{\text{ATE}} = \text{E}[\alpha_{\text{ATE}}(\mathbf{x})]$ and $\alpha_{\text{ATET}} = \text{E}[\alpha_{\text{ATET}}(\mathbf{x})]$ that average across individuals with different characteristics. For example, individual-level estimates of $\widehat{\alpha}_{\text{ATE}}(\mathbf{x}_i)$ lead to $\widehat{\alpha}_{\text{ATE}} = N^{-1} \sum_{i=1}^{N} \widehat{\alpha}_{\text{ATE}}(\mathbf{x}_i)$.

A critical simplifying assumption, discussed further below, is that selection is on observables only. Assuming conditional independence, outcomes are independent of treatment after conditioning on regressors, so that:

$$f(y_{ji} | \mathbf{x}_i, d_i = 1) = f(y_{ji} | \mathbf{x}_i, d_i = 0) = f(y_{ji} | \mathbf{x}_i), j = 0, 1. \tag{14.34}$$

This assumption of exogenous selection of treatment (given $\mathbf{x}$) is often written as $y_{0i}, y_{1i} \perp d_i | \mathbf{x}_i$ and has several other names, including unconfoundedness and ignorability. For some purposes it can be weakened to apply to only y_{0i} or to

apply only to conditional means (and not the entire distribution). The assumption implies that:

$$\alpha_{\text{ATET}}(\mathbf{x}) = \text{E}[y_{1i}|\mathbf{X}_i = \mathbf{x}, d_i = 1] - \text{E}[y_{0i}|\mathbf{X}_i = \mathbf{x}, d_i = 0], \tag{14.35}$$

where the second-term conditions on $d_i = 0$, rather than $d_i = 1$ as in the original definition (14.33).

The matching approach for estimating treatment effects is based on (14.35) and compares sample averages of y_1 and y_0 for individuals with the same level of $\mathbf{x}$. This permits treatment effects to be heterogeneous and provides nonparametric estimates of their average. In practice, however, such estimates become noisy or impossible as $\mathbf{x}$ will take many values if it is continuous or high dimensional. One can instead use nonparametric methods, such as kernel weighting, that permit use of individuals with similar but not exactly the same level of $\mathbf{x}$. But more common is to match on the probability of treatment conditional on $\mathbf{x}$, known as the propensity score:

$$p(\mathbf{x}_i) = \Pr[d_i = 1|\mathbf{x}_i], \tag{14.36}$$

since Rosenbaum and Rubin (1983) showed that the conditional independence assumption carries over to conditioning on the propensity score (that is, $y_{0i}, y_{1i} \perp d_i \mid p(\mathbf{x}_i)$). For example, nearest-neighbor propensity score matching uses:

$$\widehat{\alpha}_{\text{ATET}} = N_1^{-1} \sum_{i:d_i=1} (y_{1i} - y_{0j}),$$

where $N_1 = \sum_{i=1}^{N} d_i$ and y_{0j} is the outcome for the nearest neighbor, the untreated observation with propensity score closest to that for y_{1i}. Other propensity score matching methods included kernel and stratification methods that average over several outcomes with similar propensity score. By estimating the propensity score using a flexible model, such as a semiparametric binary model or a logit model with interactions, it is more likely that observables only may determine selection. The propensity scores must have suitable common support over treatment and controls in order for matching to be feasible. For ATET it must be that $p(\mathbf{x}_i) < 1$, that is, for any value of the regressors it is possible to not receive treatment, while for ATE the requirement is that $0 < p(\mathbf{x}_i) < 1$. Note that if treatment effects are heterogeneous and matching is valid, the estimates obtained are very problem specific and not necessarily generalizable to other settings.

An alternative method is to specify and estimate a more restrictive regression model for the outcome. An obvious model is:

$$y_i = \alpha d_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{14.37}$$

which imposes the constraint that the treatment effect α is homogeneous. OLS estimation of (14.37) yields a consistent estimate of the treatment effect α, assuming conditional independence and that (14.37) is correctly specified. This is called the control function approach, as the regressors $\mathbf{x}$ here include regressors that control for selection into treatment (that is, explain d) as well as regressors that directly

explain y in the absence of treatment. This more parametric method has the advantage over matching of not requiring common support for the propensity score and permitting extrapolation beyond just the sample at hand.

The preceding methods rely on the untestable assumption of conditional independence, and presume that the dataset is rich with many control variables, since observables alone are assumed sufficient to control for treatment selection. Should these conditions fail, which will be the case in many potential applications, the previous methods are invalid. For example, the OLS estimator in the simple homogeneous effects model (14.37) is inconsistent if the treatment indicator variable is correlated with the error term even after conditioning on regressors $\mathbf{x}$. It is then necessary to allow for treatment to additionally depend on unobservable individual heterogeneity, with different methods used to control or eliminate the unobservables.

Panel data fixed effects estimators, possible if data are available for more than one period, control for unobserved heterogeneity by assuming that only the time-invariant component is correlated with treatment. A panel data version of the homogeneous effects model (14.37) is:

$$y_{it} = \alpha d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \phi_i + \delta_t + \varepsilon_{it}, \tag{14.38}$$

where here $\mathbf{x}_{it}$ does not include a constant and the intercept has both an individual-specific component ϕ_i and a time-specific component δ_t. Assume that treatment d_{it} is correlated with the unobservable ϕ_i, so OLS is inconsistent, but is uncorrelated with u_{it}. Then α can be consistently estimated by OLS estimation of the first differences model:

$$\Delta y_{it} = \alpha \Delta d_{it} + \Delta \delta_t + \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta \varepsilon_{it}, \tag{14.39}$$

or by estimation of a mean differences model (the fixed effects estimator), since ϕ_i has been eliminated. This standard method presumes panel data are available and is restricted to homogeneous treatment effects.

The related differences-in-differences method is applicable to repeated cross-sections, as well as panel data. For simplicity, suppose there are just two periods, say $t = a$ (after) and $t = b$ (before), and that all individuals are untreated in the first period and some are treated in the second period. Let $\bar{y}_{jt}$ denote the average outcome for treatment group $j = 0, 1$ in period $t = a, b$. The outcome changes over time by $(\bar{y}_{1a} - \bar{y}_{1b})$ in the treated group and by $(\bar{y}_{0a} - \bar{y}_{0b})$ in the untreated group. The differences in these differences provides an estimate of ATET, called the differences-in-differences estimator. This estimator is the OLS estimator of α in the model:

$$y_{it} = \gamma + \alpha d_i + \beta e_t + u_{it}, \ t = a, b,$$

where d_i is a binary treatment indicator and e_t is a binary time period indicator. Consistency of this estimator requires strong assumptions regarding the role of unobservables. In terms of (14.38) it is assumed that treatment selection does not depend on ε_{it} and that, while it may depend on ϕ_i, on average $\text{plim}(\bar{\phi}_{ja} - \bar{\phi}_{jb}) = 0$. The method can be extended to estimate heterogeneous effects $\alpha_{\text{ATET}}(\mathbf{x})$ by

grouping on $\mathbf{x}$ and then calculating within each group the four relevant averages of y.

Sample selection models explicitly specify a distribution for the unobservables. These introduce a latent variable to explain treatment choice, where the latent variable includes an unobserved component (or error) that is correlated with the error in the outcome equation. A linear model that permits heterogeneous effects and selection on unobservables is:

$$\begin{aligned} y_{1i} &= \mathbf{x}_i'\boldsymbol{\beta}_1 + u_{1i} \\ y_{0i} &= \mathbf{x}_i'\boldsymbol{\beta}_0 + u_{0i} \\ d_i^* &= \mathbf{z}_i'\boldsymbol{\gamma} + v_i, \end{aligned} \tag{14.40}$$

where $d_i = 1$ if the latent variable $d_i^* > 0$, and $d_i = 0$ otherwise. A homogeneous effects version restricts $y_{1i} = y_{0i}$ aside from a difference of α in the intercept. Under the assumption that (u_{0i}, u_{1i}, v_i) are joint normal (with $\sigma_v^2 = 1$), some algebra yields:

$$\begin{aligned} \mathrm{E}[y_{1i}|\mathbf{x}, d^* > 0] &= \mathbf{x}_i'\boldsymbol{\beta}_1 + \sigma_{1v}\lambda(\mathbf{z}_i'\boldsymbol{\gamma}), \\ \mathrm{E}[y_{0i}|\mathbf{x}, d^* \leq 0] &= \mathbf{x}_i'\boldsymbol{\beta}_0 - \sigma_{0v}\lambda(-\mathbf{z}_i'\boldsymbol{\gamma}), \end{aligned} \tag{14.41}$$

where $\lambda(\mathbf{z}'\boldsymbol{\gamma}) = \phi(\mathbf{z}'\boldsymbol{\gamma})/\Phi(\mathbf{z}'\boldsymbol{\gamma})$ is an inverse Mills ratio term, with $\phi()$ and $\Phi()$ denoting the standard normal density and distribution functions, and $\sigma_{jv} = \mathrm{Cov}[u_{ji}, v_i]$. From (14.41) consistent estimates of $\boldsymbol{\beta}_1$ and σ_{1v} can be obtained by OLS estimation for the treated sample of y_1 on $\mathbf{x}$ and $\lambda(\mathbf{z}'\widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\gamma}}$ is obtained by probit regression of d on $\mathbf{z}$. Similarly, OLS regression for the untreated sample of y_0 on $\mathbf{x}$ and $-\lambda(-\mathbf{z}'\widehat{\boldsymbol{\gamma}})$ gives consistent estimates of $\boldsymbol{\beta}_0$ and σ_{0v}. These estimates can then be used to estimate:

$$\alpha_{\mathrm{ATET}}(\mathbf{x}, \mathbf{z}) = \mathbf{x}_i'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\sigma_{0v} - \sigma_{1v})\lambda(\mathbf{z}_i'\boldsymbol{\gamma}).$$

The fundamental weakness of this sample selection approach is its reliance on distributional assumptions. These assumptions can be modified and relaxed, but even then the assumptions are still felt to be too strong.

Yet another method for control for selection on unobservables is instrumental variables estimation. Returning to the homogeneous effects model (14.37), the problem is that the regressor d is correlated with the error u. Assuming there is an instrument z that does not belong in the model, so $\mathrm{E}[u|\mathbf{x}, z] = 0$, but is correlated with the treatment indicator d, the treatment effect α can be consistently estimated by IV regression of y on $\mathbf{x}$ and d with instruments $\mathbf{x}$ and z.

A related method is the local average treatment effect (LATE) estimator. Begin with the homogeneous effects model with dependence on $\mathbf{x}$ dropped for simplicity, so that:

$$y_i = \beta + \alpha d_i + u_i. \tag{14.42}$$

Assume there is an instrument z with $\mathrm{E}[u|z] = 0$ and define $p(z) = \Pr[d = 1|Z = z] = \mathrm{E}[d|Z = z]$. Then:

$$\mathrm{E}[y|z] = \beta + \alpha p(z).$$

Evaluating at two points z and z' and subtracting yields the LATE:

$$\alpha_{\text{LATE}}(z) = \frac{\text{E}[y|z] - \text{E}[y|z']}{p(z) - p(z')}. \tag{14.43}$$

This can be estimated by comparing averages of the outcome y and treatment indicator d at two different values of the instrument z. If z is binary then this estimate is the same as the IV estimate. The estimate can be extended to heterogeneous effects, provided $p(z)$ is monotonic in z. Then it differs from IV and will vary with the points of evaluation z and z'. A more general treatment effect is the marginal treatment effect (MTE):

$$\alpha_{\text{MTE}}(\mathbf{x}, z) = \left. \frac{\partial \text{E}[y|\mathbf{x}, Z]}{\partial \Pr[d = 1|\mathbf{x}, Z]} \right|_{Z=z},$$

which gives the mean treatment effect for those at the margin of choosing treatment. ATE, ATET and LATE can be shown to be different weighted averages of MTE.

A final method is regression discontinuity (RD) design. Suppose treatment occurs when a variable s crosses a threshold $\bar{s}$, so that $d = 1(s > \bar{s})$, and the outcome y also depends on s. For example, a government program to improve school outcomes may be applied to schools in low-income areas. A method is developed to calculate a score, and schools with a score below a certain threshold receive the government program while those with a higher score do not. A complication is that school outcome will directly depend on this score. The obvious approach is to compare y for those with s just less than $\bar{s}$ to those with s just greater than $\bar{s}$, but this will use only a small fraction of the data. Instead use $\widehat{\alpha}$ from the least squares regression:

$$y_i = \beta + \alpha d_i + \gamma h(s_i) + u_i, \tag{14.44}$$

where $h(\cdot)$ is a flexible function that is specified (for example, polynomial) or is estimated by nonparametric methods. Given the discrete nature of the discontinuity at $\bar{s}$ it is clear that the method can also be used when effects are heterogeneous and will estimate ATE $= \text{E}[\alpha_i|s_i]$ under mild additional assumptions. Another extension is to fuzzy designs where the threshold $\bar{s}$ is not sharp, as some individuals with $s < \bar{s}$ are treated and some with $s > \bar{s}$ are untreated. Intuitively, if a fraction f of the population in the immediate vicinity of $\bar{s}$ switch from untreated to treated then ATE is estimated by f times the estimated OLS coefficient of d in (14.44). This adaptation is qualitatively similar to that for LATE in (14.43).

The literature on treatment effects is vast. Econometricians have contributed to the literature on all the preceding methods, and the sample selection, IV and LATE methods originated in econometrics. Early econometrics papers, that generally did not explicitly use the current treatment effects framework, include Ashenfelter (1978), Heckman (1978, 1979), Heckman and Robb (1985), Lalonde (1986) and Björklund and Moffitt (1982). Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (1999) emphasize matching methods. Abadie and Imbens (2006) provide results for inference. Bertrand, Duflo and Mullainathan (2004) demonstrate

that when difference-in-difference methods are used with panel data it is critical that one uses cluster-robust standard errors that cluster on the treatment unit, often the state, as the treatment regressor is highly correlated over time. Hausman and Kuersteiner (2008) consider more efficient GLS estimation in this setting. Imbens and Angrist (1994) introduce LATE and Björklund and Moffitt (1982) and Heckman and Vytlacil (2005) introduce MTE. Hahn, Todd and Van der Klaauw (2001) provide theory for RD methods; Ludwig and Miller (2007) provide a detailed application, and Imbens and Lemieux (2008) provide a survey. More recent research provides distribution theory when a nonparametric component is used and seeks to extend methods to nonlinear models (for example, Athey and Imbens, 2006), and to multiple treatments. Brief surveys include Smith (2000), Blundell and Dias (2002) and Angrist (2008), while lengthier surveys include Heckman, Lalonde and Smith (1999) and Angrist and Krueger (1999). The forthcoming book by Angrist and Pischke (2009) focuses on treatment effects methods.

14.5.2 Instrumental variables methods

Consider the linear model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{14.45}$$

where $\text{Cor}[\mathbf{x}_i, u_i] \neq \mathbf{0}$ so that OLS is inconsistent. Assume there exists an instrument $\mathbf{z}_i$ such that $\text{Cor}[\mathbf{z}_i, u_i] = \mathbf{0}$. The IV estimator for a just-identified model, considered for simplicity, is:

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{14.46}$$

If $\text{Cor}[\mathbf{z}_i, u_i] = \mathbf{0}$ then $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and:

$$\text{V}[\widehat{\boldsymbol{\beta}}_{\text{IV}}] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\Sigma\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}, \tag{14.47}$$

where $\Sigma = \text{E}[\mathbf{u}\mathbf{u}'|\mathbf{Z}]$. This estimator is easily extended to overidentified models, and to nonlinear models as a special case of GMM.

The applied literature has included many creative examples of instrument use. For example, in earnings–schooling regression a proposed instrument for schooling is distance to college, as this may be related to college attendance but may not directly effect earnings. Another possible instrument is birth month, which may be related to years of schooling as it determines age of school entry and hence years of schooling before a person reaches the minimum school leaving age.

This interest in the use of IV methods has been somewhat diminished by recognition of the problems that arise when instruments are weakly correlated with the regressor(s) being instrumented.

A weak instrument is one for which $\text{Cor}[\mathbf{z}_i, \mathbf{x}_i]$ is small. More precisely, suppose there is one endogenous regressor and several exogenous regressors. Then the instrument for the endogenous regressor is weak if the correlation between the endogenous regressor and the instrument is low after partialing out the effect of the other exogenous regressors. Then it is well known that $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ will be imprecisely estimated. Two other complications can arise.

First, suppose that $\text{Cor}[\mathbf{z}_i, u_i]$ is close to zero rather than exactly zero. Then not only is the IV estimator inconsistent, but it can be more inconsistent than the

OLS estimator. For example, in the simple case of a scalar regressor x and scalar instrument z, suppose the correlation between x and z is 0.1. Then IV becomes more inconsistent than OLS if the correlation between z and u exceeds a mere 0.1 times the correlation between x and u. This result, emphasized by Bound, Jaeger and Baker (1995), has led to increased scrutiny of assumptions regarding the validity of an instrument in any particular application.

Second, even if Cor$[\mathbf{z}_i, u_i]$ equals zero, regular asymptotic theory performs poorly in finite samples if the instrument is weak. Theoreticians established key results early. Applied researchers to subsequently highlight the problem were Nelson and Startz (1990) and Bound, Jaeger and Baker (1995). Staiger and Stock (1997) provided influential theory.

Third, regular asymptotic theory performs poorly in finite samples when there are many instruments, so that the model is greatly overidentified. This situation can arise for estimators based on conditioning on a large information set, in panel settings where regressors from other periods are valid instruments in the current period, or if an underlying instrument is interacted with exogenous regressors to generate many instruments.

There is a large theoretical literature on inference with weak instruments, including new estimators and new testing procedures (see Andrews, Moreira and Stock, 2007). Andrews and Stock (2007) provide a recent survey, and Flores-Luganes (2007) compares many of the different methods by Monte Carlo simulation and use of actual data.

14.5.3 Panel data

Panel data are repeated observations on the same cross-section units, typically individuals or firms, for several time periods. The cross-section units are usually assumed to be independent, though this assumption may be less appropriate if the cross-section units are states or countries.

An obvious advantage of panel data is that they permit increased precision in estimation, due to an increased number of observations. It is important, however, that one control for likely correlation of observations over time for a given cross-section unit. The usual method is to use cluster-robust standard errors described in section 14.4.1.

The microeconometrics literature has focused on a second advantage of panel data, that it provides a way to identify causation even if there is selection on unobservables, provided the unobservables are time invariant.

The fixed effects linear panel model specifies:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{14.48}$$

where α_i and ε_{it} are unobserved. It is assumed that the idiosyncratic error ε_{it} is uncorrelated with $\mathbf{x}_{it}$, but the individual-specific error α_i is potentially correlated with $\mathbf{x}_{it}$. Note that, while α_i is called a "fixed effect" in the literature, this term is misleading as it is being treated as random. Microeconometricians focus on short panels, with $N \to \infty$ but T permitted to be small (for a static linear model it is sufficient that $T \geq 2$).

To relate this to the treatment effects literature, $\mathbf{x}_{it}$ may include a binary treatment d_{it} that is correlated with the error term $\alpha_i + \varepsilon_{it}$ (selection on unobservables), but only with the component α_i of the error term that is time invariant. For example, an individual may self-select into a training program due to unobserved high ability, but this high ability is assumed to be time invariant.

Pooled OLS regression of y_{it} on $\mathbf{x}_{it}$ will lead to inconsistent estimation of $\boldsymbol{\beta}$, due to correlation of regressors with the error. The random effects estimator of $\boldsymbol{\beta}$, the feasible GLS estimator of (14.48) under the assumption that both α_i and ε_{it} are i.i.d., is also inconsistent if, in fact, α_i is correlated with $\mathbf{x}_{it}$. For this reason many microeconometric studies shy away from random effects models that are widely used in other fields.

Estimation of transformed models that eliminate α_i can lead to consistent estimation. The fixed-effects or within estimator is obtained by OLS estimation of the within-model:

$$(y_{it} - \bar{y}_i) = (\mathbf{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (u_{it} - \bar{u}_i). \tag{14.49}$$

A standard procedure is to use cluster-robust standard errors from this regression, assuming T is small and $N \to \infty$. Hansen (2007a) presents asymptotic theory that additionally allows $T \to \infty$ and Hansen (2007b) considers more efficient GLS estimation. The first differences estimator is obtained by OLS estimation of the first differences model:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta u_{it}. \tag{14.50}$$

Note that in both cases only the coefficients of time-varying regressors can be identified.

Extension to nonlinear models is possible only for some specific models, as there is an incidental parameters problem. The asymptotics rely on $N \to \infty$, so the number of parameters (k regression coefficients plus N fixed effects α_i) is going to infinity with the sample size. Some models permit transformations that eliminate α_i, while others do not. For nonlinear models with additive error the within and first differences transformations can again be used. For binary outcomes fixed effects estimation is possible for the logit model (see Chamberlain, 1980), but not the probit model. For count data, Hausman, Hall and Griliches (1984) presented fixed effects estimation for the Poisson model and a particular parameterization of the negative binomial model. The Poisson fixed effects estimator does not require that the data be Poisson distributed, as it is consistent provided the conditional mean is correctly specified. An active area of research is developing methods for general nonlinear fixed effects panel models that, while inconsistent due to the incidental parameters problem, are less inconsistent than existing methods (see, for example, Arellano and Hahn, 2007).

Panel data also provide the opportunity to model individual-level dynamic behavior, since the individual is observed at more than one point in time. A simple dynamic linear fixed effects model includes a lagged dependent variable, so that:

$$y_{it} = \rho y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{14.51}$$

An important result is that the fixed effects and first differences estimators of this model are inconsistent. Instrumental variables estimation of the first differences estimator is possible, using $y_{i,t-2}$ as an instrument for $(y_{i,t} - y_{i,t-1})$. Holtz-Eakin, Newey and Rosen (1988) and Arellano and Bond (1991) proposed using additional lags as instruments and estimating by GMM using an unbalanced set of instruments.

For nonlinear dynamic models fixed effects estimation is possible for the logit model (see Chamberlain, 1985; Honore and Kyriazidou, 2000), for the Poisson model (see Blundell, Griffiths and Windmeijer, 2002), and for some duration models (see Chamberlain, 1985; Van den Berg, 2001).

14.5.4 Structural models

The classic linear simultaneous equations model (SEM) has deliberately not been discussed in this section on causation, as the SEM is rarely used in microeconometric studies. Many causal studies are interested in the marginal effect of a single regressor on a single dependent variable. In that case 2SLS regression of the single equation of interest is simply instrumental variables estimation, already discussed, which itself has deficiencies, leading to increased use of the other methods given in this section. Finally, the linear SEM does not extend readily to nonlinear models and, in cases where it does, such as simultaneous equation tobit models, the distributional assumptions are very strong.

Another type of structural modeling is microeconometric models based on economic models of utility or profit maximization. Early references include Heckman (1974), MaCurdy (1981) and Dubin and McFadden (1984). The more recent labor literature most commonly uses structural economic models to explain employment dynamics (see, for example, Keane and Wolpin, 1997). Structural modeling is more often used in industrial organization (Reiss and Wolak, 2007, provide a survey).

14.6 Heterogeneity

A loose definition of heterogeneity is that data differs across observations. In a regression context this heterogeneity may be due to regressors (observables) or due to unobservables.

To begin with, consider heterogeneity due directly to observed regressors. For the linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $\mathrm{E}[u_i|\mathbf{x}_i] = 0$, $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ so that heterogeneity induces heterogeneity in the conditional means, though not in the marginal effects $\partial \mathrm{E}[y_i|\mathbf{x}_i]/\partial \mathbf{x}_i = \boldsymbol{\beta}$. Nonlinearity in the conditional mean, e.g. $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, will induce marginal effects that differ across individuals. Even simple parametric nonlinear models such as probit and tobit have this feature. The standard method is to present a single summary statistic. Often the marginal effect is evaluated at $\mathbf{x} = \bar{\mathbf{x}}$, but for most purposes a better single measure is the sample average of the individual marginal effects. Single index models, that is, $\mathrm{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i'\boldsymbol{\beta})$, have the advantage that the ratio of marginal effects for two different regressors equals the ratio of the corresponding parameters, and does not depend on the regressor values. Thus if one coefficient is twice another then the corresponding

marginal effect is twice as large. Quite flexible modeling of heterogeneity in $\mathrm{E}[y_i|\mathbf{x}_i]$ and the associated marginal effects is possible using nonparametric regression of y on $\mathbf{x}$. This yields very noisy estimates for high dimension $\mathbf{x}$, leading to the use instead of semiparametric methods such as those given in section 14.3.6.

More challenging is controlling for unobserved heterogeneity that is due to factors other than the regressors. Different individuals then have different responses even if the individuals have the same value of $\mathbf{x}$. Failure to control for this unobserved heterogeneity can lead to inconsistent parameter estimates and associated marginal effects. A simple example is omitted variables bias in the linear regression model, where the omitted variables form part of the unobserved heterogeneity. The source of the unobserved heterogeneity can also matter. In particular, in structural models of economic behavior a distinction is made as to whether or not the unobserved (to the econometrician) heterogeneity is known to the decision maker.

Meaningful discussion of unobserved heterogeneity requires the statement of an underlying structural relationship to be estimated in the presence of unobserved heterogeneity. Wooldridge (2005, 2008) provides a fairly general framework. Suppose that interest lies in a conditional mean $m(\mathbf{x}, u) = \mathrm{E}[y|\mathbf{x}, u]$, or more formally:

$$m(\mathbf{x}, u) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}, U = u],$$

where u is unobserved and for simplicity is a scalar. Ideally $m(\mathbf{x}, u)$ would be estimated but, instead, analysis is restricted to what Blundell and Powell (2004) call the average structural function (ASF):

$$m(\mathbf{x}) = \mathrm{E}_U[m(x, U)],$$

which integrates out the unobserved heterogeneity. Often, interest lies in how ASF changes as the jth regressor, say, changes. This is the average partial effect (APE):

$$\frac{\partial m(\mathbf{x})}{\partial x_j} = \mathrm{E}_U\left[\frac{\partial m(x, U)}{\partial x_j}\right].$$

Unobserved heterogeneity poses a problem because the ASF $m(\mathbf{x})$ in general differs from the conditional mean $\mathrm{E}[y|\mathbf{x}] = \mathrm{E}_{U|\mathbf{x}}[m(x, U)]$, and hence APE differs from $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j$, but it is only $\mathrm{E}[y|\mathbf{x}]$ that is identified from the observed data.

The simplest assumption, and one commonly made, is that u is independent of $\mathbf{x}$, as then $\mathrm{E}[y|\mathbf{x}] = m(\mathbf{x})$. In a model with additive heterogeneity, analysis is particularly straightforward. If $m(\mathbf{x}, u) = g(\mathbf{x}, \boldsymbol{\beta}) + u$ then $\mathrm{E}[y|\mathbf{x}] = g(\mathbf{x},\boldsymbol{\beta})$ given u independent of $\mathbf{x}$ with mean zero. Analysis is more complicated if unobserved heterogeneity enters in a nonlinear manner. For example, if $m(\mathbf{x}, u) = g(\mathbf{x}'\boldsymbol{\beta} + u)$ then $\mathrm{E}[y|\mathbf{x}] = \mathrm{E}_u[g(\mathbf{x}'\boldsymbol{\beta} + u)]$ will typically require specification of the distribution of u and integration over this. In some cases analytical expressions can be obtained. In other cases numerical methods are used. If u is low dimensional (in many applications it is a scalar) then Gaussian quadrature methods work well. Otherwise the simulation methods given in sections 14.3.3 and 14.3.4 can be used. Examples include negative binomial models for counts obtained by a Poisson–gamma mixture, Weibull–gamma mixtures for durations, random utility models for binary and

multinomial data (where u is now a vector) and normal mixtures for linear and nonlinear panel data. While often the unobserved heterogeneity is interpreted as a random intercept, this can be generalized to random slopes (a random coefficients model). An alternative is finite mixtures models, used particularly in duration and count data analysis.

Panel data offer the opportunity to permit u to be dependent on $\mathbf{x}$. In that case u_{it} is decomposed into a time-varying component that is independent of $\mathbf{x}_{it}$ and a time-invariant component that may be correlated with $\mathbf{x}_{it}$. Fixed effects estimators for these models have been discussed in section 14.5.3. It is important to note that in nonlinear models these methods identify $\boldsymbol{\beta}$ but not ASF, so that the APEs are only estimated up to scale.

Panel data also offer the possibility of distinguishing between persistence in behavior over time due to unobserved heterogeneity and persistence in behavior over time due to true state dependence. For example, rather than the static linear model $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$, where correlation of u_i with $\mathbf{x}_{it}$ causes problems, a more appropriate model may be a dynamic model $y_{it} = \rho y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$, where there is now no complication of unobserved heterogeneity. These models have quite different structural interpretations with quite different policy consequences. For example, high persistence of unemployment given regressors may be due to stigma attached to being unemployed (state dependence) or may be due to unobserved low ability (unobserved heterogeneity).

The treatment effects literature allows for unobserved heterogeneity. By assuming that selection is on observables it is possible to estimate ATET($\mathbf{x}$), which is the APE for the treatment variable.

Wooldridge (2005, 2008) proposes the use of proxy variables to identify the ASF and APE. For simplicity, consider the linear model $y = \mathbf{x}'\boldsymbol{\beta} + u$. If $\mathrm{E}[u|\mathbf{x}] = 0$, then $m(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ so unobserved heterogeneity causes no problem. Now consider an omitted variables situation where $u = \mathbf{z}'\boldsymbol{\gamma} + \varepsilon$ with $\mathrm{E}[\varepsilon|\mathbf{x}] = 0$ but $\mathrm{E}[\mathbf{z}|\mathbf{x}] \neq \mathbf{0}$. The ASF is $m(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \mathrm{E}[\mathbf{z}]'\boldsymbol{\gamma}$, whereas the conditional mean $\mathrm{E}[y|\mathbf{x}] = \boldsymbol{\beta} + \mathrm{E}[\mathbf{z}|\mathbf{x}]'\boldsymbol{\gamma}$. These terms differ unless $\mathrm{E}[\mathbf{z}|\mathbf{x}] = \mathrm{E}[\mathbf{z}]$, the case where the unobserved heterogeneity is independent of $\mathbf{x}$. A weaker assumption than independence is to assume that there is a proxy variable $\mathbf{w}$ for $\mathbf{z}$ with the properties that (i) $\mathbf{x}$ and $\mathbf{z}$ are independent conditional on $\mathbf{w}$ so $\mathrm{E}[\mathbf{z}|\mathbf{x}, \mathbf{w}] = \mathrm{E}[\mathbf{z}|\mathbf{x}]$, and (ii) $\mathrm{E}[y|\mathbf{x}, \mathbf{w}, \mathbf{z}, \varepsilon] = \mathrm{E}[y|\mathbf{x}, \mathbf{w}, \varepsilon]$, so that $\mathbf{z}$ is redundant in the original model. Then $\mathrm{E}[y|\mathbf{x}, \mathbf{w}] = \mathbf{x}'\boldsymbol{\beta} + \mathrm{E}[\mathbf{z}|\mathbf{w}]'\boldsymbol{\gamma}$, which can be identified by regression of y on $\mathbf{x}$ and $\mathbf{z}$. Taking the expected value with respect to $\mathbf{w}$ then gives the desired ASF. Wooldridge (2005) generalizes this approach to nonlinear models and argues that, even though failure to control for unobserved heterogeneity may lead to inconsistent parameter estimates, it is still possible in some cases to consistently estimate the ASF and APE.

There is also a growing literature on heterogeneity in nonparametric models: see, for example, Matzkin (2008). A simple approach is to start with the conditional cumulative distribution function (c.d.f.) $F(y|\mathbf{x})$, which can be nonparametrically estimated. Define $u = F(y|\mathbf{x})$, then u is uniformly distributed on $(0, 1)$ and hence uncorrelated with $\mathbf{x}$. Inverting yields $y = F^{-1}(u|\mathbf{x}) = G(\mathbf{x}, u)$. This provides a decomposition into observables $\mathbf{x}$ and unobservables u that is independent of $\mathbf{x}$,

a separable model. But this is not a structural model in the sense of the ASF given earlier.

Controlling for unobserved heterogeneity is an active area in microeconometrics, as much of the variation in the outcome is due to unobserved factors since, typically, $R^2 < 0.5$. It is particularly important when there is sample selection or self-selection. For example, in OLS regression we essentially require only that $\mathrm{E}[u|\mathbf{x}] = 0$, whereas if the sample is truncated or censored, much stronger assumptions on u are needed even if semiparametric methods are used. Heckman (2000, 2005) and related papers explicitly consider heterogeneity and structural estimation (see also Blundell and Powell, 2004; Wooldridge, 2005).

14.7 Data issues

Microeconometric data are often survey data that come from sampling schemes more complicated than simple random sampling, and key variables can be mismeasured or even missing due to nonresponse. These issues are generally ignored in applied work. Ignoring the sampling scheme is reasonable in the many cases where the sampling scheme or nonresponse mechanism leads to a sample that is nonrepresentative only of the regressors, while maintaining representativeness of the dependent variable conditional on regressors. It is also reasonable to ignore measurement error if it is classical measurement error in the dependent variable in a linear model. In other cases standard estimators are often inconsistent and alternative estimators are needed.

14.7.1 Sampling schemes

Survey data often use stratified and clustered sampling to lower interview costs and to provide more precise estimates for population sub-groups, such as regions with relatively few people, than would otherwise be the case. The extensive sample survey literature, initially focused on estimation of population means but then extended to the regression case, has generally been ignored by the econometrics literature.

The first issue raised by survey sampling schemes is that the sample is no longer representative of the population. For inference on a single variable it is necessary to adjust for this. For example, average earnings in a nonrepresentative sample will be an inconsistent estimate of population mean earnings. For regression analysis, adjustment is necessary if the sample is nonrepresentative for the dependent variable after conditioning on regressors (endogenous stratification), but may not be necessary if the sample is nonrepresentative only for the regressors (exogenous stratification).

For endogenous stratification, where stratification is on the dependent variable in a regression, standard estimation methods lead to inconsistent parameter estimates.

Consider stratification in a likelihood framework. Let the conditional distribution of y given $\mathbf{x}$ be denoted $f(y|\mathbf{x}, \theta)$. Usually the joint density of y and $\mathbf{x}$ is $g(y, \mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}) \times g(\mathbf{x})$, where the parameters in $g(\mathbf{x})$ are suppressed. Under

exogenous sampling $g(\mathbf{x})$ does not involve $\boldsymbol{\theta}$, so that inference on $\boldsymbol{\theta}$ can be based on the conditional log-likelihood based only on $f(y|\mathbf{x}, \boldsymbol{\theta})$.

Under endogenous stratification, however, it can be shown that $g(y, \mathbf{x}|\boldsymbol{\theta})$ takes a more complicated form, and ML estimation needs to be based on the joint log-likelihood based on $g(y, \mathbf{x}|\boldsymbol{\theta})$. Standard estimators that instead continue to use $f(y|\mathbf{x}, \boldsymbol{\theta})$ are inconsistent. Examples include truncated regression (for example, hours of work are modeled and only workers are surveyed), choice-based sampling (for example, commute mode choice is modeled and bus-riders are deliberately oversampled as there are relatively few bus-riders), on site sampling and case-control studies. Much of the econometrics literature has focused on choice-based sampling in discrete choice models, with estimation by weighted MLE (see Manski and Lerman, 1977), or more efficient GMM methods, (see Imbens, 1992). A more general presentation for endogenous stratification is given by Imbens and Lancaster (1996). Wooldridge (2001) considers inverse-probability weighted estimators for m-estimators.

Stratified surveys usually provide sample weights that can be used to obtain population representative statistics. Under exogenous stratification, these sample weights need not be used in the typical situation where correct specification of a regression model is assumed. For example, assume that the regression function is linear in $\mathbf{x}$, so that $y = \mathbf{x}'\boldsymbol{\beta} + u$, $\mathrm{E}[u|\mathbf{x}] = 0$ and $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. Then OLS is consistent even if the regressors $\mathbf{x}$ are not representative of the population in $\mathbf{x}$. The reason for using these weights in estimation is if we wish to relax the assumption that $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, due to nonlinearity or because $\boldsymbol{\beta}$ varies across strata. Then weighted OLS should be used as it provides an estimate of the so-called census coefficient $\boldsymbol{\beta}^*$ that has probability limit equal to the regression coefficient that would be obtained by regression of y on $\mathbf{x}$ using the entire population (see DuMouchel and Duncan, 1983). For example, a weighted OLS regression of earnings on years of schooling provides a consistent estimate of the population marginal effect on earnings of one more year of schooling, without assuming that the model is linear. Note that even if unweighted estimation is appropriate, weights may still be used in making predictions from the model. For example, if $\mathrm{E}[y|\mathbf{x}]$ is nonlinear in $\mathbf{x}$ then marginal effects vary with evaluation point $\mathbf{x}$, so that weights should be used to compute an estimate of the population marginal effect.

A big reason for stratification is to improve efficiency of estimates of the population mean of a single variable, such as earnings or unemployment, when the mean of that variable differs across strata. This efficiency gain can carry over to regression, and some regression packages include commands to do so. These are widely used in biostatistics but not in econometrics, in part because the efficiency gains are felt to be small and in part because not all datasets provide the necessary information on the strata. Bhattacharya (2005) presents results for m-estimation and a good discussion of the issues.

In addition to stratification, survey methods often induce dependence for sub-groups of observations. For example, several households on the same block may be interviewed. Then data in that sub-group are likely to be positively correlated and, even after controlling for regressors, model errors are likely to be

positively correlated. The standard procedure in econometrics is to use estimators that ignore the clustering and base inference on cluster-robust standard errors, presented in section 14.4.1. More efficient estimation is possible using feasible GLS estimators that model the error correlation. In particular, hierarchical linear models or multilevel models are often used in other social science disciplines but are seldom used in econometrics. If clustering is felt to induce correlation of errors with regressors, then cluster-specific fixed effects, analogous to an individual-specific fixed effect in a panel data model, may also be used.

14.7.2 Missing data

The starting point for analysis of missing data is the terminology and assumptions made about the nature of the process leading to missing data on w_i, say, due to Rubin (1976). These have many similarities with the potential outcomes model, where the unknown counterfactual can also be viewed as a missing data problem. If the probability of w_i being missing depends on neither its own value or on other data in the data set then w_i is missing completely at random (MCAR), and missing data on w_i causes no problems aside from efficiency loss. If the probability of w_i being missing depends on other data in the data set, but not its own value, then w_i is missing at random (MAR), and missing data may lead to estimator inconsistency. If w_i is MAR, then it is possible to adjust for missingness if the missing data mechanism is ignorable, meaning that the parameters of the missing data mechanism are unrelated to the parameters that we estimate, similar to weak exogeneity.

Simple corrections for missing data include dropping an observation if any variable is missing (listwise deletion or case deletion) and simple imputation methods such as using the sample average or predictions from a fitted regression model. These corrections are valid if data are MCAR or the missing data are regressors only that are MAR with probability that is independent of the dependent variable.

The modern approach is to use multiple imputation methods that regard missing data as random variables and replaces with draws from an assumed underlying distribution. Let $\mathbf{W} = (\mathbf{W}_{obs}, \mathbf{W}_{miss})$ denote the data partitioned into observed and missing observations, and suppose $\mathbf{W}$ has density $f(\mathbf{W}|\boldsymbol{\theta})$. The multiple imputation method imputes $\mathbf{W}_{miss}$ under the assumption of MAR with ignorable missingness. There are several ways to make imputations. A preferred, though computationally expensive, method is to use data augmentation and MCMC methods. Given an sth round estimate of $\boldsymbol{\theta}^{(r)}$, we impute $\mathbf{W}_{miss}^{(r+1)}$ by making a draw from $f(\mathbf{W}_{miss}|\mathbf{W}_{obs}, \boldsymbol{\theta}^{(r)})$. Then a new estimate $\boldsymbol{\theta}^{(r+1)}$ is obtained by drawing from $f(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{miss}^{(r+1)})$. The chain is continued to convergence, giving an imputed value for $\mathbf{W}_{miss}$. Suppose we obtain imputed value $\mathbf{W}_{miss}^{(I)}$ and then obtain the MLE based on $f(\mathbf{W}_{obs}, \mathbf{W}_{miss}^{(I)}|\boldsymbol{\theta})$. This will overstate estimator precision as it fails to account for the uncertainty created by imputation of $\mathbf{W}_{miss}^{(I)}$. Multiple imputation overcomes this by obtaining m different imputed values for $\mathbf{W}_{miss}$ and hence m estimates $\widehat{\boldsymbol{\theta}}_r$, $r = 1, \ldots, m$, with associated variance matrices $\widehat{\mathbf{V}}_r = \widehat{\mathbf{V}}[\widehat{\boldsymbol{\theta}}_r]$. For further details see Little and Rubin (1987), Rubin (1987) and Schafer (1997).

Microeconometrics applications rarely use multiple imputation methods, in part due to concern that missingness may be for nonignorable reasons, such as endogenous stratification discussed in section 14.7.1. Wooldridge (2007) considers use of inverse-probability weighting estimators when data are missing and provides a link with the framework of Rubin (1976).

14.7.3 Measurement error

Standard results for measurement error consider the linear regression model with classical measurement error in regressors. OLS coefficients are then inconsistent and understate the magnitude of the true coefficient. More recent work has considered nonlinear regression models and, in some cases, nonclassical measurement error.

Suppose $y = \beta x^* + u$, with error u uncorrelated with x^*, but we observe x rather than x^* and regress y on x. Then, from Angrist and Krueger (1999), the OLS estimator $\widehat{\beta} = [\sum_i x_i^2]^{-1} \sum_i x_i y_i = [\sum_i x_i^2]^{-1} \sum_i x_i(\beta x_i^* + u_i)$ is in general inconsistent as:

$$\text{plim}\widehat{\beta} = [\text{V}[x]]^{-1}\text{Cov}[x, x^*]\beta = \lambda\beta, \tag{14.52}$$

where $\lambda = \text{Cov}[x, x^*]/\text{V}[x]$ is the reliability ratio of x as a measure of x^*, and we have assumed that $\text{plim}N^{-1} \sum_i x_i u_i = 0$. This assumption that x is uncorrelated with u requires the additional assumption that u is uncorrelated with the measurement error $v = x - x^*$, in addition to the usual assumption that the model error u is uncorrelated with x^*.

The size of the inconsistency depends on the size of the reliability ratio, which has been measured in various survey validation studies. Angrist and Kruger (1999, p. 1346) present a summary table with reliability ratios for log annual earnings, annual hours and years of schooling ranging from 0.71 to 0.94. Bound, Brown and Mathiowetz (2001, pp. 3749–830) summarize many validation studies for labor-related data that also indicate that measurement error is large enough to lead to appreciable bias in OLS coefficients.

Result (14.52) makes few assumptions beyond independence of measurement error and model error. Textbook treatments of measurement error emphasize the classical measurement error model, a more restrictive model that assumes:

$$y = \beta x^* + u,\ u \sim iid\left[0, \sigma_u^2\right]$$

$$x = x^* + v,\ v \sim iid\left[0, \sigma_v^2\right] \quad \text{and} \quad x^* \sim iid\left[0, \sigma_{x^*}^2\right]. \tag{14.53}$$

Then $\text{plim}\widehat{\beta} = \lambda\beta$, where $\lambda = \sigma_{x^*}^2/(\sigma_{x^*}^2 + \sigma_v^2) = 1/(1 + s)$ and where $s = \sigma_v^2/\sigma_{x^*}^2$ is the noise-to-signal ratio. Since $s \geq 0$, $\widehat{\boldsymbol{\beta}}$ is downward biased asymptotically towards zero, a bias called attenuation bias. The attenuation bias is reduced if additional (correctly measured) regressors are included, and is increased if panel data are used with estimation by differencing methods such as the within-estimator.

There are several ways to secure identification of $\boldsymbol{\beta}$. These include instrumental variables methods (assuming availability of an instrument z that is correlated with x^* but not with the model error u), use of replicated data or validation sample data to estimate key sample cross-moments, and use of additional distributional assumptions, such as symmetry of the error. Bounds on $\boldsymbol{\beta}$ can also be obtained using reverse regression. Wansbeek and Meijer (2000) review many identification methods. Few studies correct for measurement error, however, in part due to lack of necessary data or reluctance to make strong assumptions about the nature of the measurement error.

The preceding methods do not generalize easily and in a systematic way to nonlinear models. Carroll, Ruppert and Stefanski (1995) summarize the statistics literature and Hausman (2001) considers the econometrics literature. For measurement error in regressors in nonlinear regression with additive error, an early reference is Y. Amemiya (1985) and a more recent reference is Schennach (2004). For nonlinear models with nonadditive errors, such as discrete outcome and count models, measurement error in the dependent variable can also cause problems. For example, Hausman, Abrevaya and Scott-Morton (1998) consider mismeasurement in the dependent variable in binary outcome models, taking a parametric approach with strong assumptions.

The classical measurement error model maintains that the measurement error is i.i.d. Some work relaxes this. An early example is that, for a binary regressor, the measurement error is necessarily correlated with the true value, since the only way to mismeasure a value of 0 is as a 1, and vice versa. Mahajan (2006) gives a quite general treatment for binary regressors. Kim and Solon (2005) consider standard linear panel estimators when measurement error in a regressor is negatively correlated with the true value.

14.8 Conclusion

Microeconometricians are very ambitious in their desire to obtain marginal effects that can be given a causative interpretation, permit individual heterogeneity and are obtained under minimal assumptions. The associated statistical inference should also rely on minimal assumptions. This has led to a literature and toolkit that is quite advanced for an area of applied statistics.

This survey has of necessity been selective. The methods used in labor economics and public economics have been emphasized. General approaches have been presented, with specialization usually to the linear model. For econometrics methods for specific types of data – binary, multinomial, durations and counts – good starting points are the specialized monographs by, respectively, Maddala (1983), Train (2003), Lancaster (1990), and Cameron and Trivedi (1998), as well as the more general texts cited in the introduction and Cameron and Trivedi (2008). The chapters by Greene and Jones in this volume are also highly relevant.

Acknowledgments

This chapter draws considerably on Cameron and Trivedi (2005). Earlier versions were presented at the Japanese Statistical Society 75th Anniversary Symposium on Applied Microeconometrics, University of Tokyo, September 2006, and at the 23rd Annual Summer Meeting of the Society for Political Methodology, UC Davis, July 2006. The author benefitted from comments of conference participants, from sabbatical leave at UC Berkeley, and from discussions with Bryan Graham, Michael Jansson, Oscar Jorda, Guido Kuersteiner, Jim Powell and Paul Ruud.

References

Abadie, A. and G.W. Imbens (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–67.

Abadie, A. and G.W. Imbens (2008) On the failure of the bootstrap for matching estimators, *Econometrica*. Forthcoming.

Abowd, J.M. and D. Card (1987) On the covariance of earnings and hours changes. *Econometrica* **57**, 411–45.

Abrevaya, J. and J. Huang (2005) On the bootstrap of the maximum score estimator. *Econometrica* **73**, 1175–204.

Altonji, J.G. and L.M. Segal (1996) Small sample bias in GMM estimation of covariance structures. *Journal of Business and Economic Statistics* **14**, 353–66.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Amemiya, Y. (1985) Instrumental variable estimator for the nonlinear error in variables model. *Journal of Econometrics* **28**, 273–89.

Andrews, D.W.K., M.J. Moreira and J.H. Stock (2007) Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics* **139**, 116–32.

Andrews, D.W.K. and J.H. Stock (2007) Inference with weak instruments. In R. Blundell, W. Newey and T. Persson (eds.), *Advances in Economic Theory: Ninth World Congress, Volume 3*. Cambridge: Cambridge University Press.

Angrist, J. (2008) Treatment effects. In S.N. Durlauf and E. Blume (eds.), *The New Palgrave Dictionary of Economics* (second edition). Basingstoke: Palgrave Macmillan.

Angrist, J., V. Chernozhukov and I. Ferandez-Val (2006) Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74**, 539–63.

Angrist, J.D. and A.B. Krueger (1999) Empirical strategies in labor economics, in O.C. Ashenfelter and D.E. Card (eds.), *Handbook of Labor Economics, Volume 3A*, pp. 1277–397. Amsterdam: North-Holland.

Angrist, J.D. and J.-S. Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. Forthcoming.

Anselin, L. (2001) Spatial Econometrics. In B. Baltagi (ed.), *A Companion to Theoretical Econometrics*, pp. 310–30. Oxford: Blackwell.

Arellano, M. (1987) Computing robust standard errors for within-group estimators. *Oxford Bulletin of Economics and Statistics* **49**, 431–4.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277–98.

Arellano, J. and J. Hahn (2007) Understanding bias in nonlinear panel models: some recent developments. In R. Blundell, W. Newey and T. Persson (eds.), *Advances in Economic Theory: Ninth World Congress, Volume 3*. Cambridge: Cambridge University Press.

Ashenfelter, O. (1978) Estimating the effect of training programs on earnings. *Review of Economics and Statistics* **60**, 47–57.

Athey, S. and G.W. Imbens (2006) Identification and inference in nonlinear difference-in-difference models. *Econometrica* **74**, 431–97.

Bellemare, C., B. Melenberg and A. van Soest (2002) Semiparametric models for satisfaction with income. *Portuguese Economic Journal* **1**, 181–203.
Bertrand, M., E. Duflo and S. Mullainathan (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 249–75.
Bhattacharya, D. (2005) Asymptotic Inference from Multi-stage Samples. *Journal of Econometrics* **126**, 145–71.
Björklund, A. and R. Moffitt (1982) The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics* **69**, 42–9.
Blundell, R. and M.C. Dias (2002) Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal* **1**, 91–115.
Blundell, R., A. Gosling, H. Ichimura and C. Meghir (2007) Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica* **75**, 323–63.
Blundell, R., R. Griffith and F. Windmeijer (2002) Individual effects and dynamics in count data models. *Journal of Econometrics* **102**, 113–31.
Blundell, R.W. and J.L. Powell (2004) Endogeneity in semiparametric binary response models. *Review of Economic Studies* **71**, 655–79.
Bound, J., C. Brown and N. Mathiowetz (2001) Measurement error in survey data. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics, Volume 5*. Amsterdam: North-Holland.
Bound, J., D.A. Jaeger and R.M. Baker (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–50.
Buchinsky, M. (1994) Changes in the U.S. wage structure 1963–1987: application of quantile regression. *Econometrica* **62**, 405–58.
Cameron, A.C., Gelbach, J. and D.L. Miller (2006) Robust inference with multi-way clustering. NBER Technical Working Paper No. 327.
Cameron, A.C., Gelbach, J. and D.L. Miller (2008) Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* **90**(3), 414–27.
Cameron, A.C. and P.K. Trivedi (1998) *Regression Analysis for Count Data*. Econometric Society Monograph No. 30. Cambridge: Cambridge University Press.
Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.
Cameron, A.C. and P.K. Trivedi (2008) *Microeconometrics using Stata*. College Station, Texas: Stata Press.
Carroll, R.J.D. Ruppert and L.A. Stefanski (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.
Chamberlain, G. (1980) Analysis of covariance with qualitative data. *Review of Economic Studies* **47**, 225–38.
Chamberlain, G. (1982) Multivariate regression models for panel data. *Journal of Econometrics* **18**, 5–46.
Chamberlain, G. (1984) Panel Data. In Z. Griliches and M. Intrilligator (eds.), *Handbook of Econometrics, Volume 2*, pp. 1247–318. Amsterdam: North-Holland.
Chamberlain, G. (1985) Heterogeneity, omitted variable bias and duration dependence. In J.J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data*, pp. 3–38. Cambridge: Cambridge University Press.
Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–34.
Chernozhukov, V. and C. Hansen (2005) An IV model of quantile treatment effects. *Econometrica* **73**, 245–62.
Chernozhukov, V., H. Hong and E. Tamer (2007) Estimation and confidence regions for parameter sets in econometric models, *Econometrica* **75**, 1243–84.

Chib, S. (1992) Bayes regression for the tobit censored regression model. *Journal of Econometrics* **58** 79–99.
Chib, S. (2001) Markov chain Monte Carlo methods: computation and inference. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics, Volume 5*, pp. 3570–649. Amsterdam: North–Holland.
Conley, T.G. (1999) GMM estimation with cross sectional dependence. *Journal of Econometrics* **92**, 1–45.
Davidson, R. and J.G. MacKinnon (1993) *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Deb, P. and P.K. Trivedi (2002) The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics* **21**, 601–25.
Dehejia, R.H. and S. Wahba (1999) Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–62.
Donald, S.G. and K. Lang (2007) Inference with differences in differences and other panel data. *Review of Economics and Statistics* **89**, 221–33.
Driscoll, J.C. and A.C. Kraay (1998) Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, **80**, 549–60.
Dubin, J.A. and D.L. McFadden (1984) An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* **52**, 345–62.
DuMouchel, W.K. and G.J. Duncan (1983) Using sample survey weights in multiple regression analyses of stratified samples, *Journal of the American Statistical Association* **78**, 535–43.
Efron, B. (1979) Bootstrapping methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.
Flores-Lagunes, A. (2007) Finite sample evidence of IV estimators under weak instruments. *Journal of Applied Econometrics* **22**, 677–94.
Gelfland, A.E. and A.F.M. Smith (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–39.
Geweke, J., G. Gowrisankaran and R.J. Town (2003) Bayesian inference for hospital quality in a selection model. *Econometrica* **71**, 1215–38.
Greene, W.H. (2003) *Econometric Analysis* (fifth edition). Upper Saddle River, NJ: Prentice-Hall.
Gouriéroux, C. and A. Monfort (1996) *Simulation Based Econometrics Methods*. New York, Oxford University Press.
Hahn, J., P. Todd and W. Van der Klaauw (2001) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69**, 201–9.
Haile, P.A. and E.T. Tamer (2003) Inference with an incomplete model of English auctions. *Journal of Political Economy* **111**(1), 1–51.
Hall, P. and J.L. Horowitz (1996) Bootstrap critical values for tests based on generalized method of moments estimators. *Econometrica* **64**, 891–916.
Hansen, C.B. (2007a) Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics* **141**, 597–620.
Hansen, C.B. (2007b) Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics* **140**, 670–94.
Hansen, L.P. (1982) Large sample properties of generalized methods of moments estimators. *Econometrica* **50**, 1029–54.
Hansen, L.P., J. Heaton and A. Yaron (1996) Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* **14**, 262–80.
Hausman, J.A. (1978) Specification tests in econometrics. *Econometrica* **46**, 1251–71.
Hausman, J.A. (2001) Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic Perspectives* **15**, 57–68.

Hausman, J.A., J. Abrevaya and F.M. Scott-Morton (1998) Misclassification of the dependent variable in a discrete response setting. *Journal of Econometrics* **87**, 239–69.
Hausman, J.A., B.H. Hall and Z. Griliches (1984) Econometric models for count data with an application to the patents-R&D relationship. *Econometrica* **52**, 909–38.
Hausman, J.A. and G. Kuersteiner (2008) Difference in difference meets generalized least squares: higher order properties of hypotheses tests. *Journal of Econometrics* **144**, 371–91.
Heckman, J.J. (1974) Shadow prices, market wages, and labor supply. *Econometrica* **42**, 679–94.
Heckman, J.J. (1978) Dummy endogenous variables in a simultaneous equations system. *Econometrica* **46**, 931–60.
Heckman, J.J. (1979) Sample selection as a specification error. *Econometrica* **47**, 153–61.
Heckman, J.J. (2000) Causal parameters and policy analysis in economics: a twentieth century perspective. *Quarterly Journal of Economics* **115**, 45–98.
Heckman, J.J. (2005) The scientific model of causality. *Sociological Methodology* **35**, 1–97.
Heckman, J.J., H. Ichimura and P. Todd (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Review of Economic Studies* **64**, 605–54.
Heckman, J.J., R.J. Lalonde and J.A. Smith (1999) The economics and econometrics of active labor market programs. In O. Ashenfelter, and D. Card (eds.), *Handbook of Labor Economics, Volume 3A*, pp. 1865–2097. Amsterdam: North–Holland.
Heckman, J.J. and R. Robb (1985) Alternative methods for evaluating the impact of interventions. In J.J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.
Heckman, J.J. and E. Vytlacil (2005) Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **73**, 669–738.
Holtz-Eakin, D., W. Newey and H.S. Rosen (1988) Estimating vector autoregressions with panel data. *Econometrica* **56**, 1371–95.
Honore, B.E. and L. Hu (2004) Estimation of cross-sectional and panel data censored regression models with endogeneity. *Journal of Econometrics* **122**, 293–316.
Honore, B.E. and E. Kyriazidou (2000) Panel data discrete choice models with lagged dependent variables, *Econometrica* **88**, 839–74.
Horowitz, J.L. (1994) Bootstrap-based critical values for the information matrix test. *Journal of Econometrics* **61**, 395–411.
Horowitz, J.L. (2001) The bootstrap. In J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Volume 5*, pp. 3159–228. Amsterdam: North-Holland.
Imbens, G.W. (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60**, 1187–214.
Imbens, G.W. (2002) Generalized method of moments and empirical likelihood, *Journal of Business and Economic Statistics* **20**, 493–506.
Imbens, G.W. and J. Angrist (1994) Identification and estimation of local average treatment effect, *Econometrica* **62**, 467–75.
Imbens, G.W. and T. Lancaster (1996) Efficient estimation and stratified sampling. *Journal of Econometrics* **74**, 289–318.
Imbens, G.W. and T. Lemieux (2008) Regression discontinuity designs: a guide to practice. *Journal of Econometrics* **141**, 615–35.
Imbens, G.W. and J. Wooldridge (2007) What's new in econometrics? Summer Institute Minicourse, National Bureau of Economic Research, http://www.nber.org/minicourse3.html.
Keane, M.P. and K.I. Wolpin (1997) The career decisions of young men. *Journal of Political Economy* **105**(3), 473–522.
Kim, B. and G. Solon (2005) Implications of mean-reverting measurement error for longitudinal studies of wages and employment. *Review of Economics and Statistics* **87**, 193–6.
Kitamura, Y. (2006) Empirical likelihood methods in econometrics: theory and practice. Cowles Foundation Discussion Paper No. 1569.

Kloek, T. and H.K. van Dijk (1978) Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* **46**, 1–19.
Koenker, R. (2005) *Quantile Regression*. Econometric Society Monograph No. 38. Cambridge: Cambridge University Press.
Koenker, R. and G. Bassett (1978) Regression quantiles. *Econometrica* **46**, 33–50.
Koenker, R. and K.F. Hallock (2001) Quantile regression. *Journal of Economic Perspectives* **15**, 143–56.
Koop, G.M. (2003) *Bayesian Econometrics*. New York: Wiley.
Koop, G.M., D.J. Poirier and J.L. Tobias (2007) *Bayesian Econometric Methods*. Cambridge: Cambridge University Press.
Lalonde, R. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 604–20.
Lancaster, T. (1990) *The Econometric Analysis of Transitional Data*. Econometric Society Monograph No. 17. Cambridge: Cambridge University Press.
Lancaster, T. (2004) *An Introduction to Modern Bayesian Econometrics*, Oxford: Blackwell.
Lee, L.-F. (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**, 1899–926.
Lerman, S.R. and C.F. Manski (1981) On the use of simulated frequencies to approximate choice probabilities. In C.F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 305–19. Cambridge, Mass.: MIT Press.
Li, Q. and J. Racine (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.
Liang, K.-Y. and S.L. Zeger (1986) Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
Little, R.J.A. and D. Rubin (1987) *Statistical Analysis with Missing Data*. New York: John Wiley.
Ludwig, J. and D.L. Miller (2007) Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics* **122**(1), 159–208.
MacKinnon, J.G. (2002) Bootstrap inference in econometrics. *Canadian Journal of Economics* **35**, 615–45.
MaCurdy, T.E. (1981) An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* **89**, 1059–85.
Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Economics*. Cambridge: Cambridge University Press.
Mahajan, A. (2006) Identification and estimation of regression models with misclassification. *Econometrica* **74**(3), 631–65.
Manski, C.F. (1975) The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* **3**, 205–28.
Manski, C.F. (1988) *Analog Estimation Methods in Econometrics*. London: Chapman and Hall.
Manski, C.F. (1995) *Identification Problems in the Social Sciences*. Cambridge, Mass.: Harvard University Press.
Manski, C.F. (2008) Partial identification in econometrics. In S.N. Durlauf and L.E. Blume (eds.), *New Palgrave Dictionary of Economics* (second edition). Basingstoke: Palgrave Macmillan.
Manski, C.F. and S.R. Lerman (1977) The estimation of choice probabilities from choice-based samples. *Econometrica* **45**, 1977–88.
Manski, C.F. and J.V. Pepper (2000) Monotone instrumental variables with an application to the returns to schooling. *Econometrica* **68**, 997–1010.
Matzkin, R.L. (2008) Identification in nonparametric simultaneous equations. *Econometrica* **58**, 757–82.
McCullagh, P. and J.A. Nelder (1983) *Generalized Linear Models* (first edition). London: Chapman and Hall (second edition 1989).

McCulloch, R.E., N.G. Polson and P.E. Rossi (2000) A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–93.
McFadden, D. (1989) A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57**, 995–1026.
Meyer, B.D. (1990) Unemployment insurance and unemployment spells. *Econometrica* **58**, 757–82.
Mittelhammer, R.C., G.G. Judge and R. Schoenberg (2005) Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods. In D.W.K. Andrews and J.H. Stock (eds.), *Identification and Inference for Econometric Models*. Cambridge: Cambridge University Press.
Moreira, M.J., J.R. Porter and G.A. Suarez (2004) Bootstrap and higher-order expansion validity when instruments may be weak. NBER Technical Working Paper No. 302.
Moulton, B.R. (1990) An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* **72**, 334–8.
Nelson, C.R. and R. Startz (1990) The distribution of the instrumental variables estimator and its t-ratio when the instrument is poor one. *Journal of Business* **63**, S125–40.
Newey, W.K. (1985) Maximum likelihood specification testing and conditional moment tests. *Econometrica* **53**, 1047–70.
Newey, W.K. (1990) Semiparametric efficiency bounds, *Journal of Applied Econometrics* **5**, 99–135.
Newey, W.K. and D. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle and D. McFadden (eds.), *Handbook of Econometrics, Volume 4*, pp. 2111–245. Amsterdam: North-Holland.
Newey, W.K. and R.J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–55.
Newey, W.K. and K.D. West (1987) Hypothesis testing with efficient method of moments estimators. *International Economic Review* **28**, 777–87.
Owen, A.B. (1988) Empirical likelihood ratios confidence intervals for a single functional. *Biometrika* 75, 237–49.
Pagan, A.R. and A. Ullah (1999) *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
Pakes, A.S. and D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027–57.
Pesaran, M.H. and M. Weeks (2001) Non-nested hypothesis testing: an overview. In B. Baltagi (ed.), *A Companion to Theoretical Econometrics*, pp. 279–309. Oxford: Blackwell.
Politis, D.N. and J.P. Romano (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Annals of statistics* **22**, 2031–50.
Powell, J.L. (1984) Least squares absolute deviations estimation for the censored regression model, *Journal of Econometrics* **25**, 303–25.
Powell, J.L. (1986) Censored regression quantiles. *Journal of Econometrics* **32**, 143–55.
Powell, J.L., J.H. Stock and T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403–30.
Qin, J. and J. Lawless (1994) Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300–25.
Racine, J.S. and Q. Li (2004) Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**(1), 99–130.
Reiss, P.C. and F.A. Wolak (2007) Structural econometric modeling: rationales and examples from industrial organization. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics, Volume 6A*, pp. 4277–416. Amsterdam: North-Holland.
Robinson, P.M. (1988) Root-N-consistent semiparametric regression. *Econometrica* **56**, 931–54.
Rosenbaum, P. and D.B. Rubin (1983) The central role of propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D.B. (1976) Inference and missing data. *Biometrika* **63**, 581–92.
Rubin, D.B. (1978) Bayesian inference for causal effects. *Annals of Statistics* **6**, 34–58.
Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
Schennach, S. (2004) Estimation of nonlinear models with measurement error. *Econometrica* **72**, 33–75.
Severini, T.A. and G. Tripathi (2001) A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics* **102**, 23–66.
Smith, J. (2000) A critical survey of empirical methods for evaluating active labor market policies. *Schweizerische Zeitschrift fur Volkswirtschaft und Statistik* **136**(3), 1–22.
Staiger, D. and J. Stock (1997) Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–86.
Tauchen, G. (1985) Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* **30**, 415–43.
Train, K.E. (2003) *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
Van den Berg, G. (2001) Duration models: specification, identification, and multiple durations. In J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Volume 5*, pp. 3381–460. Amsterdam: North-Holland.
Wansbeek, T. and E. Meijer (2000) *Measurement Error and Latent Variables in Econometrics*. Amsterdam: North-Holland.
White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–38.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
White, H. (1984) *Asymptotic Theory for Econometricians*. San Diego, Calif.: Academic Press.
Windmeijer, F. (2005) A finite sample correction for the variance of linear two-step GMM estimators. *Journal of Econometrics* **126**, 25–51.
Wooldridge, J.M. (2001) Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory* **17**, 451–70.
Wooldridge, J.M. (2008) *Econometric Analysis of Cross Section and Panel Data* (second edition). Cambridge, Mass.: MIT Press.
Wooldridge, J.M. (2003) Cluster-sample methods in applied econometrics. *American Economic Review* **93**, 133–38.
Wooldridge, J.M. (2005) Unobserved heterogeneity and estimation of average partial effects. In D.W.K. Andrews and J.H. Stock (eds.), *Identification and Inference for Econometric Models*. Cambridge: Cambridge University Press.
Wooldridge, J.M. (2006) Cluster sample methods in applied econometrics: an extended analysis. Department of Economics, Michigan State University.
Wooldridge, J.M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* **141**, 1281–301.

15

Computational Considerations in Empirical Microeconometrics: Selected Examples

David T. Jacho-Chávez and Pravin K. Trivedi

Abstract

The substance and style of modern microeconometrics is shaped by its role in analyses of public policy issues. Computational considerations have proved to be an important influence on the methodology and scope of empirical analyses that address these issues. To be convincing to a wide readership the empirical analyses need to be based on representative data and flexible modeling approaches. In this chapter we illustrate, through a variety of empirical examples, how modelers handle the complexities that arise from the richness of survey data and the heterogeneity in behavior of market participants. After introductory sections on data and programming languages, the remainder of the chapter covers many leading computationally intensive econometric techniques. These are illustrated by means of specific numerical examples. An algorithmic format is used to describe the computational features.

15.1 Introduction

Since the 1980s there has been a huge growth in the availability of both census and survey data, in part due to the expansion of electronic recording and collection of data. As data, computational power, and modeling opportunities have grown, so have the variety and complexity of the modeling objectives of empirical researchers. These developments have created modeling opportunities and challenges that were largely absent when only aggregated market-level data were available.

Explosive growth in the volume and types of data has also given rise to numerous methodological issues. Certain features of micro-data are also ultimately responsible for added computational complexity. Sample survey data, the raw material of microeconometrics, are often subject to problems of sample selection, measurement errors, incomplete and/or missing data, all of which generate additional uncertainty about the econometric specifications used by empirical researchers and impede the empirical generalizations from sample to population. One response to these issues is that empirical researchers often explore several modeling strategies, leading to additional computational complexity.

Model specification, estimation, testing, and then revision are essential components of a modeling cycle. Whether econometric models are intended to be exploratory and descriptive, or whether they aim to quantify structural relationships, computation is central to every step in the modeling cycle. Important features of modern applied econometrics include the following:

- **Disaggregation**. Microeconometrics is about regression-based modeling of economic relationships using data at the levels of individuals, households, and firms. The low level of aggregation in the data has immediate implications for the functional forms used to model the relationships of interest. Disaggregation

brings out heterogeneity of individuals, firms, and organizations. Modeling such heterogeneity is often essential for making valid inferences. While aggregation usually reduces noise and leads to smoothing, disaggregation leads to loss of continuity and smoothness, and increases the range of variation in the data. For example, household average weekly consumption of (say) meat is likely to vary smoothly, while that of an individual household in a given week may frequently be zero, and may also switch to positive values from time to time. As Pudney (1989) has pointed out, micro-data exhibit "holes, kinks and corners." Discreteness and nonlinearity of response are intrinsic to microeconometrics, and they contribute to an increase in computational complexity relative to linear models.

- **Model complexity**. Empirical microeconometrics is closely tied to issues of public policy. Attempts to strengthen the policy relevance of models typically increase the complexity of the models. For example, one may be interested not only in the average impact of a policy change, but also in its distribution. The greater the heterogeneity in response, the greater the relevance of the latter. One feature of such complexity is that potentially models have high dimension. A multiple equation nonlinear regression model with scores of variables is not at all unusual.
- **Restrictions**. Strong functional form and distributional restrictions are not favored. Instead, non- and semiparametric models, or flexible parametric models, are preferred.
- **Robustness**. Models whose conclusions are not very sensitive to modeling assumptions are preferred. Robustness comparisons between competing specifications are an important part of the modeling cycle.
- **Testing and evaluation**. Model testing and evaluation, generally requiring additional computational effort, are integral parts of the search for robust specifications.
- **Asymptotic approximations**. Traditionally, inference about economic relations relied heavily on large sample approximations, but with additional computational effort it is often possible to improve the quality of such approximations. Increasingly, applied econometricians engage in such improvements.
- **Computing advances**. Breathtaking advances in the speed and scope of computing encourage applied econometricians to experiment with computationally more ambitious projects. Avoiding a technique purely because it is computationally challenging is no longer regarded as a serious argument.

In this chapter we selectively and illustratively survey how computational advances have influenced, and in turn are influenced by, the style and direction of modern applied microeconometrics. Although it is fascinating to do so, we do not systematically take a historical perspective on econometric computation. The interested reader is directed to Berndt (1991, pp. 1–10) and the articles by Renfro (2004b) and Slater (2004). To limit the potentially vast scope of this article, we restrict the coverage to a few selected topics whose discussion is illustrated with specific data-based examples. We do not cover the important developments in Bayesian computation

(see, e.g., Chib, 2004; Geweke, 2005). The preliminary issues are tackled in sections 15.2 and 15.3. Section 15.4 discusses the advantages of flexible functional forms generated using non- and semiparametric methods. Section 15.5 illustrates some of the ways in which heterogeneous responses are modeled and how resampling methods are used in the related econometric inference. Section 15.6 illustrates the use of simulation assisted estimation in handling endogenous regressors in a simultaneous nonlinear model of choice and outcomes. Section 15.7 concludes.

15.2 Preliminary

A computer has become the standard tool of the trade for most applied econometricians. It is unimaginable how today's empirical analyses and advances could have been performed without one. However, a byproduct of this matrimony between computing and applied econometrics is today's necessity that empirical researchers should know and understand many definitions and specialized computer science jargon. In this section, we provide a quick introduction to essential concepts, terminology and practical considerations that most researchers come into contact with while working with a computer.

15.2.1 Programming languages

A computer programming language can be thought of as a set of characters, along with rules to combine them into words and symbols, that can be used to express detailed instructions to a computer. Most programming languages used by economists are *high-level languages*, in the sense that they are not generally specific to a computer central processing unit (CPU), but instead they are portable across different computer hardware. These high-level languages can generally be divided into compiled languages and interpreted languages.

Compiled programming languages require a computer program, called *compiler*, to translate human-readable text instructions, called *source code*, into a *low-level language* such as assembly language or machine language. Most *third-generation programming languages* (3GL), such as C, C++ and Fortran, are compiled languages. Computer programs written in these languages need to be compiled first, and then executed.

Interpreted programming languages are implemented by a computer program called *interpreter*. The interpreter executes instructions written in these languages. Popular object-oriented *fourth-generation programming languages* (4GL), such as GAUSS™ (Aptech Systems, Inc., Black Diamond, Washington), MATLAB® (The MathWorks™, Natick, Massachusetts), R (R Development Core Team, Vienna, Austria), and Stata® (StataCorp LP, College Station, Texas) are interpreted languages.

The interested reader is recommended to consult articles by Nerlove (2004) and Renfro (2004a) for a comprehensive list and history of programming languages used in econometrics up to 2004.

15.2.1.1 Characteristics

Researchers are often faced with the dilemma of what programming language to use. Although the final decision is mostly a matter of taste and time constraints, the following should serve as guidelines:

- **Portability**. Source code written in 4GLs requires a local installation of the relevant compiler for their implementation. Since these languages are often proprietary, this could make sharing and replicability inconvenient.[1] On the other hand, after compilation, 3GL's executable files do not require any kind of formal installation onto a computer's permanent storage device to be executed and can be sent to others by e-mail, enabling it to be used on multiple computers.
- **Complexity**. Compiled programming languages are general purpose languages. They often only include primitive mathematical functions and, with the exception of Fortran, they solely provide uniform pseudo-random number generators. On the other hand, interpreted languages are conceived with a specific purpose in mind. They admit a richer set of data types than integer and double precision numbers. Multidimensional arrays are also native objects in some of these languages. Standard econometrics and mathematical techniques are seldom implemented in compiled languages, but they are plentiful in interpreted languages.
- **Expandability**. Both 3GL and 4GL can readily be expanded by the inclusion of *libraries* and *packages*. Libraries are collections of special purpose algorithms. Examples include the IMSL libraries of Visual Numerics Inc. for C and Fortran, and NAG Fortran Library subroutines of the Numerical Analysis Group. Packages are collections of programs linked together. Examples include packages in R, and modules in Stata.
- **Efficiency**. A main disadvantage of interpreted languages is that applications run slower than if they had been compiled, especially iterative algorithms that require loops iterated many times, e.g., bootstrapping and cross-validation. This is because the interpreter must analyze each statement in the source code each time it is executed, and then perform the desired action, whereas the compiled code simply performs the action. Furthermore, 3GL such as C and Fortran permit better control of dynamic memory allocation that can potentially speed up many implementations.[2]

However, it is no longer the case that researchers must choose between 3GL and 4GL when facing a programming task. Modern computing environments allow the inclusion of C or Fortran subroutines, via *foreign language interface*, into 4GL such as GAUSS, MATLAB, R, Stata and vice versa, as well as the possibility of *parallelization*.

15.2.2 Foreign language interface

Parts or chunks of source code written in an interpreted language can be rewritten in C or Fortran, and then compiled into a library (*dynamic link library* in MS Windows®, or *share object* in *NIX[3] platforms). This compiled library can then be

dynamically loaded by the original program to speed up overall execution. Furthermore, since some interpreted languages are written in C, it is not surprising that native C code can make use of libraries in local installations of these interpreted languages' compilers.[4]

15.2.3 Parallelization

Standard 4GL performs *serial* computations, i.e., instructions are executed one after another on a single computer having a single CPU. However, the advent of *clusters*,[5] workstations and single computers with multiple processors have made parallel computing a tangible possibility to most economists. *Parallel computing* is the simultaneous use of more than one CPU to execute a program or solve a computational problem. The problem is broken into parts that can be solved concurrently. Each part is further broken down to a series of instructions that are executed simultaneously on different CPUs. This is achieved by using a language-independent communications protocol known as *Message Passing Interface* (MPI).

Parallel computing means that computations that would otherwise take hours or days could be performed in minutes or hours. Computational algorithms that allow parallelization commonly involve iterative loops that can be performed independently of each other. Examples include most resampling methods, such as the bootstrap, or set search algorithms often used in cross-validation methods in non/semiparametric techniques (see section 15.4.1.1 for an illustration). Creel (2005) discusses many other econometric examples that allow for parallel computation, such as Monte Carlo simulation, maximum likelihood (ML) and generalized method of moments (GMM) estimation.

Most 4GL used in econometrics either allow the parallelization of procedures or make full use of multiprocessor computers. Examples include `MatlabMPI` in MATLAB, `Rmpi`, and `Snow` in R for the former, and Stata/MP® for the latter.

15.3 Computing and modeling

There are many ways in which more memory and computing power can potentially improve the quality of empirical analysis in microeconometrics.

The first context is that of the storage and manipulation of large complex datasets, and in providing numerical, graphical and visual displays of data in ways that provide valuable insights into the pattern and structure within such datasets.

The second context is that of solving (especially) high-dimensional optimization problems that arise in model estimation. Estimation of many standard microeconometric models involves solution of nonlinear equations by iterative methods, which are generically referred to as optimizers. Efficient optimizers that can handle high-dimensional problems are essential in microeconometric modeling.

Increasingly, computer-intensive methods such as Monte Carlo simulators are an important tool for studying the finite sample properties of estimators and tests. Simulation is also an essential component in estimating model parameters, as in the case of simulation-assisted optimization and Markov chain Monte Carlo (MCMC) methods used in Bayesian modeling.

Fourth, computer-intensive resampling methods, such as the bootstrap and jackknife, are increasingly used as substitutes for analytically complex computations such as sample estimates of asymptotic variances.

Finally, cross-validation is another computer-intensive tool that is useful not only for parameter tuning (as illustrated in section 15.4) but also for model evaluation and comparison.

In the remainder of this section we define and outline each type of application.

15.3.1 Data summary and visualization

A starting point of almost any empirical microeconometric application involves providing data summaries. The most common manifestation of this takes the form of a table of sample moments, such as means, variances, skewness and kurtosis. However, visual data summaries, such as histograms and kernel (marginal) density plots, are often a more efficient way of providing information (see, e.g., Huynh and Jacho-Chávez, 2007). The kernel density estimator is a generalization of the histogram estimate using *kernel* weights, $k(\cdot)$, that integrate to 1. These weights depend on a smoothing parameter, h, called the *bandwidth*, and $2h$ is the *window width* (see section 15.4.1 for definitions and examples). Given $k(\cdot)$ and h the estimator is easy to implement, and if the estimator is evaluated at r distinct values, then computation of the kernel estimator requires at most nr operations when the kernel has unbounded support. However, as is evident even in the most elementary computation of a histogram, constructing such visual displays involves the choice of bin size, and the results may be sensitive to the choice of h. This motivates the search for another estimator that treats h like an unknown parameter, i.e., cross-validation. Such a method is inherently more computer-intensive, as will be shown in section 15.4. An extension of this concept, also considered in the next section, is the estimator of the conditional probability density function, which also involves considerations similar to those in the estimation of marginal densities.

15.3.2 Numerical optimization

Microeconometrics frequently employs an estimator $\widehat{\boldsymbol{\theta}}$ that maximizes a stochastic objective function $Q_n(\boldsymbol{\theta})$, where usually $\widehat{\boldsymbol{\theta}}$ solves the first-order conditions $\partial Q_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$; n being the sample size. The objective function may be a likelihood for parametric models, a weighted sum of squares function for semiparametric models, or a linear function subject to inequality restrictions when the objective function has an L_1 (e.g., least absolute deviations) rather than an L_2 (e.g., sum of squared residuals) norm. For many nonlinear models there is no closed-form solution of the first-order conditions, only a nonlinear system of equations in the unknown $\boldsymbol{\theta}$. Estimation algorithms use iterative methods to solve the first-order conditions. Iterative methods involve an updating rule for obtaining a new estimate, $\widehat{\boldsymbol{\theta}}_{s+1}$, given a current estimate $\widehat{\boldsymbol{\theta}}_s$. Historically, iterative procedures constituted a computational challenge, but now they are standard. When the objective function is in the L_2 norm, gradient methods are most common. Non-gradient

methods are used in the L_1 norm case, quantile regression (QR) being a leading example.

Gradient methods use an iterative updating rule:

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \mathbf{A}_s \mathbf{g}_s, \qquad s = 1, \ldots, S, \tag{15.1}$$

where $\mathbf{A}_s = \mathbf{A}(\widehat{\boldsymbol{\theta}}_s)$ is a $q \times q$ matrix, and $\mathbf{g}_s = \partial Q_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\big|_{\widehat{\boldsymbol{\theta}}_s}$ is the $q \times 1$ gradient vector evaluated at $\widehat{\boldsymbol{\theta}}_s$. The iterative process continues until it converges. Convergence is usually defined in terms of an arbitrarily small value of either $|Q_n(\widehat{\boldsymbol{\theta}}_{s+1}) - Q_n(\widehat{\boldsymbol{\theta}}_s)|$ or $|\widehat{\boldsymbol{\theta}}_{s+1} - \widehat{\boldsymbol{\theta}}_s|$. Gradient methods differ by their choice of matrix $\mathbf{A}_s$, as summarized in Table 15.1. Some methods, most notably Newton–Raphson and the method of scoring, use second derivatives of the objective function, whereas others, most notably BHHH (Berndt–Hall–Hall–Hausman), is based on the gradient function only. The derivatives may be computed using either the analytical expressions, which are then programmed in the algorithm, or numerical derivatives, which is often a default option. The numerical derivatives are computed using:

$$\frac{\Delta Q_n(\widehat{\boldsymbol{\theta}}_s)}{\Delta \theta_j} = \frac{Q_n(\widehat{\boldsymbol{\theta}}_s + \tau \mathbf{e}_j) - Q_n(\widehat{\boldsymbol{\theta}}_s - \tau \mathbf{e}_j)}{2\tau}, \quad j = 1, \ldots, q, \tag{15.2}$$

Table 15.1 Some standard gradient-based iteration methods

$\mathbf{A}(\widehat{\boldsymbol{\theta}}_s)$	*Reference*
$-\left.\dfrac{\partial^2 Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right\vert_{\widehat{\boldsymbol{\theta}}_s}$	Newton–Raphson
$-\mathrm{E}\left[\dfrac{\partial^2 Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]\Bigg\vert_{\widehat{\boldsymbol{\theta}}_s}$	Method of scoring
$-\sum_{i=1}^n \dfrac{\partial q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \dfrac{\partial q_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big\vert_{\widehat{\boldsymbol{\theta}}_s}$	Berndt–Hall–Hall–Hausman (BHHH)
$\mathbf{I}_q$	Steepest descent
$\mathbf{A}_s = \mathbf{A}_{s-1} + \dfrac{\delta_{s-1}\delta_{s-1}^\top}{\delta_{s-1}^\top \gamma_{s-1}} + \dfrac{\mathbf{A}_{s-1}\gamma_{s-1}^\top \gamma_{s-1}\mathbf{A}_{s-1}}{\gamma_{s-1}^\top \mathbf{A}_{s-1}\gamma_{s-1}}$, where $\delta_{s-1} = \mathbf{A}_{s-1}\mathbf{g}_{s-1}$; $\gamma_{s-1} = \mathbf{g}_s - \mathbf{g}_{s-1}$	Davidon–Fletcher–Powell (DFP)
$\mathbf{A}_s = \mathbf{A}_{s-1} + \dfrac{\delta_{s-1}\delta_{s-1}^\top}{\delta_{s-1}^\top \gamma_{s-1}} + \dfrac{\mathbf{A}_{s-1}\gamma_{s-1}^\top \gamma_{s-1}\mathbf{A}_{s-1}}{\gamma_{s-1}^\top \mathbf{A}_{s-1}\gamma_{s-1}}$ $-(\gamma_{s-1}^\top \mathbf{A}_{s-1}\gamma_{s-1})\eta_{s-1}\eta_{s-1}^\top$, where $\eta_{s-1} = (\delta_{s-1}/\delta_{s-1}^\top \gamma_{s-1}) - (\mathbf{A}_{s-1}\gamma_{s-1}/\gamma_{s-1}^\top \mathbf{A}_{s-1}\gamma_{s-1})$	Boyden–Fletcher–Goldfarb–Shannon (BFGS)

where τ is small and $\mathbf{e}_j = [0, \ldots, 0, 1, 0, \ldots, 0]^\top$ is a K-vector with unity in the jth row and zeros elsewhere. In theory, τ should be chosen such that $\partial Q_n(\boldsymbol{\theta})/\partial\theta_j = \lim_{\tau\to 0} \Delta Q_n(\boldsymbol{\theta})/\Delta\theta_j$. Although, in principle, analytical derivatives are preferred, numerical approximations produce virtually identical results in many regular cases.

Whether the application of a gradient based optimizer turns out to be computationally demanding depends upon many factors, including the following: (i) dimension of $\boldsymbol{\theta}$; (ii) whether the objective function is well approximated by a quadratic expansion in the neighborhood of the optimum; (iii) whether $\boldsymbol{\theta}$ is robustly identified in the sample; and (iv) whether the iterative algorithm starts sufficiently close to the optimum. Conversely, convergence can be slow when the dimension of $\boldsymbol{\theta}$ is high, some components of $\boldsymbol{\theta}$ are weakly identified in the data, and/or the objective function admits multiple maxima or a flat region around the maximum, and the starting values for the update equation are poor. If an objective function is difficult to maximize, e.g., due to multiple local optima, non-gradient methods may be used, such as simulated annealing and genetic algorithms. A leading example is kernel-based cross-validation (see, e.g., algorithms 15.4.1.1.1 and 15.4.1.2.1 below), where Powell's direction set search algorithm is also a viable alternative (see Press *et al.*, 1992, section 10.5, pp. 412–20).

Another example of a case where a gradient-based method is inappropriate is the least absolute deviation (LAD) regression or the quantile regression. In the case of LAD regression, the objective function $Q_n(\boldsymbol{\theta}) = n^{-1}\sum_i |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|$ has no derivative. In the case of quantile regression, the objective function is minimized over $\boldsymbol{\beta}_q$; i.e.:

$$Q_n(\boldsymbol{\beta}_q) = \sum_{i:\, y_i \geq \mathbf{x}_i^\top \beta}^{n} q|y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_q| + \sum_{i:\, y_i < \mathbf{x}_i^\top \beta}^{n} (1-q)|y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_q| \tag{15.3}$$

$$= \sum \rho_q(u_q), \tag{15.4}$$

where $0 < q < 1$, $\rho_q(\lambda) = (q - \mathbb{I}(q < 0))\lambda$ denotes the check function, $\mathbb{I}(\cdot)$ represents the indicator function that equals 1 if its argument is true and 0 otherwise, and the notation $\boldsymbol{\beta}_q$ emphasizes that, for different choices of q, different values of $\boldsymbol{\beta}$ are obtained.

The estimator defined by the minimand $\min_{\boldsymbol{\beta}_q} Q_n(\boldsymbol{\beta}_q)$ is an M-estimator and, as such, its asymptotic properties are well established (see Amemiya, 1985). The optimization problem has an interpretation in the GMM framework as well as in a linear programming (LP) framework (see Buchinsky, 1995). To see the LP representation, the QR is written thus:

$$\begin{aligned} y_i &= \mathbf{x}_i^\top \boldsymbol{\beta}_q + u_{iq} \\ &= \mathbf{x}_i^\top (\boldsymbol{\beta}_q^{[1]} - \boldsymbol{\beta}_q^{[2]}) + (\varepsilon_{iq}^{[1]} - \varepsilon_{iq}^{[2]}), \end{aligned}$$

where $\beta_{q,j}^{[1]} \geq 0$, $\beta_{q,j}^{[2]} \geq 0$, $j = 1, \ldots, K$, and $\varepsilon_{iq}^{[1]} \geq 0$, $\varepsilon_{iq}^{[2]} \geq 0$, $i = 1, \ldots, n$. The optimization problem can be expressed as that of minimizing a linear objective

function subject to linear equality constraints:

$$\min_{\mathbf{c}^\top \mathbf{z}} \mathbf{c}^\top \mathbf{z}, \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y},\ \mathbf{z} \geq \mathbf{0},$$

where $\mathbf{A} = [\mathbf{X}, -\mathbf{X}, \mathbf{I}_n, -\mathbf{I}_n]$, $\mathbf{y} = [y_1, \ldots, y_n]^\top$, $\mathbf{z} = [\boldsymbol{\beta}_q^{[1]}, \boldsymbol{\beta}_q^{[2]}, \varepsilon_q^{[1]}, \varepsilon_q^{[2]}]^\top$, $\mathbf{c}^\top = [\mathbf{0}^\top, \mathbf{0}^\top, q\iota^\top, (1-q)\iota^\top]$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$, $\mathbf{0}$ is a vector of zeros, ι is a vector of 1s, and $\mathbf{I}_n$ is the identity matrix of order n.

The classic method for solving the linear program is the simplex method, which is guaranteed to yield a solution in a finite number of simplex iterations. When n is of the order of several thousands, an efficient computational algorithm is essential. A number of alternative algorithms have been proposed, including a computationally efficient one due to Barrodale and Roberts (1973), thus making QR a suitable method for practical application in large samples. Point estimation is actually a lesser computational challenge than the calculation of variances. Bootstrap methods for variance estimation, often used in preference to analytical expressions, contribute to computational intensity; an example is given in section 15.2.

15.3.3 Simulation-assisted estimation

The role of simulation in theoretical econometrics is well established (see Doornik, 2006). Its use in empirical microeconometrics is more recent but growing rapidly (see McFadden and Ruud, 1994; Gouriéroux and Monfort, 1996). First, MCMC methodology, which is simulation-based, is now standard in modern Bayesian estimation. Second, estimation of several leading microeconometric models, e.g., the multinomial probit (MNP), involve calculation of probability integrals that can be efficiently estimated using simulation methods. Third, empirical microeconometric models often include a latent variable to capture the effects of unobserved heterogeneity (UH). These too lead naturally to simulation-based estimation. We consider two leading applications of simulation-based estimation.

15.3.3.1 MNP example

Consider the MNP model with m choices and with the utility of the jth choice given by:

$$U_j = V_j + \varepsilon_j, \qquad j = 1, 2, \ldots, m, \tag{15.5}$$

where the errors $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_m]^\top$ are joint normally distributed, $\boldsymbol{\varepsilon} \sim \mathcal{N}[\mathbf{0}, \Sigma]$. Usually the linear specifications $V_j = \mathbf{x}_j^\top \boldsymbol{\beta}$ or $V_j = \mathbf{x}^\top \boldsymbol{\beta}_j$ are used. This is an additive random utility model. The covariance matrix Σ is subject to normalization and identification restrictions because U_j is latent. The standard practice is to choose U_1 as the benchmark alternative and place one restriction on Σ.

In estimating the model by ML, a problem is that there is no closed form expression for the choice probabilities. For an m-choice MNP model the choice probabilities are $(m-1)$-fold integrals, e.g.:

$$p_1 = \Pr[y = 1] = \int_{-\infty}^{-\widetilde{V}_{m1}} \cdots \int_{-\infty}^{-\widetilde{V}_{21}} f(\widetilde{\varepsilon}_{21}, \ldots \widetilde{\varepsilon}_{m1}) d\widetilde{\varepsilon}_{21} \ldots d\widetilde{\varepsilon}_{m1},$$

where the $\sim$ denotes normalization relative to the first alternative. When $m > 3$, this integral is difficult to evaluate numerically. An alternative is to use simulation methods. One well-known simulator is the GHK simulator due to Geweke (1992), Hajivassiliou, McFadden and Ruud (1994) and Keane (1994) (see Train, 2003, for details). In this context the maximum simulated likelihood (MSL) estimator maximizes:

$$\widehat{\mathcal{L}}_n(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij},$$

where the $\widehat{p}_{ij}$ are obtained using the GHK or other simulator. For consistent estimation we require that the number of draws in the simulator $S \to \infty$ as well as $n \to \infty$. Because the method is computationally burdensome, a great deal of work has been done on improving the simulator and on developing other models that are good substitute specifications but are easier to estimate (e.g., McFadden and Train, 2000).

15.3.3.2 Heterogeneity example

Suppose the conditional density $h(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i)$ for an observation involves a continuous separable UH term $\mathbf{u}_i$, assumed to be independent of $\mathbf{x}_i$, with density $g(\mathbf{u}_i)$. Then the marginal density is defined by the integral:

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int h(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i) g(\mathbf{u}_i) d\mathbf{u}_i, \tag{15.6}$$

which needs to be estimated numerically if there is no closed form solution.

An unbiased and consistent estimator $\widehat{f}_i$ of f_i is the direct Monte Carlo simulator:

$$\widehat{f}(y_i|\mathbf{x}_i, \mathbf{u}_{iS}, \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} h(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s), \tag{15.7}$$

which averages $h(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_i^s)$ over the S draws, where $\mathbf{u}_i^s$, $s = 1, \ldots, S$, are independent draws from $g(\mathbf{u}_i)$. Estimators superior to this direct estimator are also available.

The MSL estimator $\widehat{\boldsymbol{\theta}}_{\text{MSL}}$ maximizes:

$$\widehat{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{u}_{iS}). \tag{15.8}$$

If $\widehat{f}(\cdot)$ is differentiable in $\boldsymbol{\theta}$ then $\widehat{\boldsymbol{\theta}}_{\text{MSL}}$ can be computed using the standard gradient methods mentioned above. This MSL estimator is asymptotically equivalent to the ML estimator if $S, n \to \infty$ and $\sqrt{n}/S \to 0$, and has a limit normal distribution.

The MSL method is one of the simplest simulation-based estimators. Just as the MSL method parallels ML estimation, the method of simulated moments (MSM) parallels the method of moments. For space reasons we do not elaborate on the distinctions between these methods.

A direct application of the MSL method sometimes leads to very slow convergence of the simulated likelihood function. Simulation-acceleration techniques are

available. An example is the method that uses quasi-random draws based on Halton sequences described in Bhat (2001) and Train (2003).[6] Halton sequences have two desirable properties vis-á-vis the standard pseudo-random draws. First, they are designed to give more even coverage over the domain of the mixing distribution. Second, the simulated probabilities are negatively correlated over observations. This negative correlation reduces the variance in the simulated likelihood function. Under suitable regularity conditions (see Bhat, 2001), the integration error using pseudo-random sequences is in the order of n^{-1} as compared to pseudo-random sequences where the convergence rate is $n^{-1/2}$. An example of an MSL estimator using Halton quasi-random draws is given in section 15.6.

15.3.4 Resampling methods

Empirical microeconometrics depends heavily on asymptotic theory to justify point and interval estimation procedures because exact results are generally not available. Sometimes an investigator may be interested in some function of estimates and data for which even the asymptotic result may not be available. In some cases only an approximation to the asymptotic approximation may be available. The availability of an asymptotic approximation is no guarantee that it provides a good approximation to the sampling distribution of an estimator. Motivated by these difficulties, econometricians increasingly use computer-intensive resampling methods to obtain estimates of the moments of the asymptotic distribution. The bootstrap and jackknife are two leading examples, but only the first is sketched here.

There exists a wide range of bootstrap methods (see, e.g., Davidson and MacKinnon, 2006). The simplest bootstrap methods can support statistical inference when conventional methods, such as variance estimation, are difficult to implement, either because a formula is not available or because it is computationally intractable. Another, more complicated, bootstrap attempts to provide asymptotic refinements that can lead to an improvement over the asymptotic results. Applied researchers are most often interested in the first, simpler, application of the bootstrap.

The basic idea behind the bootstrap is to approximate the distribution of a statistic by a simulation in which one samples from the empirical or the fitted distribution of the data. That is, one uses a given sample repeatedly to derive the sampling properties of statistics of interest. Bootstrap methods rely on asymptotic theory for their validity.

Consider, in the context of the regression of $y_i|\mathbf{x}_i$, $i = 1, \ldots, n$, the problem of inference on a parameter $\boldsymbol{\theta}$, $\boldsymbol{\theta} = \phi(\boldsymbol{\beta})$, where $\phi(\boldsymbol{\beta})$ is a continuous function of the regression parameters $\boldsymbol{\beta}$. The bootstrap algorithm for the variance of $\boldsymbol{\theta}$ is explained in algorithm 15.3.4.0.1.

A computationally simpler alternative to the bootstrap variance is an estimate obtained by the so-called delta method, based on a first-order Taylor approximation of $\phi(\boldsymbol{\beta})$. In some applications this method can yield very poor results, whereas the bootstrap may be more robust.

Algorithm 15.3.4.0.1 Bootstrap variance estimation – implementation

1. Given data $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, draw (with replacement) a bootstrap sample of size n, denoted $(\mathbf{y}_1^*, \mathbf{x}_1^*), \ldots, (\mathbf{y}_n^*, \mathbf{x}_n^*)$.
2. Calculate the estimate $\widehat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}$.
3. Repeat steps 1–2 B independent times, where B is a large number, obtaining B bootstrap replications of the statistic of interest, such as $\widehat{\boldsymbol{\theta}}_1^*, \ldots, \widehat{\boldsymbol{\theta}}_B^*$.
4. Use these B bootstrap replications to obtain the bootstrap variance $\mathrm{V}_B(\widehat{\boldsymbol{\theta}}) = B^{-1} \sum (\widehat{\boldsymbol{\theta}}_j^* - \overline{\widehat{\boldsymbol{\theta}}^*})^2$, where $\overline{\widehat{\boldsymbol{\theta}}^*} = B^{-1} \sum \widehat{\boldsymbol{\theta}}_j^*$. (It is assumed that the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ exists.)

In implementing a bootstrap, details generally vary with the specific application. For example, bootstrap samples may be drawn differently, the value of B may vary, and the target statistic of interest may or may not involve asymptotic refinement.

Section 15.5.1 provides an example that compares standard errors of quantile regression parameters obtained using an analytical formula with those from a bootstrap.

15.3.5 Structural models based on dynamic programming

Dynamic programming (DP) models represent a relatively new strand in structural microeconometric models. The special appeal of the approach comes from the potential of this class of models to address issues relating to new policies or old policies in a new environment. Further, the models are dynamic in the sense that they can incorporate expectational factors and intertemporal dependence between decisions. From a computational viewpoint these models are an order of magnitude more complex than most other methods. What follows is merely a bare-bones sketch of the approach.

The DP approach is rooted in detailed structural specifications derived from strong theoretical specifications and contrasts sharply with the looser "atheoretical" models. The distinctive characteristics of this approach include: a close integration with underlying theory; an assumption of rational optimizing agents; extensive use of assumptions and restrictions necessary to support the close integration of the empirical model with the underlying theory; a high level of parameterization of the model; concentration on causal parameters that play a key role in policy simulation and evaluation; and an approach to the estimation of model parameters that is substantially different from the standard approaches used in estimating moment condition models.

There are many studies that follow the dynamic programming approach. Representative examples are Rust (1987); Hotz and Miller (1993); Keane and Wolpin (1994). Some key features of DP models can be studied using a model due to Rust and Phelan (1997), which provides an empirical analysis of how the incentives and constraints of the US social security and Medicare insurance system affects the labor supply of older workers.

The main components of the DP model are as follows. State variables are denoted by s_t, control variables by d_t. β is the intertemporal discount factor. In implementation all continuous state variables are discretized – a step which greatly expands the dimension of the problem. Hence all continuous choices become discrete choices, d_t is a discrete choice sequence, and the choice set is finite. There is a single period utility function $u_t(s_t, d_t, \boldsymbol{\theta}_u)$ and $p_t(s_{t+1}|s_t, d_t, \boldsymbol{\theta}_p, \alpha)$ denotes the probability density of transitions from s_t to s_{t+1}. The optimal decision sequence is denoted by $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_T)^\top$, where $d_t = \delta_t(s_t)$, and is the optimal solution that maximizes the expected discounted utility:

$$V_t(s) = \max_{\boldsymbol{\delta}} E_{\boldsymbol{\delta}}\{\textstyle\sum_{j=t}^{T} \beta^{j-t} u_j(s_j, d_j, \boldsymbol{\theta}_u)|s_t = s\}. \tag{15.9}$$

Estimation of the model typically uses the likelihood function:

$$\begin{aligned} L(\boldsymbol{\theta}) &= L(\beta, \boldsymbol{\theta}_u, \boldsymbol{\theta}_p) \\ &= \textstyle\prod_{i=1}^{n} \prod_{t=1}^{T} P_t(d_t^i|x_t^i, \boldsymbol{\theta}_u) p_t(x_t^i|x_{t-1}^i, d_{t-1}^i, \boldsymbol{\theta}_p), \end{aligned} \tag{15.10}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_u^\top, \boldsymbol{\theta}_p^\top)^\top$, and assumptions are invoked to separate the parameters in the transition probability $p_t(s_{t+1}|s_t, d_t, \boldsymbol{\theta}_p, \alpha)$ from those in the utility function $u_t(s_t, d_t, \boldsymbol{\theta}_u)$.

DP models are typically high dimensional. To handle the dimensionality issue a popular estimation strategy has two steps: (1) estimate $\boldsymbol{\theta}_p$ using a first-stage partial likelihood function involving only products of the p_t terms; (2) estimate $\boldsymbol{\theta}_p$ by solving the DP problem numerically, using a nested fixed point (NFXP) algorithm, applied to the partial likelihood function consisting of only products of p_t; (see, e.g., Rust, 1986, 1987, 1994, 1997). Given the enormous computational burden of a high-dimensional dynamic discrete choice model, a great deal of recent research seeks ways of making it more manageable; e.g., Aguirregabiria and Mira (2002) provide survey methods which avoid repeated full solution of the structural model in estimation, and the application of simulation and approximation methods (see, e.g., Aguirregabiria and Mira, 2007).

15.4 Non/semiparametric methods

In many situations of interest in economics, we are interested in identifying and estimating particular aspects of the joint distribution of a random vector $[y, \mathbf{x}^\top]$ based on a sample $\{y_i, \mathbf{x}_i^\top\}$, $i = 1, \ldots, n$, where $y \in \mathbb{R}$ and $\mathbf{x}$ is a mixture of continuous variables, $\mathbf{x}^c = [x^1, \ldots, x^{q_1}] \in \mathbb{R}^{q_1}$, and discrete, $\mathbf{x}^d = \left[x^{q_1+1}, \ldots, x^q\right]^\top \in S^d$, where S^d is the support of $\mathbf{x}^d$, and $q_2 = q - q_1$. For example, our aim could be to analyze the relationship between y and $\mathbf{x}$ encapsulated in the conditional mean function $E[y|\mathbf{x}] = m(\mathbf{x})$, the conditional density function $f(y|\mathbf{x})$, or on a finite-dimensional vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^q$, leaving other aspects of the distribution unspecified. The available methods, such as kernels, rely on averaging observations which are closer to those we want to make inference about. A weighting function,

called a kernel, provides the necessary weights, and a smoothing parameter, known as the bandwidth, defines how close observations are to each other. A very comprehensive introduction to such methods can be found in Pagan and Ullah (1999), Li and Racine (2007), Yatchew (2003) and Racine and Ullah (2006). Kernel methods are by nature computationally intensive. The number of operations necessary for their application grows exponentially with the number of data points and variables used. In this section we illustrate with real data the computational issues arising from the implementation of kernel smoothing in applied non/semiparametric analysis of cross section information. Unless otherwise stated, all calculations were performed using the `np` and `gam` libraries of Hayfield and Racine (2007) and Hastie (2006) respectively, written in `R` version 2.4.1 or above. We use a Pentium IV (HT) processor, running at 3.20 GHz.

15.4.1 Nonparametric estimation

For two particular points, $\mathbf{x}_i = [\mathbf{x}_i^c, \mathbf{x}_i^d]$ and $\mathbf{x}_j = [\mathbf{x}_j^c, \mathbf{x}_j^d]$, let us define the functions:

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) = \prod_{l=1}^{q_1} \frac{1}{h_l} k\left(\frac{x_i^l - x_j^l}{h_l}\right), \tag{15.11}$$

$$L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda) = \prod_{l=1}^{q_2} l(x_i^l, x_j^l; \lambda_l), \tag{15.12}$$

where i indexes the "estimation data" and j the "evaluation data," which are typically the same. The kernel function $k(\cdot)$ for continuous variables satisfies $\int k(u)\, du = 1$ and some other regularity conditions depending on its order, p, and $h = [h_1, \ldots, h_{q_1}]^\top$ is a vector of smoothing parameters satisfying $h_s \to 0$ as $n \to \infty$ for $s = 1, \ldots, q_1$. Similarly, the kernel function $l(\cdot)$ for discrete variables lies between 0 and 1, and $\lambda = \left[\lambda_1, \ldots, \lambda_{q_2}\right]^\top$ is a vector of smoothing parameters such that $\lambda_s \in [0, 1]$, and $\lambda_s \to 0$ as $n \to \infty$ for $s = 1, \ldots, q_2$ (see, e.g., Li and Racine, 2003).

Functions (15.11) and (15.12) are the building blocks of kernel smoothing. For example, the Nadaraya–Watson estimator of $m(\cdot)$ evaluated at $\mathbf{x}_j$ is $\widehat{m}(\mathbf{x}_j) = \sum_{i=1, i\neq j}^n y_i K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda) / \sum_{i=1, i\neq j}^n K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) \times L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)$.

15.4.1.1 Example: kernel density estimation

The marginal density function of $\mathbf{x}$ can be estimated at evaluation point $\mathbf{x}_i$ as:

$$\widehat{f}(\mathbf{x}_j) = (n-1)^{-1} \sum_{i=1}^n K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) \times L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda). \tag{15.13}$$

We are interested in estimating the density of the average annual earnings in 1988 (measured in 1982 US dollars) of a random sample from two age groups: 19–26 (1,109 observations) and 30–32 (1,479 observations). This sample is a sub-set of a larger dataset considered in Mills and Zandvakili (1997). In this case, $q_1 = 1$ and $q_2 = 0$, and algorithm 15.4.1.1.1 illustrates the necessary steps.

Algorithm 15.4.1.1.1 Conditional density estimation – implementation

1. Select function $k(\cdot)$ in (15.11), and choose the smoothing parameter h by cross-validation maximum likelihood, i.e., select h to minimize:
$$\mathcal{L}_{[h]} = \sum_{i=1}^{n} \log \widehat{f}_{-i}(\mathbf{x}_i^c),$$
where $\widehat{f}_{-i}(\cdot)$ equals (15.13) after replacing $\sum_{i=1}^{n}$ by $\sum_{i=1;\, i\neq j}^{n}$.
2. Using h, found in the previous step, calculate $\widehat{f}(\mathbf{x}_i^c)$ in (15.13) for each $i = 1, \ldots, n$.
3. Plot (if feasible) or present summary statistics.

Parallelization

Since the above algorithm relies on local averaging, the computational order of step 2 is $O\left(n^2 q\right)$. Data-driven methods, such as bandwidth selection by cross-validation, performed in step 1, add an additional order of computational magnitude. The burden increases as the amount of available data rises. These numerical demands can potentially overwhelm the computational resources of modern day desktop workstations. For these reasons, although approximations are available (see, e.g., Silverman, 1982; Scott and Sheather, 1985) parallelization, as discussed in section 15.2.3, is an attractive alternative (see, e.g., Racine, 2002).

We proceed to implement algorithm 15.4.1.1.1 for each dataset using a high-performance computer cluster. Each implementation uses 10 multiple starts to numerically minimize $\mathcal{L}_{[h]}$ in step 1, and 1,000 bootstrap replications of step 2 to calculate a 95% confidence interval. The first experiment uses a single node with its 2×2.0 GHz quad-core Intel Xeon 5335 processor, i.e., two processors, and takes 27 minutes and 19 seconds of CPU time. The second experiment uses eight nodes with two of the above processors each, i.e., 16 processors in total, and takes 2 minutes and 52 seconds for completion. Computational time is reduced roughly 10 times by parallelization.[7] Both experiments provide the exact same result, i.e., Figure 15.1.

Both distributions are left-skewed, but earnings of young individuals show a larger right tail. Unlike their older counterparts, the distribution of average annual earnings for 19–26-year-old individuals seems to be a mixture of at least three heterogeneous groups.

15.4.1.2 Example: conditional density estimation

To illustrate the computational intensity of fully nonparametric methods, we consider the estimation of the conditional probability density function (p.d.f.) of vector $\mathbf{x}^c$ given another vector $\mathbf{x}^d$, i.e.:

$$\widehat{f}(\mathbf{x}_i^c | \mathbf{x}_i^d) \equiv \widehat{f}(\mathbf{x}_i^c, \mathbf{x}_i^d) / \widehat{p}(\mathbf{x}_i^d), \tag{15.14}$$
$$= \frac{\sum_{j=1}^{n} K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{j=1}^{n} L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}.$$

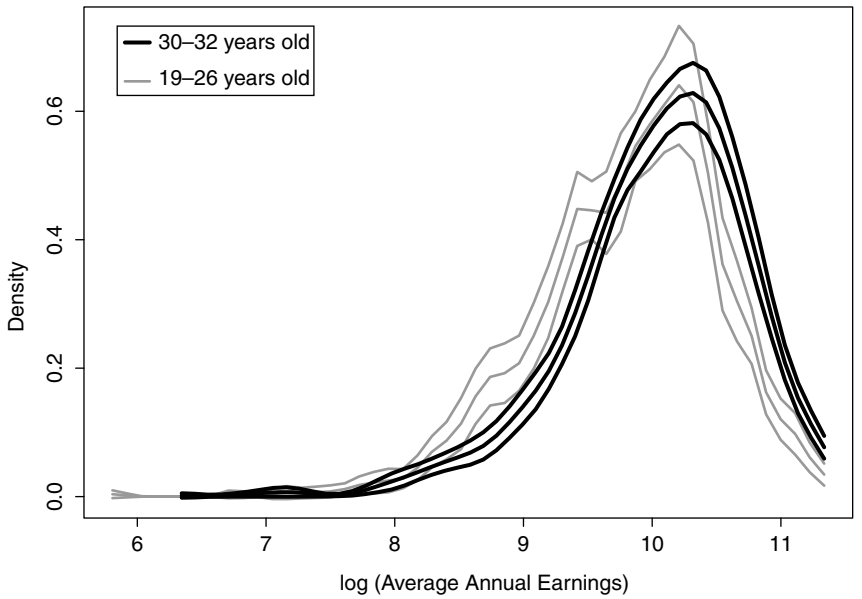


Figure 15.1 Marginal probability density functions

Algorithm 15.4.1.2.1 implements the necessary steps.

Algorithm 15.4.1.2.1 Conditional density estimation – implementation

1. Select functions $k(\cdot)$ and $l(\cdot)$ in (15.11) and (15.12), and vectors of smoothing parameters h and λ by cross-validation maximum likelihood, i.e., select $[h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{q_2}]$ to minimize:

$$\mathcal{L}_{[h,\lambda]} = \sum_{j=1}^{n} \log \widehat{f}_{-i}(\mathbf{x}_i^c | \mathbf{x}_i^d),$$

   where $\widehat{f}_{-i}(\cdot)$ equals (15.14) after replacing $\sum_{j=1}^{n}$ by $\sum_{j=1;\, j\neq i}^{n}$.
2. Using h and λ found in the previous step, calculate $\widehat{f}(\mathbf{x}_i^c | \mathbf{x}_i^d)$ in (15.14) for each $i = 1, \ldots, n$.
3. Plot (if feasible) or present summary statistics.

Using 453 observations of individual transportation mode choices in Croissant (2006) (dataset `Mode`), we estimate the conditional joint density of preferred transportation mode's cost, x_1, and time, x_2, given observed choices: car ($x^d = 1$), carpool ($x^d = 2$), bus ($x^d = 3$), and rail ($x^d = 4$). The results are shown in Figure 15.2. The estimated joint p.d.f.s are multi-modal. This might indicate that other characteristics besides cost influence transportation mode choices.

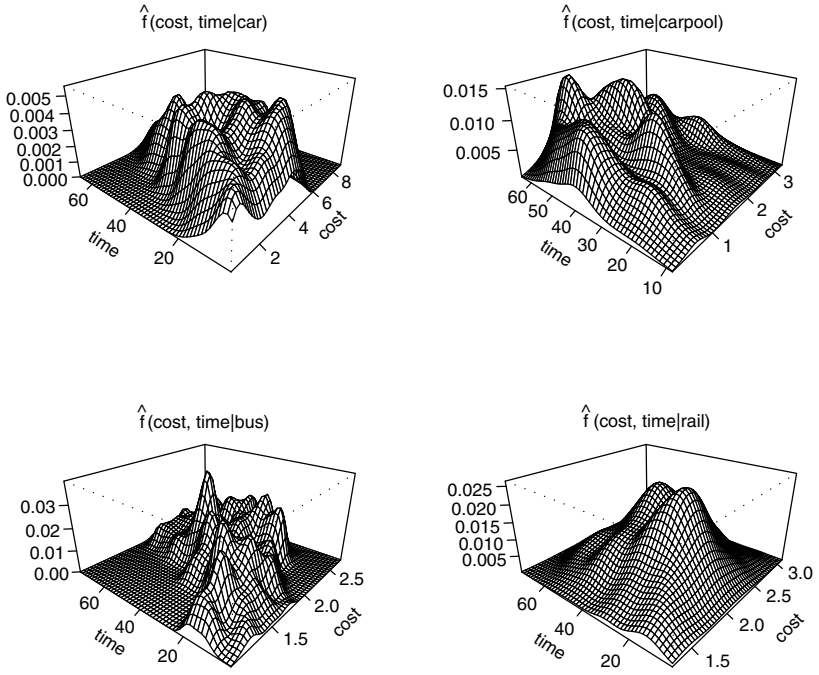


Figure 15.2 Multivariate conditional p.d.f.

15.4.1.3 Example: additive models

Kernels, as well as many other nonparametric methods, do not perform well when q is large. The sparseness of data in this setting inflates the variance of the estimates, and the numerical accuracy of estimates rapidly decreases with the number of regressors. This problem is sometimes referred to as the "curse of dimensionality." Computationally, the curse of dimensionality means that for kernel methods, as q becomes large relative to a fixed sample size n, division by 0 becomes more frequent in the calculation of $\widehat{m}(\mathbf{x})$.

To overcome these difficulties, Stone (1985) proposed additive models, i.e.:

$$m(\mathbf{x}) = m_1\left(x_1^c\right) + \cdots + m_{q_1}\left(x_{q_1}^c\right),$$

where the x_l^cs are all univariate continuous variables, and the $m_l(\cdot)$ are unknown functions for $l = 1, \ldots, q_1$. A benefit of an additive model is that estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables. Suppose $q_1 = 4$, then the following identities:

$$E[y - m_2(x_2^c) - m_3(x_3^c) - m_4(x_4^c)|x_1^c] = m_1(x_1^c),$$
$$E[y - m_1(x_1^c) - m_3(x_3^c) - m_4(x_4^c)|x_2^c] = m_2(x_2^c),$$

$$E[y - m_1(x_1^c) - m_2(x_2^c) - m_4(x_4^c)|x_3^c] = m_3(x_3^c), \text{ and}$$
$$E[y - m_1(x_1^c) - m_2(x_2^c) - m_3(x_3^c)|x_4^c] = m_4(x_4^c),$$

imply the backfitting algorithm 15.4.1.3.1.

Algorithm 15.4.1.3.1 Standard backfitting algorithm – implementation

1. Select initial estimates $\widehat{m}_1^{[0]}(x_1^c), \ldots, \widehat{m}_4^{[0]}(x_4^c)$, say $\widehat{m}_l^{[0]}\left(x_l^c\right) = 0$ for all $l = 1, \ldots, 4$.
2. For step $s = 1, 2, \ldots$, obtain:

$$m_1^{[s]}(x_1^c) = \widehat{E}[y_i - \widehat{m}_2^{[s-1]}(x_2^c) - \widehat{m}_3^{[s-1]}(x_3^c) - \widehat{m}_4^{[s-1]}(x_4^c)|x_{1;i}^c = x_1^c],$$
$$\vdots = \vdots$$
$$m_4^{[s]}(x_4^c) = \widehat{E}[y_i - \widehat{m}_1^{[s-1]}(x_1^c) - \widehat{m}_2^{[s-1]}(x_2^c) - \widehat{m}_3^{[s-1]}(x_3^c)|x_{4;i}^c = x_4^c],$$

   where, for a random variable a_i, $\widehat{E}[a_i|x_{l;i}^c = x_l^c]$ is the univariate kernel estimator of $E[a_i|x_{l;i}^c = x_l^c]$ using bandwidth h_l.
3. Continue iterations in step 2. until a pre-specified convergence criterion is reached.

Suppose it takes d iterations for the above algorithm to converge, then the rate of convergence of $\sum_{l=1}^4 \widehat{m}_l^{[d]}\left(x_l^c\right)$ is the same as if the regression model were a function of a single continuous variable instead.

We use a cross-section of 872 observations from Switzerland (see Gerfin, 1996) in order to illustrate the methodology. Whether or not a woman participates in the labor market, $E[y = 1|\mathbf{x}] \equiv \Pr[\mathbf{x}]$, is modeled as:

$$\log\left\{\frac{\Pr[\mathbf{x}]}{1 - \Pr[\mathbf{x}]}\right\} = \sum_{l=1}^{4} m_l\left(x_l\right),$$

where x_1 is the log of non-labor income (LNNLINC), x_2 is age in years divided by 10 (AGE), x_3 is the number of years of formal education (EDUC), and x_4 is the number of children (NC). Figure 15.3 illustrates the results. Each panel shows the regressors' individual effects on their entire observed support. Both LNNLINC and AGE certainly have quadratic effects, while the estimated effects of EDUC and NC might suffer from boundary effects.

Marginal integration is an alternative to backfitting (see, e.g., Linton and Nielsen, 1995; Linton and Härdle, 1996).

15.4.2 Semiparametric estimation

Another way to alleviate the curse of dimensionality is to finitely parameterize certain aspects of the joint distribution of y and $\mathbf{x}$, e.g., its mean, while allowing others to remain unknown, e.g., the conditional variance $\text{var}(y|\mathbf{x})$. The object of

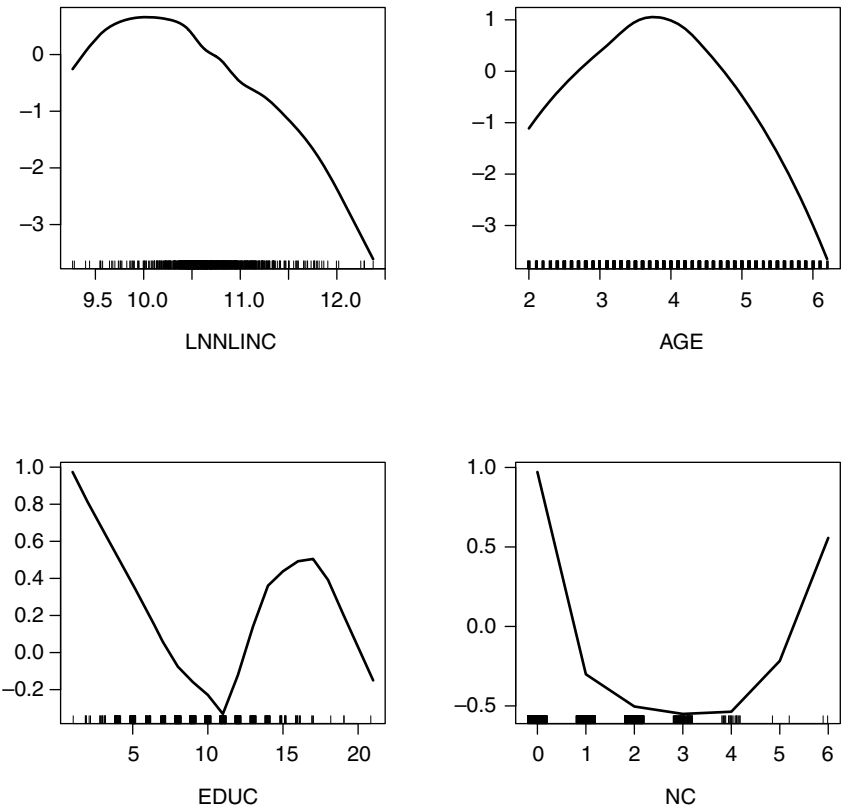


Figure 15.3 Generalized additive model: labor force participation

interest then becomes a finite dimensional parameter, whose estimator is numerically calculated in multiple stages with fully nonparametric calculations in the first stages. This is known as semiparametric estimation, and its use is becoming increasingly popular with the advent of faster and cheaper computing. Although we only discuss a few of these estimators, a more complete exposition can be found in, e.g., Horowitz (1998), Pagan and Ullah (1999), and Li and Racine (2007).

15.4.2.1 Example: efficient estimation with heteroskedasticity of unknown form

The heteroskedastic linear model specifies $E[y|\mathbf{x}] = \mathbf{x}^\top \boldsymbol{\beta}$ and $\text{var}(y|\mathbf{x}) = \sigma^2(\mathbf{x})$, where the variance function $\sigma^2(\cdot)$ is left unspecified. Robinson (1987) proposed a semiparametric feasible generalized least squares (GLS) estimator $\widehat{\boldsymbol{\beta}}$ that requires the nonparametric estimation of $\sigma_i^2 \equiv \sigma^2(\mathbf{x}_i)$ in:

$$\widehat{\boldsymbol{\beta}} = (\textstyle\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \widehat{\sigma}_i^{-2})^{-1} (\sum_{i=1}^n \mathbf{x}_i y_i \widehat{\sigma}_i^{-2}), \tag{15.15}$$

by k-nearest neighbor methods. In particular, let $R_{x_s} \equiv R_n(x_s)$ denote the Euclidean distance between x_s and its k_s-th nearest neighbor among $x_{s;i}$, for $i = 1, \ldots, n$ and

$s = 1, \ldots, q_1$, then:

$$\widehat{\sigma}_i^2 = \frac{\sum_{j=1}^n \widehat{u}_j^2 K(\mathbf{x}_i^c, \mathbf{x}_j^c; R_{\mathbf{x}^c}) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{j=1}^n K(\mathbf{x}_i^c, \mathbf{x}_j^c; R_{\mathbf{x}^c}) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}, \tag{15.16}$$

where $\widehat{u}_j^2$ is the residual from first-stage ordinary least squares (OLS) regression of y_i on $\mathbf{x}_i$. Robinson (1987) showed that (15.15) is adaptive when $k_1 = \cdots = k_{q_1}$, $k(u) = (1/2) \times \mathbb{I}(|u| \leq 1)$, and $(\mathbf{x}_i^d, \mathbf{x}_j^d)$ is empty in (15.11), (15.12), and (15.16). Its asymptotic variance can be estimated as:

$$\widehat{\text{asy. var}}(\widehat{\boldsymbol{\beta}}) = (\textstyle\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \widehat{\sigma}_i^{-2})^{-1}. \tag{15.17}$$

For illustration we will fit a linear hedonic pricing model of house attributes using data collected by Ho (1995) of 92 detached homes in the Ottawa area that were sold in 1987. The data file contains continuous variables such as sale price (SALEPRIX), average neighborhood income (AVGINC), distance to highway (DISTHWY), lot size (LOTAREA), square footage of usable space (USESPACE), location coordinate in the south (SOUTH), and west (WEST). It also contains discrete variables such as indicators for the presence of fire place (FIREPLAC), garage (GARAGE), luxurious bath (LUXBATH), and the number of bedrooms (NRBED). Using SALEPRIX as y and the remaining variables as regressors $\mathbf{x}$, algorithm 15.4.2.1.1 provides implementation details.

Algorithm 15.4.2.1.1 Semiparametric feasible GLS – implementation

1. Regress y on $\mathbf{x}$ by simple OLS and save the fitted squared residuals, $\{\widehat{u}_i^2\}_{i=1}^n$.
2. Select functions $k(\cdot)$ and $l(\cdot)$ in (15.11) and (15.12), and choose vectors of smoothing parameters k and λ by leave-one-out cross-validation, i.e., select $[h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{q_2}]$ to minimize:

$$CV_{[k,\lambda]} = \textstyle\sum_{i=1}^n \left[\widehat{u}_i^2 - \widehat{\sigma}_{-i}^2\right]^2,$$

   where $\widehat{\sigma}_{-i}^2$ equals (15.16) after replacing $\sum_{j=1}^n$ by $\sum_{j=1;\, j\neq i}^n$.
3. Using k and λ found in the previous step, calculate $\widehat{\sigma}_i^2$ in (15.16) for each $i = 1, \ldots, n$.
4. Calculate (15.15) and (15.17).

The results are presented in Table 15.2. The first column shows the results from running simple OLS with corrected standard errors. The second column presents results using algorithm 15.4.2.1.1. Although there is not much difference between parameter point estimates, the estimated standard errors have been reduced dramatically.

Table 15.2 Hedonic prices of housing attributes: estimated models

	(1) OLS		*(2) Semip. FGLS*		*(3) Partially linear*	
	Coef.	*Robust std. error*	*Coef.*	*Efficient std. error*	*Coef.*	*Robust std. error*
Intercept	73.978	18.367	77.821	5.821	–	–
FIREPLAC	11.795	5.916	9.634	1.759	9.548	3.792
GARAGE	11.838	4.439	10.870	1.821	3.826	3.849
LUXBATH	60.736	10.351	58.821	4.205	44.329	10.820
AVGINC	0.477	0.199	0.502	0.059	−0.414	1.037
DISTHWY	−15.277	5.596	−12.954	1.479	-23.079	32.358
LOTAREA	3.243	1.937	2.542	0.625	3.978	1.928
NRBED	6.586	4.484	4.276	1.439	1.668	3.222
USESPACE	21.128	12.070	27.125	4.043	17.767	7.059
WEST	3.206	1.884	2.931	0.702	–	–
SOUTH	−7.527	1.782	−7.254	0.527	–	–
s	21.95		1.006			14.603
Adjusted R^2	0.569		0.947			0.847
Comp. time	0.64 seconds		196.62 seconds		127.17 seconds	

Notes: s represents the squared root of the estimated variance of the regression. FGLS = feasible generalized least squares.

15.4.2.2 Example: partially linear model

The partially linear model specifies $E[y|\mathbf{z},\mathbf{x}] = \mathbf{z}^\top\boldsymbol{\beta} + g(\mathbf{x})$, where $g(\mathbf{x})$ is left unspecified, and $\mathbf{z} = [z_1,\ldots,z_p] \in \mathbb{R}^p$. In particular, assuming that, for $u = y - \mathbf{z}^\top\boldsymbol{\beta} - g(\mathbf{x})$, $E[u|\mathbf{z},\mathbf{x}] = 0$, and $\text{var}(u|\mathbf{z},\mathbf{x}) = \sigma^2(\mathbf{z},\mathbf{x})$, and after observing that $E[y|\mathbf{x}] = E[\mathbf{z}|\mathbf{x}]^\top\boldsymbol{\beta} + g(\mathbf{x})$, Robinson (1988) proposed an estimator $\widehat{\boldsymbol{\beta}}$ that requires the nonparametric estimation of $E[y_i|\mathbf{x}_i]$, and $E[\mathbf{z}_i|\mathbf{x}_i]$, $i = 1,\ldots,n$, in:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n (\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])(\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])^\top\right)^{-1}\left(\sum_{i=1}^n (\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])(y_i - \widehat{E}[y_i|\mathbf{x}_i])\right), \tag{15.18}$$

by the Nadaraya–Watson estimator, i.e., for vectors of smoothing parameters h and λ:

$$\widehat{E}[y_i|\mathbf{x}_i] = \frac{\sum_{j=1}^n y_i K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{j=1}^n L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}, \tag{15.19}$$

$$\widehat{E}[z_{l;i}|\mathbf{x}_i] = \frac{\sum_{j=1}^n z_{l;i} K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{j=1}^n L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}, \text{ for } l = 1,\ldots,p. \tag{15.20}$$

Its asymptotic variance can be estimated consistently by:

$$\widehat{\text{asy. var}}(\widehat{\boldsymbol{\beta}}) = n^{-1}\widehat{\Phi}^{-1}\widehat{\Omega}\widehat{\Phi}^{-1}, \tag{15.21}$$

where $\widehat{\Phi} = n^{-1}\sum_{i=1}^{n}(\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])(\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])^{\top}$, and $\widehat{\Omega} = n^{-1}\sum_{i=1}^{n}\widehat{u}_i^2(\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])(\mathbf{z}_i - \widehat{E}[\mathbf{z}_i|\mathbf{x}_i])^{\top}$, with $\widehat{u}_i = y_i - \mathbf{z}_i^{\top}\widehat{\boldsymbol{\beta}} - \widehat{g}(\mathbf{x}_i)$, and:

$$\widehat{g}(\mathbf{x}_i) = \frac{\sum_{j=1}^{n}(y_i - \mathbf{z}_i^{\top}\widehat{\beta})K(\mathbf{x}_i^c, \mathbf{x}_j^c; h)L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{j=1}^{n} L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}. \tag{15.22}$$

Algorithm 15.4.2.2.1 makes precise the necessary steps for the above calculations.

Algorithm 15.4.2.2.1 Partially linear model – implementation

1. For each $l = 1, \ldots, p$, select functions $k(\cdot)$ and $l(\cdot)$ in (15.11) and (15.12), and choose vectors of smoothing parameters h and λ by leave-one-out cross-validation, i.e., select $[h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{q_2}]$ to minimize:

$$CV_{[h,\lambda]} = \sum_{i=1}^{n}\left[y_i - \widehat{E}_{-i}[y_i|\mathbf{x}_i]\right]^2$$
$$CV_{[h,\lambda]} = \sum_{i=1}^{n}\left[z_{l;i} - \widehat{E}_{-i}[z_{l;i}|\mathbf{x}_i]\right]^2,$$

   where $\widehat{E}_{-i}$ equals (15.19) and (15.20) after replacing $\sum_{j=1}^{n}$ by $\sum_{j=1;\, j\neq i}^{n}$.
2. Using the bandwidths found in the previous step, calculate (15.19) and (15.20) for each data point $i = 1, \ldots, n$.
3. Calculate (15.18) and (15.21).
4. Select functions $k(\cdot)$ and $l(\cdot)$ in (15.11) and (15.12), and choose vectors of smoothing parameters h and λ by leave-one-out cross-validation, i.e., select $[h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{q_2}]$ to minimize:

$$CV_{[h,\lambda]} = \sum_{i=1}^{n}\left[y_i - \mathbf{z}_i^{\top}\widehat{\beta} - \widehat{E}_{-i}[y_i - \mathbf{z}_i^{\top}\widehat{\beta}|\mathbf{x}_i]\right]^2,$$

   where $\widehat{E}_{-i}$ equals (15.19) after replacing y_i by $y_i - \mathbf{z}_i^{\top}\widehat{\beta}$ and $\sum_{j=1}^{n}$ by $\sum_{j=1;\, j\neq i}^{n}$.
5. Calculate (15.22) using the bandwidths found in the previous step for each $i = 1, \ldots, n$.

Since house location (WEST and SOUTH) has no natural parametric effect in housing prices, we proceed to include a two-dimensional nonparametric effect, $g(\cdot)$, in the hedonic pricing model of housing attributes in section 15.4.2.2. The results are presented in column 3 of Table 15.2.

15.4.2.3 Example: binary choice model

The binary choice model specifies the following relationship:

$$y = \mathbb{I}(\mathbf{x}^{\top}\boldsymbol{\beta} + u > 0), \tag{15.23}$$

where $\mathbb{I}$ is the indicator function, and u represents unobserved characteristics with continuous symmetric distribution $F(\cdot)$ which is assumed independent of $\mathbf{x}$. It then

follows that:

$$\Pr[y = 1|\mathbf{x}] = E[y|\mathbf{x}] = F(\mathbf{x}^\top \boldsymbol{\beta}), \tag{15.24}$$

where $\mathbf{x}^\top \boldsymbol{\beta}$ is known as a *single index*. In parametric probit and logit models, the unknown parameter $\boldsymbol{\beta}$ in (15.23) is obtained by numerically maximizing the log-likelihood function:

$$\sum_{i=1}^{n} \{y_i \log[F(\mathbf{x}_i^\top \boldsymbol{\beta})] + (1 - y_i) \log[1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})]\}, \tag{15.25}$$

where $F(\cdot)$ is assumed to be a normal or logistic distribution function respectively. Column 1 in Table 15.3 shows results from fitting a probit model to female labor participation decisions in Portugal, using 2,339 observations taken from Martins (2001). The regressors include variables such as the number of children younger than 18 living in the family (CHILD), the number of children younger than three years of age (YCHILD), the number of years of formal schooling (EDU), the log of the husband's monthly wages (LNHW), women's age divided by 10 and 100 respectively (AGE and AGE2). The model does not include a constant and the coefficient multiplying AGE has been normalized to be equal to one for comparison purposes.

Notice that all regressors, except the number of years of formal schooling, have a negative effect on the probability that a woman works for wages.

If $F(\cdot)$ is unknown, then $F(\mathbf{x}^\top \boldsymbol{\beta})$ is observationally equivalent to $\overline{F}\left(\mathbf{x}^\top \boldsymbol{\beta} - a\right)$, and $\widetilde{F}\left(\mathbf{x}^\top (\boldsymbol{\beta}/c)\right)$, with $\overline{F}(u) = F(u + a)$ and $\widetilde{F}(u) = F(c \cdot u)$ respectively $\forall a, c \in \mathbb{R}$. Under these circumstances, restrictions are needed to identify the unknown vector of parameters $\boldsymbol{\beta}$. For example, the so called location-scale normalization sets any intercept in $\mathbf{x}^\top \boldsymbol{\beta}$ equal to 0, and the coefficient of a continuous regressor (say the first regressor x_1) equal to 1.

Let us define $\widetilde{\mathbf{x}}^c = [x_2^c, \ldots, x_{q_1}^c]$, $\widetilde{\mathbf{x}} = [\widetilde{\mathbf{x}}^c, \mathbf{x}^d]$, $\mathbf{x} = [x_1, \widetilde{\mathbf{x}}]$, $\widetilde{\boldsymbol{\beta}} = [\beta_2, \ldots, \beta_{q_1}, \beta_{q_1+1}, \ldots, \beta_{q_2}]^\top$, and $\boldsymbol{\beta} \equiv (1, \widetilde{\boldsymbol{\beta}}^\top)$. Semiparametric methods have been shown to

Table 15.3 Female labor participation: estimated models

	(1) Probit		*(2) Klein–Spady (1993)*		*(3) Ichimura (1993)*	
	Coef.	*Std. error*	*Coef.*	*Std. error*	*Coef.*	*Std. error*
CHILD	−0.079	0.071	−0.012	0.029	−0.145	0.123
YCHILD	−0.129	0.026	−0.083	0.012	−0.096	0.076
EDU	0.144	0.009	0.076	0.003	0.134	0.013
LNHW	−0.181	0.009	−0.043	0.004	−0.138	0.017
AGE2	−0.148	0.004	−0.145	0.002	−0.158	0.014
AGE	1	–	1	–	1	–
Part. prob.	1408.00		1400.11		1399.17	
Comp. time	0.421 seconds		491.072 seconds		2462.740 seconds	

Note: Part. prob. = participation probability.

identify and consistently estimate $\widetilde{\boldsymbol{\beta}}$ when $F(\cdot)$ is only known to be a non-constant smooth function on the support of $\mathbf{x}^\top\boldsymbol{\beta}$; and varying the values of $\mathbf{x}^d$ does not divide the support of $\mathbf{x}^\top\boldsymbol{\beta}$ into disjoint subsets (see Horowitz, 1998).

Note that, for given $\boldsymbol{\beta}$, a consistent nonparametric estimator of (15.24) is:

$$\widehat{F}(\mathbf{x}_i^\top\boldsymbol{\beta}) = \frac{\sum_{j=1}^n y_i k((\mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_j^\top\boldsymbol{\beta})/h)}{\sum_{j=1}^n k((\mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_j^\top\boldsymbol{\beta})/h)}. \tag{15.26}$$

Klein and Spady's estimator

Klein and Spady (1993) proposed estimating $\boldsymbol{\beta}$ by maximizing (15.25) after replacing $F(\mathbf{x}_i^\top\boldsymbol{\beta})$ by $\widehat{F}_{-i}(\mathbf{x}_i^\top\boldsymbol{\beta})$, where $\widehat{F}_{-i}(\mathbf{x}_i^\top\boldsymbol{\beta})$ is like $\widehat{F}(\mathbf{x}_i^\top\boldsymbol{\beta})$ but with $\sum_{j=1}^n$ replaced by $\sum_{j=1;\, j\neq i}^n$ in (15.26). Computationally, the maximization must be performed numerically by solving the first order conditions of the problem. Under regularity conditions, Klein and Spady showed that the resulting estimator $\widehat{\boldsymbol{\beta}}$ has an asymptotic normal distribution. Furthermore, they showed that the asymptotic variance of their estimator attains the semiparametric efficiency bound, and that it can be consistently estimated by:

$$\widehat{\text{asy. var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^n \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top [\widehat{F}_{-i}^{(1)}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]^2 / [\widehat{F}_{-i}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})(1-\widehat{F}_{-i}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}))]\right)^{-1}, \tag{15.27}$$

where $\widehat{F}_{-i}^{(1)}(\cdot)$ is the first derivative of $\widehat{F}_{-i}(\cdot)$. Algorithm 15.4.2.3.1 describes the necessary steps.

Algorithm 15.4.2.3.1 Klein and Spady (1993) – implementation

1. Select function $k(\cdot)$ in (15.26), and numerically find jointly the bandwidth h and vector of coefficients $\boldsymbol{\beta}$ that minimizes the leave-one-out estimated log-likelihood:

$$\sum_{i=1}^n \{y_i \log[\widehat{F}_{-i}(\mathbf{x}_i^\top\boldsymbol{\beta})] + (1-y_i)\log[1-\widehat{F}_{-i}(\mathbf{x}_i^\top\boldsymbol{\beta})]\},$$

   where $\widehat{F}_{-i}(\cdot)$ equals (15.26) after replacing $\sum_{j=1}^n$ by $\sum_{j=1;\, j\neq i}^n$.
2. Using the bandwidth and vector of coefficients found in the previous step, calculate $\widehat{F}_{-i}^{(1)}(\cdot)$ and $\widehat{F}_{-i}(\cdot)$ at each data point $i = 1, \ldots, n$.
3. Calculate (15.27).

Column 2 in Table 15.3 shows results from applying the above methodology to the modeling of female labor force participation decisions. Although signs coincide with those of the fully parametric model (column 1), their magnitudes and standard errors are remarkably different. Martins (2001) pointed out that this considerable efficiency gain may occur because the semiparametric model is sufficiently perturbed from the usual probit specification for women with low index values.

Ichimura's estimator

The semiparametric least squares (SLS) estimator of Ichimura (1993) numerically minimizes:

$$\sum_{i=1}^{n}[y_i - \widehat{F}_{-i}(\mathbf{x}_i^\top \boldsymbol{\beta})]^2, \tag{15.28}$$

with respect to $\boldsymbol{\beta}$, where, as before, $\widehat{F}_{-i}(\cdot)$ is the kernel estimator for the unknown link function $F(\cdot)$. The resulting estimator, $\widehat{\boldsymbol{\beta}}$, is asymptotically normally distributed, and its asymptotic variance can be consistently estimated by:

$$\widehat{\text{asy. var}}(\widehat{\boldsymbol{\beta}}) = n^{-1}\widehat{\Gamma}^{-1}\widehat{\Sigma}\widehat{\Gamma}^{-1}, \tag{15.29}$$

where:

$$\widehat{\Gamma} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top[\widehat{F}_{-i}^{(1)}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]^2, \text{ and} \tag{15.30}$$

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top[\widehat{F}_{-i}^{(1)}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]^2[y_i - \widehat{F}_{-i}(\mathbf{x}_i^\top\boldsymbol{\beta})]^2. \tag{15.31}$$

Column 3 in Table 15.3 presents the results of applying algorithm 15.4.2.3.2 to our example.

Algorithm 15.4.2.3.2 Ichimura (1993) – implementation

1. Select function $k(\cdot)$ in (15.26), and numerically find jointly the bandwidth h and vector of coefficients $\boldsymbol{\beta}$ that minimizes the semiparametric least squares objective function (15.28).
2. Using the bandwidth and vector of coefficients found in the previous step, calculate $\widehat{F}_{-i}^{(1)}(\cdot)$ and $\widehat{F}_{-i}(\cdot)$ at each data point $i = 1, \ldots, n$.
3. Calculate (15.29), (15.30) and (15.31).

The results differ mainly in the magnitude of the effects of each regressor as well as in their precision. Both semiparametric estimators provide estimates of the participation probabilities (Part. prob.) closer to the actual 1,400 women in the sample that are observed to work for wages. The parametric probit specification overestimates this quantity.

It should be emphasized that, unlike Klein and Spady's estimator, Ichimura's is applicable to a wider range of problems where the response y_i is not only binary, but takes on different discrete or continuous values.

15.4.2.4 Further considerations

The numerical stability of likelihood cross-validated bandwidths in step 1 in algorithms 15.4.1.1.1 and 15.4.1.2.1 might be affected by outliers. However, the suggested likelihood cross-validated method may work well for thin-tailed distributions.

In the semiparametric models discussed above, the first-order asymptotic distribution of the normalized and centered estimators does not depend on the

smoothing parameter(s). Therefore, at least asymptotically, any sequence of smoothing parameters is going to give a consistent estimator as long as it satisfies certain conditions. However, this observation does not necessarily imply that such an estimator is going to be optimal in the mean squared error (MSE) sense (see, e.g., Linton, 1995; Härdle, Hall and Ichimura, 1993).

Finally, the asymptotic theory for the above estimators requires trimming those observations arbitrarily close to the boundary of their observed supports. We did not address this in the above discussion, partly because of the lack of guidance on how to select trimming parameters with fixed sample sizes, and partly because no trimming amounts to assuming that the actual support of the variables involved is larger than that observed in the data.

15.5 Modeling heterogeneity

In the introductory section we noted that heterogeneity is a pervasive feature in microeconometrics. The most tractable way of handling observed heterogeneity is to control for it by including sociodemographic variables such as family composition, age, gender, location, etc. The potential of this approach is limited by the scope of the available data. There still remains variation that is induced by unobserved factors, referred to in econometrics as unobserved heterogeneity.

Neglecting individual-level unobserved heterogeneity can have potentially serious consequences, analogous to those of omitted regressors. Even when heterogeneity is explicitly accommodated, the precise manner and assumptions under which it enters the model can have important consequences. Hence, not surprisingly, every proposed econometric specification is routinely scrutinized for the manner in which it accommodates heterogeneous behavior.

Not only are models with heterogeneity more flexible, and hence generally fit the data better, but they also lead to relaxation of strong constraints. For example, the multinomial logit (MNL) discrete choice model is subject to the strong restriction of independence of irrelevant alternatives (IIA), whereas the random parameter version of the MNL does not have the IIA restriction.

There are a number of distinctive ways of allowing for unobserved heterogeneity.

1. Treat heterogeneity either as an additive or a multiplicative random effect (uncorrelated with included regressors) or as a fixed effect (potentially correlated with included regressors). Within the class of random effects models, heterogeneity distributions may be treated as continuous or discrete. Examples include a random intercept in linear regression and linear panel models, fixed effects in linear and nonlinear panels, and multiplicative heterogeneity in models of counts and durations.
2. Allow both intercept and slope parameters to vary randomly and parametrically. Examples include random parameter discrete choice and outcome models and finite mixture models.
3. Model heterogeneity explicitly in terms of both observed and unobserved variables using mixed models, hierarchical models and/or models of clustering.

Which approach one takes to modeling heterogeneity and how it is combined with other assumptions, e.g., functional forms, has important consequences for computation. One researcher may impose strong functional forms but achieve modeling flexibility by allowing for behavior heterogeneity. Another may make very flexible functional form assumptions but not allow explicitly for unobserved heterogeneity. Currently both strategies coexist in the literature.

We now consider computational aspects of the different modeling strategies.

1. Under the fixed effects specification, the most popular approach is to treat heterogeneity in the nuisance parameter framework and eliminate it by a suitable transformation such as: within transformation; first-difference transformation; differences-in-differences transformation. In some parametric models with fixed effects, the nuisance parameters can be eliminated by applying the conditional likelihood approach that replaces the fixed effect by a sufficient statistic, but such a statistic is not always available.
2. Dummy variables can be introduced to capture individual-specific heterogeneity. This approach is familiar in linear panel models, but due to the incidental parameter problem, this formulation may not lead to consistent estimates. In panel models with a large cross-section dimension (n), the estimation dimension also increases and in nonlinear models this may lead to computational challenges, though there are examples where ingenious algorithms can address the issues (e.g., Greene, 2004a, 2004b).
3. In random effects models the standard and time-honored approach involves integrating out the distribution of unobserved heterogeneity, as was discussed in section 15.3.3. Such integration is often implemented numerically, leading to nontrivial computational challenges, especially if parameter dimension is large. Random parameter multinomial logit and multinomial probit models are two outstanding examples of this approach (see Train, 2003).

All the foregoing discussion has been carried out by reference to individual level heterogeneity. Heterogeneity may also exist at the level of groups, which leads to issues of clustering and interdependence that create additional computational challenges. A group can be a geographical, social, ethnic, or merely a sampling unit that exhibits some form of interdependent behavior induced by common environment or culture. Group membership may be modeled as an observable or a latent variable, analogous to fixed and random effects formulations in panel data.

In the remainder of this section we illustrate, using two data-based examples, how modeling heterogeneity can be computationally demanding but empirically informative.

15.5.1 Example: quantile regression

Whereas the standard linear regression is a useful tool for summarizing the average relationship between the outcome variable of interest and a set of regressors, i.e., the conditional mean function, $E[y|\mathbf{x}]$, QR can provide a more complete picture

of the relationship between outcome y and regressors $\mathbf{x}$ at different points in the distribution of y.

As stated in section 15.3.2, the objective function for QR is in the L_1-norm. Unlike the squared error loss function of the OLS framework, in the case of QR the loss is expressed through an asymmetric absolute loss function. The special case of median (LAD) regression corresponds to $q = 0.5$.

QRs have a certain robustness property and they also permit more informative modeling of the data. Specifically, the median regression is more robust to outliers compared with the mean regression. QR facilitates a richer interpretation of the data because it permits the study of the impact of regressors on both the location and scale parameters of the model, while avoiding parametric assumptions about data.

Other attractive properties of QR are as follows: (i) it is consistent under weaker stochastic assumptions than least squares estimation; (ii) it is based on weaker distributional assumptions; (iii) it has an equivariance to monotone transformations property, which implies that it does not run into the retransformation problem. The quantiles of a transformed variable y, denoted $h(y)$, where h is a monotonic function, equal the transforms of the quantiles of $y : Q_{h(y)}(\tau) = h(Q_y(\tau))$. Hence, if the quantile model is expressed in terms of $h(y)$, e.g., $\ln(y)$, then one can use the inverse transformation to translate the results back to y. This is not possible for the least squares estimator, i.e., if $E[h(y)] = \mathbf{x}^\top \boldsymbol{\beta}$, then $E[y|\mathbf{x}] \neq h^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$.

Impediments to the use of QR are twofold. First, there are computational hurdles. Because the objective function is not differentiable, the gradient optimization methods mentioned in section 15.3 cannot be used and, instead, linear programming methods are applied. There is no closed-form solution for $\widehat{\boldsymbol{\beta}}_q$ and hence the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_q$ cannot be obtained using standard methods. An analytical expression for the asymptotic variance of $\widehat{\boldsymbol{\beta}}_q$ comes from the result that:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, q(1-q)\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right], \tag{15.32}$$

where $n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top \xrightarrow{p} \mathbf{A}$, $n^{-1}\sum_{i=1}^{n} f_{u_q}(\xi(q)|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top \xrightarrow{p} \mathbf{B}$, $\xrightarrow{p}$ represents convergence in probability, and $f_{u_q}(\xi(q)|\mathbf{x})$ is the conditional density of the error term $u_q = y - \mathbf{x}^\top \boldsymbol{\beta}_q$ evaluated at $\xi(q)$, i.e., the q-quantile of u_q. Estimation of the variance of $\widehat{\boldsymbol{\beta}}_q$ is complicated by the need to estimate $f_{u_q}(\xi(q)|\mathbf{x})$. It is easier to obtain standard errors for $\widehat{\boldsymbol{\beta}}_q$ using the computationally more intensive bootstrap method, as is done in the example that follows.

Buchinsky (1995) evaluated a number of variance estimators for the QR in a Monte Carlo setting and recommended the use of a "design matrix" (or paired) bootstrap estimator. Application of the bootstrap variance is shown in algorithm 15.5.1.0.1.

Computational advances have made the application of QR more accessible to users and many popular packages include it. As an illustration we report a regression analysis of total medical expenditure (TOTEXP) by the Medicare elderly. The data

Algorithm 15.5.1.0.1 QR–bootstrap standard errors

1. Draw $(y_i^b, \mathbf{x}_i^{\top b})$, $b = 1, \ldots, B$, $i = 1, \ldots, n$, a (paired) bootstrap sample drawn from the empirical distribution of $(y_i, \mathbf{x}_i^{\top})$.
2. Estimate the conditional quantile function $\mathbf{x}_i^{\top b} \widehat{\boldsymbol{\beta}}_q^b$, where $\widehat{\boldsymbol{\beta}}_q^b$ is the bootstrap estimate of $\boldsymbol{\beta}_q$.
3. The bootstrap estimate of the variance, $\text{var}(\widehat{\boldsymbol{\beta}}_q)$, is given by:

$$\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}_q] = \frac{n}{B} \sum (\widehat{\boldsymbol{\beta}}_q^b - \overline{\boldsymbol{\beta}}_q^b)(\widehat{\boldsymbol{\beta}}_q^b - \overline{\boldsymbol{\beta}}_q^b)^{\top},$$

where $\overline{\boldsymbol{\beta}}_q^b = B^{-1} \sum \widehat{\boldsymbol{\beta}}_q^b$.

are derived from the Medical Expenditure Panel Survey, and consists of 2,873 observations. The dependent variable is log(TOTEXP), so that zero values are omitted. The explanatory variables are supplementary private insurance (SUPPINS), which is a dummy variable, the number of chronic conditions (TOTCHR) as a measure of health status, and two demographic variables (AGE, FEMALE), and log(income) (LINC).

The numerical results are displayed in Table 15.4, which reports the OLS results with Eicker–White heteroskedasticity robust standard errors, and QR estimates for three values of q, 0.25, 0.50, 0.75. The OLS results in column 1 can be compared with the median regression results in column 3; the two should not be too different if the conditional distribution of the dependent variable is symmetric. The standard errors were computed using a paired bootstrap with 499 bootstrap replications. Because of the relatively small number of regressors in the model, this computation is manageable even on a desktop PC. Figure 15.4 displays the graphs of the quantile regression coefficients at different values of q. Such a visual representation is potentially very informative as it reveals the heterogeneity in response to variables at different quantiles of expenditure. For example, the impact of SUPPINS at $q = 0.25$ is nearly three times as large as at $q = 0.75$. The graphs also include the constant least squares coefficient as a benchmark.

Table 15.4 OLS and bootstrapped quantile regressions

	(1) OLS		*(2) QR: $q = 0.25$*		*(3) QR: $q = 0.50$*		*(4) QR: $q = 0.75$*	
	Coef.	*Std. error*	*Coef.*	*Std. error*	*Coef.*	*Std. error*	*Coef.*	*Std. error*
SUPPINS	0.224	0.0484	0.336	0.0631	0.246	0.0550	0.118	0.0658
AGE	0.0150	0.00364	0.0190	0.00481	0.0178	0.00420	0.0212	0.00500
FEMALE	−0.0512	0.0469	0.0252	0.0565	−0.0493	0.0507	−0.145	0.0548
TOTCHR	0.445	0.0176	0.461	0.0239	0.391	0.0195	0.371	0.0208
LINC	0.0593	0.0275	0.0774	0.0304	0.0716	0.0329	0.0566	0.0391
Intercept	5.876	0.298	4.609	0.408	5.726	0.347	6.445	0.415

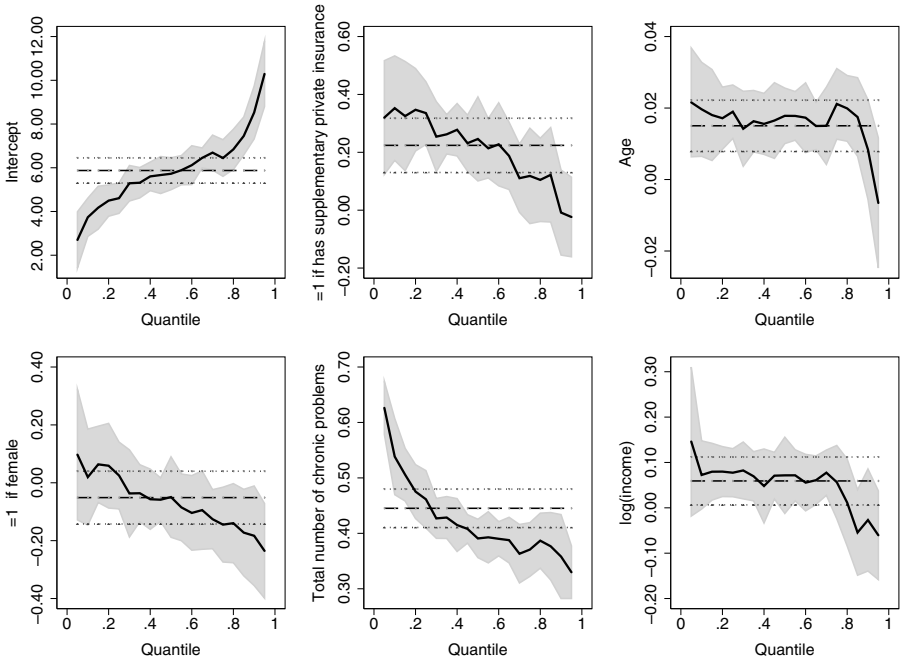


Figure 15.4 Coefficients of regressors at various quantiles

15.5.2 Example: finite mixture model

Fully parametric models are popular in microeconometrics even though a parametric distributional assumption is an important simplification. There are many ways of replacing or relaxing this assumption, which may lead to additional computation. Replacing the assumption of a given parametric distribution by the assumption that the data distribution is a discrete mixture, also called a finite mixture (FM), of two or more distributions, not necessarily from the same family, can provide additional flexibility (Frühwirth-Schnatter, 2006). The FM representation is an intuitively attractive representation of heterogeneity in terms of a number of latent classes, each of which may be regarded as a "type" or a "group." It has found numerous applications in health and labor economics and in models of discrete choice. The FM model is related to latent class analysis (Aitken and Rubin, 1985; McLachlan and Peel, 2000).

In an FM model a random variable is a draw from an additive mixture of C distinct populations in proportions $\pi_1, \ldots, \pi_C$, where $\sum_{j=1}^{C} \pi_j = 1$, $\pi_j \geqslant 0$ $(j = 1, \ldots, C)$, denoted as:

$$f(y_i|\Theta) = \sum_{j=1}^{C-1} \pi_j f_j(y_i|\boldsymbol{\theta}_j) + \pi_C f_C(y_i|\boldsymbol{\theta}_C), \tag{15.33}$$

where each term in the sum on the right-hand side is the product of the mixing probability π_j and the component (sub-population) density $f_j(y_i|\boldsymbol{\theta}_j)$. Sometimes such models are referred to as models of permanent unobserved heterogeneity. In general the π_j are unknown and hence need to be estimated along with all the other parameters, denoted Θ. Also $\pi_C = (1 - \sum_{j=1}^{C-1} \pi_j)$. For identifiability the (label switching) restriction $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_C$ is imposed; this can always be satisfied by rearrangement, post-estimation. Therefore, it plays no role in estimation. The parameter π_j may be further parameterized in terms of observed covariates using, e.g., the logit function.

For given C, maximum likelihood is a natural estimator for the FM model (see McLachlan and Peel, 2000, Ch. 2). Lindsay (1983) showed that finding the MLE involved a standard convex maximization problem in which a concave function is maximized over a convex set. An implication that follows is that if the likelihood is bounded, there exists a distribution, in the class of discrete distribution functions G with n or fewer points of support, that maximizes the likelihood.

There are two commonly used computational approaches – direct gradient-based optimization based on the roots of the likelihood equations, or the expectation maximization (EM) algorithm described below (see McLachlan and Peel, 2000). If the likelihood is bounded, then under correct specification of the model it has a global maximum. But it may be difficult to locate the global maximum if the component distributions are not well separated, or the convergence may be sensitive to the starting values. One way to guard against such a possibility is to check for robustness of convergence from different starting values. In some cases, such as the mixture of normals, the likelihood is unbounded and no global maximizer exists, so convergence will be to a local maximum. In practice, especially when the sample is small, the presence of local maxima cannot be ruled out.

In practice C is unknown. For a given sample size n, the standard way of selecting C is to treat this as a model selection problem and to use information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) (see Deb and Trivedi, 2002, for a detailed application).

15.5.2.1 EM algorithm for model estimation

If C is given, the problem is to maximize the log-likelihood $\mathcal{L}(\pi, \Theta|C, \mathbf{y})$. Let $\mathbf{d}_i = (d_{i1}, \ldots, d_{iC})^\top$ define an indicator (dummy) variable such that $d_{ij} = 1$, $\sum_j d_{ij} = 1$, indicating that y_i was drawn from the jth (latent) group or class for $i = 1, \ldots, n$. That is, each observation may be regarded as a draw from one of the C latent classes or "types," each with its own distribution. The FM model specifies that $(y_i|\mathbf{d}_i, \boldsymbol{\theta}, \boldsymbol{\pi})$ are independently distributed with densities $\prod_{j=1}^{C} f(y_i|\boldsymbol{\theta}_j)^{d_{ij}}$, and $(d_{ij}|\ \boldsymbol{\theta}, \boldsymbol{\pi})$ are independent and identically distributed (i.i.d.) with multinomial distribution $\prod_{j=1}^{C} \pi_j^{d_{ij}}$, $0 < \pi_j < 1$, $\sum_{j=1}^{C} \pi_j = 1$. Hence the likelihood

function is:

$$\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{\pi}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{C} \pi_j^{d_{ij}} [f_j(\mathbf{y}; \boldsymbol{\theta}_j)]^{d_{ij}}, \quad 0 < \pi_j < 1, \quad \sum_{j=1}^{C} \pi_j = 1. \tag{15.34}$$

If π_j, $j = 1, \ldots, C$, is given, the posterior probability that observation y_i belongs to the population j, $j = 1, 2, \ldots, C$, is denoted z_{ij}, and $E\left[z_{ij}\right] = \pi_j$.

Estimation is implemented using the EM algorithm explained below. This may be slow to converge, especially if the starting values are not good. Gradient-based methods, such as Newton–Raphson or BFGS, are also used (see Böhning, 1995). The application reported below uses Newton–Raphson with gradients estimated using analytical formulae.

The EM algorithm is structured as in algorithm 15.5.2.1.1.

Algorithm 15.5.2.1.1 EM – implementation

---

1. Given an initial estimate $[\boldsymbol{\pi}^{(0)}, \boldsymbol{\theta}^{(0)}]$, the likelihood function (15.34) may be maximized using the EM algorithm in which the variable d_{ij} is replaced by its expected value, $E\left[d_{ij}\right] = \widehat{z}_{ij}$, yielding the expected log-likelihood:

$$E[\mathcal{L}(\Theta|\mathbf{y}, \boldsymbol{\pi})] = \sum_{i=1}^{n} \sum_{j=1}^{C} \widehat{z}_{ij} \left[\ln f_j(y_i; \boldsymbol{\theta}_j) + \ln \pi_j\right]. \tag{15.35}$$

2. The M-step of the EM algorithm maximizes (15.35) by solving the first-order conditions:

$$\widehat{\pi}_j - n^{-1} \sum_{i=1}^{n} \widehat{z}_{ij} = 0, \quad j = 1, \ldots, C$$

$$\sum_{i=1}^{n} \sum_{j=1}^{C} \widehat{z}_{ij} \frac{\partial \ln f_j(y_i; \boldsymbol{\theta}_j)}{\partial \theta_j} = 0.$$

3. Evaluate the marginal posterior probability $z_{ij}|\widehat{\pi}_j, \widehat{\theta}_j$, $j = 1, \ldots, C$,

$$z_{ij} \equiv \Pr[y_i \in \text{population } j] = \frac{\pi_j f_j(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j)}{\sum_{j=1}^{C} \pi_j f_j(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j)}.$$

   The E-step of the EM procedure obtains new values of $E[d_{ij}]$ using $E\left[z_{ij}\right] = \pi_j$.
4. Repeat steps 1–3 until $|\mathcal{L}(\widehat{\Theta}(k+1)) - \mathcal{L}(\widehat{\Theta}(k))| <$ tol, where tol denotes the selected tolerance level.

For estimating $\mathrm{var}(\widehat{\Theta})$, one can use either the observed information matrix or the robust Eicker–White sandwich formula. Though asymptotically valid given

regularity conditions, in practice such an application is subject to qualifications, especially in small samples or if the model is overfitted, so that one or more mixture components are small. This increases the appeal of variance calculation using the bootstrap (see McLachlan and Peel, 2000, Ch. 2.16).

15.5.2.2 FM of count models

Finite mixtures can be applied to continuous, discrete, or censored data (see McLachlan and Peel, 2000). Although the methods have not yet become common in widely used econometrics software, there are many examples in the literature. In the remainder of this section, we illustrate the potential richness of this modeling approach using an application to a count data regression (see Deb and Trivedi, 1997, 2002).

The dataset consists of 3,064 observations, where the dependent variable is number of doctor visits (DOCVIS). The predictors are age (AGE), squared age (AGE2), years of education (EDUC), a dichotomous indicator of activity limitation (ACTLIM), total number of chronic conditions (TOTCHR), and a dichotomous indicator of private health insurance status (PRIVATE). Table 15.5 shows results obtained using standard gradient methods of maximizing the likelihood. Poisson regression results appear in the first column. Typically this model gives a poor fit to the data because it imposes the assumption of equidispersion ($E[y|\mathbf{x}] = \text{var}(y|\mathbf{x})$),

Table 15.5 Poisson NB2 mixture models for DOCVIS

	(1) POISSON		*(2) POISSON (FM2)*		*(3) NB2 (FM2)*	
	Coef.	*Std. error*	*Coef.*	*Std. error*	*Coef.*	*Std. error*
Component 1						
AGE	0.299	0.0662	0.352	0.0953	0.446	0.162
AGE2	−0.198	0.0441	−0.232	0.0638	−0.290	0.107
EDUC	0.0288	0.00539	0.0299	0.00712	0.0417	0.0119
ACTLIM	0.160	0.0414	0.0701	0.0582	−0.0554	0.147
TOTCHR	0.259	0.0130	0.338	0.0174	0.527	0.0693
PRIVATE	0.154	0.0374	0.225	0.0582	0.400	0.114
Intercept	−10.34	2.474	−13.18	3.537	−17.59	6.141
Component 2						
AGE			0.219	0.107	0.272	0.0878
AGE2			−0.145	0.0720	−0.181	0.0586
EDUC			0.0204	0.00747	0.0228	0.00707)
ACTLIM			0.137	0.0592	0.201	0.0500
TOTCHR			0.203	0.0249	0.197	0.0278
PRIVATE			0.139	0.0610	0.108	0.0454
Intercept			−6.337	3.961	−8.799	3.290
π_1			0.878	0.0860	0.405	0.099
$\ln \alpha_1$					−1.186	0.307
$\ln \alpha_2$					−0.828	0.0811
log-like	−12148		−9311		−8711	

which is usually inconsistent with the data. The negative binomial regression is often the next step because it can accommodate overdispersion in the data. As a common and plausible explanation of overdispersion comes from the presence of heterogeneity in the data, another approach to the problem is to allow for the possibility that a better specification is a two-component Poisson mixture, with each component corresponding to one type of individual. The results for this specification are presented in the next two columns. A further generalization is to allow for the possibility that the distribution is a two-component mixture of negative binomial (NB2) distributions with a quadratic variance function. The advantage here is that the NB2 specification allows for within-group heterogeneity as well. Thus each component represents the behavior of one group of individuals, but also allows for within-group heterogeneity.

Because the three models are nested, the log-likelihoods are comparable and the likelihood ratio can be used to test the restrictive models against the general 2-component mixture of NB2 distributions. Clearly, the FM2 specification of the NB2 distribution is the best-fitting model. The two components correspond to low users (around 40.5% of the population) and high users (around 59.5% of the population) of doctor visits. Evaluating the conditional means of distributions at the sample average of the predictors, the average is 3.92 for the first group and 8.62 for the second group. The groups also differ in their sensitivity to variations in the predictors. Of course, we have deliberately used a very simple specification so the exact numbers are only illustrative. The important point is that, although the mixture models are harder to estimate, especially when the number of components is increased, they are much more informative about the heterogeneity in the population.

15.6 Simulation-based maximum likelihood

In this section we consider an application of the MSL estimator to a nonlinear model with discrete outcomes and endogenous dummy regressors. There are two well-established approaches, limited information and full information, for handling endogeneity in linear models. Implementation of full information methods, based on the joint distribution of all endogenous variables, is often harder to implement because closed-form expressions for the joint distribution are rarely available. Thus there is strong motivation for limited-information methods based on instrumental variables, such as the GMM and two-stage sequential estimation. These have been extended to nonlinear models in a number of special cases, sometimes on an *ad hoc* computationally feasible basis, though not always with supporting formal justification. However, the consistency property of the two-step estimator may depend on particular assumptions about the structure of dependence. For example, in discrete outcome models sequential two-step estimation, based on the replacement of an endogenous variable by a fitted value, yields a consistent estimator only if the causal structure is recursive (Blundell and Powell, 2004; Chesher, 2005).

We consider computational issues of full information estimation in a model where the outcome of interest is a non-negative count which depends on a set of variables that includes dummy variables generated by a multinomial choice model (see Deb and Trivedi, 2006a, 2006b). The following section formally describes the nature of dependence.

15.6.1 Model specification

Consider a selectivity model in which individual i chooses a treatment from a set of three or more choices. This implies a multinomial choice model with a benchmark choice. $E[\mathrm{V}_{ij}^*]$ denotes the indirect utility of selecting the jth treatment, $j = 0, 1, 2, \ldots, J$, and:

$$E[\mathrm{V}_{ij}^*] = \mathbf{z}_i^\top \boldsymbol{\alpha}_j + \delta_j l_{ij} + \eta_{ij}, \tag{15.36}$$

where $\mathbf{z}_i$ denotes exogenous covariates, with associated parameters $\boldsymbol{\alpha}_j$, and η_{ij} are i.i.d. error terms. In addition, $E[\mathrm{V}_{ij}^*]$ includes a latent factor l_{ij} which incorporates unobserved characteristics common to individual i's treatment choice and outcome. The l_{ij} are assumed to be independent of η_{ij}. Without loss of generality, let $j = 0$ denote the control group and $E[\mathrm{V}_{i0}^*] = 0$.

Let d_j be binary (dummy) variables representing the observed treatment choice and $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{iJ})^\top$. In addition, let $\mathbf{l}_i = (l_{i1}, l_{i2}, \ldots, l_{iJ})^\top$. Then the probability of treatment can be represented as:

$$\Pr(\mathbf{d}_i|\mathbf{z}_i, \mathbf{l}_i) = \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 l_{i1}, \mathbf{z}_i^\top \boldsymbol{\alpha}_2 + \delta_2 l_{i2}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J l_{iJ}), \tag{15.37}$$

where $\mathbf{g}$ is an appropriate multinomial probability distribution. Specifically, we specify a mixed multinomial logit structure (MMNL) defined as:

$$\Pr(\mathbf{d}_i|\mathbf{z}_i, \mathbf{l}_i) = \frac{\exp(\mathbf{z}_i^\top \boldsymbol{\alpha}_j + \delta_j l_{ij})}{1 + \sum_{k=1}^{J} \exp(\mathbf{z}_i^\top \boldsymbol{\alpha}_k + \delta_k l_{ik})}. \tag{15.38}$$

For the count variable the expected outcome equation is:

$$E(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{J} \gamma_j d_{ij} + \sum_{j=1}^{J} \lambda_j l_{ij}, \tag{15.39}$$

where $\mathbf{x}_i$ is a set of exogenous covariates with associated parameter vector $\boldsymbol{\beta}$ and the γ_j denote the treatment effects relative to the control. $E(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i)$ also depends on latent factors l_{ij}, i.e., the outcome is affected by unobserved characteristics that also affect selection into treatment. When λ_j, the factor loading parameter, is positive (negative), treatment and outcome are positively (negatively) correlated through unobserved characteristics, i.e., there is positive (negative) selection, with $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ the associated parameter vectors respectively.

Assume that f is the negative binomial-2 density:

$$f(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \frac{\Gamma(y_i + \psi)}{\Gamma(\psi)\Gamma(y_i + 1)} \left(\frac{\psi}{\mu_i + \psi}\right)^{\psi} \left(\frac{\mu_i}{\mu_i + \psi}\right)^{y_i}, \tag{15.40}$$

where $\mu_i = E(y_i|\ \mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \mathbf{l}_i^\top \boldsymbol{\lambda})$ and $\psi \equiv 1/\alpha$ ($\alpha > 0$) is the overdispersion parameter.

As in the standard MNL model, the parameters in the MMNL are only identified up to a scale. Therefore, a normalization for the scale of the latent factors is required. We assume $\delta_j = 1$ for each j, without loss of generality. Although the model is identified when $\mathbf{z}_i = \mathbf{x}_i$, for robust identification it would be preferable to include some variables in $\mathbf{z}_i$ that are not included $\mathbf{x}_i$, i.e., identification via exclusion restrictions is the preferred approach.

15.6.2 Estimation algorithm

The joint distribution of treatment and outcome variables, conditional on the common latent factors, can be written as the product of the marginal density of treatment and the conditional density:

$$\begin{aligned} \Pr(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{l}_i) &= f(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) \times \Pr(\mathbf{d}_i|\mathbf{z}_i, \mathbf{l}_i) \\ &= f(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \mathbf{l}_i^\top \boldsymbol{\lambda}) \\ &\quad \times \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 l_{i1}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J l_{iJ}). \end{aligned} \tag{15.41}$$

The problem in estimation arises because the l_{ij} are unknown. Assume that the l_{ij} are i.i.d. draws from a standard normal distribution so their joint distribution $\mathbf{h}$ can be integrated out of the joint density, i.e.:

$$\begin{aligned} \Pr(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) &= \int \Big[f(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \mathbf{l}_i^\top \boldsymbol{\lambda}) \\ &\quad \times \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 l_{i1}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J l_{iJ}) \Big] \mathbf{h}(\mathbf{l}_i) d\mathbf{l}_i. \end{aligned} \tag{15.42}$$

The main computational problem, given suitable specifications for $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{h}_j$, is that the integral (15.42) does not have, in general, a closed-form solution. As was explained in section 15.3.3, this difficulty can be tackled using simulation-based estimation (Gouriéroux and Monfort, 1996). Note that:

$$\begin{aligned} \Pr(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) &= E\Big[f(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \mathbf{l}_i^\top \boldsymbol{\lambda}) \\ &\quad \times \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 l_{i1}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J l_{iJ}) \Big] \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \Big[f(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \tilde{\mathbf{l}}_{is}^\top \boldsymbol{\lambda}) \ , \\ &\quad \times \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 \tilde{l}_{i1s}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J \tilde{l}_{iJs}) \Big], \end{aligned} \tag{15.43}$$

where $\tilde{\mathbf{l}}_{is}$ is the sth draw (from a total of S draws) of a pseudo-random number from the density $\mathbf{h}$. The simulated log-likelihood function for the data is given by:

$$\begin{aligned} \ln l(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) &\approx \sum_{i=1}^{N} \ln \left(\frac{1}{S} \sum_{s=1}^{S} \Big[f(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{d}_i^\top \boldsymbol{\gamma} + \tilde{\mathbf{l}}_{is}^\top \boldsymbol{\lambda}) \tilde{l}_{is} \ , \right. \\ &\quad \left. \times \mathbf{g}(\mathbf{z}_i^\top \boldsymbol{\alpha}_1 + \delta_1 \tilde{l}_{i1s}, \ldots, \mathbf{z}_i^\top \boldsymbol{\alpha}_J + \delta_J \tilde{l}_{iJs}) \Big] \right). \end{aligned} \tag{15.44}$$

Provided that S is sufficiently large, maximization of the simulated log-likelihood is equivalent to maximizing the log-likelihood. The covariance of the MSL estimates may be obtained using the robust Eicker–White formula.

Table 15.6 Endogenous NB, MSL with $S = 50$ or 100

	(1) $S = 50$		*(2)* $S = 100$	
	Coef.	*Std. error*	*Coef.*	*Std. error*
HMO				
AGE	0.173	0.018	0.172	0.0184
FIRMSIZE	0.0224	0.00120	0.0224	0.00121
Intercept	−1.449	0.0735	−1.446	0.0737
OMC				
AGE	0.229	0.0301	0.231	0.0301
FIRMSIZE	0.0190	0.00194	0.0190	0.00193
Intercept	−3.505	0.125	−3.514	0.125
DOCVIS				
HMO	1.161	0.0573	1.080	0.254
OMC	0.511	0.173	0.850	0.424
AGE	0.225	0.0125	0.224	0.0123
Intercept	−0.519	0.0760	−0.529	0.103
LNALPHA				
Intercept	0.197	0.0625	0.163	0.0716
λ_{HMO}				
	−0.976	0.0498	−0.866	0.306
λ_{OMC}				
	−0.107	0.180	−0.473	0.474
log-like	−41725		−41722	
Comp. time	932.11 seconds		1165.27 seconds	

15.6.3 Example: MSL estimation

To illustrate the method, we use pooled data from the Medical Expenditure Panel Surveys 1996–2003. The sample consists of 13,469 observations on persons between the ages of 19 and 64. The outcome variable is the number of doctor visits in a year (DOCVIS) and the multinomial treatment variable describes the type of health insurance plan (INSTYPE) and takes three values: (i) fee-for-service (FFS)–the control; (ii) health maintenance organizations (HMO); (iii) other managed care organizations (OMC). Exogenous covariates are AGE and FIRMSIZE; the latter serves as an exclusion restriction, i.e., as an instrument. The specification is deliberately (over)simplified by excluding many variables that would appear both in the choice and outcome models. Our objective is to demonstrate the feasibility of computation. MSL estimates, obtained using Halton sequences with $S = 50$ and 100, are given in Table 15.6.

The results show that even for such a simplified example the computational time is nontrivial. The sample size in this example is fairly large, which is expected to improve the precision of the estimation. Although the log-likelihood values are similar for $S = 50$ and $S = 100$, we see some differences in the coefficient estimates.

Train (2003, Ch. 9) gives examples in which 100 Halton draws have efficiency that exceeds that of 1,000 random draws. However, even $S = 100$ may not be high enough. The penalty for setting S too low is potential bias of the estimator, but having to determine the appropriate value of S by trial and error is a disadvantage. One expects longer computational time when additional regressors appear in the model, as would be the case with a more realistic model that adds further sociodemographic and health status factors. In this example, one factor-loading coefficient is estimated to be significantly different from zero, which confirms that endogeneity of the HMO variable is an empirically important consideration.

15.7 Concluding remarks

In applied microeconometrics computational matters have always featured prominently, and computational convenience has often been a criterion for choosing a methodology. Historically, advances in the quality and scope of research have moved in tandem with computational advances. During the 1960s and 1970s, the computational treatment of sample selection, discrete choice, nonlinearities, and limited dependent variable models remained central to the research agenda in microeconometrics. For example, in his 1976 survey paper on quantal choice analysis, McFadden mentioned computation of the multinomial probit model as an important unsolved problem, and remarked that "It would be particularly useful to achieve a computational breakthrough on the multinomial normal model." This remained an important research topic for close to two decades. Although such topics have not disappeared altogether, their importance is now smaller. Taking advantage of raw computing power, simulation-based estimation and inference methods based on resampling have emerged as feasible and practical approaches to many computational problems. Although this chapter did not survey the Bayesian approaches, the advances in this area have also been revolutionary; indeed, there are numerous cases in which the Bayesian MCMC computational approaches have proved more attractive than their frequentist counterparts. However, a major persistent computational challenge remains. It arises from the goal of constructing empirical models that can address important and detailed issues of public policy without resorting to excessive use of parametric restrictions. Such models are inherently structural, dynamic and high dimensional, and they often attempt to accommodate the heterogeneity in tastes, constraints and objectives of decision makers. These models face both the conceptual problems of identification and computational problems of implementation. It seems safe to predict that such challenges will remain with us for the foreseeable future.

Notes

1. Notable exceptions are source codes written in GAUSS, MATLAB and S-PLUS® that can be often interpreted by OxGAUSS, Octave and R respectively. The latter are non-proprietary languages.

2. A direct application of this capability is in storing a *sparse* matrix (a matrix populated primarily with zeros). By only storing the non-zero entries as opposed to storing all entries, it can yield huge savings in memory that can potentially speed up running time.
3. *NIX is often used to describe UNIX and other UNIX-like platforms, i.e., UNIX, BSD, and GNU/Linux distributions.
4. For example, if a local installation of R exists, then a C program has access to all R's pseudo-random number generators via the inclusion of `#include<R.h>` at the beginning of the C pseudo-code.
5. A cluster is a group of servers and other resources that act like a single system.
6. An implementation in Stata is described in Drukker and Gates (2006).
7. We would like to thank Jeffrey S. Racine for providing us with the necessary software to run these experiments at Indiana University's High Performance Clusters.

References

Aguirregabiria, V. and P. Mira (2002) Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models. *Econometrica* **70**(4), 1519–43.

Aguirregabiria, V. and P. Mira (2007) Dynamic discrete choice structural models: a survey. Working Papers Tecipa-297, University of Toronto, Department of Economics.

Aitken, M. and D.B. Rubin (1985) Estimation and hypothesis testing in finite mixture models. *Journal of Royal Statistical Society, Series B* **47**(1), 67–75.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Mass.: Harvard University Press.

Barrodale, J. and F.D.K. Roberts (1973) An improved algorithm for discrete $l - 1$ linear approximation. *SIAM Journal on Numerical Analysis* **10**(5), 839–48.

Berndt, E.R. (1991) *The Practice of Econometrics: Classic and Contemporary* (first edition). Reading, Mass.: Addison-Wesley.

Bhat, C.R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological* **35**(17), 677–93.

Blundell, R.W. and J.L. Powell (2004) Endogeneity in semiparametric binary response models. *Review of Economic Studies* **71**(7), 655–79.

Böhning, D. (1995) A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference* **47**(1–2), 5–28.

Buchinsky, M. (1995) Quantile regression, Box–Cox transformation model, and the U.S. wage structure, 1963—1987. *Journal of Econometrics* **65**(1), 109–54.

Chesher, A. (2005) Nonparametric identification under discrete variation. *Econometrica* **73**(5), 1525–50.

Chib, S. (2004) MCMC technology. In J. Gentle, W. Härdle and Y. Mori (eds.), *Handbook of Computational Statistics: Concepts and Fundamental, Volume* I, pp. 71–102. Heidelberg: Springer-Verlag.

Creel, M. (2005) User-friendly parallel computations with econometric examples. *Computational Economics* **26**(2), 107–28.

Croissant, Y. (2006) *Ecdat: Data Sets for Econometrics*. R package version 0.1–5.

Davidson, R. and J. MacKinnon (2006) Bootstrap methods in econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics: Econometric Theory, Volume 1*, pp. 812–38. Basingstoke: Palgrave Macmillan.

Deb, P. and P.K. Trivedi (1997) Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**(3), 313–26.

Deb, P. and P.K. Trivedi (2002) The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics* **21**(4), 601–25.

Deb, P. and P.K. Trivedi (2006a) Maximum simulated likelihood estimation of a negative-binomial regression model with multinomial endogenous treatment. *Stata Journal* **6**(2), 1–10.

Deb, P. and P.K. Trivedi (2006b) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econometrics Journal* **9**(2), 307–31.

Doornik, J. (2006) The role of simulation in econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Economics: Econometric Theory, Volume 1*, pp. 787–811. Basingstoke: Palgrave Macmillan.

Drukker, D. and R. Gates (2006) Generating Halton sequences using Mata. *Stata Journal* **6**(2), 214–28.

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models* (first edition). Springer Series in Statistics. Heidelberg: Springer.

Gerfin, M. (1996) Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics* **11**(3), 321–39.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In J. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics, Volume 4*, pp. 169–93. Oxford: Oxford University Press.

Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics* (first edition). New York: Wiley.

Gouriéroux, C. and A. Monfort (1996) *Simulation Based Econometrics Methods* (first edition). New York: Oxford University Press.

Greene, W.H. (2004a) The behaviour of the fixed effects estimator in nonlinear models. *Econometrics Journal* **7**(1), 98–119.

Greene, W.H. (2004b) Fixed effects and the incidental parameters problem in the tobit model. *Econometric Reviews* **23**(2), 125–48.

Hajivassiliou, V., D. McFadden and P. Ruud (1996) Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results. *Journal of Econometrics* **72**(1–2), 85–134.

Härdle, W., P. Hall and H. Ichimura (1993) Optimal smoothing in single-index models. *Annals of Statistics* **21**(1), 157–78.

Hastie, T. (2006) *GAM: Generalized Additive Models*. R package version 0.98.

Hayfield, T. and J.S. Racine (2007) *NP: Nonparametric Kernel Smoothing Methods for Mixed Datatypes*. R package version 0.13–1.

Ho, M. (1995) Essays on the housing market. Unpublished Ph.D. dissertation, University of Toronto.

Horowitz, J.L. (1998) *Semiparametric Methods in Econometrics*. Volume 131 of Lectures Notes in Statistics. New York: Springer-Verlag.

Hotz, V. and R. Miller (1993) Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies* **60**(3), 497–529.

Huynh, K. and D. Jacho-Chávez (2007) Conditional density estimation: an application to the Ecuadorian manufacturing sector. *Economics Bulletin* **3**(62), 1–6.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics*. **58**, 71–120.

Keane, M.P. (1994) A computationally practical simulation estimator for panel data. *Econometrica* **62**(1), 95–116.

Keane, M. and K. Wolpin (1994) The solutions and estimation of discrete choice dynamic programming models by simulation and interpretation: Monte Carlo evidence. *Review of Economic Studies* **76**(4), 648–72.

Klein, R.W. and R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica* **61**(2), 387–421.

Li, Q. and J. Racine (2003) Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* **86**(2), 266–92.

Li, Q. and J.S. Racine (2007) *Nonparametric Econometric: Theory and Practice* (first edition). Princeton: Princeton University Press.

Lindsay, B.G. (1983) The geometry of mixture likelihoods: a general theory. *Annals of Statistics* **11**(1), 86–94.

Linton, O.B. (1995) Second order approximation in the partially linear regression model. *Econometrica* **63**(3), 1079–112.
Linton, O.B. and W. Härdle (1996) Estimating additive regression with known links. *Biometrika* **83**, 529–40.
Linton, O.B. and J.P. Nielsen (1995) A kernel model of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93–100.
Martins, M.F.O. (2001) Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. *Journal of Applied Econometrics* **16**(1), 23–39.
McFadden, D. (1976) Quantal choice analysis: a survey. *Annals of Economic and Social Measurement* **5**(4), 363–90.
McFadden, D. and P.A. Ruud (1994) Estimation by simulation. *Review of Economic Studies* **76**(4), 591–608.
McFadden, D. and K. Train (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics* **15**(5), 447–70.
McLachlan, G. and D. Peel (2000) *Finite Mixture Models*. New York: John Wiley.
Mills, J.A. and S. Zandvakili (1997) Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* **12**(2), 133–50.
Nerlove, M. (2004) Programming languages: a short history for economists. *Journal of Economic and Social Measurement* **29**, 189–203.
Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics* (first edition). Themes in Modern Econometrics. Cambridge: Cambridge University Press.
Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1992) *Numerical Recipes in C: The Art of Scientific Computing* (second edition). New York: Cambridge University Press.
Pudney, S. (1989) *Modelling Individual Choice: The Econometrics of Corners, Kinks and Holes*. Oxford: Blackwell.
Racine, J. (2002) Parallel distributed kernel estimation. *Computational Statistics & Data Analysis* **40**(2), 293–302.
Racine, J. and A. Ullah (2006) Nonparametric econometrics In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1*, pp. 1001–34. Basingstoke: Palgrave Macmillan.
Renfro, C.G. (2004a) A compendium of existing econometric software packages. *Journal of Economic and Social Measurement* **29**, 359–409.
Renfro, C.G. (2004b) Some milestones in econometric computing. *Journal of Economic and Social Measurement* **29**, 111–14.
Robinson, P.M. (1987) Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55**(4), 875–91.
Robinson, P.M. (1988) Root n-consistent semiparametric regression. *Econometrica* **56**, 931–54.
Rust, J. (1986) Structural estimation of Markov decision processes. In R.F. Engle and D. McFadden (eds.), *Handbook of Econometrics, Volume 4*, pp. 3081–143. Amsterdam: North-Holland.
Rust, J. (1987) Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. *Econometrica* **55**(5), 999–1033.
Rust, J. (1994) Estimation of dynamic structural models, problems and prospects: discrete decision processes. In C.A. Sim (ed.), *Advances in Econometrics: Sixth World Congress*, pp. 5–33. Volume II of Econometric Society Monographs. Cambridge: Cambridge University Press.
Rust, J. (1997) Using randomization to break the curse of dimensionality. *Econometrica* **65**(3), 487–516.
Rust, J. and C. Phelan (1997) How social security and medicare affect retirement behaviour in a world of incomplete markets. *Econometrica* **65**(4), 781–831.
Scott, D. and S. Sheather (1985) Kernel density estimation with binned data. *Communications in Statistics Theory and Methods* **14**, 1353–9.

Silverman, B.W. (1982) Algorithm AS176. kernel density estimation using the fast Fourier transform. *Applied Statistics* **31**, 166–72.

Slater, L.J. (2004) EDSAC: recollections on early days in the Cambridge computing laboratory. *Journal of Economic and Social Measurement* **29**(1–3), 119–22.

Stone, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* **13**(2), 689–705.

Train, K. (2003) *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Yatchew, A. (2003) *Semiparametric Regression for the Applied Econometrician* (first edition). Themes in Modern Econometrics. Cambridge: Cambridge University Press.

This page intentionally left blank

Part VI
Applications of Econometrics to Economic Policy

This page intentionally left blank

16

The Econometrics of Monetary Policy: An Overview

Carlo A. Favero

Abstract

This chapter concentrates on the econometrics of monetary policy. We describe the evolution of models estimated to evaluate the macroeconomic impact of monetary policy. We argue that the main challenge for the econometrics of monetary policy is in combining theoretical models and information from the data to construct empirical models. The failure of the large econometric models at the beginning of the 1970s might be explained by their incapability of taking proper account of both these aspects. The critiques by Lucas and Sims have generated an alternative approach which, at least initially, has been almost entirely dominated by theory. The LSE approach has instead concentrated on the properties of the statistical models and on the best way of incorporating information from the data into the empirical models, paying little attention to the economic foundation of the adopted specification. The realization that the solution of a DSGE model can be approximated by a restricted VAR, which is also a statistical model, has generated a potential link between the two approaches. The open question is which type of VARs are most appropriate for the econometric analysis of monetary policy.

16.1 The econometrics of monetary policy: what have we learnt?

Econometric evaluation of monetary policy started with large simultaneous equation models, in the tradition of the Cowles Commission. This first generation of models was largely driven by the IS/LM framework, in which the supply side was

left virtually unmodeled and relative price movements were not considered (see Fukac and Pagan, 2006). Large-scale models were obtained by specifying equations that described the determinants of variables in the national accounting identity for gross domestic product (GDP), for example, investment and consumption. This approach was aimed at the quantitative evaluation of the effects of modification in the variables controlled by the monetary policy maker (the instruments of monetary policy) on the macroeconomic variables which represent the final goals of the policy maker. The analysis was performed in three stages: *specification and identification* of the theoretical model, *estimation* of the relevant parameters and assessment of the dynamic properties of the model, with particular emphasis on the long-run properties, and *simulation* of the effects of monetary policies.

The crucial feature of the identification-specification stage was that the specified empirical model was usually loosely related to theoretical models and that identification was achieved by imposing numerous a priori restrictions attributing exogeneity status to a number of variables. As a consequence, identification was usually achieved within Cowles Commission models with a large number of overidentifying restrictions.

Interestingly, traditional modeling was aware of the presence of some misspecification in the estimated equations. This resulted in a departure from the conditions which warrant that ordinary least squares (OLS) estimators are best linear unbiased estimators (BLUE). The solution proposed was not re-specification but, instead, a modification of the estimation techniques. This is well reflected in the structure of the traditional textbooks (see, for example, Goldberger, 1991, where the OLS estimator is introduced first and then different estimators are considered as solutions to different pathologies in the model residuals). Pathologies are identified as departures from the assumptions which guarantee that OLS estimators are BLUE.

Stagflation condemned the first-generation models in the late 1970s, as they "did not represent the data, . . . did not represent the theory . . . [and] were ineffective for practical purposes of forecasting and policy evaluation" (Pesaran and Smith, 1995). Different explanations of the failure of these models were proposed. We classify them into diagnoses related to the solution of the structural identification problem and diagnoses related to the (lack of a) solution of the statistical identification problem.

The distinction between structural and statistical identification has been introduced by Spanos (1990). Structural models can be viewed statistically as a reparameterization, possibly (in the case of overidentified models) with restrictions, of the reduced form. Structural identification refers to the uniqueness of the structural parameters, as defined by the reparameterization and restriction mapping from the statistical parameters in the reduced form, while statistical identification refers to the choice of a well-defined statistical model as the reduced form.

The Lucas (1976) and Sims (1980) critiques are the diagnoses related to the solution of the identification problem. Lucas questions the superexogeneity status of the policy variables. and criticizes the identification scheme proposed by the Cowles Commission by pointing out that these models do not take expectations into account explicitly. Therefore, the identified parameters within the Cowles

Commission approach are a mixture of "deep parameters," describing preference and technology in the economy, and expectational parameters which, by their nature, are not stable across different policy regimes. The main consequence of such instability is that traditional structural macro-models are useless for the purpose of policy simulation.

Sims reinforced Lucas' point by labeling the Cowles Commission restrictions as "incredible"; in fact, no variable can be deemed as exogenous in a world of forward-looking agents whose behavior depends on the solution of an intertemporal optimization model. Optimality of monetary policy requires its endogeneity. Note also that, by invalidly imposing exogeneity of monetary policy, the model might induce a spuriously significant effectiveness of policy in the determination of macroeconomic variables. Endogeneity of policy does generate correlations between macroeconomic and policy variables, which, by invalidly assuming policy as exogenous, can be interpreted as a causal relation running from policy to the macroeconomic variables.

The diagnosis related to the specification of the statistical model explains the ineffectiveness of the Cowles Commission models for the practical purposes of forecasting and policy as being due to their incapability of representing the data. The root of the failure of the traditional approach lies in the inadequate attention paid to the statistical model implicit in the estimated structure. The diagnosis related to the specification of the statistical model gave rise to the LSE approach to macroeconometric modeling[1] and to the "structural cointegrating VAR" approach. The LSE approach has greatly emphasized the importance of a correct dynamic specification of the reduced form model and has placed very little emphasis on the explicit modeling of the economy based on intertemporal optimization. Recently the link between theory and dynamic specification has been re-established by a research approach based on the belief that economic theory is most informative about the long-run relationships between the relevant variables, proposed by Hashem Pesaran and a number of co-authors (see, for example, Pesaran and Shin, 2002; Garratt *et al*., 2006) in the so-called "structural cointegrating VAR approach." This approach is based on testing theory-based overidentifying restrictions on the long-run relations to provide a statistically coherent framework for the analysis of the short run.

The Lucas and Sims critiques have instead generated a totally new approach to econometric policy evaluation. These great critiques made clear that questions like "How should a central bank respond to shocks in macroeconomic variables?" are to be answered within the framework of quantitative monetary general equilibrium models of the business cycle. So the answer should rely on a theoretical model rather than on an empirical *ad hoc* macroeconometric model. Initially, this approach led to the construction of real business cycle (RBC) models where monetary policy played no role in explaining macroeconomic fluctuations. Moreover, these models depended on a limited numbers of structural parameters that were not estimated but calibrated. This period has been labeled by John Taylor (2005) as the "dark age" of the econometrics of monetary policy. This "dark age" came to an end

as a consequence of developments in macroeconomic theory and empirical modeling. On the theory side, the realization of the importance of price stickiness and of slow adjustment to the forward-looking rational expectations equilibria led to the "renaissance" of the role of monetary policy in understanding macroeconomic fluctuations. At the same time a new role was attributed to empirical analysis of providing evidence on the stylized facts to include in the theoretical model adopted for policy analysis and deciding between competing general equilibrium monetary models. This new role emerged with the realization that the solution of a dynamic stochastic general equilibrium (DSGE) model can be well approximated by a vector autoregressive (VAR) model, and VARs have become the natural tool for model evaluation.

The use of VARs led to the establishment of a number of facts and features to be included in models for monetary policy evaluation, well described by Christiano, Eichenbaum and Evans (2005) and Sims (2007).

1. Since VAR models are not estimated to yield advice on the best policy but rather to provide empirical evidence on the response of macroeconomic variables to policy impulses in order to discriminate between alternative theoretical models of the economy, it then becomes crucial to identify policy actions using restrictions independent from the theoretical models of the transmission mechanism under empirical investigation, taking into account the potential endogeneity of policy instruments.
2. Most of the monetary actions are systematic responses to the state of the economy, so there is very little in the way of random fluctuations in policy to produce business cycles.
3. Money supply is close to a random walk and monetary aggregate shocks do not look like monetary policy shocks in their effect. The foundation of the way people think about monetary policy is based on interest rate adjustments.

The main results of the VAR-based evaluation model is that, in order to match fluctuations in the data, any model must feature some attrition that causes temporary but rather persistent deviations from the long-run equilibrium defined by a frictionless neoclassical economy.

Adding frictions implies increasing the number of parameters, especially along the dimension of parameters little related to theory. As a consequence, calibration became impractical for attributing numerical values to the DSGE parameters and estimation came back into fashion. However, estimating DSGE models by classical maximum likelihood methods proved to be very hard, as the convergence of the estimates to values that ensure a unique stable solution turned out to be practically impossible to achieve when implementing unconstrained maximum likelihood estimation. Note that three types of solution are possible for a DSGE model, depending on its parameterization: no stable rational expectations solution exists, the stable solution is unique (determinacy), or there are multiple stable solutions (indeterminacy). Determinacy is a prerequisite in order to use a model to simulate the effects of economic policy.[2]

The practical impossibility of applying the classical maximum likelihood principle to estimate DSGE models paved the way for Bayesian methods. These methods have been used both for parameter estimation (see, for example, Smets and Wouters, 2003) and model evaluation (Del Negro and Schorfheide, 2004). As clearly pointed out by Sims (2007), this practice leads to a new interaction between theory and empirical analysis where the theoretical DSGE model should not be considered as a model for the data but as a generator of a prior distribution for the empirical model.

16.2 The econometrics of monetary policy in large econometric models

Consider a model designed to evaluate the effect of monetary policy. A first-generation structural model can be represented as follows:

$$\mathbf{A}\begin{pmatrix}\mathbf{Y}_t\\ \mathbf{M}_t\end{pmatrix} = \mathbf{C}_1(L)\begin{pmatrix}\mathbf{Y}_{t-1}\\ \mathbf{M}_{t-1}\end{pmatrix} + \mathbf{B}\begin{pmatrix}\boldsymbol{\nu}_t^Y\\ \boldsymbol{\nu}_t^M\end{pmatrix}, \tag{16.1}$$

$$\begin{pmatrix}\boldsymbol{\nu}_t^Y\\ \boldsymbol{\nu}_t^M\end{pmatrix} | I_{t-1}\Bigg) \sim (\mathbf{0}, \mathbf{I}).$$

The vector of n variables of interest is partitioned into two sub-sets: $\mathbf{Y}$, which represents the vector of macroeconomic variables of interest, and $\mathbf{M}$, the vector of monetary policy variables determined by the interaction between the policy maker and the economy.

The probabilistic structure for the variables of interest is determined by the implied reduced form. This statistical model has the following representation:

$$\begin{pmatrix}\mathbf{Y}_t\\ \mathbf{M}_t\end{pmatrix} = D_1(L)\begin{pmatrix}\mathbf{Y}_{t-1}\\ \mathbf{M}_{t-1}\end{pmatrix} + \mathbf{u}_t, \tag{16.2}$$

$$\mathbf{u}_t = \begin{pmatrix}\mathbf{u}_t^Y\\ \mathbf{u}_t^M\end{pmatrix},$$

$$u_t \mid I_{t-1} \sim n.i.d.\left(\mathbf{0}, \sum\right),$$

$$\begin{pmatrix}\mathbf{Y}_t\\ \mathbf{M}_t\end{pmatrix} | I_{t-1}\Bigg) \sim \left(D_1(L)\begin{pmatrix}\mathbf{Y}_{t-1}\\ \mathbf{M}_{t-1}\end{pmatrix}, \sum\right).$$

This system specifies the statistical distribution for the vector of variables of interest conditional upon the information set available at time $t-1$.[3] In relating the structure of interest to the statistical model a crucial identification problem has to be solved, since there is more than one structure of economic interest which can give rise to the same statistical model for our vector of variables.

Any given structure (16.1) will give rise to the observed reduced form (16.2) when the following restrictions are satisfied:

$$\mathbf{A}^{-1}\mathbf{C}_1(L) = D_1(L), \quad \mathbf{A}\begin{pmatrix} \mathbf{u}_t^Y \\ \mathbf{u}_t^M \end{pmatrix} = \mathbf{B}\begin{pmatrix} \boldsymbol{\nu}_t^Y \\ \boldsymbol{\nu}_t^M \end{pmatrix}.$$

There exists a whole class of structures which produce the same statistical model (16.2) under the same class of restrictions:

$$F\mathbf{A}\begin{pmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \end{pmatrix} = F\mathbf{C}_1(L)\begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \end{pmatrix} + F\mathbf{B}\begin{pmatrix} \boldsymbol{\nu}_t^Y \\ \boldsymbol{\nu}_t^M \end{pmatrix}, \tag{16.3}$$

where F is an admissible matrix, that is, it is conformable by product with $\mathbf{A}$, $\mathbf{C}_1(L)$, and $\mathbf{B}$, and $F\mathbf{A}$, $F\mathbf{C}_1(L)$, $F\mathbf{B}$ feature the same restrictions as $\mathbf{A}$, $\mathbf{C}_1(L)$, $\mathbf{B}$.

The identification problem is solved in the Cowles Commission approach by imposing restrictions on the $\mathbf{A}$, $\mathbf{C}_1(L)$ and $\mathbf{B}$ matrices so that the only admissible F matrix is the identity matrix. This is typically achieved by attributing an exogeneity status to the policy variables. Engle, Hendry and Richard (1983) illustrate that estimation requires weak exogeneity ($\mathbf{A}$ and $\mathbf{B}$ lower triangular), forecasting requires strong exogeneity ($\mathbf{A}$, $\mathbf{C}_1(L)$ and $\mathbf{B}$ lower triangular), while policy simulation requires superexogeneity, that is, strong exogeneity plus invariance of the parameters of interest to changes in the distribution of the policy variables.

Having identified the model and estimated the parameters of interest, the effect of monetary policy can be simulated. For given values of the parameters and the exogenous variables, values for the endogenous variables are recovered by finding the dynamic solution of the model.

Dynamic simulation is used to evaluate the effect of different policies, defined by specifying different patterns for the exogenous variables. Policy evaluation is implemented by examining how the predicted values of the endogenous variables change after some exogenous variables are modified. This implies simulating the model twice. First, a baseline, *control*, simulation is run. Such simulations can be run within the sample, in which case observed data are available for the exogenous variables, or outside the available sample, and values are assigned to the exogenous variables. The results of the baseline simulations are then compared with those obtained from an alternative, *disturbed*, simulation, based on the modification of the relevant exogenous variables. Policy evaluation was usually based on *dynamic multipliers,* which show the effect over time of the modification in the exogenous variables.

The construction of diagnostics for model evaluation is related to the solution of the identification problem. In fact, in the (very common) case of overidentified models, a test of the validity of the overidentifying restrictions can be constructed by comparing the restricted reduced form implied by the structural model with the reduced form implied by the just-identified model in which each endogenous variable depends on all exogenous variables with unrestricted coefficients. The statistics are derived in Anderson and Rubin (1949) and Basmann (1960). The logic of the test attributes a central role to the structural model. The statistical model of

reference for the evaluation of the structural model is derived from the structural model itself.

16.3 The different diagnoses of the failure of large econometric models

The monetary policies based on first generation models failed to prevent stagflation in the late 1970s. There are different explanations of this failure, which focused either on structural identification or on statistical identification.

16.3.1 Diagnoses related to structural identification

Lucas (1976) questions the superexogeneity status of the policy variables. He attacks the identification scheme proposed by the Cowles Commission by pointing out that these models do not take expectations explicitly into account and, therefore, the identified parameters within the Cowles Commission approach are a mixture of "deep parameters," describing preferences and technology in the economy, and expectational parameters which, by their nature, are not stable across different policy regimes. The main consequence of such instability is that traditional structural macro-models are useless for the purpose of policy simulation. To illustrate the point, assume the following data-generating process (DGP), in which expected monetary policy matters for the determination of macroeconomic variables in the economy:

$$\begin{pmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \end{pmatrix} = \begin{pmatrix} c_{01} \\ c_{02} \end{pmatrix} + \begin{pmatrix} c_{11} & c_{12} \\ 0 & c_{22} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma \\ 0 \end{pmatrix} \left(\mathbf{M}^e_{t+1}\right) + \begin{pmatrix} \mathbf{u}^Y_t \\ \mathbf{u}^M_t \end{pmatrix}. \quad (16.4)$$

A Cowles Commission model is estimated without explicitly including expectations and it will have the following specification:

$$\begin{pmatrix} 1 & a_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \end{pmatrix} = \begin{pmatrix} d_{01} \\ c_{02} \end{pmatrix} + \begin{pmatrix} d_{11} & d_{12} \\ 0 & c_{22} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{u}^Y_t \\ \mathbf{u}^M_t \end{pmatrix}. \quad (16.5)$$

Under the assumed DGP the restrictions $a_{12} = \gamma c_{22}$ and $d_{01} = \gamma c_{02}$ apply and simulation of alternative policy regimes, that is, alternative values of c_{02} and c_{22}, cannot be implemented by keeping the estimated parameters constant.

Sims (1980) reinforced the Lucas critique by emphasizing a point originally made by Liu (1960), labeling the traditional restrictions as "incredible." In fact, no variable can be deemed exogenous in a world of forward-looking agents whose behaviour depends on the solution of an intertemporal optimization model. Optimality of policy cannot be consistent with the restrictions that $\mathbf{A}$, $\mathbf{C}_1(L)$, and $\mathbf{B}$ are lower triangular. Note also that, by invalidly imposing such restrictions, the model might induce a spurious statistical effectiveness of policy in the determination of macroeconomic variables. Endogeneity of policy does generate correlations between macroeconomic and policy variables, which, by invalidly assuming policy as exogenous, can be wrongly interpreted as a causal relation running from policy to the macroeconomic variables.

16.3.2 Diagnoses related to statistical identification

The diagnosis related to the specification of the statistical model explains the ineffectiveness of the first-generation models for the practical purposes of forecasting and policy as due to their incapability of representing the data. The root of the failure of the traditional approach lies in the lack of attention paid to the statistical model implicit in the estimated structure. Any identified structure is bound to fail if the implied reduced form, that is, the statistical model, is not an accurate description of the data. The accuracy of the description of the data is to be measured by evaluating the properties of the residuals of the statistical model: "congruent" models should feature residuals that are normally distributed, free of autocorrelation and homoskedastic. Spanos (1990) considers the case of a simple demand and supply model to show how the reduced form is ignored in the traditional approach. The example is based on the market for commercial loans discussed in Maddala (1988). Most of the widely used estimators allow the derivation of numerical values for the structural parameters without even seeing the statistical models represented by the reduced form. Following this tradition, the estimated (by two-stage least squares (2SLS)) structural model is a static model that relates the demand for loans to the average prime rate, to the Aaa corporate bond rate and to the industrial production index, while the supply of loans depends on the average prime rate, the three-month bill rate and total bank deposits. The quantity of commercial loans and the average prime rate are considered as endogenous while all other variables are taken as, at least, weakly exogenous variables and no equation for them is explicitly estimated. Given that there are two omitted instruments in each equation, one overidentifying restriction is imposed in both the demand and supply equations. The validity of the restrictions is tested via the Anderson–Rubin (1949) tests, and leads to the rejection of the restrictions at the 5% level in both equations, although in the second equation the restrictions cannot be rejected at the 1% level. Estimation of the statistical model, that is, the reduced form implied by the adopted identifying restrictions, yields a model for which the underlying statistical assumptions of linearity, homoskedasticity, absence of autocorrelation and normality of residuals are all strongly rejected. On the basis of this evidence the adopted statistical model is not considered as appropriate. An alternative model is then considered which allows for a richer dynamic structure (two lags) in the reduced form. Such dynamic specification is shown to provide a much better statistical model for the data than the static reduced from. Of course, the adopted structural model implies many more overidentifying restrictions than the initial one. When tested, the validity of these restrictions is overwhelmingly rejected for both the demand and supply equations. This evidence leads Spanos to conclude that statistical identification should be distinguished from structural identification. Statistical identification refers to the choice of a well-defined statistical model, structural identification refers to the uniqueness of the structural parameters as defined by the reparameterization and restriction mapping from the statistical parameters. Lucas and Sims concentrate on model failure related to structural identification problems, but models can fail

independently from structural identification problems as a consequence of lack of statistical identification.

16.4 Model specification and model diagnostics when statistical identification matters

The diagnosis related to the specification of the statistical model gave rise to the LSE approach to macroeconometric modeling and to the "structural cointegrating VAR" approach.

There are several possible causes for the inadequacy of the statistical models implicit in structural econometric models: omission of relevant variables, or of the relevant dynamics for the included variables, or invalid assumptions of exogeneity. The LSE solution to the specification problem is the theory of reduction. Any econometric model is interpreted as a simplified representation of the unobservable DGP. For the representation to be valid or "congruent," to use Hendry's own terminology, the information lost in reducing the DGP to its adopted representation, given by the reduced form, must be irrelevant to the problem at hand. Adequacy of the statistical model is evaluated by analyzing the reduced form, that is, by checking statistical identification. Therefore, the prominence of the structural model, with respect to the reduced form representation, is reversed. The LSE approach starts its specification and identification procedure with a general dynamic reduced form model. The congruency of such a model cannot be directly assessed against the true DGP, which is unobservable. However, model evaluation is made possible by applying the general principle that congruent models should feature truly random residuals; hence, any departure of the vector of residuals from a random normal multivariate distribution should signal a misspecification. Stationarity of the statistical model is a crucial feature when the model has to be simulated. Nonstationarity in macroeconomic time series is treated in the LSE methodology by reparameterizing the reduced form VAR as a cointegrated VAR. This is achieved by imposing rank reduction restrictions on the matrix determining the long-run equilibria of the system and by solving the identification problem of cointegrating vectors (see Johansen, 1995). Once the baseline model has been validated, the reduction process begins by simplifying the dynamics and reducing the dimensionality of the model by omitting the equations for those variables for which the null hypothesis of exogeneity is not rejected. Different tests are proposed for the different concepts of exogeneity by Engle, Hendry and Richard (1983) and even the validity of the Lucas critique becomes a testable concept (Engle and Hendry, 1993; Hendry, 1988). The product of the process of reduction is a statistical model for the data, possibly discriminating between short-run dynamics and long-run equilibria. Only after this validation procedure can the structural model be identified and estimated. A just-identified specification does not require any further testing, as its implicit reduced form does not impose any further restrictions on the baseline statistical model. The validity of the overidentified specification is, instead, tested by evaluating the restrictions implicitly imposed on the general reduced form. The most popular applications of the general-to-specific specification strategy are in the

area of money demand (Baba, Hendry and Starr, 1992) and aggregate consumption expenditure (see, for example, Hendry, Muellbauer and Murphy, 1990). As is well discussed in Fukac and Pagan (2006), the LSE approach was influential in the development of a second generation of large equation models, such as the Canadian model RDX2 (Helliwell *et al.* 1991) and the MPS model at the Fed (Gramlich, 2004), which, apart from introducing much stronger supply-side features with respect to traditional IS/LM models, paid considerable attention to dynamic specification and to the implementation of error correction models. In this type of specification the static solution represented a target to which the decision variable adjusted.

In practice, the LSE approach has almost exclusively concentrated on the statistical diagnosis of the failure of large structural models and has brought more attention to the dynamic specification and the long-run properties of models built in the Cowles Commission tradition and used by policy makers. It has paid much less attention to the possibility of specifying a forward-looking microeconomically founded model consistent with the theory-based diagnosis for the failure of traditional Cowles Commission models (an interesting example of this approach can be found in Juselius and Johansen, 1999). In a recent paper, Juselius and Franchi (2007) propose formulating all the basic assumptions underlying a theoretical model as a set of hypotheses on the long-run structure of a cointegrated VAR. They also argue in favor of using an identified cointegrated VAR as a way of structuring the data that offers a number a "sophisticated" stylized facts to be matched by empirically relevant theoretical models.

The idea of constructing empirical models based on the belief that economic theory is most informative about the long-run relationships between the relevant variables has been further developed by Hashem Pesaran and a number of co-authors (see, for example, Pesaran and Shin, 2002; Garratt *et al.,* 2006) in the so-called "structural cointegrating VAR approach." This approach is based on testing theory based overidentifying restrictions on the long-run relations to provide a statistically coherent framework for the analysis of the short-run. In practice, the implementation is based on a log-linear VARX model, where the baseline VAR model is augmented with weakly exogenous variables, such as oil prices or country specific foreign variables. Theory-based cointegrating relationships are tested and, whenever not rejected, imposed on the specification. No restrictions are imposed on the short-run dynamics of the model except for the, inevitable, choice of lag length for the VARX. Models are then used to evaluate the effect of policies via generalized impulse response functions (see Pesaran and Shin, 1998) and for forecasting.

16.5 Model specification and model diagnostics when structural identification matters

The great critiques made clear that questions like "How should a central bank respond to shocks in macroeconomic variables?" are to be answered within the framework of a quantitative monetary general equilibrium model of the business cycle, a DSGE model. The general linear (or linearized around equilibrium) DSGE model takes the following form (see Sims, 2002):

$$\mathbf{\Gamma}_0\mathbf{Z}_t = \mathbf{\Gamma}_1\mathbf{Z}_{t-1} + C + \Psi\epsilon_t + \Pi\eta_t, \tag{16.6}$$

where C is a vector of constants, ϵ_t is an exogenously evolving random disturbance, and η_t is a vector of expectations errors, $\left(E_t\left(\eta_{t+1}\right) = \mathbf{0}\right)$, not given exogenously but to be treated as part of the model solution. The forcing processes here are the elements of the vector ϵ_t, which contains processes like total factor productivity or policy variables that are not determined by an optimization process. Policy variables set by optimization, typically included in $\mathbf{Z}_t$, are naturally endogenous as optimal policy requires some response to current and expected developments of the economy.[4] Expectations at time t for some of the variables of the system at time $t+1$ are also included in the vector $\mathbf{Z}_t$ whenever the model is forward-looking. Models like (16.6) can be solved using standard numerical techniques (see, for example, Sims, 2002), and the solution can be expressed as:

$$\mathbf{Z}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{Z}_{\mathbf{t-1}} + \mathbf{R}\epsilon_t,$$

where the matrices $\mathbf{A}_0$, $\mathbf{A}_1$,and $\mathbf{R}$ contain convolutions of the underlying structural model parameters. Note that the solution is naturally represented as a VAR. In fact, it is a VAR potentially with stochastic singularity, as the dimension of the vector of shocks is typically smaller than that of the vector of variables included in the VAR. However, this problem is promptly solved by adding the appropriate number of measurement errors.[5] Canonical RBC models (see, for example, Kydland and Prescott, 1982; King, Plosser and Rebelo, 1988) contain a limited number of parameters, and within this class of models the role of estimation was clearly de-emphasized and parameters have often been calibrated rather than estimated.

Calibration is extensively described in Cooley (1997). The aim of calibration is not to provide a congruent representation of the data, but simply to find values for the structural parameters of the model that are jointly compatible with the theory and the data in particular well-specified dimensions. The main difference between calibration and standard econometrics lies in the bidirectional relationship between theory and measurement that characterizes the former (see Favero, 2001). Cooley (1995, p. 60) states very clearly that in the calibration approach, data and measurement are concepts determined by the features of the theory. The empirics of calibration proceeds in several stages.

First, a preliminary, non-theoretical inspection of the data identifies some general stylized facts that any economic model should internalize. The theoretical framework at hand, then, integrated by these observed stylized facts, generates the parametric class of models to be evaluated. Once a particular model has been developed, it precisely defines the quantities of interest to be measured, and suggests how available measurements have to be reorganized if they are inconsistent with the theory.

Then, measurements are used to give empirical content to the theory, and in particular to provide empirically based values for the unknown parameters. They are chosen, according to Cooley (1997, p. 58), by specifying first some features of the data for the model to reproduce[6] and then by finding some one-to-one relationship

between these features and the deep parameters of the model. Finally, this relationship is inverted to determine the parameter values that make the model match the observed features.

From this point of view, calibration can be interpreted as a method of moments estimation procedure that focuses on a limited parameter sub-set, setting only the discrepancy between some simulated and observed moments to zero. Christiano and Eichenbaum (1992) generalize this idea and propose a variant of Hansen's (1982) generalized method of moments (GMM) procedure to estimate and assess stochastic general equilibrium models using specific moments of the actual data. These procedures are formal developments of the basic methodological approach, and share with standard calibration the focus on a limited set of previously selected moments, while standard econometric methods use, in principle, the whole available information set, weighting different moments exclusively according to how much information on them is contained in the actual data, as, for example, in maximum likelihood methods.

Generally, not all parameters can be calibrated, simply because there are more unknown parameters than invertible relationships. A sub-set of them has to be left to more standard econometric techniques.

Once a parameterization is available, the model is simulated and different kinds of numerical exercises are performed. At this stage model evaluation can also be implemented. Model evaluation was initially conducted by assessing the ability of the model to reproduce some particular features (of course, ones that are different from those used to calibrate it) of the data. The metric chosen to compare the observed properties and the simulated ones is a critical issue. In the traditional calibration procedure, an informal, "aesthetic" metric is used, based on the comparison between simulated and observed moments of the relevant variables (see, for example, Kydland and Prescott, 1996, p. 75). Moreover, as DSGE models are usually solved by linearizing them around equilibrium, raw data cannot be used to generate the set of statistics relevant for model evaluation. Raw data contain trends, so they are usually detrended using filtering techniques before using them to generate the relevant statistics.[7]

Model evaluation in DSGE models became much more sophisticated when the practice started to exploit the fact that a solved DSGE model is a VAR.

If we repartition the vector of variables included in the VAR into macroeconomic and policy variables $[\mathbf{Y}_t \mathbf{M}_t]$, the solved DSGE model could be represented as a structural VAR (SVAR):

$$\mathbf{A} \begin{pmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \end{pmatrix} = \mathbf{C}(L) \begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \end{pmatrix} + \mathbf{B} \begin{pmatrix} \boldsymbol{\nu}_t^Y \\ \boldsymbol{\nu}_t^M \end{pmatrix}. \tag{16.7}$$

Within this framework a new role for empirical analysis emerges: to provide evidence on the stylized facts to include in the theoretical model adopted for policy analysis and to decide between competing DSGE models. The operationalization of this research program in the analysis of monetary policy is very well described

in a paper by Christiano, Eichenbaum and Evans (1998). There are three relevant steps:

1. monetary policy shocks are identified in actual economies, that is, in a VAR without theoretical restrictions;
2. the response of relevant economic variables to monetary shocks is then described;
3. finally, the same experiment is performed in the model economies to compare actual and model-based responses as an evaluation tool and a selection criterion for theoretical models.

The identification of the shocks of interest is the first and most relevant step in VAR-based model evaluation. VAR modeling recognizes that identification and estimation of structural parameters is impossible without explicitly modeling expectations. Therefore, a structure like (16.7) can only be used to run special experiments that do not involve simulating different scenarios for the parameters of interest. A natural way to achieve these results is to experiment with the shocks $\boldsymbol{\nu}_t^M$. *Facts* are then provided by looking at impulse response analysis, variance decompositions and historical decompositions. Impulse response analysis describes the effect over time of a policy shock on the variables of interest, variance decomposition illustrates how much of the variance of the forecasting errors for macroeconomic variables at different horizons can be attributed to policy shocks, and historical decomposition allows the researcher to evaluate the effect of zeroing policy shocks on the variables of interest. All these experiments are run by keeping estimated parameters unaltered. Importantly, running these experiments is easier if shocks to the different variables included in the VAR are orthogonal to each other, otherwise it would not be possible to simulate a policy shock by maintaining all the other shocks at zero. As a consequence, VAR models need a structure because orthogonal shocks are normally not a feature of the statistical model. This fact generates an identification problem. In the reduced form we have:

$$\begin{pmatrix} \mathbf{Y}_t \\ \mathbf{M}_t \end{pmatrix} = \mathbf{A}^{-1}\mathbf{C}(L)\begin{pmatrix} \mathbf{Y}_{t-1} \\ \mathbf{M}_{t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{u}_t^Y \\ \mathbf{u}_t^M \end{pmatrix},$$

where $\mathbf{u}$ denotes the VAR residual vector, normally and independently distributed with full variance-covariance matrix $\boldsymbol{\Sigma}$. The relation between the residuals in $\mathbf{u}$ and the structural disturbances in $\boldsymbol{\nu}$ is therefore:

$$\mathbf{A}\begin{pmatrix} \mathbf{u}_t^Y \\ \mathbf{u}_t^M \end{pmatrix} = \mathbf{B}\begin{pmatrix} \boldsymbol{\nu}_t^Y \\ \boldsymbol{\nu}_t^M \end{pmatrix}. \tag{16.8}$$

Undoing the partitioning, we have:

$$\mathbf{u}_t = \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\nu}_t,$$

from which we can derive the relation between the variance-covariance matrices of $\mathbf{u}_t$ (observed) and $\boldsymbol{v}_t$ (unobserved) as follows:

$$E\left(\mathbf{u}_t\mathbf{u}_t'\right) = \mathbf{A}^{-1}\mathbf{B}E\left(\boldsymbol{v}_t\boldsymbol{v}_t'\right)\mathbf{B}'\mathbf{A}^{-1}.$$

Substituting population moments with sample moments we have:

$$\widehat{\sum} = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\mathbf{I}\widehat{\mathbf{B}}'\widehat{\mathbf{A}}^{-1}, \tag{16.9}$$

$\widehat{\sum}$ contains $n(n+1)/2$ different elements, so this is the maximum number of identifiable parameters in matrices **A** and **B**. Therefore, a necessary condition for identification is that the maximum number of parameters contained in the two matrices equals $n(n+1)/2$, and such a condition makes the number of equations equal to the number of unknowns in system (16.9). As usual, for such a condition also to be sufficient for identification, no equation in (16.9) should be a linear combination of the other equations in the system (see Amisano and Giannini, 1996; Hamilton, 1994). As for traditional models, we have the three possible cases of underidentification, just-identification and overidentification. The validity of overidentifying restrictions can be tested via a statistic distributed as a χ^2 with the number of degrees of freedom equal to the number of overidentifying restrictions. Once identification has been achieved, the estimation problem is solved by applying generalized method of moments estimation.

Since VAR models are used to discriminate between alternative theoretical models of the economy, it then becomes crucial to identify policy actions using restrictions independent from the theoretical models of the transmission mechanism under empirical investigation, taking into account the potential endogeneity of policy instruments. Restrictions based on the theoretical predictions of models are clearly inappropriate, and so are the Cowles Commission type of restrictions, as they do not acknowledge the endogeneity of systematic policy. The recent literature on the monetary transmission mechanism (see Bernanke and Mihov, 1998; Christiano, Eichenbaum and Evans, 1996a; Leeper, Sims and Zha, 1996) offers good examples on how these kind of restrictions can be derived. VARs of the monetary transmission mechanism are specified on six variables, with the vector of macroeconomic non-policy variables including GDP, the consumer price index (P) and the commodity price level (Pcm), while the vector of policy variables includes the federal funds rate (FF), the quantity of total bank reserves (TR) and the amount of non-borrowed reserves (NBR). Given the estimation of the reduced form VAR for the six macro and monetary variables, a structural model is identified by: (i) assuming orthogonality of the structural disturbances; (ii) requiring that macroeconomic variables do not simultaneously react to monetary variables, while simultaneous feedback in the other direction is allowed, and (iii) imposing restrictions on the monetary block of the model reflecting the operational procedures implemented by the monetary policy maker. All identifying restrictions satisfy the criterion of independence from specific theoretical models. In fact, within the class of models estimated on monthly data, restrictions (ii) are consistent with a wide spectrum

of alternative theoretical structures and imply a minimal assumption on the lag of the impact of monetary policy actions on macroeconomic variables, whereas restrictions (iii) are based on institutional analysis. Restrictions (ii) are made operational by setting to zero an appropriate block of elements of the **A** matrix. Note that restrictions on the contemporaneous feedbacks among variables is not the only way of imposing restrictions consistent with a wide spectrum of theoretical models. In fact, such an aim could be achieved by imposing restrictions on the long run effects of shocks (for example, there is a clear consensus among macroeconomists that demand shocks have zero effect on output in the long run) or on the shape of some impulse response functions. These types of restrictions are easily imposed on the SVAR (see, for example, Blanchard and Quah, 1989; Uhlig, 1999), although one must always be aware of the effect of imposing invalid restrictions on parameter estimates (Faust and Leeper, 1997). Finally, note that partial identification can easily be implemented in a VAR model. If the relevant dimension for model comparison is the response of the economy to monetary policy shocks, then there is no need to identify the non-monetary structural shocks in the model.

16.5.1 VAR-based model evaluation: an assessment

VAR-based model evaluation can be assessed by first discussing the results achieved and their impact on model building, and then offering some considerations on the specification of the VAR and on the evaluation of the statistical model adopted.

The main results of the VAR-based evaluation model is that, in order to match fluctuations in the data, any model must feature some attrition that causes temporary but rather persistent deviations from the long-run equilibrium defined by a frictionless neoclassical economy. In a series of recent papers, Christiano, Eichenbaum and Evans (1996a, 1996b) apply the VAR approach to derive "stylized facts" on the effect of a contractionary policy shock, and conclude that plausible models of the monetary transmission mechanism should be consistent at least with the following evidence on price, output and interest rates: (i) the aggregate price level initially responds very little; (ii) interest rates initially rise, and (iii) aggregate output initially falls, with a *j*-shaped response and a zero long-run effect of the monetary impulse. Such evidence leads to the dismissal of traditional RBC models, which are not compatible with the liquidity effect of monetary policy on interest rates, and of the Lucas (1972) model of money, in which the effect of monetary policy on output depends on price misperceptions. The evidence seems to be more in line with alternative interpretations of the monetary transmission mechanism based on sticky price models (Goodfriend and King, 1997), limited participation models (Christiano and Eichenbaum, 1992) or models with indeterminacy–sunspot equilibria (Farmer, 1997). When models are extended to analyze the components of output, more frictions need to be added to explain the dynamics of consumption and investment: typically, some habit persistence is needed to explain fluctuations in consumption and some adjustment costs are needed to match the dynamics of investment and the stock of capital in the data.

Specification of the VAR and its statistical adequacy is an issue that has not received much explicit attention in the literature. It seems that the choice of the

variables included in the VAR is driven by the theoretical model. This is natural: if the theoretical model is a restricted VAR, the natural benchmark is the same VAR without restrictions. But what about potential misspecification of the statistical model?

Statistical analysis of the unrestricted VAR is rather rare, although some implicit consideration has clearly been devoted to this issue. Think, for example, of the "liquidity puzzle" and the "price puzzle" for models of the monetary transmission mechanism.

VAR models of the monetary transmission mechanism were initially estimated on a rather limited set of variables, that is, prices, money and output, and identified by imposing a diagonal form on the matrix **B** and a lower triangular form on the matrix **A**, with money coming last in the ordering of the variables included in the VAR (Choleski identification). The typical impulse responses obtained within this type of model show that prices slowly react to monetary policy, output responds in the short run, in the long run (from two years after the shock onwards) prices start adjusting and the significant effect on output vanishes. There is no strong evidence for the endogeneity of money. Macroeconomic variables play a very limited role in explaining the variance of the forecasting error of money, while money instead plays an important role in explaining fluctuations of both the macroeconomic variables.

Sims (1980) extended the VAR to include the interest rate on federal funds, ordered just before money as a penultimate variable in the Choleski identification. The idea was to assess the robustness of the above results after identifying the part of money which is endogenous to the interest rate. Impulse response functions and forecast error variance decomposition (FEVD) raise a number of issues.

1. Though little of the variation in money is predictable from past output and prices, a considerable amount becomes predictable when past short-term interest rates are included in the information set.
2. It is difficult to interpret the behavior of money as driven by money supply shocks. The response to money innovations gives rise to the "liquidity puzzle": the interest rate initially declines very slightly in response to a money shock and then starts increasing afterwards.
3. There are also difficulties with interpreting shocks to interest rates as monetary policy shocks. The response of prices to an innovation in interest rates gives rise to the "price puzzle": prices increase significantly after an interest rate hike. An accepted interpretation of the liquidity puzzle relies on the argument that the money stock is dominated by demand rather than supply shocks. Moreover, the interpretation of money as demand shocks driven is consistent with the impulse response of money to interest rates. Note also that, even if the money stock were to be dominated by supply shocks, it would reflect both the behavior of central banks and the banking system. For both these reasons the broad monetary aggregate has been substituted by narrower aggregates, bank reserves, on which it is easier to identify shocks mainly driven by the behavior

> of the monetary policy maker. The "price puzzle" has been attributed to the misspecification of the four-variables VAR used by Sims. Suppose that there exists a leading indicator for inflation to which the Fed reacts. If such a leading indicator is omitted from the VAR, then we have an omitted variable positively correlated with inflation and interest rates. Such omission makes the VAR misspecified and explains the positive relation between prices and interest rates observed in the impulse response functions. It has been observed (see Christiano, Eichenbaum and Evans, 1996b; Sims and Zha, 1996) that the inclusion of a Commodity Price Index in the VAR solves the "price puzzle."

As a result of these developments, a consensus was reached on the specification of the VAR to provide facts on the monetary transmission mechanism (MTM) as a model including prices, output, a commodity price index, the policy rate and the narrow money indicators necessary to model the market for bank reserves.

Note that the final specification is very different from the initial one and the modifications in the specification are driven by a number of puzzles found in the impulse responses of discarded VARs. One can, of course, interpret these puzzles as signals of misspecification of the VAR, but it is not clear that puzzles are the best way to diagnose misspecification of the statistical model. Think, for example, of the recent practice of identifying shocks by imposing constraints on the shape of the impulse response functions. It might well be regarded as reasonable to assume that a monetary policy restriction has a non-positive effect on inflation. Obviously, if VARs of the MTM would have always been identified by imposing this restriction, then the price puzzle would never have been observed and one is left to wonder if the consensus specification of the VAR to analyze the MTM would have evolved differently from what it did.

Another issue of crucial importance is structural stability of the parameters estimated in the VAR. If the VAR is a reduced form of a forward-looking model it is of crucial importance to estimate its parameters on a single regime. Although this issue has been explicitly recognized in some papers, (for example, Bernanke and Mihov, 1998), the consensus VAR is normally estimated on a sample including different monetary regimes. The main justification for this practice is that monetary policy shocks are robust to the different identifications generated by the different monetary policy regimes. Some authors have been left skeptical by such robustness and some criticisms have been made of VAR-based monetary policy shocks. Rudebusch (1998) argues that VAR-based monetary shocks do not make sense as they are very weakly correlated with monetary policy shocks directly derived from asset prices (the Federal Fund future). The mainstream reaction to this criticism is that, even if the two types of shocks are very weakly correlated, the impulses responses of macroeconomic variables to VAR based and financial market-based monetary policy shocks are not significantly different from each other. Rudebusch's criticism has shared the same fate as other criticisms of the VAR approach. Lippi and Reichlin (1993) pointed out that a crucial assumption in structural VAR modeling is that structural shocks are linear combinations of the residuals in reduced form VAR models, so that modern macroeconomic models which are linearized

into dynamic systems tend to include noninvertible moving average components and structural shocks are therefore not identifiable. In fact, the linearized modern macroeconomic models of the monetary transmission mechanism deliver short VARs. In such models structural shocks are combinations of the residuals in the reduced form VARs (the Wold innovations) and the Lippi–Reichlin critique does not seem to be applicable (for a further discussion of this point see Amisano and Giannini, 1996).

To sum up, although the original idea of the Cowles Commission to use the implied unrestricted reduced form as a benchmark to evaluate the structural model is clearly reflected in the VAR-based evaluation of DSGE models, the potential importance of the formal evaluation of the adequacy of the statistical model adopted has certainly not received the same attention. However, in the practice of VAR specification some attention to the issue of potential misspecification has clearly been paid, although such attention has been more related to the economic interpretation of results than to the implementation of formal statistical criteria for model evaluation.

16.6 From VAR-based model evaluation to Bayesian analysis of DSGE models

VAR-based evaluation of early DSGE models made clear that a large number of nominal and real frictions should be added to the traditional new-classical RBC models to replicate relevant features in observed data (see, for example, Christiano, Eichenbaum and Evans, 2005). Adding frictions implies increasing the number of parameters, especially along the dimension of parameters little related to theory. As a consequence, calibration became impractical for attributing numerical values to the DSGE parameters and estimation came back into fashion. However, estimating DSGE models by classical maximum likelihood methods proved to be very hard, as the convergence of the estimates to values that ensure a unique stable solution turned out to be practically impossible to achieve when implementing unconstrained maximum likelihood estimation. A paper by Ireland (2004) was an exception and obtained convergence of numerical estimates of parameters of a DSGE model to values that allow economic policy simulation. In fact, the Ireland method consists of penalizing the likelihood function along some dimension so that the range of variation of many parameters is limited (for an interesting discussion of the estimation implemented in Ireland, see Johansen, 2004).

In practice, one can think of Ireland's method as a naive Bayesian one in which some form of (very tight) prior is imposed on (at least a sub-set of) the parameters. A natural development of Ireland's proposal was to extend the naive Bayesian framework to a proper Bayesian framework. This is what happened as soon as the use of MCMC methods to derive the relevant posterior distribution of parameters became widespread (see An and Schorfeide, 2006; Del Negro *et al.*, 2006; Ruge-Murcia, 2003, for surveys and applications).

Once adopted, the Bayesian framework naturally offered some new possibilities of integrating theoretical and empirical models. Originally this interaction

was proposed as a set of modern model evaluation tools. These were generated by pairing the tradition of model evaluation in the Bayesian approach to macroeconometrics with the VAR nature of a solved DSGE model.

The Bayesian approach made its way into applied macroeconometrics to solve the problem of the lack of parsimony of VARs. In practice, data availability from a single regime poses a binding constraint on the number of variables and the number of lags that can be included in a VAR without overfitting the data. A solution to the problem of over-parameterization is to constrain the parameters by shrinking them toward some specific point in the parameter space. The Minnesota prior, proposed by Doan, Litterman and Sims (1984), uses the Bayesian approach to shrink the estimates toward the univariate random walk representation for all variables included in the VAR. Within this framework, Bayesian methods are used to save degrees of freedom on the basis of the well established statistical evidence that no-change forecasts are known to be very hard to beat for many macroeconomic variables. DeJong, Ingram and Whiteman (1996, 2000) and Ingram and Whiteman (1994) proposed evaluating RBC models by comparing the forecasting performance of a Bayesian VAR estimated via the Minnesota prior with that of a VAR in which the atheoretical prior information in the Minnesota prior was supplanted by the information in an RBC model.

In a series of papers, Del Negro and Schorfeide (2004, 2006) and Del Negro *et al.* (2006) use this approach to develop a model evaluation method that tilts coefficient estimates of an unrestricted VAR toward the restriction implied by a DSGE model. The weight placed on the DSGE model is controlled by a hyperparameter called λ. This parameter takes values ranging from 0 (no-weight on the DSGE model) to ∞ (no weight on the unrestricted VAR). Therefore, the posterior distribution of λ provides an overall assessment of the validity of the DSGE model restrictions. To see how the approach is implemented, consider that the solved DSGE model generates a restricted moving average (MA) representation for the vector of n variables of interest, $\mathbf{Z}_t = \begin{pmatrix} \mathbf{Y}_t & \mathbf{M}_t \end{pmatrix}$, that can be approximated by a VAR of order p:

$$\begin{aligned}
\mathbf{Z}_t &= \mathbf{\Phi}_0^*(\theta) + \mathbf{\Phi}_1^*(\theta)\mathbf{Z}_{t-1} + ... + \mathbf{\Phi}_p^*(\theta)\mathbf{Z}_{t-p} + \mathbf{u}_t^* \\
\mathbf{u}_t^* &\sim N\left(\mathbf{0}, \Sigma_u^*(\theta)\right) \\
\mathbf{Z}_t' &= \mathbf{X}_t'\mathbf{\Phi}^*(\theta) + \mathbf{u}_t', \\
\mathbf{X}_t &= \left[1, \mathbf{Z}_{t-1}', ..., \mathbf{Z}_{t-p}'\right] \\
\mathbf{\Phi}^*(\theta) &= \left[\mathbf{\Phi}_0^*(\theta), \mathbf{\Phi}_1^*(\theta), ..., \mathbf{\Phi}_p^*(\theta)\right]',
\end{aligned}$$

where all coefficients are convolutions of the structural parameters in the model included in the vector θ. The chosen benchmark to evaluate this model is the unrestricted VAR derived from the solved DSGE model:

$$\begin{aligned}
\mathbf{Z}_t' &= \mathbf{X}_t'\mathbf{\Phi} + \mathbf{u}_t', \\
\mathbf{\Phi} &= \left[\mathbf{\Phi}_0, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_p\right],
\end{aligned}$$

where:

$$\mathbf{\Phi} = \mathbf{\Phi}^{*}(\theta) + \mathbf{\Phi}^{\Delta}$$
$$\Sigma_u = \Sigma_u^{*}(\theta) + \Sigma_u^{\Delta}.$$

The DSGE restrictions are imposed on the VAR by defining:

$$\Gamma_{XX}(\theta) = E_\theta^D\left[\mathbf{X}_t\mathbf{X}_t'\right]$$
$$\Gamma_{XZ}(\theta) = E_\theta^D\left[\mathbf{X}_t\mathbf{Z}_t'\right],$$

where E_θ^D defines the expectation with respect to the distribution generated by the DSGE model. Such a distribution needs to be well defined. We then have:

$$\mathbf{\Phi}^{*}(\theta) = \Gamma_{XX}(\theta)^{-1}\Gamma_{XZ}(\theta).$$

Beliefs about the DSGE model parameters θ and model misspecification matrices $\mathbf{\Phi}^{\Delta}$ and Σ_u^{Δ} are summarized in prior distributions, that, as shown in Del Negro *et al.* (2006), can be transformed into priors for the VAR parameters $\mathbf{\Phi}$ and Σ_u. In particular we have:

$$\Sigma_u|\theta \sim IW\left(\lambda T\Sigma_u^{*}(\theta), \lambda T - k, n\right)$$
$$\mathbf{\Phi}|\Sigma_u, \theta \sim N\left(\Phi^{*}(\theta), \frac{1}{\lambda T}\left[\Sigma_u^{-1}\otimes\Gamma_{XX}(\theta)\right]^{-1}\right),$$

where the parameter λ controls the degree of model misspecification with respect to the VAR: for small values of λ the discrepancy between the VAR and the DSGE-VAR is large and a sizeable distance is generated between unrestricted VAR and DSGE estimators, large values of λ correspond to small model misspecification and, for $\lambda = \infty$, beliefs about DSGE misspecification degenerate to a point mass at zero. Bayesian estimation could be interpreted as estimation based on a sample in which data are augmented by a hypothetical sample in which observations are generated by the DSGE model; within this framework λ determines the length of the hypothetical sample.

Given the prior distribution, posteriors are derived by Bayes' theorem:

$$\Sigma_u|\theta, Z \sim IW\left((\lambda+1)T\hat{\Sigma}_{u,b}(\theta), (\lambda+1)T - k, n\right)$$
$$\mathbf{\Phi}|\Sigma_u, \theta, Z \sim N\left(\hat{\mathbf{\Phi}}_b(\theta), \Sigma_u\otimes\left[\lambda T\Gamma_{XX}(\theta) + \mathbf{X}'\mathbf{X}\right]^{-1}\right)$$
$$\hat{\mathbf{\Phi}}_b(\theta) = \left(\lambda T\Gamma_{XX}(\theta) + \mathbf{X}'\mathbf{X}\right)^{-1}\left(\lambda T\Gamma_{XZ}(\theta) + \mathbf{X}'\mathbf{Z}\right)$$
$$\hat{\Sigma}_{u,b}(\theta) = \frac{1}{(\lambda+1)T}\left[\left(\lambda T\Gamma_{ZZ}(\theta) + \mathbf{Z}'\mathbf{Z}\right) - \left(\lambda T\Gamma_{XZ}(\theta) + \mathbf{X}'\mathbf{Z}\right)\hat{\mathbf{\Phi}}_b(\theta)\right],$$

which shows that the smaller λ, the closer the estimates are to the OLS estimates of an unrestricted VAR, the higher λ the closer the estimates are to the values implied by the DSGE model parameters θ.

In practice, a grid search is conducted on a range of values for λ to choose that value which maximizes the marginal data density. The typical result obtained when using DSGE-VECM(λ) to evaluate models with frictions is that "the degree of misspecification in large-scale DSGE models is no longer so large as to prevent their use in day-to-day policy analysis, yet is not small enough that it can be ignored."

16.6.1 DSGE-VAR analysis: an assessment

If we consider the DSGE-VAR approach to be a model evaluation tool, we observe that it takes the Lucas and Sims critique very seriously but it ignores the issue of specification of the statistical model. The VAR used as a benchmark is the solved DSGE model that is generalized only by relaxing restrictions on parameters. The validity of the statistical model underlying the empirical specification is never questioned. Although the models are different, the evaluation strategy in the DSGE-VAR approach is very similar to the approach of evaluating models by testing overidentifying restrictions without assessing the statistical model, as implemented in Cowles Commission models. In fact, the DSGE-VAR approach is looser than that of the Cowles Commission approach as model-based restrictions are not imposed and tested but a different question is asked: restrictions are made fuzzy by imposing a distribution on them. Within this approach the relevant question becomes "What is the amount of uncertainty that we have to add to model based restrictions in order to make them compatible not with the data but with a model-derived unrestricted VAR representation of the data?" The natural question here is "How well does this procedure do in rejecting false models?" Spanos (1990) has clearly shown that modification of the structure of the statistical model could lead to dramatic changes in the outcome of tests for overidentifying restrictions. Why is this worry so strongly de-emphasized in the DSGE-VAR literature?

What are the potential sources of model-derived VAR misspecification? An obvious candidate are all those variables that are related to the misspecification of the theoretical model, but there are also all those variables that are not theory-related but are important for modeling the actual behavior of policy makers. Think, for example, of the commodity price index and the modeling of the behavior of the monetary policy authority. We have discussed in the previous section how the inclusion of this variable in a VAR to identify monetary policy shocks has been deemed important to model correctly the information set of the monetary policy maker when forecasting inflation and then to fix the "price puzzle." DSGE models do not typically include the commodity price index in their specification, and, as a consequence, the VAR derived by relaxing the theoretical restrictions in a DSGE model is misspecified. Thus the evaluation of the effects of conducting model misspecification with a "wrong" benchmark is a practically relevant one.

As a matter of fact, DSGE models tend to produce a high number of very persistent shocks (see Smets and Wouters, 2003), and this would have certainly been taken as a signal of model misspecification by an LSE-type methodology. Still, the models

do not do too badly when judged by the metric of the λ test. It would be important to have some evaluation of phenomena like this.

Another dimension potentially relevant for evaluating the statistical model underlying the VAR-DSGE is the structural stability of the VAR parameters. If the DSGE restrictions are valid, then parameters in the VAR are convolutions of structural parameters that, by their nature, should be constant over time. It is well known that tests for structural stability have problems of power, especially in the presence of multiple breaks at unknown dates. Detecting structural breaks in parameters of interest becomes even harder when structural innovations in the DSGE are allowed to have volatilities that vary over time. Justiniano and Primiceri (2005) have extended the Bayesian framework to develop an algorithm for inferring DSGE model parameters and time-varying volatilities of structural shocks. Allowing for time-varying volatilities makes the DSGE model consistent with structural breaks while keeping the deep parameters constant. However, it is hard to distinguish empirically the case for genuine stochastic volatility against a situation in which allowing for stochastic volatility in the estimation picks up parameter instability in a VAR model with constant volatility of structural shocks (see Benati and Surico, 2007).

There are alternatives to the use of a VAR as a benchmark. The limited information problem of VARs could be solved by combining traditional VAR analysis with recent developments in factor analysis for large data sets and using a factor-augmented VAR (FAVAR) as the relevant statistical model to conduct model evaluation. A recent strand of the econometric literature (Stock and Watson, 2002; Forni and Reichlin, 1996, 1998; Forni *et al.*, 2000) has shown that very large macroeconomic datasets can be properly modeled using dynamic factor models, where the factors can be considered to be an exhaustive summary of the information in the data. This approach has been successfully employed to forecast macroeconomic time series and, in particular, inflation. As a natural extension of the forecasting literature, Bernanke and Boivin (2003), and Bernanke, Boivin and Eliasz (2005) proposed exploiting these factors in the estimation of VARs. A FAVAR benchmark for the evaluation of a DSGE model will take the following specification:

$$\begin{pmatrix} \mathbf{Z}_t \\ \mathbf{F}_t \end{pmatrix} = \begin{bmatrix} \mathbf{\Phi}_{11}(L) & \mathbf{\Phi}_{12}(L) \\ \mathbf{\Phi}_{21}(L) & \mathbf{\Phi}_{22}(L) \end{bmatrix} \begin{pmatrix} \mathbf{Z}_{t-1} \\ \mathbf{F}_{t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{u}_t^Z \\ \mathbf{u}_t^F \end{pmatrix},$$

where $\mathbf{Z}_t$ are the variables included in the DSGE model and $\mathbf{F}_t$ is a small vector of unobserved factors extracted from a large dataset of macroeconomic time series that capture additional economic information relevant to model the dynamics of $\mathbf{Z}_t$. The system reduces to the standard VAR used to evaluate DSGE models if $\mathbf{\Phi}_{12}(L) = 0$. Therefore, within this context, the relevant λ test would add to the usual DSGE model-related restrictions on $\mathbf{\Phi}_{11}(L)$ the restrictions $\mathbf{\Phi}_{12}(L) = 0$.

Consolo, Favero and Paccagnini (2007) apply this idea to find that FAVAR models dominate VAR specifications generated by adopting unrestricted versions of the solution of DSGE models. Such dominance is clearly established by analysis of

residuals and evaluation of forecasting performance. However, when the Bayesian approach is applied to the DSGE-FAVAR instead of the DSGE-VAR, some support for the DSGE model is still found in the data (the optimal λ in the DSGE-FAVAR is different from zero). Moreover, the optimal combination of the DSGE model and the statistical model based on a larger information set (the FAVAR) delivers a forecasting model (the DSGE-FAVAR) that dominates all alternatives. This evidence leads to a new interaction between theory and empirical analysis, where the theoretical DSGE model should not be considered as a model for the data but as a generator of prior distributions for the empirical model. The use of the FAVAR as an empirical model allows including in the analysis the information that is not considered in the theoretical model.

Besides this application there has been no work using FAVAR to evaluate DSGE. Interestingly, what has instead happened is that FAVAR has been interpreted as the reduced form of a DSGE model. This result has been achieved by removing the assumption that economic variables included in a DSGE are properly measured by a single indicator: variables in the theoretical model are considered as unobservable and the information in the factors is used to map them (Boivin and Giannoni, 2006). This approach makes a FAVAR the reduced form of a DSGE model, although the restrictions implied by the DSGE model on a general FAVAR are very difficult to trace and model evaluation becomes even more difficult to implement. In fact, a very tightly parameterized theory model can have a very highly parameterized reduced form if one is prepared to accept that the relevant theoretical concepts in the model are combinations of many macroeconomic and financial variables. Identification of the relevant structural parameters, which is already very hard in DSGE model with observed variables (see Canova and Sala, 2005), becomes even harder. Natural advantages of this approach are increased efficiency in the estimation of the model and improved forecasting performance. However, model evaluation becomes almost impossible to pursue and a theoretical model can only by rejected by another theoretical model, while the implied statistical model is made so general that it becomes very hard to use theory as a generator of prior distributions and it becomes impossible to use the evidence from the data to reject theory.

16.7 What's next?

The main challenge for the econometrics of monetary policy is in combining of theoretical models and information from the data to construct empirical models. The failure of the large econometric models at the beginning of the 1970s might be explained by their incapability of taking proper account of both these aspects. The great critiques by Lucas and Sims have generated an alternative approach which, at least initially, has been almost entirely dominated by theory. The LSE approach has instead concentrated on the properties of the statistical models and on the best way of incorporating information from the data into the empirical models, paying little attention to the economic foundation of the adopted specification. The realization that the solution of a DSGE model can be approximated by a restricted VAR, which is also a statistical model, has generated a potential link between the

two approaches. The open question is which type of VARs are most appropriate for the econometric analysis of monetary policy.

At the moment there are a number of alternative answers to this question. A first approach looks at theoretical DSGE models as the natural way to generate prior distributions for the empirical model, which should be an (optimal) combination of a tightly parameterized theoretical model and of a more general empirical model. This approach requires the application of Bayesian methods. A second approach looks at theory as informative only for the long-run relations between economic variables, so theory should be used to specify a cointegrated VAR in which the short-run dynamics are determined by the data but the long-run properties of the model depend on testable (and tested) theoretical assumptions. Importantly, both these answers recognize the importance of both the theoretical and the statistical model, although the relative weights can be very different. Within this framework, modeling nonlinearity and structural breaks could be an important development.

The econometrics of monetary policy is now based on models that incorporate a large number of nominal and real frictions added to the traditional neoclassical RBC models to replicate relevant features in observed data. These models typically incorporate the labour market, consumers and producers behavior and monetary and fiscal policies, so the next step is probably more accurate and explicit modeling of the interaction between financial markets and product markets.

16.8 Appendix: The Sims (2002) representation of a small macroeconomic model

Consider a small New Keynesian DSGE model of the economy which features a representative household optimizing over consumption, real money holdings and leisure, a continuum of monopolistically competitive firms with price adjustment costs and a monetary policy authority which sets the interest rate. The model is driven by three exogenous processes which determine government spending, g_t, the stationary component of technology, z_t, and the policy shock, $\epsilon_{R,t}$. A full description of the model can be found in Woodford (2003). For the purpose at hand we focus on its log-linear representation, which takes each variable as deviations from its trend. The model has a deterministic steady state with respect to the detrended variables: the common component is generated by a stochastic trend in the exogenous process for technology. The specification follows Del Negro and Schorfheide (2004)(DS) and reads:

$$\tilde{x}_t = E_t\tilde{x}_{t+1} - \frac{1}{\tau}(\tilde{R}_t - E_t\tilde{\pi}_{t+1}) + (1-\rho_G)\tilde{g}_t + \rho_z\frac{1}{\tau}\tilde{z}_t \tag{16.10}$$

$$\tilde{\pi}_t = \beta E_t\tilde{\pi}_{t+1} + \kappa\left(\tilde{x}_t - \tilde{g}_t\right) \tag{16.11}$$

$$\tilde{R}_t = \rho_R\tilde{R}_{t-1} + (1-\rho_R)(\psi_1\tilde{\pi}_t + \psi_2\tilde{x}_t) + \epsilon_{R,t} \tag{16.12}$$

$$\tilde{g}_t = \rho_g\tilde{g}_{t-1} + \epsilon_{g,t} \tag{16.13}$$

$$\tilde{z}_t = \rho_z\tilde{z}_{t-1} + \epsilon_{z,t}, \tag{16.14}$$

where $\tilde{x}_t$ is the output gap, $\tilde{\pi}_t$ is the inflation rate, $\tilde{R}_t$ is the short-term interest rate and $\tilde{g}_t$ and $\tilde{z}_t$ are two stationary AR(1) processes for government and technology, respectively.

The first equation is an intertemporal Euler equation obtained from the household's optimal choice of consumption and bond holdings. There is no investment in the model and so output is proportional to consumption up to an exogenous process that describes fiscal policy. The net effects of these exogenous shifts on the Euler equation are captured in the process $\tilde{g}_t$. The parameter $0 < \beta < 1$ is the household's discount factor and $\tau > 0$ is the inverse of the elasticity of intertemporal substitution. The second equation is the forward-looking Phillips curve, which describes the dynamics of inflation and where κ determines the degree of the short-run trade-off between output and inflation. The third equation is the monetary policy reaction function. The central bank follows a nominal interest rate rule by adjusting its instrument to deviations of inflation and output from their respective target levels. The shock $\epsilon_{R,t}$ can be interpreted as an unanticipated deviation from the policy rule or as policy implementation error. Fiscal policy is simply described by an autoregressive process. The set of structural shocks is thus $\epsilon_t = \left(\epsilon_{R,t}, \epsilon_{g,t}, \epsilon_{z,t}\right)'$, which collects technology, government and monetary shocks.

To cast the model in the form of:

$$\Gamma_0 \widetilde{\mathbf{Z}}_t = \Gamma_1 \widetilde{\mathbf{Z}}_{t-1} + C + \Psi \epsilon_t + \Pi \eta_t, \tag{16.15}$$

specify the relevant matrices as follows:

$$\widetilde{\mathbf{Z}}_t = \begin{bmatrix} \widetilde{x_t} \\ \widetilde{\pi_t} \\ \widetilde{R_t} \\ \widetilde{R_t^*} \\ \widetilde{g_t} \\ \widetilde{z_t} \\ \widetilde{E_t x_{t+1}} \\ \widetilde{E_t \pi_{t+1}} \end{bmatrix} \quad \epsilon_t = \begin{bmatrix} \epsilon_t^R \\ \epsilon_t^G \\ \epsilon_t^Z \end{bmatrix} \quad \eta_t = \begin{bmatrix} \eta_t^x = x_t - E_{t-1}(x_t) \\ \eta_t^\pi = \pi_t - E_{t-1}(\pi_t) \end{bmatrix}$$

$$\Gamma_0 = \begin{bmatrix} 1 & 0 & \frac{1}{\tau} & 0 & -(1-\rho_g) & -\frac{\rho_z}{\tau} & -1 & -\frac{1}{\tau} \\ -\kappa & 1 & 0 & 0 & \kappa & 0 & 0 & -\beta \\ 0 & 0 & 1 & -(1-\rho_R) & 0 & 0 & 0 & 0 \\ -\psi_2 & -\psi_1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Gamma_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho_R & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho_G & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho_Z & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Psi = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \Pi = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

As a solution to (16.15), we obtain the following policy function:

$$\tilde{Z}_t = T(\theta)\,\tilde{Z}_{t-1} + R(\theta)\,\epsilon_t. \tag{16.16}$$

To provide the mapping between the observable data and those computed as deviations from the steady-state of the model, we set the following measurement equations, as in DS:

$$\Delta \ln x_t = \ln \gamma + \Delta \tilde{x}_t + \tilde{z}_t \tag{16.17}$$

$$\Delta \ln P_t = \ln \pi^* + \tilde{\pi}_t \tag{16.18}$$

$$\ln R_t = 4[(\ln R^* + \ln \pi^*) + \tilde{R}_t], \tag{16.19}$$

which can also be cast into matrices as:

$$Y_t = \Lambda_0(\theta) + \Lambda_1(\theta)\,\tilde{Z}_t + \nu_t, \tag{16.20}$$

where $Y_t = \left(\Delta \ln x_t, \Delta \ln P_t, \ln R_t\right)'$, $\nu_t = 0$ and Λ_0 and Λ_1 are defined accordingly. For completeness, we write the matrices T, R, Λ_0 and Λ_1 as a function of the structural parameters in the model, $\theta = \left(\ln \gamma, \ln \pi^*, \ln r^*, \kappa, \tau, \psi_1, \psi_2, \rho_R, \rho_g, \rho_Z, \sigma_R, \sigma_g, \sigma_Z\right)'$: such a formulation derives from the rational expectations solution.

The evolution of the variables of interest, Y_t, is therefore determined by (16.15) and (16.20), which impose a set of restrictions across the parameters of the MA representation. Finally, the MA representation is approximated by a finite-order VAR representation.

Notes

1. The LSE approach was initiated by Denis Sargan but owes its diffusion to a number of Sargan's students and is extremely well described in the book by David Hendry (1995).

2. Importantly, the analysis of determinacy of the equilibria led to the discovery that a central bank can need fiscal backing; in fact, there is a class of equilibria for the economy that are invisible if one focuses entirely on money demand. These are equilibria in which the monetary authority is completely passive: it picks a nominal interest rate and agrees to accommodate any amount of debt issue by monetizing it. In conventional models this leads to an indeterminate price level, but in a model in which the fiscal authority is committed to a fixed level of primary surpluses there is a unique price level. So inflation cannot be controlled by only controlling the stock of money (see Leeper, 1991; Sims, 2007).
3. The statistical model is a VAR. When variables included in the VAR are non-stationary, the model can be reparameterized as a vector error correction model (VECM). In this case, after the solution of the identification problems of cointegrating vectors, the information set available at $t-1$ contains n lagged endogenous variables and r cointegrating vectors.
4. See the appendix for an example of this representation applied to a simple macroeconomic model.
5. Expressing the solution of a DSGE as a VAR might also involve solving some noninvertibility problems of the matrix governing the simultaneous relation among variables originally considered in the theoretical model. This problem is carefully discussed by Fabio Canova in Chapter 2 of this volume.
6. Importantly, these features ought to be different from those under examination.
7. Some abuses of this practice are present in the literature; the most common one is to compare the properties of filtered raw data with those of filtered model-generated data. Filtering model-generated data is clearly hard to justify given that model-generated data are stationary by their nature.

References

Amisano, G. and C. Giannini (1996) *Topics in Structural VAR Econometrics*. SpringerVerlag.

An, S. and F. Schorfheide (2006) Bayesian analysis of DSGE Models. *Working Paper 06-5, Federal Reserve Bank of Atlanta*.

Anderson, T.W. and H. Rubin (1949) Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20**, 46–63.

Baba, Y., D.F. Hendry and R.M. Starr (1992) The demand for M1 in the U.S.A., 1960–1988. *Review of Economic Studies* **59**, 25–61.

Basmann, R.L. (1960) On finite sample distributions of generalized classical linear identifiability test statistic. *Journal of the American Statistical Association* **55**, 650–9.

Benati, L. and P. Surico (2007) Var analysis of the Great Moderation, http://unjobs.org/authors/paolo-surico.

Bernanke, B.S., and J. Boivin (2003) Monetary policy in a data-rich environment. *Journal of Monetary Economics* **50**, 525–64.

Bernanke, B.S., J. Boivin and P. Eliasz (2005) Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics* **120**(1), 387–422.

Bernanke, B.S. and I. Mihov (1998) Measuring monetary policy. *Quarterly Journal of Economics* **113**(3), 869–902.

Blanchard, O.J. and D.T. Quah (1989) The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* **79**, 655–73.

Boivin, J. and M.P. Giannoni (2005) DSGE models in a data-rich environment. Working Paper.

Canova, F. and L. Sala (2005) Back to square one: identification issues in DSGE models. IGIER Working Paper 303, Università Bocconi.

Christiano, L.J. and M. Eichenbaum (1992) Liquidity effects and the monetary transmission mechanism. *American Economic Review* **82** (2), 346–53.

Christiano, L.J., M. Eichenbaum and C.L. Evans (1996a) The effects of monetary policy shocks: evidence from the flow of funds. *Review of Economics and Statistics* **78**, 16–34.
Christiano, L.J., M. Eichenbaum and C.L. Evans (1996b) Monetary policy shocks and their consequences: theory and evidence. Paper presented at ISOM.
Christiano, L.J., M. Eichenbaum and C.L. Evans (1998) Monetary policy shocks: what have we learned and to what end?' NBER Working Paper No. 6400.
Christiano, L., M. Eichenbaum and C. Evans (2005) Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* **113**, 1–45.
Consolo, A., C.A. Favero and A. Paccagnini (2007) On the statistical identification of DSGE models. IGIER Discussion Paper.
Cooley, T.F. (1995) *Frontiers of Business Cycle Research*. Princeton: Princeton University Press.
Cooley, T.F. (1997) Calibrated models. *Oxford Review of Economic Policy* **13**, 55–69.
Doan, T., R. Litterman and C. Sims (1984) Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews* **3**, 1–100.
DeJong, D., B. Ingram and C. Whiteman (1996) A Bayesian approach to calibration. *Journal of Business Economics and Statistics* **14**, 1–9.
DeJong, D., B. Ingram and C. Whiteman (2000) A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* **98**, 203–23.
Del Negro, M. and F. Schorfheide (2004) Priors from general equilibrium models for VARs. *International Economic Review* **45**, 643–73.
Del Negro, M. and F. Schorfheide (2006) How good is what you've got? DSGE-VAR as a toolkit for evaluating DSGE models. *Federal Reserve Bank of Atlanta Economic Review*.
Del Negro, M., F. Schorfheide, F. Smets and R. Wouters (2006) On the fit of New-Keynesian models. Working Paper.
Engle, R. and D.F. Hendry (1993) Testing superexogeneity and invariance in regression models. *Journal of Econometrics* **56**, 119–39.
Engle, R., D.F. Hendry and J.F. Richard (1983) Exogeneity. *Econometrica* **51**, 277–302.
Farmer, R. (1997) Money in a real business cycle model. *Journal of Money, Credit and Banking*, November.
Favero, C.A. (2001) *Applied Macroeconometrics*. Oxford: Oxford University Press.
Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000) The generalized factor model: identification and estimation. *Review of Economics and Statistics* **82**(4), 540–54.
Forni, M. and L. Reichlin (1996) Dynamic common factors in large cross-sections. *Empirical Economics*, **21**, 27–42.
Forni, M. and L. Reichlin (1998) Let's get real: a dynamic factor analytical approach to disaggregated business cycle. *Review of Economic Studies* **65**, 453–74.
Fukac, M. and A. Pagan (2006) Issues in adopting DSGE models for the use in policy process. CAMA Working Paper Series.
Garratt, A., K. Lee, M.H. Pesaran and Y. Shin (2006) *Global and National Macroeconometric Modelling: A Long-Run Structural Approach*. Oxford: Oxford University Press.
Goldberger, A.S. (1991). *A Course in Econometrics*. Cambridge, Mass.: Harvard University Press.
Goodfriend, M. and King, R. (1997) The Newclassical synthesis and the role of monetary policy. Paper presented at the 12th NBER Annual Macroeconomic Conference.
Gramlich, E.M. (2004) Remarks. Paper presented to the Conference on Models and Monetary Policy, Federal Reserve Bank Board of Governors.
Hamilton, J. (1994) *Time Series Analysis*. Princeton: Princeton University Press.
Hansen, L. (1982) Large sample properties of the generalized method of moments estimators. *Econometrica* **50**, 1269–91.
Helliwell, J.F., G.R. Sparks, F.W. Gorbet, H.T. Shapiro, I.A. Stewart and D.R. Stephenson (1991) The structure of RDX2. Bank of Canada Staff Study No. 7.
Hendry, D.F. (1988) The encompassing implications of feedback versus feedforward mechanisms in econometrics. *Oxford Economic Papers* **4**, 133–47.
Hendry, D.F. (1995) *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D.F., J.N.J. Muellbauer and T.A. Murphy (1990) The econometrics of DHSY. In J.D. Hey and D. Winch (eds.), *A Century of Economics*, pp. 298–334. Oxford: Blackwell.
Ingram, B. and C. Whiteman (1994) Supplanting the Minnesota prior–Forecasting macroeconomics time series using real business cycle model priors. *Journal of Monetary Economics* **34**, 497–510.
Ireland, P.N. (2004) A method for taking models to the data. *Journal of Economic Dynamics and Control* **28**, 1205–26.
Johansen, S. (1995) *Likelihood Based Inference on Cointegration in the Vector Autoregressive Model*. Oxford: Oxford University Press.
Johansen, S. (2004) How not to take a model to the data. Mimeo.
Juselius, K. and M. Franchi (2007) Taking a DSGE model to the data meaningfully. Economics Discussion Papers, http://www.economics-ejournal.org/economics/journalarticles/2007-4.
Juselius, K. and S. Johansen (1999) Macroeconomic transmission mechanism: empirical applications and econometric methods. Paper available at http://www.econ.ku.dk/okokj/.
King, R.G., C.I. Plosser and S.T. Rebelo (1988) Production, growth and business cycles. *Journal of Monetary Economics* **21**, 195–232.
Kydland, F. and E. Prescott (1982) Time to build and aggregate fluctuations. *Econometrica* **50**, 1345–70.
Kydland, F. and P. Prescott (1996) The computational experiment: an econometric tool. *Journal of Economic Perspectives* **10**, 69–85.
Leeper, E.M. (1991) Equilibria under "active" and "passive" monetary and fiscal policies. *Journal of Monetary Economics* **27**, 129–47.
Leeper, E.M., C.A. Sims and T. Zha (1996) What does monetary policy do? Available at http://eco-072399b.princeton.edu/yftp/bpea/bpeaf.pdf.
Liu, T.-C. (1960) Underidentification, structural estimation and forecasting. *Econometrica* **28**(4), 855–65.
Lucas, R.E. Jr. (1972) Expectations and the neutrality of money. *Journal of Economic Theory* **4**, 103–24.
Lucas, R.E. Jr. (1976) Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*. Amsterdam: North-Holland.
Maddala, G.S. (1988) *Introduction to Econometrics*. New York: Macmillan.
Pesaran, M.H and Y. Shin (1998) Generalized impulse response analysis in linear multivariate models. *Economics Letters* **58**, 17–29.
Pesaran, M.H and Y. Shin (2002) Long-run structural modelling. *Econometric Reviews* **21**, 49–87.
Pesaran, M.H. and R. Smith (1995) The role of theory in econometrics. *Journal of Econometrics* **67**, 71–9.
Rudebusch, G.D. (1998) Do measures of monetary policy in a VAR make sense? *International Economic Review* **39**, 907–31.
Ruge-Murcia, F.J. (2003) Methods to estimate dynamic stochastic general equilibrium models. *CIREQ, Cahier 17-2003*.
Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* **48**, 1–48.
Sims, C.A. (2002) Solving linear rational expectations models. *Computational Economics* **20**(1–2), 1–20.
Sims, C.A. (2007) Interview with Christopher Sims. Federal Reserve Bank of Minneapolis, http://www.minneapolisfed.org/pubs/region/07-06/sims.cfm.
Sims, C.A. and T. Zha (1996) Does monetary policy generate recessions? Mimeo. Available at ftp://ftp.econ.yale.edu/pub/sims/mpolicy.
Smets, F. and R. Wouters (2003) An estimated stochastic dynamic general equilibrium model of the Euro area. *Journal of the European Economic Association* **1**, 1123–75.
Spanos, A. (1990) The simultaneous-equations model revisited: statistical adequacy and identification. *Journal of Econometrics* **44**, 87–105.

Stock, J. and M. Watson (2002) Macroeconomic forecast in using diffusion indexes. *Journal of Business Economics and Statistics* **20**(2), 147–62.

Taylor, J.B. (2005) Thirty five years of model building for monetary policy evaluation: breakthroughs, dark ages, and a renaissance. Mimeo.

Uhlig, H. (1999) A toolkit for analyzing nonlinear dynamic stochastic models easily. In R. Marimon and A. Scott (eds.), *Computational Methods for the Study of Dynamic Economies*. Oxford: Oxford University Press.

Woodford, M. (2003) *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

17
Macroeconometric Modeling for Policy

Gunnar Bårdsen and Ragnar Nymoen

Abstract

The first part of this chapter sets out a coherent approach to dynamic macroeconometric model-building; the second part demonstrates the approach through building and evaluating a small econometric model; the final part demonstrates various usages of the model for policy.

> "I think it should be generally agreed that a model that does not generate many properties of actual data cannot be claimed to have any 'policy implications' ..." (Clive W.J. Granger (1992, p. 4))

17.1 Introduction

Depending upon its properties, a macroeconometric model can highlight various aspects of economic policy: communication of policy actions, structuring of economic debate, policy simulations, testing of competing theories, forecasting, stress testing, etc. From an academic perspective, the desired properties of a model are also legion, but the end result will depend upon the preferences for coherence along dimensions such as: theory foundations (microfoundations/aggregation/general/partial), econometric methods (Bayesian/frequentist), and model properties (size/robustness/nonlinearities/transparency/dynamics).

Satisfying the different needs and desires of policy making could therefore entail a collection of models. Such a model collection could include more or less calibrated theory models, structural and Bayesian vector autoregressions (VARs), simultaneous equation models (SEMs), and dynamic stochastic general equilibrium (DSGE) models (see Pagan, 2003, for an overview). A choice of model(s) for the event at hand could then be made on the basis of strengths and weaknesses of the various candidates. The inherent weaknesses of the main candidates are well known. If one were to follow Ambrose Bierce, and write a "Devil's Dictionary" of macroeconometrics, some of the entries could read: Structural VARs: how to estimate models inefficiently; SEMs: estimates of something; Bayesian estimation: *see* calibration; DSGEs: sophisticated naivety. Therefore, a choice of model(s) for the event at hand should be made on the basis of strengths and weaknesses of the various model classes.

The profession's collective understanding of the causes and possible remedies of model limitations, both in forecasting and in policy analysis, has improved

markedly over the last decades. The Lucas (1976) critique and the Clements and Hendry (1999) analysis of the sources of forecast failures with macroeconometric models are milestones in that process. Interestingly, the methodological ramifications of those two critiques are different: the Lucas critique has led to the current dominance of representative agent-based macroeconomic models. Hendry (2001a), on the other hand, concludes that macroeconometric systems of equations, despite their vulnerability to regime shifts, but because of their potential adaptability to breaks, remain the best long-run hope for progress in macroeconomic forecasting. Since monetary policy can be a function of the forecasts, as with inflation forecast targeting (cf. Svensson, 1997), the choice of forecasting model(s) is important.

The class of macroeconometric models we present in this chapter requires coherent use of economic theory, data, and mathematical and statistical techniques. This approach, of course, has a long history in econometrics, going back to Tinbergen's first macroeconometric models, and has enjoyed renewed interest in the last decades. Recent advances in econometrics and in computing mean that we now have much better tools than, say, 20 years ago for developing and maintaining macroeconometric models in this tradition (see Garratt *et al.*, 2006, for one recent approach).

Regardless of underlying theory, a common aim of macroeconometric model-building is identification of invariant relationships, if they exist at all (see Haavelmo, 1944, Ch. II). A well-specified macroeconometric model is a good starting point for such a quest, since it provides an ideal test-bed for further over-identifying restrictions of microeconomic behavior. Such a strategy is, in particular, relevant to the challenges from behavioral economics, with implications for time inconsistency (hyperbolic discounting), changing expectations (learning), asset bubbles (herd behavior), etc.

Macroeconomic models of the representative agent, intertemporal optimizing, type are said to have structural interpretations, with "deep structural parameters" that are immune to the Lucas critique. However, when the model's purpose is to describe observed macroeconomic behavior, its structural properties are conceptually different. Heuristically, we take a model to have structural properties if it is invariant and interpretable (see Hendry, 1995b). Structural properties are nevertheless relative to the history, nature and significance of regime shifts. There is always the possibility that the next shocks to the system may incur real damage to a model with hitherto high structural content. The approach implies that a model's structural properties must be evaluated along several dimensions, and the following seem particularly relevant:

1. theoretical interpretation
2. ability to explain the data
3. ability to explain earlier findings, i.e., encompass the properties of existing models
4. robustness to new evidence in the form of updated/extended data series and new economic analysis suggesting, e.g., new explanatory variables.

Economic analysis (1) is an indispensable guidance in the formulation of econometric models. Clear interpretation also helps communication of ideas and results among researchers, in addition to structuring debate. However, since economic theories are necessarily simplifying abstractions, translations of theoretical to econometric models must lead to problems such as biased coefficient estimates, wrong signs of coefficients, and/or residual properties that hamper valid inference. The main distinction seems to be between seeing theory as representing *the* correct specification (leaving parameter estimation to the econometrician) and viewing theory as a guideline in the specification of a model which also accommodates institutional features, attempts to accommodate heterogeneity among agents, addresses the temporal aspects for the dataset, etc. (see Granger, 1999).

Arguments against "largely empirical models" include sample dependency, lack of invariance, unnecessary complexity (in order to fit the data) and chance findings of "significant" variables. Yet the ability to characterize the data (2) remains an essential quality of useful econometric models and, given the absence of theoretical truisms, the implications of economic theory have to be confronted with the data in a systematic way.

We use cointegration methods on linearized and discretized dynamic systems to estimate theory-interpretable and identified steady-state relationships, imposed in the form of equilibrium correction models. We also make use of an automated model selection approach to sift out the best theory-interpretable and identified dynamic specifications. Hoover and Perez (1999), Hendry and Krolzig (1999) and Doornik (2008) have shown that automated model selection methods have a good chance of finding a close approximation to the data-generating process (DGP), and that the danger of overfitting is, in fact, (surprisingly) low. Conversely, acting *as if* the specification is given by theory alone, with only coefficient estimates left to "fill in," is bound to result in the econometric problems noted above, and to a lower degree of relevance of the model for the economy it claims to represent.

In order to develop a scientific basis for policy modeling in macroeconometrics, a new model's capability of encompassing earlier findings should be regarded as an important aspect of structure (3). There are many reasons for the coexistence of contested models for the same phenomena, some of which may be viewed as inherent (limited number of data observations, measurement problems, controversy about operational definitions, new theories). Nevertheless, the continued use of corroborative evaluation (i.e., only addressing goodness-of-fit or predicting the stylized fact correctly) may inadvertently hinder the accumulation of evidence. One suspects that there would be huge gains from a breakthrough in new standards of methodology and practice for the profession.

Ideally, empirical modeling is a cumulative process whereby models continuously become overtaken by new and more useful ones. By "useful," we mean models that are relatively invariant to changes elsewhere in the economy, i.e., they contain autonomous parameters (see Haavelmo, 1944; Johansen, 1977; Aldrich, 1989; Hendry, 1995b). Models with a high degree of autonomy represent structural properties: they remain invariant to changes in economic policies and other

shocks to the economic system, as implied by (4) above.[1] However, structure is partial in two respects. First, autonomy is a relative concept, since an econometric model cannot be invariant to every imaginable shock. Second, all parameters of an econometric model are unlikely to be equally invariant, and only the parameters with the highest degree of autonomy represent structure. Since elements of structure typically will be grafted into equations that also contain parameters with a lower degree of autonomy, forecast breakdown may frequently be caused by shifts in these non-structural parameters.[2]

A strategy for model evaluation that puts emphasis on forecast behavior, without a careful evaluation of the causes of forecast failure *ex post*, runs a risk of discarding models that actually contain important elements of structure. Hence, e.g., Doornik and Hendry (1997) and Clements and Hendry (1999, Ch. 3) show that the main source of forecast failure is location shifts (shifts in means of levels, changes, etc.), and not shifts in such coefficients that are of primary concern in policy analysis, i.e., the derivative coefficients of behavioral equations. Therefore, a rough spell in terms of forecasting performance does not, by itself, disqualify the model's relevance for policy analysis. If the cause of the forecast failure is location shifts, they can be attenuated *ex post* by intercept correction or additional differencing "within" the model (Hendry, 2004). With these add-ons, and once the break period is in the information set, the model forecast will adapt to the new regime and improve again. Failure to adapt to the new regime may then be a sign of a deeper source of forecast failure, in the form of non-constant derivative coefficients, which also undermines the models relevance for policy analysis.[3] In general, without adaptive measures, models with high structural content will lose regularly to simple forecasting rules (see, e.g., Clements and Hendry, 1999; Eitrheim, Husebø and Nymoen, 1999). Hence different models may be optimal for forecasting and for policy analysis, which fits well with the often heard recommendations of a suite of monetary policy models.

Structural breaks are always a main concern in econometric modeling but, like any hypothesis or theory, the only way to judge the significance of a hypothesized break is by confrontation with the evidence in the data. Moreover, given that an encompassing approach is followed, a forecast failure is not only destructive but represents potential for improvement, if successful respecification follows in its wake (cf. Eitrheim, Jansen and Nymoen, 2002). In the same vein, one important intellectual rationale for DSGE models is the Lucas critique. If the Lucas critique holds, any "reduced-form" equation in a model is liable to be unstable over the historical sample, due to regime shifts and policy changes that have taken place in the economy. Hence, according to the Lucas critique, parameter instability may be endemic in any model that fails to obey the rational expectations hypothesis (REH), with the possible consequence that without integration of the REH, the model is unsuited for policy analysis. However, as stated by Ericsson and Irons (1995), the Lucas critique is only a possibility theorem, not a truism, and the implications of the Lucas critique can be tested (see also, e.g., Hendry, 1988; Engle and Hendry, 1993; Ericsson and Hendry, 1999). In Bårdsen, Jansen and Nymoen

(2003) we have shown, by extensive testing of a previous version of our model, that the Lucas critique has little force for our system of equations. This finding is consistent with the international evidence presented in Ericsson and Irons (1995) and Stanley (2000). On the basis of these results, our model is more consistent with agents adopting robust forecasting rules, in line with the analysis and suggestions of Hendry and Mizon (2000). In that case the Lucas critique does not apply, although the degree of autonomy remains an issue that needs to be evaluated as fully as possible, given the information available to us.

This chapter documents the approach we use to dynamic macroeconometric model-building and policy analysis. To make the analysis applied, the approach is illustrated through a model of the Norwegian economy. Our approach is, of course, applicable to other economies, but we know more about market characteristics, policy changes and institutional development in Norway than in any other country or economic area. And since such factual knowledge is an indispensable and complementary aid to formal econometrics in the building of an empirical model, we prefer to work with the economy we have most knowledge about.

The rest of the chapter consists of three main sections. Section 17.2 sets out a coherent approach to dynamic macroeconometric model building; section 17.3 demonstrates the approach through building and evaluating a small econometric model; section 17.4 demonstrates various tools for policy analysis using the model. Section 17.2 involves three steps in going from general to specific. The first step is theoretical and establishes a framework for linearizing and discretizing an approximation to a general theory model with constant steady-state values. The second step is to estimate, and solve, the steady-state model in the form of overidentifying cointegrating relationships and common trends. The third step is to identify and estimate the dynamic structure of the model.

Section 17.3 illustrates the approach set out in section 17.2 by the construction and evaluation of a small open-economy model.

Section 17.4 demonstrates five tools for policy using the model: tractability, simulations of policy responses, optimal policy considerations, theory evaluation, and forecasting. The first use is illustrated by introducing a method to construct stylized versions of complex models. The second use is illustrated by evaluating responses to monetary policy shocks. The third use shows how important model specification is for the derivation of optimal monetary policy. The fourth use is illustrated by testing the New Keynesian Phillips curve. The final use evaluates possible sources affecting forecast performance.

17.2 A modeling framework

As the values of all major economic variables are announced regularly, it is easy to believe that a local approximation to a DGP can exist. It is an interesting philosophical question whether the true generating mechanism can (ever) be completely described, but the usefulness of the concept does not hinge on the answer to that question. The main point is that once the real economic world, in its enormous, ever-changing complexity, is accepted as a premise for macroeconomic modeling,

it follows that the main problems of macroeconometrics are model specification and model evaluation, rather than finding the best estimator under the assumption that the model is identical to the DGP.

The local DGP is changing with the evolution of the real world economy–through technical progress, changing patterns of family composition and behavior, and political reform. Sometimes society evolves gradually and econometric models are then usually able to adapt to the underlying real-life changes, i.e., without any noticeable loss in "usefulness." Often, however, society evolves so quickly that estimated economic relationships break down and cease to be of any aid in understanding the current macroeconomy and in forecasting its development even over the first couple of years. In this case we speak of a changing local approximation in the form of a regime shift in the generating process, and a structural break in the econometric model. Since the complexity of the true macroeconomic mechanism, and the regime shifts also contained in the mechanism, lead us to conclude that any model will at best be a local approximation to the DGP, judging the quality of, and choosing between, the approximations becomes central.

In the rest of this section we present our approach to finding a local approximation useful for policy.[4]

17.2.1 Linearization

Consider a very simple example of an economic model in the form of the differential equation:

$$\frac{dy}{dt} = f(y, x), \quad x = x(t), \tag{17.1}$$

in which a constant input $x = \bar{x}$ induces $y(t)$ to approach asymptotically a constant state $\bar{y}$ as $t \to \infty$. Clearly $\bar{x}$ and $\bar{y}$ satisfy $f(\bar{y}, \bar{x}) = 0$. For example, standard DSGE models usually take this form, with the models having deterministic steady-state values. The usual procedure then is to expand the differential (or difference) equation about this steady-state solution (see, e.g., Campbell, 1994; Uhlig, 1999). Employing this procedure yields:

$$f(y, x) = f(\bar{y}, \bar{x}) + \frac{\partial f(\bar{y}, \bar{x})}{\partial y}(y - \bar{y}) + \frac{\partial f(\bar{y}, \bar{x})}{\partial x}(x - \bar{x}) + R, \tag{17.2}$$

where:

$$R = \frac{1}{2!}\left(\frac{\partial^2 f(\xi, \eta)}{\partial x^2}(x - \bar{x})^2 + 2\frac{\partial^2 f(\xi, \eta)}{\partial x \partial y}(x - \bar{x})(y - \bar{y}) + \frac{\partial^2 f(\xi, \eta)}{\partial y^2}(y - \bar{y})^2\right),$$

and (ξ, η) is a point such that ξ lies between y and $\bar{y}$ while η lies between x and $\bar{x}$. Since $\bar{y}$ and $\bar{x}$ are the steady-state values for y and x respectively, then the expression for $f(y, x)$ takes the simplified form:

$$f(y, x) = a\,(y - \bar{y}) + \delta\,(x - \bar{x}) + R, \tag{17.3}$$

where $a = \partial f(\bar{y}, \bar{x})/\partial y$ and $\delta = \partial f(\bar{y}, \bar{x})/\partial x$ are constants.

If f is a linear function of y and x then $R = 0$ and so:

$$f(x, y) = a\left(y - \bar{y} + \frac{\delta}{a}(x - \bar{x})\right) = a\,(y - bx - c), \tag{17.4}$$

in which $b = -\delta/a$ and $c = \bar{y} + (\delta/a)\bar{x}$.

17.2.2 Discretization

For a macroeconometric model, a discrete representation is usually practical, and it can be worked out as follows. Let $t_1,\ t_2,\ \cdots,\ t_k,\ \cdots$ be a sequence of times spaced h apart and let $y_1,\ y_2,\ \cdots,\ y_k,\ \cdots$ be the values of a continuous real variable $y(t)$ at these times. The backward-difference operator $\triangle$ is defined by the rule:

$$\triangle y_k = y_k - y_{k-1}, \quad k \geq 1. \tag{17.5}$$

By observing that $y_k = (1 - \triangle)^0 y_k$ and $y_{k-1} = (1 - \triangle)^1 y_k$, the value of y at the intermediate point $t = t_k - sh$ $(0 < s < 1)$ may be estimated by the interpolation formula:

$$y(t_k - sh) = y_{k-s} = (1 - \triangle)^s y_k, \quad s \in [0, 1]. \tag{17.6}$$

When s is not an integer, $(1 - \triangle)^s$ should be interpreted as the power series in the backward-difference operator obtained from the binomial expansion of $(1 - x)^s$. This is an infinite series of differences. Specifically:

$$(1 - \triangle)^s = 1 - s\triangle - \frac{s(1-s)}{2!}\triangle^2 - \frac{s(1-s)(2-s)}{3!}\triangle^3 - \cdots. \tag{17.7}$$

With this preliminary background, the differential equation:

$$\frac{dy}{dt} = f(y, x), \quad x = x(t), \tag{17.8}$$

may be integrated over the time interval $[t_k, t_{k+1}]$ to obtain:

$$y(t_{k+1}) - y(t_k) = \triangle y_{k+1} = \int_{t_k}^{t_{k+1}} f(y(t), x(t))\,dt, \tag{17.9}$$

in which the integral on the right-hand side of this equation is to be estimated by using the backward-difference interpolation formula given in equation (17.7). The substitution $t = t_k + sh$ is now used to change the variable of this integral from $t \in [t_k, t_{k+1}]$ to $s \in [0, 1]$. The details of this change of variable are:

$$\int_{t_k}^{t_{k+1}} f(y(t), x(t))\,dt = \int_0^1 f(y(t_k + sh), x(t_k + sh))\,(h\,ds) = h\int_0^1 f_{k+s}\,ds,$$

where $f_{k+s} = f(y(t_k+sh), x(t_k+sh))$. The value of this latter integral is now computed using the interpolation formula based on (17.7). Thus:

$$\begin{aligned}\int_0^1 f_{k+s}\,ds &= \int_0^1 (1-\triangle)^{-s} f_k\,ds \\ &= \int_0^1 \left(f_k + s\triangle f_k + \frac{s(1+s)}{2!}\triangle^2 f_k + \frac{s(1+s)(2+s)}{3!}\triangle^3 f_k + \cdots\right) ds \\ &= f_k + \frac{1}{2}\triangle f_k + \frac{5}{12}\triangle^2 f_k + \frac{3}{8}\triangle^3 f_k + \cdots .\end{aligned}$$

The final form for the backward-difference approximation to the solution of this differential equation is therefore

$$\triangle y_{k+1} = hf_k + \frac{h}{2}\triangle f_k + \frac{5h}{12}\triangle^2 f_k + \frac{3h}{8}\triangle^3 f_k + \cdots . \tag{17.10}$$

17.2.3 Equilibrium correction representations and cointegration

The discretization scheme (17.10) applied to the linearized model (17.3), with $k = t-1$ and $h = 1$, gives the equilibrium correction model, EqCM, representation

$$\begin{aligned}\triangle y_t = a\,(y - bx - c)_{t-1} + R_{t-1} + \frac{1}{2}a(\triangle y_{t-1} - b\triangle x_{t-1}) + \frac{1}{2}\triangle R_{t-1} \\ + \frac{5}{12}a(\triangle^2 y_{t-1} - b\triangle^2 x_{t-1}) + \frac{5}{12}\triangle^2 R_{t-1} + \cdots .\end{aligned}$$

At this point two comments are in place. The first is that an econometric specification will mean a truncation of the polynomial both in terms of powers and lags. Diagnostic testing is therefore imperative to ensure a valid local approximation, and indeed to test that the statistical model is valid (see Hendry, 1995a, p. 15.1; Spanos, 2008). The second is that the framework allows for flexibility regarding the form of the steady state. The standard approach in DSGE modeling has been to filter the data, typically using the so-called Hodrick–Prescott filter, to remove trends, hopefully achieving stationary series with constant means, and then work with the filtered series. Another approach, popular at present, is to impose the theoretical balanced growth path of the model on the data, expressing all series in terms of growth corrected values. However, an alternative approach is to estimate the balanced growth paths in terms of finding the number of common trends and identifying and estimating cointegrating relationships. The present approach allows for all of these interpretations.

To illustrate the approach in terms of cointegration, consider real wages to be influenced by productivity, as in many theories.[5] Assume that the logs of the real wage rw_t and productivity z_t are each integrated of order one, but found to be cointegrated, so:

$$rw_t \sim I(1)\,,\ \triangle rw_t \sim I(0) \tag{17.11}$$

$$z_t \sim I(1)\,,\ \triangle z_t \sim I(0) \tag{17.12}$$

$$(rw - \beta z)_t \sim I(0)\,. \tag{17.13}$$

Letting $y_t \equiv (rw - \beta z)_t$ and $x_t \equiv \Delta z_t$ then gives:

$$\Delta rw_t = -ac + a(rw - \beta z)_{t-1} + \frac{a}{2}\Delta\,(rw - \beta z)_{t-1} + \beta\Delta z_t - ab\Delta z_{t-1} - \frac{ab}{2}\,\Delta^2 z_{t-1} + \cdots.$$

17.2.4 System representations

The approach easily generalizes to a system representation. For ease of exposition, we illustrate the two-dimensional case for which $y_1 \to \bar{y}_1$ and $y_2 \to \bar{y}_2$ as $t \to \infty$. Expanding with respect to y_1 and y_2 about their steady-state values yields:

$$\begin{bmatrix} f_1\,(y_1, y_2) \\ \\ f_2\,(y_1, y_2) \end{bmatrix} = \begin{bmatrix} f_1\,(\bar{y}_1, \bar{y}_2) \\ \\ f_2\,(\bar{y}_1, \bar{y}_2) \end{bmatrix} + \begin{bmatrix} \dfrac{\partial f_1(\bar{y}_1, \bar{y}_2)}{\partial y_1} & \dfrac{\partial f_1(\bar{y}_1, \bar{y}_2)}{\partial y_2} \\ \dfrac{\partial f_2(\bar{y}_1, \bar{y}_2)}{\partial y_1} & \dfrac{\partial f_2(\bar{y}_1, \bar{y}_2)}{\partial y_2} \end{bmatrix} \begin{bmatrix} y_1 - \bar{y}_1 \\ \\ y_2 - \bar{y}_2 \end{bmatrix} + \begin{bmatrix} R_1 \\ \\ R_2 \end{bmatrix},$$

where $[R_1, R_2]'$ denotes the vector:

$$\frac{1}{2!}\begin{bmatrix} \dfrac{\partial^2 f_1(\zeta, \eta)}{\partial y_1^2}\,(y_1 - \bar{y}_1)^2 + 2\dfrac{\partial^2 f_1(\zeta, \eta)}{\partial y_1 \partial y_2}\,(y_1 - \bar{y}_1)\,(y_2 - \bar{y}_2) + \dfrac{\partial^2 f_1\,(\zeta, \eta)}{\partial y_2^2}\,(y_2 - \bar{y}_2)^2 \\ \dfrac{\partial^2 f_2\,(\zeta, \eta)}{\partial y_1^2}\,(y_1 - \bar{y}_1)^2 + 2\dfrac{\partial^2 f_2\,(\zeta, \eta)}{\partial y_1 \partial y_2}\,(y_1 - \bar{y}_1)\,(y_2 - \bar{y}_2) + \dfrac{\partial^2 f_2\,(\zeta, \eta)}{\partial y_2^2}\,(y_2 - \bar{y}_2)^2 \end{bmatrix},$$

so that:

$$\begin{bmatrix} \dfrac{\partial y_1}{\partial t} \\ \dfrac{\partial y_2}{\partial t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} y_1 - \bar{y}_1 \\ \\ y_2 - \bar{y}_2 \end{bmatrix} + \begin{bmatrix} R_1 \\ \\ R_2 \end{bmatrix}.$$

The backward-difference approximation to the solution of the system of differential equations gives the system in EqCM form (see Bårdsen, Hurn and Lindsay, 2004, for details), namely:

$$\begin{aligned} \begin{bmatrix} \Delta y_1 \\ \Delta y_2 \end{bmatrix}_t &= \begin{bmatrix} -\alpha_{11} c_1 \\ -\alpha_{22} c_2 \end{bmatrix} + \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \begin{bmatrix} y_1 - \delta_1 y_2 \\ y_2 - \delta_2 y_1 \end{bmatrix}_{t-1} + \begin{bmatrix} R_1 \\ \\ R_2 \end{bmatrix}_{t-1} \\ &+ \frac{1}{2}\begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} \Delta y_1 \\ \Delta y_2 \end{bmatrix}_{t-1} + \begin{bmatrix} \Delta R_1 \\ \\ \Delta R_2 \end{bmatrix}_{t-1} \\ &+ \frac{5}{12}\begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} \Delta^2 y_1 \\ \Delta^2 y_2 \end{bmatrix}_{t-1} + \begin{bmatrix} \Delta^2 R_1 \\ \\ \Delta^2 R_2 \end{bmatrix}_{t-1} \\ &+ \frac{3}{8}\begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} \Delta^3 y_1 \\ \Delta^3 y_2 \end{bmatrix}_{t-1} + \begin{bmatrix} \Delta^3 R_1 \\ \\ \Delta^3 R_2 \end{bmatrix}_{t-1} + \cdots, \end{aligned}$$

with:

$$c_1 = \left(\bar{y}_1 + \delta_1 \bar{y}_2\right), \quad \delta_1 = \frac{\alpha_{12}}{\alpha_{11}}$$
$$c_2 = \left(\bar{y}_2 + \delta_2 \bar{y}_1\right), \quad \delta_2 = \frac{\alpha_{21}}{\alpha_{22}}.$$

As before, the variables y_1 and y_2 can be considered as stationary functions of non-stationary components – cointegration is imposed upon the system. Consider the previous example, assuming linearity so $R_i = 0$, and ignoring higher-order dynamics for ease of exposition:

$$\begin{bmatrix} \triangle y_1 \\ \triangle y_2 \end{bmatrix}_t = \begin{bmatrix} -\alpha_{11} c_1 \\ -\alpha_{22} c_2 \end{bmatrix} + \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \begin{bmatrix} y_1 - \delta_1 y_2 \\ y_2 - \delta_2 y_1 \end{bmatrix}_{t-1}$$
$$\begin{bmatrix} \triangle(rw - \beta z) \\ \Delta^2 z \end{bmatrix}_t = \begin{bmatrix} -\alpha_{11} c_1 \\ -\alpha_{22} c_2 \end{bmatrix} + \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \begin{bmatrix} (rw - \beta z) - \delta_1 \Delta z \\ \Delta z - \delta_2 (rw - \beta z) \end{bmatrix}_{t-1},$$

or multiplied out:

$$\triangle rw_t = -\alpha_{11} c_1 + \alpha_{11} (rw - \beta z)_{t-1} + \beta \Delta z_t - \alpha_{12} \Delta z_{t-1}$$
$$\Delta z_t = -\alpha_{22} \left(\bar{y}_2 + \frac{\alpha_{21}}{\alpha_{22}} \bar{y}_1\right) + \left(\alpha_{22} - 1\right) \Delta z_{t-1} - \alpha_{21} (rw - \beta z)_{t-1}.$$

If $\alpha_{21} = 0$ and $\left|\alpha_{22} - 1\right| < 1$ the system simplifies to the familiar exposition of a bivariate cointegrated system with z being weakly exogenous for β:

$$\triangle rw_t = -\alpha_{11} c_1 + \alpha_{11} (rw - \beta z)_{t-1} + \beta \Delta z_t - \alpha_{12} \Delta z_{t-1}$$
$$\Delta z_t = -\alpha_{22} \bar{z} + \left(\alpha_{22} - 1\right) \Delta z_{t-1},$$

where the common trend is a productivity trend.

17.2.5 From a discretized and linearized cointegrated VAR representation to a dynamic SEM in three steps

We will keep this section brief, as comprehensive treatments can be found in many places – e.g., Hendry (1995a), Johansen (1995, 2006), Juselius (2007), Garratt *et al.* (2006), and Lütkepohl (2006) – and only make some comments on issues in each step in the modelling process we believe merit further attention.

17.2.5.1 First step: the statistical system

Our starting point for identifying and building a macroeconometric model is to find a linearized and discretized approximation as a data-coherent statistical system representation in the form of a cointegrated VAR:

$$\Delta \mathbf{y}_t = \mathbf{c} + \Pi \mathbf{y}_{t-1} + \sum_{i=1}^{k} \Gamma_{t-i} \Delta \mathbf{y}_{t-i} + \mathbf{u}_t, \tag{17.14}$$

with independent Gaussian errors $\mathbf{u}_t$, as a basis for valid statistical inference about economic theoretical hypotheses.

The purpose of the statistical model (17.14) is to provide the framework for hypothesis testing, the inferential aspect of macroeconometric modeling. However, it cannot be postulated directly, since the cointegrated VAR itself rests on assumptions. Hence, validation of the statistical model is an essential step: is a model which is linear in the parameters flexible enough to describe the fluctuations of the data? What about the assumed constancy of parameters, does it hold over the sample that we have at hand? The assumption of Gaussian distributed error terms also needs validation, since that assumption underlies the use of (17.14) for statistical inference. The main intellectual rationale for the model validation aspect of macroeconometrics is exactly that the assumptions of the statistical model requires separate attention (Johansen, 2006; Spanos, 2006). In practice, one important step in model validation is to make the hypothesized statistical model subject to a battery of misspecification tests using the ordinary least squares (OLS) residuals $\hat{\mathbf{u}}_t$ as data.[6]

As pointed out by Garratt *et al.* (2006), the representation (17.14) does not preclude forward-looking behavior in the underlying model, as rational expectations models have backward-looking solutions. The coefficients of the solution will be defined in specific ways though, and this entails restrictions on the VAR which can be utilized for testing rational expectations (see Johansen and Swensen, 1999, 2004).

Even with a model which, for many practical purposes, is small scale, it is usually too big to be formulated in "one go" within a cointegrated VAR framework. Hence, model (17.14) is not interpretable as a rather high-dimensional VAR, with the (incredible) long lags which would be needed to capture the complicated dynamic interlinkages of a real economy. Instead, as explained in Bårdsen *et al.* (2003), our operational procedure is to partition the (big) simultaneous distribution function of markets and variables (prices, wages, output, interest rates, the exchange rate, foreign prices, and unemployment, etc.) into a (much smaller) simultaneous model of wage- and price-setting – the labor market – and several sub-models of the rest of the macro-economy. The econometric rationale for specification and estimation of single equations, or of markets, subject to exogeneity conditions, before joining them up in a complete model, is discussed in Bårdsen, Jansen and Nymoen (2003) and also in Bårdsen *et al.* (2005, Ch. 2).

17.2.5.2 Second step: the overidentified steady state

The second step of the model-building exercise will then be to identify the steady state, by testing and imposing overidentifying restrictions on the cointegration space:

$$\Delta \mathbf{y}_t = \mathbf{c} + \alpha \beta^{'} \mathbf{y}_{t-1} + \sum_{i=1}^{k} \Gamma_{t-i} \Delta \mathbf{y}_{t-i} + \mathbf{u}_t,$$

thereby identifying both the exogenous common trends, or permanent shocks, and the steady state of the model.

Even though there now exists a literature on identification of cointegration vectors, it is worthwhile to reiterate that identification of cointegrating vectors cannot be databased. Identifying restrictions have to be imposed *a priori*. It is therefore of crucial importance to have a specification of the economic model and its derived steady-state before estimation. Otherwise we will not know what model and hypotheses we are testing and, in particular, we could not be certain that it was identifiable from the available dataset.

17.2.5.3 Third step: the dynamic SEM

The final step is to identify the dynamic structure:

$$\mathbf{A}_0 \Delta \mathbf{y}_t = \mathbf{A}_0 \mathbf{c} + \mathbf{A}_0 \alpha \beta^{'} \mathbf{y}_{t-1} + \sum_{i=1}^{k} \mathbf{A}_0 \Gamma_{t-i} \Delta \mathbf{y}_{t-i} + \mathbf{A}_0 \mathbf{u}_t,$$

by testing and imposing overidentifying restrictions on the dynamic part – including the forward-looking part – of the statistical system.

The estimated parameters and, therefore, the interpretation of the model dynamics are dependent upon the dating of the steady-state solution. However, the steady-state multipliers are not. The economic interpretations of the derived paths of adjustment are not invariant to the identification of the dynamic part of the model, whereas the steady-state parts of the model are (see Bårdsen and Fisher, 1993, 1999).

17.2.6 Example: the supply side of a medium-term macroeconomic model

One main focus of an empirical macro-model is always going to be the supply side. We end the section on methodology by giving an extended example of the theoretical and econometric specification of a labor market model that we later include in a macro-model, intended for medium-term analysis and forecasting. The first step is the specification of the relevant economic theory to test. We next develop the theoretical relationships into hypotheses about cointegration that can be tested in a statistical model and identified as steady-state relationships, Steps 1 and 2 above. We also go through Step 3 in detail. Throughout the rest of the chapter we let lower-case letters denote natural logarithms of the corresponding upper-case variable names, so $x_t \equiv \ln(X_t)$.

17.2.6.1 Economic theory

A main advance in the modeling of labor markets rests on the perception that firms and their workers are engaged in a partly cooperative and partly conflicting sharing of the rents generated by the operation of the firm. In line with this assumption, *nominal wages* are modeled in a game theoretic framework which fits the comparatively high level of centralization and coordination in Norwegian wage-setting (see, e.g., Nymoen and Rødseth, 2003; Barkbu, Nymoen and Røed, 2003, for a discussion of the degree of coordination).

The modeling of nominal wage-setting in a game theoretic framework is a theoretical advance with several implications. Linked with an assumption of monopolistically competitive firms, it represents an incomplete competition model of the

supply side, which we refer to as ICM in the following (see Bårdsen *et al.*, 2005, Chs. 5 and 6). In applications, the gap between the formal relationships of the theory and the empirical relationships that may be present in the data must be closed. The modeling assumption about $I(1)$-ness introduced above is an important part of the bridge between theory and data. This is because $I(1)$-ness allows us to interpret the theoretical wage and price equations as hypothesized cointegration relationships. From that premise, a dynamic model of the supply side in equilibrium-correction form follows logically.

There is a number of specialized models of "non-competitive" wage-setting. Our aim here is to represent the common features of these approaches by extending the model in Nymoen and Rødseth (2003) with monopolistic competition among firms.

We start with the assumption of a large number of firms, each facing downward-sloping demand functions. The firms are price-setters and equate marginal revenue to marginal costs. With labor being the only variable factor of production (and constant returns to scale), we have the price-setting relationship:

$$Q_i = \frac{El_Q Y}{El_Q Y - 1} \frac{W_i(1 + T1_i)}{Z_i},$$

where $Z_i = Y_i/N_i$ is average labor productivity, Y_i is output and N_i is labor input. W_i is the wage rate in the firm, and $T1_i$ is a payroll tax rate. $El_Q Y > 1$ denotes the absolute value of the elasticity of demand facing each firm i with respect to the firm's own price. In general, $El_Q Y$ is a function of relative prices, which provides a rationale for the inclusion of, e.g., the real exchange rate in aggregate price equations. However, it is a common simplification to assume that the elasticity is independent of other firms prices and is identical for all firms. With constant returns technology, aggregation is no problem, but for simplicity we assume that average labor productivity is the same for all firms and that the aggregate price equation is given by:

$$Q = \frac{El_Q Y}{El_Q Y - 1} \frac{W(1 + T1)}{Z}. \tag{17.15}$$

The expression for real profits (Π) is therefore

$$\Pi = Y - \frac{W(1 + T1)}{Q} N = \left(1 - \frac{W(1 + T1)}{Q} \frac{1}{Z}\right) Y.$$

We assume that the wage W is set in accordance with the principle of maximizing the Nash product:

$$(V - V_0)^{\upsilon} \Pi^{1-\upsilon}, \tag{17.16}$$

where V denotes union utility and V_0 denotes the fallback utility or reference utility. The corresponding break-point utility for the firms has already been set to zero in (17.16), but for unions the utility during a conflict (e.g., strike or work-to-rule) is non-zero because of compensation from strike funds. Finally, υ represents the relative bargaining power of unions.

Union utility depends on the consumer real wage of an unemployed worker and the aggregate rate of unemployment, thus $V(\frac{W}{P}, U, A_v)$ where P denotes the consumer price index.[7] The partial derivative with respect to wages is positive, and negative with respect to unemployment ($V'_W > 0$ and $V'_U \leq 0$). A_v represents other factors in union preferences. The fallback or reference utility of the union depends on the overall real wage level and the rate of unemployment, hence $V_0 = V_0(\frac{\bar{W}}{P}, U)$ where $\bar{W}$ is the average level of nominal wages, which is one of the factors determining the size of strike funds. If the aggregate rate of unemployment is high, strike funds may run low, in which case the partial derivative of V_0 with respect to U is negative ($V'_{0U} < 0$). However, there are other factors working in the other direction, for example that the probability of entering a labor market program, which gives laid-off workers higher utility than open unemployment, is positively related to U.

With these specifications of utility and break-points, the Nash product, denoted $\mathcal{N}$, can be written as:

$$\mathcal{N} = \left\{ V\left(\frac{W}{P}, U, A_v\right) - V_0\left(\frac{\bar{W}}{P}, U\right) \right\}^{\upsilon} \left\{ \left(1 - \frac{W(1+T1)}{Q}\frac{1}{Z}\right) Y \right\}^{1-\upsilon},$$

or:

$$\mathcal{N} = \left\{ V\left(\frac{RW}{P_q(1+T1)}, U, A_v\right) - V_0\left(\frac{\bar{W}}{P}, U\right) \right\}^{\upsilon} \left\{ \left(1 - RW\frac{1}{Z}\right) Y \right\}^{1-\upsilon},$$

where $RW = W(1+T1)/Q$ is the producer real wage, and $P_q(1+T1) = P(1+T1)/Q$ is the *wedge* between the consumer and producer real wage.

Note also that, unlike many expositions of the so-called "bargaining approach" to wage modeling (e.g., Layard, Nickell and Jackman, 1991, Ch. 7), there is no aggregate labor demand function – employment as a function of the real wage – subsumed in the Nash product. In this we follow Hahn and Solow (1997, Ch. 5.3), who point out that bargaining is usually over the nominal wage and not over employment.

The first-order condition for a maximum is given by $\mathcal{N}_{RW} = 0$, or:

$$\upsilon \frac{V'_W\left(\frac{RW}{P_q(1+T1)}, U, A_v\right)}{V\left(\frac{RW}{P_q(1+T1)}, U, A_v\right) - V_0\left(\frac{\bar{W}}{P}, U\right)} = (1-\upsilon)\frac{\frac{1}{Z}}{\left(1 - RW\frac{1}{Z}\right)}. \tag{17.17}$$

In a symmetric equilibrium, $W = \bar{W}$, leading to $\frac{RW}{P_q(1+T1)} = \frac{\bar{W}}{P}$ in equation (17.17). The aggregate bargained real wage RW^b is defined implicitly as:

$$RW^b = F(P_q(1+T1), Z, \upsilon, U), \tag{17.18}$$

or, using the definition:

$$RW^b \equiv W^b(1+T1)/Q,$$

we obtain the solution for the bargained nominal wage:

$$W^b = \frac{Q}{(1+T1)} F(P_q(1+T1), Z, \mho, U). \tag{17.19}$$

Letting lower-case letters denote logs of variables, a log-linearization of (17.19) gives:

$$w^b = m_w + q_t + (1-\delta_{12})(p-q) + \delta_{13} z - \delta_{15} u - \delta_{16} T1. \tag{17.20}$$
$$0 \le \delta_{12} \le 1,\ 0 < \delta_{13} \le 1,\ \delta_{15} \ge 0,\ \ 0 \le \delta_{16} \le 1.$$

The elasticity of the wedge variable $(p-q)$ is $(1-\delta_{12})$ in (17.20). In econometric models of wage-setting in manufacturing, the hypothesis of $\delta_{12} = 1$ is typically not rejected, meaning that the wedge variable drops out and the bargained nominal wage is linked one-to-one with the producer price q (see, e.g., Nymoen and Rødseth, 2003). However, at the aggregate level, a positive coefficient of the wedge is typically reported. This may be due to measurement problems: since gross domestic product (GDP) is an income variable, the price deflator q is not a good index of "producer prices." That said, the estimated importance of the wedge may also reflect that the economy-wide average wage is influenced by the service sector, where wage claims are linked to cost of living considerations, implying that $(1-\delta_{12})$ is different from zero.

Irrespective of the split between q and p, productivity z is found to be a main determinant of the secular growth in wages in bargaining-based systems, so we expect the elasticity δ_{13} to be close to one. The impact of the rate of unemployment on the bargained wage is given by the elasticity $-\delta_{15} \le 0$. Blanchflower and Oswald (1994) provide evidence for the existence of the empirical law that the value of $-\delta_{15}$ is 0.1, which is the slope coefficient of their *wage curve*. Other authors instead emphasize that the slope of the wage-curve is likely to depend on the level of aggregation and on institutional factors. For example, one influential view holds that economies with a high level of coordination and centralization are expected to be characterized by a higher responsiveness to unemployment (a higher $-\delta_{15}$) than uncoordinated systems that give little incentive to solidarity in wage-bargaining (Layard, Nickell and Jackman, 2005, Ch. 8). Finally, from the definition of the wedge, one could set $\delta_{16} = \delta_{12}$, but we keep δ_{16} as a separate coefficient to allow for separate effects of the payroll tax on wages.

Equation (17.20) is a general proposition about the bargaining outcome and its determinants, and can serve as a starting point for describing wage formation in any sector or level of aggregation of the economy. In the following we regard equation (17.20) as a model of the average wage in the total economy and, as explained above, we therefore expect $(1-\delta_{12}) > 0$, meaning that there is a wedge effect in the long-run wage equation.

Equation (17.15) already represents a price-setting rule based upon so-called normal cost pricing. Upon linearization we have:

$$q^f = m_q + (w + T1 - z), \tag{17.21}$$

where we use q^f as a reminder that this is a theoretical equation for firms' price-setting.

17.2.6.2 Cointegration and long-run identification

At this point we show how the two theoretical relationships (17.20) and (17.21) can be transformed into hypothesized relationships between observable time series. As explained in section 17.2.3, our maintained modeling assumption is that the real wage and productivity are $I(1)$ series. The rate of unemployment is assumed to be $I(0)$, possibly after removal of deterministic shifts in the mean.

Using subscript t to indicate period t variables, equation (17.20) defines w_t^b as an $I(1)$ variable. Next define:

$$ecm_t^b = rw_t - rw_t^b \equiv w_t - w_t^b.$$

Under the null hypothesis that the theory is correct, the "bargained wage" w_t^b cointegrates with the actual wage, hence $ecm_t^b \sim I(0)$, which is a testable hypothesis. We can write the long-run wage equation following from bargaining theory as:

$$w_t = m_w + q_t + (1 - \delta_{12})(p_t - q_t) + \delta_{13} z_t - \delta_{15} u_t - \delta_{16} T1_t + ecm_t^b. \qquad (17.22)$$

With reference to equation (17.21), a similar argument applies to price-setting. The "firm-side" real wage can be defined as:

$$rw_t^f \equiv w_t + T1_t - q_t^f = -m_q + z_t,$$

and the difference between the actual real wage and the real wage implied by price-setting becomes:

$$ecm_t^f = rw_t - rw_t^f = w_t + T1_t - q_t - \{-m_q + z_t\}.$$

Hence, the implied long-run price-setting equation becomes:

$$q_t = m_q + (w_t + T1_t - z_t) - ecm_t^f, \qquad (17.23)$$

where $ecm_t^f \sim I(0)$ for the equation to be consistent with the modeling assumptions.

The two cointegrating relationships (17.22) and (17.23) are not identified in general, but in several cases of relevance, identification is unproblematic (see Bårdsen *et al.*, 2005, p. 81). Here we consider a case which is relevant for an aggregate model of the supply side in an open economy. Equations (17.22) and (17.23) can then be combined with a definition of the consumer price index p_t,

$$p_t = (1 - \zeta) q_t + \zeta pi_t + \eta T3_t, \; 0 < \zeta < 1, \quad 0 < \eta \leq 1, \qquad (17.24)$$

where the import price index pi_t naturally enters. The parameter ζ reflects the openness of the economy.[8] Also, the size of the parameter η will depend on how much of the retail price basket is covered by the indirect tax-rate index $T3_t$. By substitution of (17.24) in (17.22), and of (17.23) in (17.24), the system can be

specified in terms of w_t and p_t:

$$w_t = m_w + \left\{1 + \zeta \frac{\delta_{12}}{(1-\zeta)}\right\} p_t \tag{17.25}$$
$$- \frac{\delta_{12}\zeta}{(1-\zeta)} pi_t - \frac{\delta_{12}\eta}{(1-\zeta)} T3_t + \delta_{13} z_t - \delta_{15} u_t - \delta_{16} T1_t + ecm_t^b$$

$$p_t = (1-\zeta) m_q + (1-\zeta)\left\{w_t + T1_t - z_t\right\} + \zeta pi_t + \eta T3_t - (1-\zeta) ecm_t^f. \tag{17.26}$$

By simply viewing (17.25) and (17.26) as a pair of simultaneous equations, it is clear that the system is unidentified in general. However, for the purpose of modeling the aggregate economy, we choose the consumer price index p_t as the representative domestic price index by setting $\delta_{12} = 0$. In this case, (17.26) is unaltered, while the wage equation becomes:

$$w_t = m_w + p_t + \delta_{13} z_t - \delta_{15} u_t - \delta_{16} T1_t + ecm_t^b. \tag{17.27}$$

The long-run price equation (17.26) and the long-run wage equation (17.27) are identified by the order condition.

17.2.6.3 VAR and identified equilibrium correction system

The third stage in the operationalization is the equilibrium-correction system, where we follow Bårdsen and Fisher (1999). In brief, we allow wage growth Δw_t to interact with current and past price inflation, changes in unemployment, changes in tax rates, and previous deviations from the desired wage level consistent with (17.27):

$$\begin{aligned} \Delta w_t - \alpha_{12,0} \Delta q_t = {} & c_1 + \alpha_{11}(L)\,\Delta w_t + \alpha_{12}(L)\,\Delta q_t + \beta_{12}(L)\,\Delta z_t \\ & - \beta_{14}(L)\,\Delta u_t - \beta_{15}(L)\,\Delta T1_t \\ & - \gamma_{11} ecm_{t-r}^b + \beta_{18}(L)\,\Delta p_t + \epsilon_{1t}, \end{aligned} \tag{17.28}$$

where Δ is the difference operator, and the $\alpha_{1j}(L)$ and $\beta_{1j}(L)$ are polynomials in the lag operator L:

$$\alpha_{1j}(L) = \alpha_{1j,1} L + \cdots + \alpha_{1j,(r-1)} L^{r-1},\ j = 1, 2,$$
$$\beta_{1j}(L) = \beta_{1j,0} + \beta_{1j,1} L + \cdots + \beta_{1j,(r-1)} L^{r-1},\ j = 2, 4, 5, 6.$$

The β-polynomials are defined so that they can contain contemporaneous effects. The order r of the lag polynomials may, of course, vary between variables and is to be determined empirically. This specification is a generalization of the typical European wage curve, where the American version is derived by setting $\gamma_{11} = 0$ (see Blanchard and Katz, 1999).

Any increase in output above the optimal trend exerts a (lagged) positive pressure on prices, measured by the output gap_t, as in Phillips curve inflation models (see Clarida, Gali and Gertler, 1999). In addition, product price inflation interacts with

wage growth and productivity gains and with changes in the payroll tax rate, as well as with corrections from an earlier period's deviation from the equilibrium price (as a consequence of, e.g., information lags; see Andersen, 1994, Ch. 6.3):

$$\Delta q_t - \alpha_{21,0}\Delta w_t = c_2 + \alpha_{22}(L)\,\Delta q_t + \alpha_{21}(L)\,\Delta w_t + \beta_{21}(L)\,gap_t - \beta_{22}(L)\,\Delta z_t + \beta_{25}(L)\,\Delta T1_t - \gamma_{22} ecm^f_{t-r} + \epsilon_{2t}, \tag{17.29}$$

where:

$$\alpha_{2j}(L) = \alpha_{2j,1}L + \cdots + \alpha_{2j,(r-1)}L^{r-1},\ j = 1, 2,$$

$$\beta_{2j}(L) = \beta_{2j,0} + \beta_{2j,1}L \cdots + \beta_{2j,(r-1)}L^{r-1}, j = 1, 2, 5.$$

Solving equation (17.24) for Δq_t (i.e., the equation is differenced first), and then substituting out in equations (17.28) and (17.29), the theoretical model condenses to a wage–price model suitable for estimation and similar to the early equilibrium-correction formulation of Sargan (1980):

$$\begin{aligned}
&\begin{bmatrix} 1 & -a_{12,0} \\ -a_{21,0} & 1 \end{bmatrix}\begin{bmatrix} \Delta w \\ \Delta p \end{bmatrix}_t \\
&= \begin{bmatrix} \alpha_{11}(L) & -a_{12}(L) \\ -a_{21}(L) & \alpha_{22}(L) \end{bmatrix}\begin{bmatrix} \Delta w \\ \Delta p \end{bmatrix}_t \\
&+ \begin{bmatrix} 0 & \beta_{12}(L) & -\zeta\dfrac{\alpha_{12}(L)}{1-\zeta} & -\beta_{14}(L) & -\beta_{15}(L) & -\eta\dfrac{\alpha_{12}(L)}{1-\zeta} \\ b_{21}(L) & -b_{22}(L) & \zeta\alpha_{22}(L) & 0 & b_{25}(L) & \eta\alpha_{22}(L) \end{bmatrix} \\
&\times \begin{bmatrix} gap \\ \Delta z \\ \Delta pi \\ \Delta u \\ \Delta T1 \\ \Delta T3 \end{bmatrix}_t - \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{bmatrix} \\
&\times \begin{bmatrix} 1 & -(1+\zeta d_{12}) & -\delta_{13} & \zeta d_{12} & \delta_{15} & \delta_{16} & \eta d_{12} \\ -(1-\zeta) & 1 & (1-\zeta) & -\zeta & 0 & -(1-\zeta) & -\eta \end{bmatrix} \\
&\times \begin{bmatrix} w \\ p \\ z \\ pi \\ u \\ T1 \\ T3 \end{bmatrix}_{t-r} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}_t,
\end{aligned} \tag{17.30}$$

where we have omitted the intercepts to save space, and have substituted the equilibrium-correction terms using (17.25) and (17.26) above. The mapping from the theoretical parameters in (17.28) and (17.29) to the coefficients of the model

(17.30) is given by:

$$
\begin{aligned}
a_{12,0} &= \frac{\alpha_{12,0}}{1-\zeta} + \beta_{18,0}, \\
a_{21,0} &= (1-\zeta)\,\alpha_{21,0}, \\
a_{12}(L) &= \frac{\alpha_{12}(L)}{1-\zeta} + \beta_{18}(L), \\
a_{21}(L) &= (1-\zeta)\,\alpha_{21}(L), \\
b_{2j}(L) &= (1-\zeta)\,\beta_{2j}(L), \; j = 1, 2, 5, \\
d_{12} &= \frac{\delta_{12}}{1-\zeta}, \\
e_1 &= \epsilon_1, \\
e_2 &= (1-\zeta)\,\epsilon_2.
\end{aligned}
\tag{17.31}
$$

The model (17.30) contains the different channels and sources of inflation discussed so far: imported inflation Δpi_t, and several relevant domestic variables – the output gap, and changes in the rate of unemployment, in productivity, and in tax rates. Finally, the model includes deviations from the two cointegration equations associated with wage-bargaining and price-setting, which have equilibrium-correction coefficients γ_{11} and γ_{22} respectively. Consistency with assumed cointegration implies that the joint hypothesis of $\gamma_{11} = \gamma_{22} = 0$ can be rejected.

17.2.6.4 Economic interpretation of the steady state

The dynamic model in (17.30) can be rewritten in terms of real wages $(w-p)_t$ and real exchange rates $(pi-p)_t$. Using a specification with first-order dynamics, Bårdsen *et al.* (2005, Ch. 6) discuss several different aspects of this model. Most importantly, the dynamic system is asymptotically stable under quite general assumptions about the parameters, including, e.g., dynamic homogeneity in the two equilibrium-correction equations. The steady state is conditional on any given rate of unemployment, which amounts to saying that our core supply-side model does not tie down an equilibrium rate of unemployment. Instead, there is a stalemate in the dynamic "tug-of-war" between workers and firms that occurs for, in principle, any given rate of unemployment (see Kolsrud and Nymoen, 1998; Bårdsen and Nymoen, 2003, for proofs). Since there are no new unit roots implied by the generalized dynamics in equation (17.30) above, asymptotic stability holds also for this, extended, version of the model. We therefore have the following important results: the dynamics of the supply side are asymptotically stable in the usual sense that, if all stochastic shocks are switched off, then $(pi_t - q_t) \to rex_{ss}(t)$, and $(w_t + T1_t - q_t) = wq_{ss}(t)$, where $rex_{ss}(t)$ and $wq_{ss}(t)$ represent deterministic steady-state growth paths of the real exchange rate and the producer real wage.

Generally, the steady-state growth paths depend on the steady-state growth rate of import prices, and of the mean of the logarithm of the rate of unemployment, denoted u_{ss}, and the expected growth path of productivity $z(t)$. However, under the

condition that $\delta_{13} = 1$, homogeneity of degree one with respect to productivity, which we have seen is implied theoretically by assuming bargaining power on the part of unions, $z(t)$ has a zero coefficient in the expression for rex_{ss}, which therefore is constant in the steady state. Moreover, assuming $\delta_{13} = 1$, the implied steady-state wage share, $wq_{ss}(t) - z(t) = ws_{ss}$, which is also a constant in steady state.

With $\delta_{13} = 1$, the implied steady-state inflation rate therefore follows immediately: since $\Delta(pi_t - q_t) = 0$ in steady state, and $\Delta p_t = (1-\zeta)\,\Delta q_t + \zeta\,\Delta pi_t$, domestic inflation is equal to the constant steady-state rate of imported inflation:

$$\Delta p_t = \Delta pi_t = \pi. \tag{17.32}$$

The above argument implicitly assumes an exogenous and, for simplicity, constant, nominal exchange rate. For the case of an endogenous nominal exchange rate, as with a floating exchange rate regime, it might be noted that, since:

$$pi_t = v_t + p_t^*,$$

where v_t is the nominal exchange rate and the index of import prices in foreign currency is denoted p_t^*, the stability of inflation requires stability of Δv_t. This condition can only be verified by the use of a more complete model representation of the economy, which is what we do when we consider the steady state of a complete econometric model in section 17.3.2 below. However, to anticipate events slightly, the complete model that we document below meets the requirement in the sense that $\Delta^2 v_t \to 0$ in the long run. But our results also indicate that π in (17.32) is affected by the rate of change in the nominal exchange rate, which might be non-zero in an asymptotically stable steady state.

The supply-side determined steady state has a wider relevance as well. For example, what does the model say about the dictum that the existence of a steady-state inflation rate requires that the rate of unemployment follows the law of the natural rate or non-accelerating inflation rate of unemployment (NAIRU)? The version of this natural rate/NAIRU view of the supply side that fits most easily into our framework is the one succinctly expressed by Layard, Nickell and Jackman (1994, p. 18; emphasis added): "Only if the real wage (W/P) desired by wage-setters is the same as that desired by price setters will inflation be stable. *And, the variable that brings about this consistency is the level of unemployment.*" Translated to our conceptual framework, this view corresponds to setting $ecm_t^b = ecm_t^f = 0$ in (17.22) and (17.23), with $\delta_{13} = 1$, and solving for the rate of unemployment that reconciles the two desired wage shares, call it u^w:[9]

$$u^w = \frac{m_w + m_q}{-\delta_{15}} + \frac{1-\delta_{12}}{-\delta_{15}}(p-q) + \frac{1-\delta_{16}}{-\delta_{15}}T1,$$

which can be expressed in terms of the real exchange rate $(p - pi)$, and the two tax rates as:

$$u^w = \frac{-(m_w + m_q)}{\delta_{15}} + \frac{1-\delta_{12}}{\delta_{15}(1-\zeta)}\zeta(p-pi) + \frac{1-\delta_{12}}{\delta_{15}(1-\zeta)}\eta T3 + \frac{1-\delta_{16}}{-\delta_{15}}T1. \tag{17.33}$$

This is one equation in two endogenous variables, u^w and the wedge $(p - pi)$, so it appears that there is a continuum of u^w values depending on the size of the wedge, in particular of the value of the real exchange rate. It is, however, customary to assume that the equilibrium value of the wedge is determined by the requirement that the current account is in balance in the long run. Having thus pinned down the long-run wedge as a constant equilibrium real exchange rate $(\overline{p - pi})$, it follows that NAIRU u^w is determined by (17.33). If the effect of the wedge on wage claims is not really a long-run phenomenon, then $\delta_{12} = 1$ and u^w is uniquely determined from (17.33), and there is no need for the extra condition about balanced trade in the long run (see Layard, Nickell and Jackman, 2005, p. 33).

Compare this to the asymptotically stable equilibrium consisting of $u_t = u_{ss}$, $\Delta p_t = \pi$ and $w_t + T1 - q_t - z_t = ws_{ss}$. Clearly, inflation is stable, even though u_{ss} is determined "from the outside" and is not determined by the wage- and price-setting equations of the model. Hence the (emphasized) second sentence in the above quotation has been disproved: it is not necessary that u_{ss} corresponds to the NAIRU u^w in equation (17.33) for inflation to be stable with a well-defined value in steady state.

Figure 17.1 illustrates the different equilibria. Wage-setting and price-setting curves correspond to (deterministic versions) of equations (17.22) and (17.23). The NAIRU u^w is given by the intersection of the curves, but the steady-state rate of unemployment u_{ss} may be lower than u^w, the case shown in the graph, or higher. The figure further indicates (by a •) that the steady-state wage share will reside at a point on the line segment A–B: heuristically, this is a point where price-setters

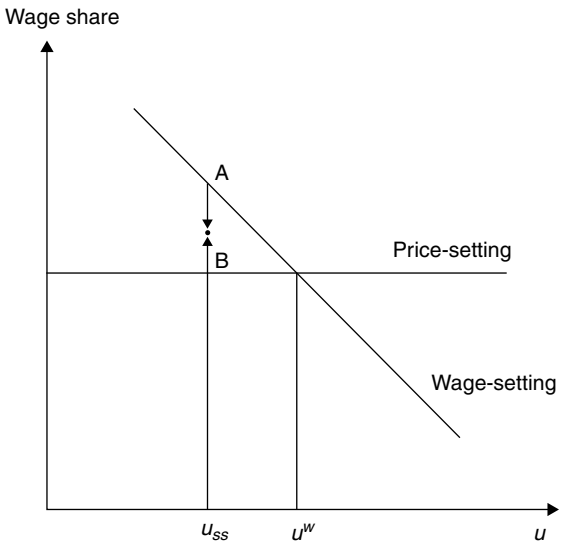


Figure 17.1 Real wage and unemployment determination, NAIRU and the steady-state rate of unemployment u_{ss}

are trying to attain a lower real wage by nominal price increases, at the same time as the wage bargain is delivering nominal wage increases that push the real wage upwards.

Bårdsen *et al.* (2005, Ch. 6) show which restrictions on the parameters of the system (17.30) are necessary for $u_t \to u_{ss} = u^w$ to be an implication, so that the NAIRU corresponds to the stable steady state. In brief, the model must be restricted in such a way that the nominal wage- and price-setting adjustment equations become two conflicting dynamic equations for the real wage. Because of the openness of the economy, this is not achieved by imposing dynamic homogeneity. What is required is to purge the model (17.30) of all nominal rigidity, which seems to be unrealistic on the basis of both macro- and micro-evidence.

We have seen that the Layard–Nickell version of the NAIRU concept corresponds to a set of restrictions on the dynamic model of wage- and price-setting. The same is true for the natural rate of unemployment associated with a vertical Phillips curve, which still represents the baseline model for the analyses of monetary policy. This is most easily seen by considering a version of (17.28) with first-order dynamics and where we simplify the equation by setting the short-run effects of productivity, unemployment and taxes equal to zero ($\beta_{12} = \beta_{14} = \beta_{15} = 0$). With first-order dynamics we have:

$$\Delta w_t - \alpha_{12,0}\Delta q_t = c_1 - \gamma_{11} ecm^b_{t-1} + \beta_{18}\Delta p_t + \epsilon_{1t},$$

and using (17.22) we can then write the wage equation as:

$$\begin{aligned}\Delta w_t = {} & k_w + \alpha_{12,0}\Delta q_t + \beta_{18}\Delta p_t - \mu_w u_{t-1} \\ & - \gamma_{11}(w_{t-1} - q_{t-1}) + \gamma_{11}(1-\delta_{12})(p_{t-1} - q_{t-1}) + \gamma_{11}\delta_{16} T1_{t-1} + \epsilon_{1t},\end{aligned} \tag{17.34}$$

where $k_w = c_1 + \gamma_{11} m_w$, and the parameter μ_w is defined in accordance with Kolsrud and Nymoen (1998) as:

$$\mu_w = \gamma_{11}\delta_{13} \text{ when } \gamma_{11} > 0 \text{ or } \mu_w = \varphi \text{ when } \gamma_{11} = 0. \tag{17.35}$$

The notation in (17.35) may seem cumbersome at first sight, but it is required to secure internal consistency: note that if the nominal wage rate is adjusting towards the long-run wage curve, $\gamma_{11} > 0$, the only logical value of φ in (17.35) is zero, since u_{t-1} is already contained in the equation, with coefficient $\gamma_{11}\delta_{15}$. Conversely, if $\gamma_{11} = 0$ so the model of collective wage-bargaining fails, it is nevertheless possible that there is a wage Phillips curve relationship consistent with the assumed $I(0)$-ness of the rate of unemployment, hence $\mu_w = \varphi \geq 0$ in this case.

Subject to the restriction $\gamma_{11} = 0$, and assuming an asymptotically stable steady-state inflation rate π, (17.34) can be solved for the Phillips curve NAIRU u^{phil}:

$$u^{phil} = \frac{k_w}{\varphi} + \frac{(\alpha_{12,0} + \beta_{18} - 1)}{\varphi}\pi,$$

which becomes a natural rate of unemployment, independent of inflation subject to dynamic homogeneity $\alpha_{12,0} + \beta_{18} = 1$.

However, the claim that u_t^{phil} represents an asymptotically stable solution must be stated with some care. As shown in, e.g., Bårdsen and Nymoen (2003), $\gamma_{11} = 0$ is a necessary but not a sufficient condition. The sufficient conditions include $\gamma_{22} = 0$ in addition to $\gamma_{11} = 0$ and, instead of equilibrium correction in wages and prices, dynamic stability requires equilibrium correction in the unemployment equation or in a functionally equivalent part of the model.

The result that the steady-state level of unemployment is generally undetermined by the wage–price sub-model is a strong case for building larger systems of equations, even if the main objective is to model inflation. Conversely, in general, no inconsistencies or issues about overdetermination arise from enlarging the wage/price-setting equations with a separate equation for the rate of unemployment, where demand-side variables may enter.

Looking ahead, in section 17.4.3 we show how the specification of the supply side, either as a Phillips curve model (PCM) or as an incomplete competition model (ICM) given by equations (17.28) and (17.29) above, gains economic significance though the implications of the chosen specification for optimal interest rate-setting.

17.2.6.5 Implementation in the Norwegian aggregate model

We have implemented the above model of the supply side in our quarterly model of the Norwegian economy, called the Norwegian aggregate model (NAM).[10]

The estimated versions of (17.30) are given in section 17.3.1, equations (17.38) and (17.39). The equilibrium correction terms are defined consistently with the two long-run equations (17.25) and (17.26). For example, $\delta_{12} = 0$, $\zeta = 0.7$, $\delta_{15} = 0.1$ and $\delta_{16} = 1$ are taken as known parameters from the cointegration analysis documented in Bårdsen *et al.* (2005, Ch. 9.2). With this parameterization, the estimated equilibrium correction coefficients $\hat{\gamma}_{11}$ and $\hat{\gamma}_{22}$ are jointly and individually significant (the t-values are 8.6 and 3.8).

The estimated short-run dynamics can also be interpreted in the light of the theoretical model (17.30). For example, the estimated wage equation (17.39) shows that $\hat{a}_{12}(1) = 1$, saying that dynamic homogeneity with respect to consumer price is a valid restriction on wage dynamics in the wage equation. Identification of the short-run wage price model is in terms of zero-restrictions on the GDP growth variable in the Δw_t equation, and on the change in the rate of unemployment in the Δp_t equation. There are overidentifying restrictions as well though.

17.3 Building a model for monetary policy analysis

Monetary policy now plays a dominant role in stabilization policy in general and in managing inflation in particular. As economists have recognized for a long time, inflation is a many-faceted phenomenon. In particular, in open economies, a proper understanding of the inflation mechanism requires the construction of a

model that separates the internal dynamics of the domestic wage price spiral from the factors that impinge upon it from outside. The complexity of the real-world inflation process also means that models which only include one or two dimensions typically fail to characterize the data. Our starting point is therefore that, at a minimum, foreign and domestic aspects of inflation have to be modeled jointly, and that the inflationary impetus from the labor market, the battle of markups between unions and monopolistic firms, needs to be represented in the model.

The last section ended with an example of the econometric specification of a model of wage- and price-setting that defines an integral part of the NAM model. Earlier versions of this model have been used to analyze the issues raised by the introduction of model-based monetary policy in Norway (see Bårdsen, Jansen and Nymoen, 2002, 2003; Bårdsen *et al.*, 2005). NAM is in use for forecasting as part of the Normetrics forecasting system.[11] A designated version is operational for stress testing by the financial stability division at the central bank of Norway.

17.3.1 The model and its transmission mechanisms

In the regime with inflation targeting, the policy instrument in the model is the money market interest rate, symbolized by R in Figure 17.2 (and throughout this chapter), with the estimated reaction function reported in equation (17.46).[12]

The qualitative transmission mechanisms of the model, from the perspective of monetary policy analysis, are shown in Figure 17.2. The corresponding quantitative approximate transmission mechanisms are easily worked out with the stylized

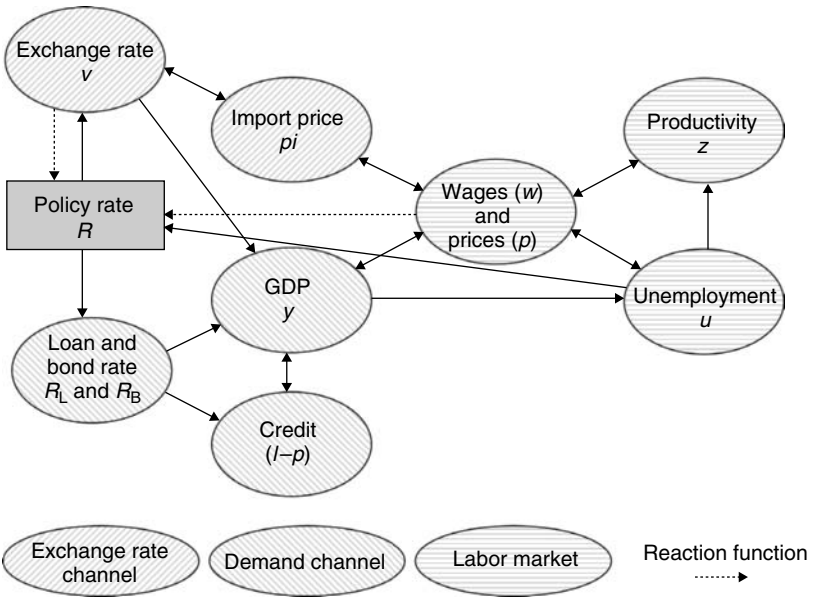


Figure 17.2 The transmission mechanisms in the model in Table 17.1

Table 17.1 The econometric model NAM

The exchange rate

$$\Delta v_t = -\underset{(0.012)}{0.042}\ \left\{\left(v + p^* - p\right) - 0.12\left[(R-\pi) - \left(R^* - \pi^*\right)\right] - \mu_v\right\}_{t-1}$$

$$-\underset{(0.00693)}{0.0361}\ IT_t \times \Delta(R - R^*)_t - \underset{(0.015)}{0.036}\ \Delta^2 po_{t-1} \tag{17.36}$$

$$\begin{array}{ll} OLS, & T = 1994\,(2) - 2007\,(2) = 56, \quad \hat{\sigma} = 1.6\% \\ F_{\mathsf{AR}(1-4)}(4, 44) = 1.60[0.19] & F_{ARCH(1-4)}\,(4, 40) = 0.99\,[0.42] \\ \chi^2_{normality}(2) = 0.07[0.97] & F_{\mathsf{x}}(6, 41) = 0.59[0.80], \end{array}$$

where $\pi_t \equiv 100\frac{\Delta_4 P_t}{P_{t-4}}$ and $\pi^*_t \equiv 100\frac{\Delta_4 P^*_t}{P^*_{t-4}}$.

Import prices

$$\Delta pi_t = -\underset{(0.0806)}{0.431}\ \left[\left(pi - v - pi^*\right) - 0.55\left(p - v - p^*\right) - \mu_{pi}\right]_{t-1}$$

$$+\underset{(0.0718)}{0.429}\ \Delta v_t + \underset{(0.211)}{1.07}\ \Delta pi^*_t \tag{17.37}$$

$$\begin{array}{ll} OLS, & T = 1990\,(1) - 2007\,(1) = 69, \quad \hat{\sigma} = 1.1\% \\ F_{AR(1-4)}(4, 59) = 1.78[0.15] & F_{ARCH(1-4)}\,(4, 55) = 0.99\,[0.42] \\ \chi^2_{Normality}(2) = 0.81[0.67] & F_{Het}(9, 53) = 1.15[0.34]. \end{array}$$

Prices, wages and productivity

$$\Delta p_t = -\underset{(0.006)}{0.052}\ \left[p_{t-3} - 0.7\,(w-z)_{t-1} - 0.3pi_{t-1} - \mu_p\right] - \underset{(0.025)}{0.06}\ \Delta z_t$$

$$+\underset{(0.049)}{0.29}\ \Delta p_{t-2} + \underset{(0.011)}{0.024}\ \Delta pi_t + \underset{(0.0038)}{0.059}\ \Delta pe_t + \underset{(0.014)}{0.042}\ \Delta y_{t-1} \tag{17.38}$$

$$\Delta w_t = -\underset{(0.019)}{0.065}\ \left[\left(w_{t-1} - p_{t-2} - z_{t-1}\right) + 0.1u_{t-4} - \mu_w\right] + \underset{(0.071)}{0.56}\ \Delta p_t$$

$$+0.43\Delta p_{t-1} - \underset{(0.0053)}{0.039}\ \left(\Delta^2 u_{t-1} + \Delta u_{t-3}\right) + \underset{(0.023)}{0.73}\ \Delta T1_t \tag{17.39}$$

$$\Delta z_t = -\underset{(0.062)}{0.64}\ \left[z_{t-3} - 0.47\left(w-p\right)_{t-1} - 0.0029Trend_t - 0.03u_{t-2} - \mu_z\right]$$

$$+\underset{(0.05)}{0.24}\ \Delta\left(w-p\right)_t - \underset{(0.036)}{0.83}\ \Delta_2 z_{t-1} \tag{17.40}$$

$$\begin{array}{l} FIML, \quad T = 1979\,(3) - 2007\,(1) = 111, \quad \hat{\sigma}_P = 0.3\%, \quad \hat{\sigma}_W = 0.6\%, \\ \hat{\sigma}_Z = 1.2\% \quad F_{vec,\mathsf{AR}(1-5)}(45, 250) = 1.07[0.36] \\ F_{vec,Het}(354, 224) = 1.11[0.20] \quad \chi^2_{vec,Normality}(6) = 7.02[0.32]. \end{array}$$

Table 17.1 (continued)

The rate of unemployment

$$\Delta u_t = -\underset{(0.019)}{0.078}\left\{u_{t-1} - 7.69\Delta\left(w-p\right)_{t-2} - 0.05\left[(R_L-\pi) - 100\Delta_4 y\right]_{t-2} - \mu_u\right\}$$

$$+\underset{(0.071)}{0.44}\ \Delta u_{t-1} + \underset{(0.059)}{0.49}\ \Delta u_{t-4} - \underset{(0.074)}{0.27}\ \Delta u_{t-5} \tag{17.41}$$

$$\begin{aligned}&\textit{OLS}, \qquad T = 1981\,(1) - 2007\,(2) = 106, \qquad \hat{\sigma} = 5.0\%\\ &F_{AR(1-5)}(5,92) = 1.04[0.25] \quad F_{ARCH(1-4)}\,(4,89) = 0.48\,[0.75]\\ &\chi^2_{Normality}(2) = 0.50[0.78] \qquad F_{Het}(13,83) = 1.44[0.16].\end{aligned}$$

Interest rates

$$\Delta R_{L,t} = -\underset{(0.024)}{0.33}\left(R_L - 0.41R_B - 0.76R - \mu_{R_L}\right)_{t-1} + \underset{(0.023)}{0.58}\ \Delta R_t \tag{17.42}$$

$$\Delta R_{B,t} = -\underset{(0.061)}{0.17}\left(R_B - 0.43R - 0.57R^*_B - \mu_{R_B}\right)_{t-1} + \underset{(0.039)}{0.43}\ \Delta R_t$$

$$+\underset{(0.067)}{0.97}\ \Delta R^*_{B,t} \tag{17.43}$$

$$\begin{aligned}&\textit{FIML}, \quad T = 1993\,(2) - 2006\,(4) = 55, \quad \hat{\sigma}_{R_L} = 0.10, \quad \hat{\sigma}_{R_B} = 0.19\\ &F_{vec,\mathsf{AR}(1-4)}(16,84) = 0.53[0.92] \quad F_{vec,Het}(54,90) = 1.30[0.13]\\ &\chi^2_{vec,Normality}(4) = 4.61[0.33].\end{aligned}$$

GDP output

$$\Delta y_t = -\underset{(0.041)}{0.21}\left[y_{t-2} - 0.9g_{t-1} - 0.16\left(v + p^* - p\right)_{t-1} + 0.06\left(R_L - \pi\right)_{t-1} - \mu_y\right]$$

$$-\underset{(0.091)}{0.74}\ \Delta y_{t-1} + \underset{(0.058)}{0.42}\ \Delta g_t + \underset{(0.11)}{0.67}\ \Delta\left(l-p\right)_{t-1} \tag{17.44}$$

$$\begin{aligned}&\textit{OLS}, \qquad T = 1986\,(2) - 2007\,(1) = 104, \qquad \hat{\sigma} = 1.4\%\\ &F_{AR(1-5)}(5,72) = 2.08[0.07] \quad F_{ARCH(1-4)}\,(4,69) = 0.30\,[0.87]\\ &\chi^2_{Normality}(2) = 1.85[0.68] \qquad F_{Het}(11,65) = 1.21[0.30].\end{aligned}$$

Credit

$$\Delta\left(l-p\right)_t = -\underset{(0.022)}{0.094}\left[\left(l-p\right)_{t-3} - 2.65y_{t-4} + 0.04(R_{L,} - i_B)_{t-4} - \mu_{l-p}\right]$$

$$+\underset{(0.043)}{0.15}\ \Delta_2 y_{t-2} + \underset{(0.087)}{0.24}\ \Delta^2\left(w-p\right)_t \tag{17.45}$$

$$\begin{aligned}&\textit{OLS}, \qquad T = 2000\,(1) - 2007\,(2) = 30, \qquad \hat{\sigma} = 0.7\%\\ &F_{AR(1-3)}(3,23) = 0.98[0.42] \quad F_{ARCH(1-3)}\,(3,20) = 0.32\,[0.81]\\ &\chi^2_{Normality}(2) = 4.73[0.09] \qquad F_{Het}(6,19) = 0.25[0.95].\end{aligned}$$

Table 17.1 (continued)

Money market interest rate

$$\Delta R_t = -\underset{(0.047)}{0.27}\left[R_{t-1} - 5.6 - 1.2\left(\pi_{Ct} - \bar{\pi}_C\right) - \left(U_{t-2} - \bar{U}\right) - 0.86\left(R^*_{t-2} - \bar{R}^*\right)\right]$$

$$+\underset{(0.088)}{0.51}\ \Delta_2 R^*_t - \underset{(0.013)}{0.051}\ \Delta_2\left(\frac{V \times P^*}{P}\right)_t \tag{17.46}$$

$$\begin{array}{ll} OLS, \quad T = 1999\,(3) - 2007\,(2) = 32, & \hat{\sigma} = 0.19 \\ F_{AR(1-3)}(3, 22) = 1.62[0.21] & F_{ARCH(1-3)}\,(3, 19) = 0.60\,[0.62] \\ \chi^2_{Normality}(2) = 0.82[0.67] & F_{Het}(12, 12) = 0.59[0.81], \end{array}$$

where $\pi_{Ct} \equiv 100\frac{\Delta_4 P_{C,t}}{P_{C,t-4}}$; $\bar{\pi}_U = 2.5$ (the inflation target); $\bar{U} = 3.4$ (average unemployment rate); $\bar{R}^* = 3.5$ (average foreign short-run interest rate).

Note: Standard errors are reported in parentheses below the coefficients. See the appendix to this chapter for information about the statistics reported below each equation.

model version in section 17.4.1. That simplified quantitative transmission mechanism is derived from the dynamic econometric model reported in Table 17.1. We report the estimated macroeconometric relationships in equilibrium-correction form, with cointegration coefficients imposed as known. The identities that complete the NAM model are not reported. To save space, seasonals and other dummies are also omitted from the equations in the table. The definitions of the variables in the equations are given in the appendix to this chapter.

Consider, for example, the analysis conducted in section 17.4.2.1 of an increase in the interest rate R. The immediate and direct effect is an appreciation of the krone, measured as an increase in the exchange rate v, defined as krone per unit foreign exchange. The multiplier is approximated in (17.60) as $\frac{\Delta v_t}{\Delta R_t} \approx -0.04$, while the complete equation is reported in (17.36).[13]

The decrease in v will affect domestic prices and wage-setting through decreased import prices pi, as reported in equations (17.37)–(17.39), which gives corresponding approximate partial multipliers as $\frac{\Delta pi_t}{\Delta v_t} \approx 0.9$ and $\frac{\Delta p_t}{\Delta pi_t} \approx 0.03$ from (17.61) and (17.62). Hence, at least for a period of time after the interest rate increase, the *exchange rate channel* will provide inflation dampening following an increase in the interest rate.

The exchange rate channel also affects wages and prices indirectly, through GDP y and unemployment u, reported in equations (17.44) and (17.41), respectively. The mechanisms are as follows. Due to nominal rigidity, the real exchange rate appreciates together with the nominal rate, causing decreased competitiveness, lower output, and higher unemployment. Together with the interaction with productivity z in equation (17.40), this constitutes the *labor market channel*. For example, the approximate partial real-wage response from a shock to unemployment is $\frac{\Delta(w-p)_t}{\Delta u_t} \approx -0.04$ from (17.63).

The interest rate effects on the real economy are first channeled through financial markets, where an increase in the money market rate leads to adjustment of the banks' interest rate R_L, and bond yield R_B (see (17.42)–(17.43)). A rise in R_L affects GDP through an increased real interest rate. This is the *demand channel* found in mainstream monetary policy models (see e.g., Ball, 1999). In the model, there is also a second, *credit channel*, whereby interest rates affect output: when interest rates are raised, the amount of available real credit is reduced, as documented in (17.45), which has a negative effect on output. The average partial multiplier is $\frac{\Delta y_t}{\Delta(l-p)_t} \approx 0.4$, using (17.68).

The transmission mechanism pictured in Figure 17.2 shows that the model contains both positive and negative feedback effects from wage and price adjustments, to GDP and unemployment. Higher inflation means that the real interest rate continues to fall in the first periods after the initial cut in the nominal rate (positive feedback). On the other hand, and again due to the raised rate of inflation, the real exchange rate will start to stabilize (negative feedback).

In the figure, the focus is on the transmission mechanisms, which may give the impression that the development of wages and prices is mainly "determined by" monetary policy. This is not the case since, e.g., the important trend component in wages is related to productivity growth through wage-bargaining – (see (17.38)–(17.40)). Having analyzed the transmission mechanism of the model, we now turn to the steady-state properties, pinned down by the overidentified cointegrated steady-state relationships of the model, which are discussed in the next section.

17.3.2 Steady state

Equations (17.47)–(17.56) represent the model's implied long-run relationships. Cointegrated combinations of non-stationary variables are on the left-hand sides of the equations, while stationary variables are evaluated at their mean values on the right.

$$\left(v + p^* - p\right)_t = -0.12\left[(R - \pi) - \left(R^* - \pi^*\right)\right] + \mu_v \tag{17.47}$$

$$\left(pi - v - pi^*\right)_t - 0.55\left(p - v - p^*\right)_t = \mu_{pi} \tag{17.48}$$

$$p_t - 0.7(w - z)_t - (1 - 0.7)\,pi_t = \mu_p \tag{17.49}$$

$$\left(w - p - z\right)_t = 0.1u + \mu_w \tag{17.50}$$

$$z_t - 0.47\left(w - p\right)_t - 0.0029Trend_t = 0.03u + \mu_z \tag{17.51}$$

$$0 = u - 7.7\Delta\left(w - p\right) - 4.5\left[0.01\,(RL - \pi) - \Delta_4 y\right] - \mu_u \tag{17.52}$$

$$0 = RL - 0.41RB - 0.76R - \mu_{RL} \tag{17.53}$$

$$0 = RB - 0.43R - 0.57RB^* - \mu_{RB} \tag{17.54}$$

$$y_t - 0.9g_t - 0.16(v + p^* - p)_t = -0.06\,(RL - \pi) + \mu_y \tag{17.55}$$

$$\left(l - p\right)_t - 2.65y_t + 0.04(RL - RB)_t = \mu_{l-p}. \tag{17.56}$$

Equation (17.47) represents the equilibrium in the market for foreign exchange in a floating exchange rate regime. The central bank's foreign currency reserves are exogenous in the floating exchange rate regime that we are modeling, and therefore are not specified in (17.47). The nominal exchange rate v equilibrates the market in each time period, specifically in the hypothetical steady state represented in (17.47). The relationship follows from the definition of the risk premium, *rp*:

$$rp_t = R_t - R_t^* - (v_{t+1}^e - v_t),$$

where $(v_{t+1}^e - v_t)$ denotes expected depreciation. In real terms, using the definition $rex = v + p^* - p$, the relationship can be written as:

$$\left(v_{t+1}^e + p_{t+1}^{*e} - p_{t+1}^e\right) - \left(v_t + p_t^* - p_t\right) = \left[R_t - \left(p_{t+1}^e - p_t\right)\right] - \left[R_t^* - \left(p_{t+1}^{*e} - p_t^*\right)\right] - rp_t$$

$$\Delta rex_t^e = \left(R_t - \pi_{t+1}^e\right) - \left(R_t^* - \pi_{t+1}^{*e}\right) - rp_t.$$

If expected depreciation of the real exchange rate is assumed to react to deviations from the equilibrium real exchange rate $\overline{rex}$,

$$\Delta rex_t^e = \alpha\left(rex_t - \overline{rex}\right),$$

the solution for the realized real exchange rate becomes:

$$rex_t = \frac{1}{\alpha}\left[\left(R_t - \pi_{t+1}^e\right) - \left(R_t^* - \pi_{t+1}^{*e}\right)\right] + \overline{rex} - \frac{1}{\alpha} rp_t.$$

Finally, replacing expected with realized inflation and assuming a constant risk premium, the steady-state real exchange rate relationship becomes:

$$rex = -.12\left[(R - \pi) - \left(R^* - \pi^*\right)\right] + \mu_v,$$

The sign of α shows whether expectations are regressive ($\alpha < 0$) or extrapolative ($\alpha > 0$), so in our case, since real interest rates are in percentage points, α is given as:

$$\frac{1}{\alpha} = -.12 \times 100,$$

implying that expected depreciation is regressive with approximately 8% adjustment per period:

$$\Delta rex_t^e = -0.083\left(rex_t - \overline{rex}\right).$$

The long-run pass-through from the exchange rate and foreign prices onto import prices in domestic currency *pi* is represented by equation (17.48). It is a homogeneous function of v and foreign producer prices pi^*, but the import price also increases if the real exchange rate (in terms of consumer prices) appreciates. This is due to pricing-to-markets in import price-setting. Equation (17.48) is written in a way that shows the long-run relationship between the two operational definitions

of the real exchange rate. Hence, a 1% real appreciation in terms of consumer prices is associated with only a 0.55% appreciation in terms of import prices.

The equations for wage formation and domestic price-setting, in steady-state form, are given by (17.49)–(17.50) and have already been discussed in section 17.2.6. Equation (17.51) models output per hour of labor input, or productivity z. Productivity is positively influenced by both the real wage $(w - p)$ and the unemployment rate u – corresponding to efficiency wage models – as well as neutral technological progress approximated by a linear trend.

The steady-state rate of unemployment in equation (17.52) is seen to depend on the growth of real wages (which is contant in the steady state), and the difference between the real interest rate and the real GDP growth rate (see also Hendry, 2001b), which can potentially change, and will then induce a change in the long-run mean of the rate of unemployment.

Equations (17.53)–(17.56) represent the steady-state relationships for the market interest rates for loans R_L and bonds R_B, (17.53) and (17.54), followed by the equations for GDP (17.55) and domestic credit (17.56). Government expenditure is seen to be important in the GDP equation, but aggegate demand is also influenced by the market for foreign exchange through the real exchange rate $(v + p^* - p)$ (dubbed *rex* above) and the domestic real interest rate. According to equations (17.55) and (17.56), secular growth in domestic real credit is conditioned by the growth of the real economy, not the other way around, but an exogenous drop in credit supply (i.e., reduced μ_{l-p}) will harm GDP growth in the short run.

From the set of long-run relationships, it is straightforward to derive the steady state of the model. Differentiation of (17.47) gives inflation as:

$$\Delta p = \Delta p^* + \Delta v, \tag{17.57}$$

and steady-state wage growth is then:

$$\Delta w = \Delta p^* + \Delta z + \Delta v, \tag{17.58}$$

while import inflation follows from the price equation (17.49) and becomes:

$$\Delta pi = \Delta p^* + \Delta v. \tag{17.59}$$

Note that these relationships represent the same qualitative conclusion that we obtained from analysis of the theoretical supply-side model; compare equation (17.32), e.g., which is therefore seen to generalize to the full set of economic steady state relationships.

In light of the above, it also becomes clear that the economic long-run relationship (17.48) represents no separate restriction on the long-run relationship between growth rates, as it is a relationship between the marginal means of the two real exchange rates. The rest of the model can be solved for the steady-state rate of productivity, and for the steady-state unemployment level using (17.58), (17.51) and (17.52). Finally, the GDP growth rate and rate of growth in credit then follow from (17.47) and (17.56).

The above steady-state properties are derived for an exogenous policy interest rate R. The nature of the solution is not changed if we instead specify either an estimated interest rate reaction function on the basis of the data, or a response function based on theories of optimal policy rules. But the behavior of the dynamics will be affected, and the policy analysis and the level of predicted long-run inflation will depend on how the interest rate is modeled.

We agree with Hendry and Mizon (2000) that there are reasons for being pragmatic about how the policy instrument is "treated" in a macroeconometric model. For some purposes it is relevant to treat the instrument as exogenous, like in the analysis of the steady state (this sub-section) and its stability (next sub-section). That analysis will answer whether there is a fundamental tendency of dynamic instability in the model that instrument use will have to counteract in order to avoid the economy taking an unstable course or, conversely, whether there is sufficient stability represented by the modeled relationships. In that case the challenges to instrument setting are more linked to timing and to meeting a specific inflation target than to "securing" overall nominal or real dynamic stability.

In the following sections we will use both the open and closed versions of the NAM model with respect to the policy instrument. In the next section, and the first section on policy use (section 17.4.2.1), we will use the open model with exogenous R_t. Later (e.g., when evaluating tracking performance in section 17.4.2.2), we will use a version with an econometrically modeled interest rate reaction function, which is documented in Table 17.1. In section 17.4.3, to discuss optimal policy we will use a version with a theoretically derived interest rate response function.

17.3.3 Stability of the steady state

In an important paper, Frisch (1936) anticipated the day when it would become common among economists to define (and measure) "normal" or natural values of economic variables by the values of the variables in a stationary state. The steady-state defined by the long-run model above corresponds to such natural values of the model's endogenous variables. But it is impractical to derive all the "natural values" using algebra, even in such a simple system as ours. Moreover, the question about the stability of the steady state, e.g., whether the steady state is globally asymptotically stable – or perhaps stability is only a saddle-path property – can only be addressed by numerical simulation of the dynamic system of equations.

We therefore follow Frisch's suggestion and simulate the dynamic NAM model. Figure 17.3 shows the "natural values" for inflation (Δp_t above), the rate of unemployment, wage inflation, GDP growth, growth in import prices and the rate of currency depreciation. The first solution period is 2007(4) and the last solution period is 2035(4). The simulation period is chosen to be so long because we want to get a clear impression about whether these variables approach constant ("natural") values or not, and whether the effect of initial conditions die out, as they should if the solution is globally asymptotically stable. The simulation is stochastic. The solid lines represent the mean of the 1,000 replications, and the 95% prediction intervals represented by the distance between the two dashed lines accommodate only residual uncertainty, not coefficient uncertainty.

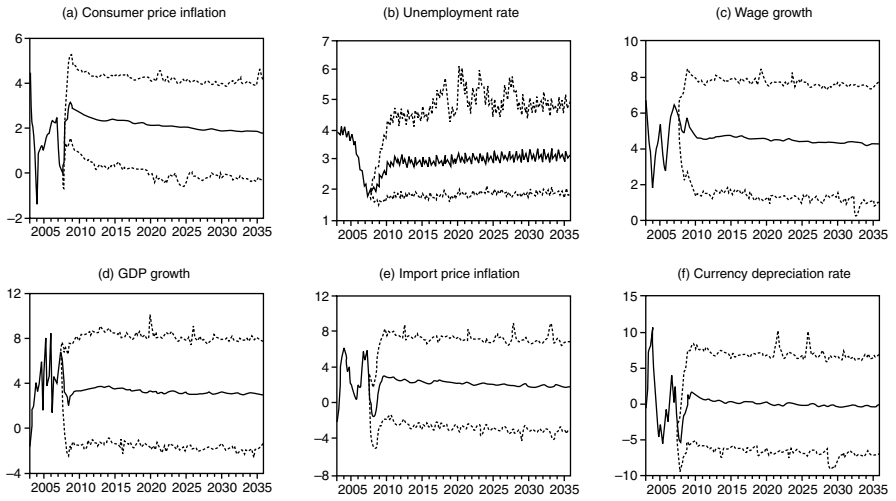


Figure 17.3 Stochastic dynamic simulation of NAM for the period 2007(4)–2035(4)
The money market interest rate is kept constant at the 2007(3) level for the length of the simulation period. The distance between the dotted lines indicates 95% prediction intervals.

In order to match the theoretical steady state above as closely as possible, the solution in Figure 17.3 is based on a constant domestic money market interest rate ($R_t = R_{2007(3)}$). The exogenous $I(1)$ variables have also been given constant growth rates for the length of the simulation period, e.g., foreign inflation is fixed at 2%. The impression is clearly that NAM has an asymptotically stable steady state, even under the assumption of a constant interest rate (the money market is in equilibrium by an endogenous money supply in the counterfactual "regime" defined by the constancy of the interest rate). The annual rate of inflation is seen to stabilize at a level just above 2.5%, and the natural rate of unemployment (in the Frisch sense) appears to be 3%. Clearly, in this scenario an inflation rate of 2.5% is attainable with very moderate instrument use. The seasonality of unemployment, modeled by dummies, is clearly visible and represents no problem in terms of stability.

17.4 Macroeconometric models as tools for policy analysis

In this last section we discuss several aspects of model usage that are of relevance for policy analysis. We begin with a practical problem, namely that congruent modeling, or a high degree of "data coherence" to cite the influential Pagan (2003) report on monetary policy modeling, gives rise to dynamic specifications that are too complicated to be of any help when the task is to explain the basic policy channels and lags between instruments and target. Nevertheless, section 17.4.1 shows that the need for simplicity in communication is not an argument for compromising empirical validity at the modeling stage, since it is always possible and convenient to work from the empirically valid and complicated model to a simple and stylized model that contains the essential dynamics of the full model. In

the other sub-sections, we use dynamic simulation, the main tool of model usage, to elucidate the strength of the policy instrument, to check that the model solution generates the properties of the actual data, and whether simulation over a long forecast horizon gives the steady state we expect from the theoretical analysis above. We also discuss optimal policy response, forecast properties, and strategies to reduce the damage that structural breaks have on forecasts from equilibrium correction models. Finally in this section, we raise the more fundamental question of testing non-nested hypotheses about the supply-side of the economy; our example will be the New Keynesian Phillips curve against the model of wage and price adjustment that has been presented above.

17.4.1 Tractability: stylized representations

There is a marked difference between the intricate and complex dynamics often found in empirical models – at least if they model the data – and the very simplified dynamics typically found in theoretical models. The purpose of this section is to enhance the understanding of the properties of a model through the use of stylized representations. A dynamic model, e.g.:

$$\begin{aligned}\Delta y_t = {} & 2 - 0.4\Delta y_{t-1} - 0.6\Delta y_{t-2} \\ & + 0.2\Delta x_t - 0.5\Delta x_{t-1} + 3\Delta x_{t-2} - 1\Delta x_{t-3} - 0.5\left(y_{t-3} - 4x_{t-4}\right) + v_t,\end{aligned}$$

can be approximated by a stylized model with simplified dynamics, in this example:

$$\Delta y_t = 1 + 0.85\Delta x_t - 0.25\left(y - 4x\right)_{t-1}.$$

This is achieved by using the mean of the dynamics of the variables.[14]

In the same way as above, we let lower cases of the variables denote natural logarithms, so $\Delta z_t \approx \frac{Z_t - Z_{t-1}}{Z_{t-1}} = g_{Z_t}$. If we assume that, on average, the growth rates are constant – the variables could be "random walks with drift" – the expected values of the growth rates are constants:

$$\begin{aligned}E\Delta y_t &= g_y \forall t \\ E\Delta x_t &= g_x \forall t.\end{aligned}$$

If the variables also are cointegrated, the expectation of the linear combination in the equilibrium correction term is also constant, so:

$$E\left(y_{t-3} - 4x_{t-4}\right) = E\left(y_{t-1} - 4x_{t-1}\right) = \mu \forall t.$$

Under these assumptions, the mean dynamics of the model becomes:

$$\begin{aligned}E\Delta y_t = {} & 2 - 0.4E\Delta y_{t-1} - 0.6E\Delta y_{t-2} \\ & + 0.2E\Delta x_t - 0.5E\Delta x_{t-1} + 3E\Delta x_{t-2} - 1E\Delta x_{t-3} \\ & - 0.5E\left(y_{t-3} - 4x_{t-4}\right) + Ev_t \\ g_y\left(1 + 0.4 + 0.6\right) = {} & 2 + \left(0.2 - 0.5 + 3 - 1\right)g_x - 0.5\mu\end{aligned}$$

$$g_y = \frac{2}{2} + \frac{1.7}{2} g_x - \frac{0.5}{2}\mu$$
$$g_y = 1 + 0.85 g_x - 0.25\mu.$$

We can therefore write the mean-approximated, or stylized, dynamic model as:

$$\Delta y_t = 1 + 0.85\Delta x_t - 0.25\left(y - 4x\right)_{t-1}.$$

To illustrate, the dynamic behavior of the model and its mean approximation are shown in Figure 17.4. The upper panel shows the dynamic, or period, responses in y_t to a unit change in x_{t-i}. The lower panel shows the cumulative, or interim, response. The graphs illustrate how the cyclical behavior – due to complex roots – is averaged out in the stylized representation.

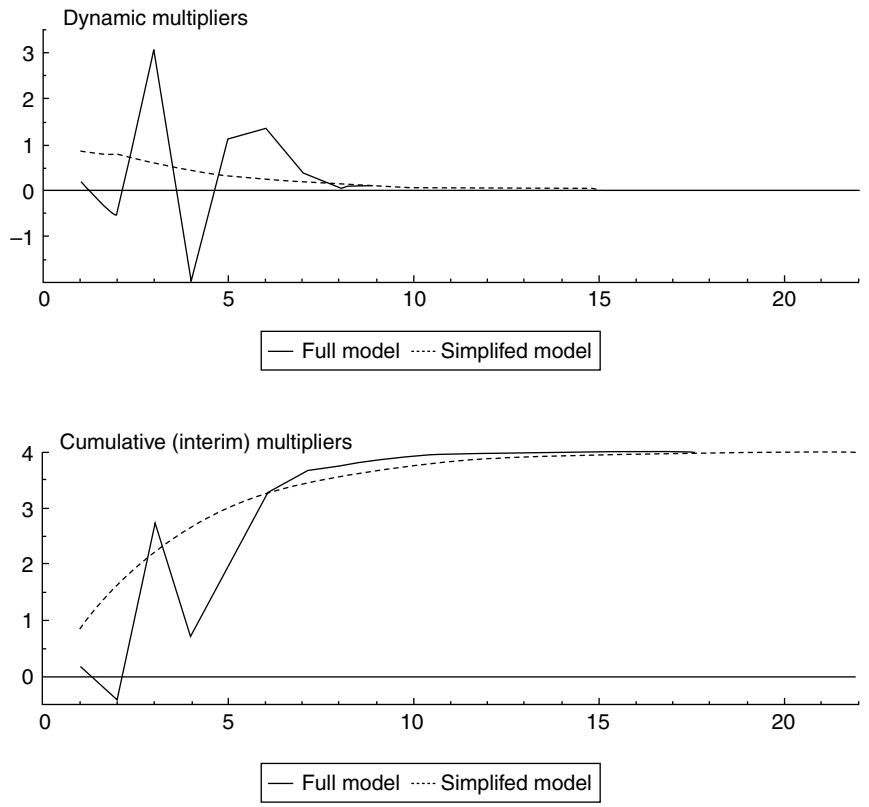


Figure 17.4 The dynamic responses of the example model and its mean approximation

Note that all that is done is to exploit so-called growth coefficients (see Patterson and Ryding, 1984; Patterson, 1987). The steady-state growth:

$$g_y = 4g_x$$

implies the steady-state mean μ:

$$(4 - 0.85)\, g_x = 1 - 0.25\mu$$
$$\mu = 4 - 12.6 g_x,$$

so the steady-state relationship between the variables is:

$$y_t = (4 - 12.6 g_x) + 4 x_t,$$

which will hold both for the complete as well as the stylized dynamic representation of the model.

Using this approach, the full econometric model reported in section 17.3.1 can be given the following stylized representation:

$$\Delta v_t = -0.04\Delta(R - R^*)_t - 0.04\{\left(v + p^* - p\right) - 0.12 \times \left[(R - \pi) - \left(R^* - \pi^*\right)\right]\}_{t-1} \tag{17.60}$$

$$\Delta\left(pi - pi^* - v\right)_t = -0.1\Delta v_t - 0.43\left[\left(pi - pi^* - v\right) - 0.55\left(p - p^* - v\right)\right]_{t-1} \tag{17.61}$$

$$\Delta p_t = -0.09\Delta z_t + 0.03\Delta pi_t + 0.08\Delta pe_t + 0.06\Delta y_t - 0.07\left[p - 0.7(w - z) - 0.3pi\right]_{t-1} \tag{17.62}$$

$$\Delta\left(w - p\right)_t = -0.04\Delta u_t + 0.73\Delta T1_t - 0.07\left[(w - p - z) + 0.1u\right]_{t-1} \tag{17.63}$$

$$\Delta z_t = 0.09\Delta\left(w - p\right)_t - 0.24\left[z - 0.47\left(w - p\right) - 0.003Trend - 0.03u\right]_{t-1} \tag{17.64}$$

$$\Delta u_t = -0.23\left\{u - 7.65\Delta\left(w - p\right) - 4.46\left[0.01\left(R_L - \pi\right) - 4\Delta y\right]\right\}_{t-1} \tag{17.65}$$

$$\Delta R_{L,t} = 0.58\Delta R_t - 0.33\left(R_L - 0.41R_B - 0.76R\right)_{t-1} \tag{17.66}$$

$$\Delta\left(R_B - R_B^*\right)_t = 0.43\Delta R_t - 0.17\left(R_B - 0.43R - 0.57R_B^*\right)_{t-1} \tag{17.67}$$

$$\Delta y_t = 0.16\Delta g_t + 0.38\Delta\left(l - p\right)_t - 0.12\left[y - 0.9g_{t-1} - 0.16(v + p^* - p) + 0.06\left(R_L - \pi\right)\right]_{t-1} \tag{17.68}$$

$$\Delta\left(l - p\right)_t = 0.3\Delta y_t - 0.09\left[\left(l - p\right) - 2.65y + 0.04(R_{L,} - R_B)\right]_{t-1}, \tag{17.69}$$

where the constants are omitted for ease of exposition. This representation reproduces the same steady state as the full model, but with stylized dynamics. The averaged transmission mechanisms can be traced through the interrelationships of the mean dynamic effects of shocks to the model – in contrast to the steady-state effects described above (see the discussion in section 17.3.1 for an example).

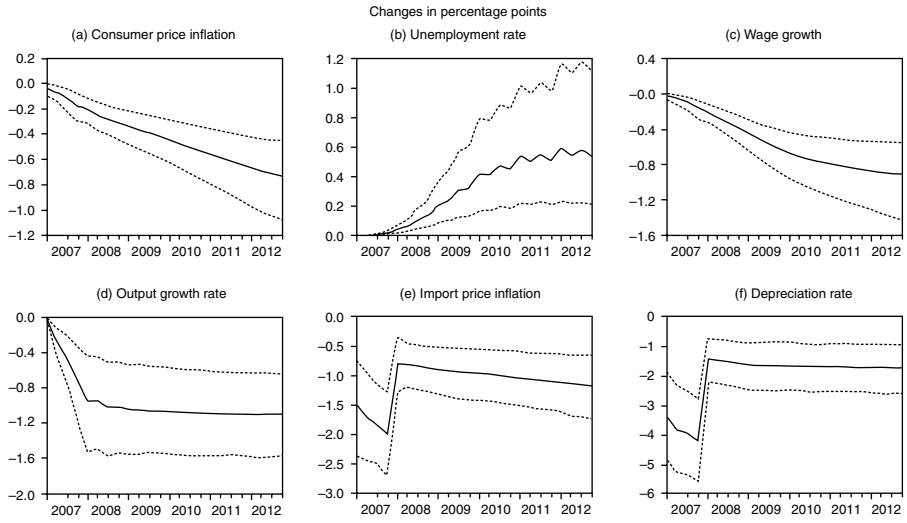


Figure 17.5 Dynamic multipliers from a permanent increase in the money market interest rate of 100 basis points, 2007(1)–2012(4)
The multipliers are shown as deviations from the baseline simulation in Figure 17.3. The distance between the two dotted lines represents the 95% confidence intervals.

17.4.2 Shock analysis: dynamic simulations

It seems obvious that a model for policy analysis should be empirically adequate to a high degree, and that care must be taken not to compromise this property for other desirable properties. However, in the same way as stability and invariance, "data fit" is a relative concept. Models that are outperformed in terms of fit might still be useful for policy analyses because they are highly relevant for the purpose, as noted by, e.g., Pesaran and Smith (1985). This sub-section will discuss relative model adequacy – also compared to optimal monetary policy.

17.4.2.1 How strong is the policy instrument?

As a background to policy analysis it is important to obtain a quantitative view of the transmission mechanisms, specifically to see whether the policy instrument has a sizeable effect on the variables that are subject to central bank targeting (in formal or more informal ways). The open version of the model, with exogenous interest rate (R_t), is then relevant, since among the parameters of the model are the dynamic multipliers of the endogenous variables with respect to the policy instrument.

Since the main channel of interest rate transmission is through the exchange rate, output and the level of unemployment, the interest rate is actually quite effective in counteracting demand shocks. However, shocks on the supply side, e.g., in wage-setting (such as permanently increased wage claims), or in foreign inflation, can be more difficult to curb by anything but huge increases in the interest rate.

Hence, stabilization of inflation after a supply shock may represent a formidable challenge to monetary policy.

The recent monetary history of Norway has demonstrated the relevance of this analysis: since mid 2003, core inflation rate has been consistently below the target of 2.5%, for long periods less than 1%, despite determined interest rate cuts in 2003, and no interest rate increases before July 2005, and then only gradually and in small steps.

The interest rate multiplier on inflation is shown in panel (a) of Figure 17.5. Panel (f) shows that the effect first goes through the *exchange rate channel*: the impact multiplier is a 3% appreciation, and in steady state there is an appreciation of 1.2%, thus illustrating equation (17.57). Panel (e) shows that import price growth is affected in the same way as the exchange rate (see equation (17.59)), but the impact effect is smaller. Panels (b), (c) and (d) show the *labor market* and *demand channels*. Wage growth is reduced, as predicted by (17.58), leaving real wage growth unaffected in steady state, as this is determined by productivity growth. Unemployment increases, since real interest rates increase, while negative GDP growth follows from increasing real interest rates and appreciation of the real exchange rate.

17.4.2.2 Fitting the facts

In this section we document how well the econometric model NAM explains the evolution of important endogenous variables over the 17-quarter period 2003(3)–2007(3).

The solutions are conditional upon the actual values of the non-modeled variables.[15] Experience has shown that particularly important explanatory variables are foreign interest rates (R^* and R_B^*) and consumer (p^*) and producer prices (pi^*). Domestic government expenditure (g) and electricity prices (pe) are also very important for the overall fit of the model. Finally, the oil price (in \$US) is a highly relevant explanatory variable, mainly through the market for foreign exchange.

The interest rate has been the instrument of monetary policy during the solution period. We therefore use the (more) closed version of the model, where the domestic money market interest rate is estimated as a function of the core inflation gap and the unemployment rate gap in equation (17.46).

Figure 17.6 shows that NAM generates the features of the macroeconomic development. Inflation in panel (a), the rate of unemployment in panel (b), GDP growth in panel (d), and the money market interest rate in panel (g) are very well explained by the model solutions. The graph of actual and simulated nominal currency depreciation in panel (i) shows that movements in the exchange rate are also well explained, brought about by the interest rate differential and equilibrium correction with respect to the real exchange rate – which is shown in panel (f).

17.4.2.3 Shock analysis with dynamic multipliers

Next, we investigate how the economy, according to the model, is likely to respond to shocks. We use the same model version as in section 17.4.2.2. Amongst the many shocks of interest to a small open economy, we here consider a negative foreign price shock. The deflationary 5% shock occurs in 2007(1), and Figure 17.7, panels

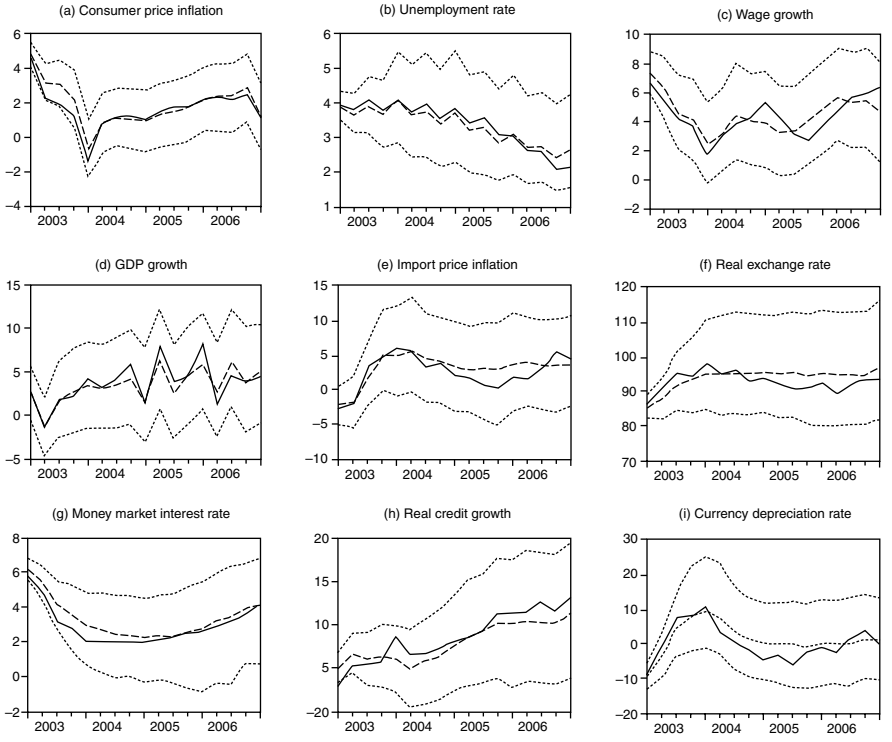


Figure 17.6 Dynamic simulation of NAM 2003(1)–2007(1)
Actual values are indicated by solid lines, and model solution by dashed lines. The distance between the two dotted lines represents 95% prediction intervals. The units on all the vertical axes are percentage points, and time is along the horizontal axes.

(a)–(f), reports the dynamic multipliers for inflation, wage increases, the rate of unemployment, GDP growth, the increases in the import price index, and the rate of currency depreciation. Since all variables in the graph are measured in percentage points, the multipliers shown are absolute deviations from the baseline. In the same way as in Figure 17.5, parameter uncertainties are indicated by the dotted lines, representing 95% confidence intervals.

Since it is a temporary shock to foreign inflation, there is no reason to expect a permanently lower rate of domestic inflation. Panels (a) and (c) of Figure 17.7 confirm that presumption, but also show that the initial multipliers of price and wage inflation are negative and significant. Part of the adjustment to a lower nominal path of the economy involves a higher unemployment rate and lower GDP growth (cf. panels (b) and (d)), which are explained by the initial appreciation of the real exchange rate. The nominal depreciation in panel (f), due to a lower domestic interest rate (not shown), is not enough to offset the loss of competitiveness, initiated by the deflationary price shock.

The nominal rigidities in the model, which transform the nominal shock into a change in the real exchange rate, are an important property in explaining the

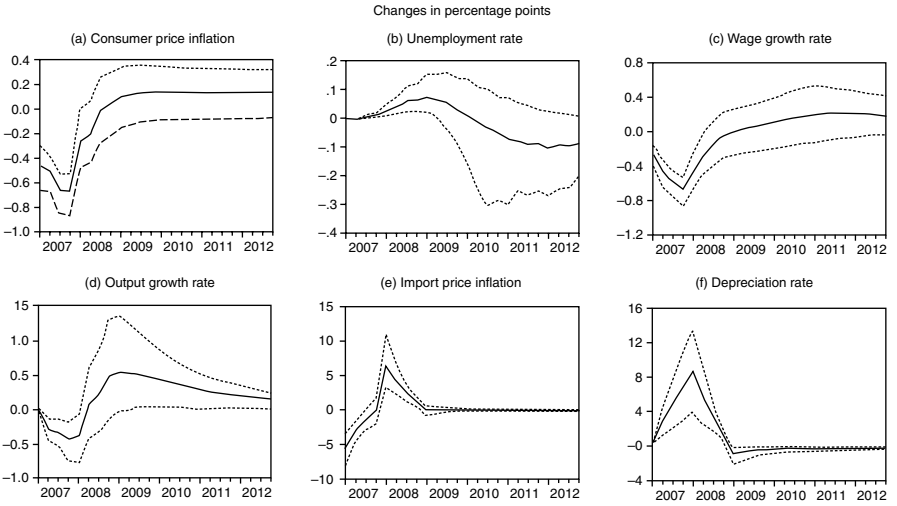


Figure 17.7 Dynamic multipliers from a permanent reduction in exogenous (foreign) prices by 5%, 2007(1)–2012(4)
The distance between the two dotted lines represent the 95% confidence intervals. The units on the vertical axes are percentage points in all panels.

multipliers. The rigidities are not due to "incredibly long" adjustment lags in price- and wage-setting: domestic inflation is seen to return to its baseline path after two and a half years. Instead, there is a second wave of effects due to the interaction between product and labor markets. Hence the experiment demonstrates rather well the important theoretical point made in section 17.2.6.4, namely that the steady-state inflation rate does not imply a unique equilibrium rate of unemployment, since the rate of inflation reaches its steady state long before the multipliers of the rate of unemployment have died out.

17.4.3 Aspects of optimal policy: the impact of model specification on optimal monetary policy

As noted above, the version of NAM with an econometrically modeled interest rate makes no claims of representing optimal interest rate-setting under inflation targeting or optimal policy response to a shock. Instead, the multipliers of the last paragraph should be interpreted as counterfactuals: they show the response that would occur if the interest rate reacted *as if* it followed the econometrically estimated interest rate equation.

However, in a recent paper, Akram and Nymoen (2008) show how optimal monetary policy can be implemented in NAM, and how the predicted economic outcome depends on the specification of the supply side of the model. For that purpose, they replace the econometrically modeled interest rate equation with a theoretically derived interest rate rule due to Akram (2007):

$$R_{t+m} = R_0 + (1 - \varrho_H)\frac{\beta_\varepsilon}{(1-\phi)}\varepsilon_t + \varrho_H(R_{t+m-1} - R_0),\ m = 0, 1, 2, \ldots, H. \quad (17.70)$$

This rule defines an interest rate path corresponding to a specific policy horizon H. The response coefficient $\beta_{\varepsilon,H} \equiv (1-\varrho_H)\frac{\beta_\varepsilon}{(1-\phi)}$ determines by how much the interest rate must deviate initially from the neutral rate R_0 to counteract the inflationary effects of a shock ε_t. A high value of the smoothing parameter ϱ_H can be associated with a strategy of gradualism in interest rate-setting. Thus two preference parameters $\beta_{\varepsilon,H}$ and ϱ_H depend on the policy horizon, H. The last parameter in the theoretical interest rate equation is ϕ, which represents the (objective) degree of persistence in the inflation shock.

In addition to the preferences about policy horizon and gradualism captured by (17.70), the optimal interest rate response is influenced by the user's choice of macroeconomic model. Akram and Nymoen (2008) focus on the labor market channel, since this part of the transmission mechanism has been the focus of most model controversy. In one model version, used to derive optimal policy, the incumbent model of the supply side, with equilibrium correction in both wage- and price-setting, is used. As explained in section 17.2.6, this model is called the incomplete model of wage- and price-setting and is referred to as ICM in Figure 17.8. The two other versions are the model with wage and price Phillips curves, PCM in the figure, and a version with a vertical Phillips curve, PCMr in the figure. As noted in section 17.2.6.4, among these three specifications, it is the PCMr that represents the consensus view in modern monetary economics.

Akram and Nymoen show that econometric encompassing tests favor the ICM model of the supply side, but also that the PCM and the PCMr appear to be well specified on their own terms, e.g., the residual properties of the two Phillips curve models do not signal any problems. Hence there is a question about whether the outcome of the encompassing test has any practical (or "economic") significance, or whether this test of model adequacy has only an academic interest. The analysis suggests that the econometric test result contains valuable information.

Consider, e.g., Figure 17.8, which presents the economic performance of (optimal and sub-optimal) policies employed in response to the supply shock. The left column of the figure shows that there is a trade-off between price and output stabilization for different ranges of policy horizons. Specifically, in the case of ICM and PCM there is a trade-off in the range of 0–8 quarters. Policy horizons that are longer than 8 quarters appear inefficient as both price and output stabilization can be improved by shortening the policy horizon. The opposite is the case for PCMr. In this case, the trade-off curve is associated with policy horizons that are longer than 6 quarters, while policy horizons shorter than 6 seem inefficient. In the right-hand column of the figure, it transpires that the three models recommend substantially different policy horizons. Even though the efficiency frontiers for ICM and PCM are defined by almost the same policy horizon, the optimal horizon is 3 quarters conditional on ICM, but 6 quarters in the case of PCM. In the case of PCMr the policy horizon is 11 quarters.

Based on the above and several other simulation experiments, Akram and Nymoen (2008) find that econometric differences bear heavily on (model-based) policy recommendations and are thus not merely of academic interest. Their analysis quite strongly suggests that differences in model specifications and even in

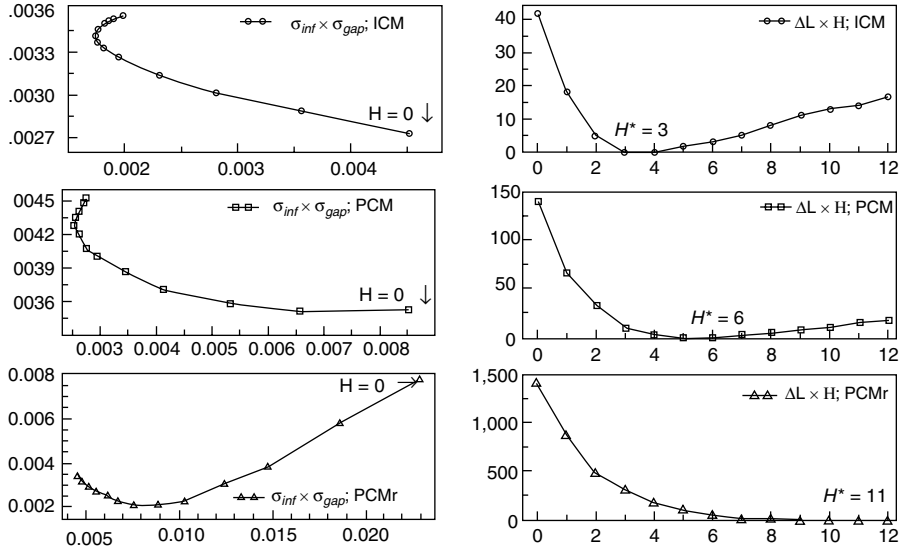


Figure 17.8 Economic performance and optimal policy suggested by three specifications of the supply side in the face of a supply shock

Left column: trade-offs between standard deviations of inflation gap, σ_{inf}, and output gap, σ_{gap} (horizontal axis) associated with different (policy) horizon-specific rules in response to the demand shock. The trade-offs are plotted for rules associated with policy horizons (H) in the range of 0–12 quarters. The trade-offs associated with different horizons follow each other, where the one for $H = 0$ is indicated. *Right column*: values of the relative loss function, denoted ΔL (in %), at the different policy horizons (horizontal axis) (Akram and Nymoen, 2008).

parameter values across models can lead to widely different policy implications. Interestingly, it appears that imposing a set of parameter restrictions may have stronger influence on policy implications than choosing a different functional form of the model. Monetary policy based on a model that turns out to be an invalid characterization of the economy and its transmission mechanism may lead to substantial losses in terms of economic performance, even when policy is guided by gradualism, e.g., in the form of a long policy horizon.

17.4.4 Theory evaluation: the New Keynesian Phillips curve

Macroeconomics is an evolving science. New hypotheses and theories are put forward, sometimes with far-reaching consequences, for policy, teaching and, of course, also for model-building. A macroeconomic model project that lasts for some time is therefore bound to face the need to adapt to new theoretical developments. However, new ideas in macroeconomics are usually partial, and can claim superiority with respect to existing ones only by replacing old *ceteris paribus* clauses by new ones. In a macroeconometric modeling context it therefore makes sense to test the new theories before they are implemented. If the model is in operational use, say for policy recommendations, this step may be as much of a virtue

as a necessity, since unverified changes in policy response and forecasting may critically damage beliefs in the relevance of model analysis.

In this section we discuss some approaches to how the model of the supply side presented in section 17.2.6 can be tested against a new and important development represented by the New Keynesian Phillips curve.

As shown by Bårdsen *et al.* (2005, Chs. 4–6), three important and much used models of inflation and unemployment are consistent with the view that wages and prices are equilibrium correcting $I(1)$ variables, while the rate of unemployment is $I(0)$. They are the incomplete competition model in equilibrium correction form, the standard open economy wage Phillips curve, and the wage Phillips curve with homogeneity restrictions; in other words, ICM, PCM and PCMr of section 17.4.3. *Qua* equilibrium correction models, these theories can readily be identified as special cases of a VAR.

Cointegration is thus a common feature of the three models. They can be identified by their different theories about the main adjustment mechanisms at work. In the case of the ICM, wages equilibrium correct with respect to deviations from the wage level predicted by the theoretical bargaining model. In the case of the PCM, by definition, wages do not equilibrium correct with respect to lagged wages, so in this model equilibrium correction has to be *indirect* and via the reaction of the rate of unemployment (see Bårdsen and Nymoen, 2008). In section 17.4.3 these properties were shown to be relevant for the assessment of different supply-side models for policy analysis.

We will show below that the same insight also applies to the New Keynesian Phillips curve, meaning that its equilibrium correction implications can be tested against the ICM or (any version) of the conventional Phillips curves. However, we first give a brief summary of the current empirical status of the New Keynesian Phillips curve.

17.4.4.1 The New Keynesian Phillips curve

The New Keynesian Phillips curve, hereafter NKPC, has become regarded by many as the new standard model of the supply side in the macro-models used for monetary policy analysis. This position is due to its theoretical underpinnings, laid out in Clarida, Gali and Gertler (1999), and to the supportive empirical results in the studies of Gali and Gertler (1999) (henceforth GG) on US data, and Gali, Gertler and López-Salido (henceforth GGL) (2001) on euro-area data.

The hybrid NKPC is given as:

$$\pi_t = \underset{\geq 0}{a^f} E_t[\pi_{t+1}] + \underset{\geq 0}{a^b}\pi_{t-1} + \underset{\geq 0}{b}\, s_t, \tag{17.71}$$

where π_t is the rate of inflation, $E_t\left[\pi_{t+1}\right]$ is expected inflation one period ahead and s_t is a time series of firms' real marginal costs. The "pure" NKPC is (17.71) with $a^b = 0$, and represents the case where all firms (that are aggregated over) form rational expectations. Both the pure and hybrid forms are usually presented as "exact," i.e., without an error term. When $E_t[\pi_{t+1}]$ is replaced by π_{t+1} for estimation, a moving average error term is implied. This has motivated "robust" estimation with,

e.g., pre-whitening switched on in generalized method of moments (GMM), and leading papers downplay the relevance of congruency for the evaluation of the NKPC.

After assessing some of the critiques that have been directed towards the NKPC, Gali, Gertler and López-Salido (2005) assert that the NKPC, in particular the dominance of forward-looking behavior, is robust to the choice of estimation procedure and to possible specification bias. They conclude that the following three results are proven characteristics of NKPC for all datasets:

1. The two null hypotheses $a^f = 0$ and $a^b = 0$ are rejected both individually and jointly.
2. The coefficient on expected inflation exceeds the coefficient on lagged inflation substantially. The hypothesis of $a^f + a^b = 1$ is typically not rejected at conventional levels of significance, although the estimated sum is usually a little less than one numerically.
3. When real marginal costs are proxied by the log of the wage share, the coefficient b is positive and significantly different from zero at conventional levels of significance.

As mentioned, critics of the NKPC have challenged the robustness of all three, but with different emphases and from different perspectives. The inference procedures and estimation techniques used by GG and GGL (2001) have been criticized by Rudd and Whelan (2005, 2007) and others, but GGL (2005) show that their initial Results 1 and 2 remain robust to these objections.

When it comes to Result 3, GGL (2005) overlook that several researchers have been unable to confirm their view that the wage share is a robust explanatory variable in the NKPC. Bårdsen, Jansen and Nymoen (2004) showed that the significance of the wage share in the GGL (2001) model is fragile, as it depends on the exact implementation of the GMM estimation method used, thus refuting that Result 3 is a robust feature of NKPC estimated on euro-area data.

Fanelli (2008), using a vector autoregressive regression model on the euro-area dataset, finds that the NKPC is a poor explanatory model. On US data, Mavroeidis (2006) has shown that real marginal costs appears to be an irrelevant determinant of inflation, confirming the view in Fuhrer (2006) about the difficulty of obtaining a sizeable coefficient on the forcing variable in the US NKPC. Already the studies cited represent evidence that refutes the claim that Result 3 is robust. Instead it is to be expected that, depending on the operational definition of real marginal costs, the estimation method and the sample, the numerical and statistical significance of b will vary across different studies.

Of course, Result 3 is just as important as Results 1 and 2 for the status of the NKPC as an adequate model, so if that part of the model is non-structural, it might be that that Results 1 and 2 have another explanation than the intended, which is that there is a good match between the NKPC and the true inflation process. Bårdsen, Jansen and Nymoen (2004) (euro-area) and Bjørnstad and Nymoen (2008)

(OECD panel data) demonstrated that the significance of a^f can be explained by a linear combination of better forcing variables which reside among the overidentifying instruments. Their presence is revealed by the significance of the Sargan (1964) specification test. Importantly, the re-specified models in the two studies lend themselves directly to interpretation either as conventional Phillips curves, or as an equilibrium-correction price equation consistent with the theory of monopolistic competition in the product market and a certain element of coordination in wage-bargaining (see Sargan, 1980; Nymoen, 1991; Bårdsen *et al.*, 2005, Chs. 4–6). Hence, the NKPC fails to parsimoniously encompass these models.

17.4.4.2 The equilibrium correction implications of the NKPC

The original NKPC makes no reference to open economy issues. Batini, Jackson and Nickell (2005) have shown that the main theoretical content of the NKPC generalizes, but that consistent estimation of the parameters a^f, a^b and b requires that the model is augmented by variables which explain inflation in the open economy case. Hence, the open economy NKPC is:

$$\Delta p_t = \underset{\geq 0}{a^f}\,\Delta p^e_{t+1} + \underset{\geq 0}{a^b}\,\Delta p_{t-1} + \underset{\geq 0}{b}\,s_t + c\,x_t, \tag{17.72}$$

where x_t, in most cases a vector, contains the open-economy variables, and c denotes the corresponding coefficient vector. The change in the real import price, $\Delta(pi_t - p_t)$ in our notation, is the single most important open economy augmentation of the NKPC. The results in Batini, Jackson and Nickell are, broadly speaking, in line with GG's and GGL's Results 1–3 above, but, as noted, those properties are not robust when tested against the existing UK model in Bårdsen, Fisher and Nymoen (1998).

We follow GG and measure s_t by the log of the labor share:

$$s_t = ulc_t - q_t, \tag{17.73}$$

where *ulc* denotes unit labor costs (in logs) and q is the log of the price level on domestic goods and services, compare section 17.2.6. Next, define the aggregate price level as:

$$p_t = \zeta q_t + (1-\zeta)\,pi_t, \tag{17.74}$$

with $(1-\zeta)$ as the import share. If we solve for q_t, insert in (17.73) and rewrite, we obtain the following equation for the wage share:

$$s_t = -\frac{1}{\gamma}\left[p_{t-1} - \gamma\; ulc_{t-1} - (1-\gamma)\,pi_{t-1}\right] + \Delta ulc_t - \frac{1}{\gamma}\Delta p_t + \frac{1-\gamma}{\gamma}\Delta pi_t. \tag{17.75}$$

We can then rewrite the open economy NKPC as:

$$\begin{aligned}\Delta p_t = {} & \frac{a^f}{\left(1+\frac{b}{\gamma}\right)}\Delta p^e_{t+1} + \frac{a^b}{\left(1+\frac{b}{\gamma}\right)}\Delta p_{t-1} - \frac{b}{(\gamma+b)}\left[p_{t-1} - \gamma\; ulc_{t-1} - (1-\gamma)\,pi_{t-1}\right] \\ & + \frac{\gamma\; b}{(\gamma+b)}\Delta ulc_t + \frac{b\,(1-\gamma)}{(\gamma+b)}\Delta pi_t + \frac{\gamma\; c}{(\gamma+b)}\,x_t,\end{aligned}$$

or:

$$\Delta p_t = \alpha^f \Delta p^e_{t+1} + \alpha^b \Delta p_{t-1} + \beta(ulc_{t-1} - p_{t-1}) - \beta(1-\gamma)(ulc_{t-1} - pi_{t-1}) \quad (17.76)$$
$$+ \beta\gamma \Delta ulc_t + \beta(1-\gamma)\Delta pi_t + \psi x_t,$$

where we have defined α^f, α^b, β and ψ as new coefficients for simplification.

Equation (17.76) brings out that the NKPC implies an equilibrium-correction price equation which is very similar to the "incumbent" model in NAM (cf. equation (17.30) in section 17.2.6). However, there are two notable differences. First and foremost, the forward-looking term Δp^e_{t+1} with an expected high coefficient α^f. The incumbent model implicitly restricts this coefficient to zero. Second, in the NKPC, there are parameter restrictions on the coefficients of the following variables: Δulc_t, Δpi_t , $(ulc_{t-1} - p_{t-1})$ and $(ulc_{t-1} - pi_{t-1})$; in fact, they are functions of the underlying parameters β and γ.

It follows that, for the purpose of testing the NKPC, we can start with an equilibrium correction model with a lead in the inflation term:

$$\Delta p_t = \alpha^f \Delta p^e_{t+1} + \alpha^b \Delta p_{t-1} + \beta_1(ulc_{t-1} - p_{t-1}) + \beta_2(ulc_{t-1} - pi_{t-1}) \quad (17.77)$$
$$+ \beta_3 \Delta ulc_t + \beta_4 \Delta pi_t + \psi x_t,$$

and test the following hypotheses: H^a_0: $\alpha^f = 0$, H^b_0: $\beta_3 = \beta_1 + \beta_2$ and H^c_0: $\beta_4 = -\beta_2$. Rejection of H^a_0 together with non-rejection of H^b_0 and H^c_0 constitute evidence that support the NKPC, while non-rejection of H^a_0 is telling evidence against the NKPC.

As noted above, NKPC models are usually specified with the rate of change in the real import price as one of the elements in x_t. Equation (17.77) is consistent with that interpretation, the only caveat applies to β_4 and H^b_0, since $\beta_4 = -\beta_2$ no longer follows logically from the NKPC. This is because β_4 is a composite parameter also when the NKPC is the valid model.

17.4.4.3 Testing the equilibrium-correction implications of the NKPC

Consider the hybrid NKPC of the form (17.72) where we allow for two lags of inflation as well as three deterministic seasonals. This is because NAM makes use of seasonally unadjusted quarterly data. The wage share variable s_t is treated as an endogenous variable, but we also include electricity prices (Δpe_t) and import prices (Δpi_t) in the x_t vector of exogenous variables. As already noted, inflation is measured by the consumer price index. Instrumental variable (IV) estimation gives:

$$\begin{aligned}\Delta p_t \;=\;& \underset{(0.107)}{0.3659}\ \Delta p_{t+1} + \underset{(0.0228)}{0.04122}\ s_t + \underset{(0.0772)}{0.08759}\ \Delta p_{t-1} \\ &+ \underset{(0.0653)}{0.2676}\ \Delta p_{t-2} + \underset{(0.00597)}{0.06385}\ \Delta pe_t + \underset{(0.0169)}{0.04024}\ \Delta pi_t \\ &+ \textit{constant and seasonals} \\ &\text{IV, } T = 111 \text{ (1979(3) - 2007(1))} \\ &\chi^2_S(9) = 11.664[0.2329].\end{aligned} \quad (17.78)$$

The results shows a significant coefficient on the forward term of the same magnitude as the sum of the coefficients on the two lagged inflation rates. The wage

share s_t has the correct positive sign and is significant at the 10% level. By and large, these results are in line with the typical NKPC Results 1–3 discussed above.

The instruments include $ulc_{t-1} - p_{t-1}$ and $ulc_{t-1} - pi_{t-1}$, as explained above, together with lags of productivity growth, lagged electricity price growth, the same wage dummy as in NAM and lagged unemployment. The Sargan (1964) test of instrument validity ($\chi^2_S(n)$) is insignificant. However, residual misspecification tests reveal that (17.78) is not a congruent model, since there is substantial heteroskedasticity, autocorrelation and non-normality. In order to obtain a more congruent NKPC model we may use the dummy saturation technique in Autometrics (see Doornik, 2008).

$$\begin{aligned}\Delta p_t = {} & \underset{(0.102)}{0.622}\ \Delta p_{t+1} + \underset{(0.013)}{0.02509}\ s_t + \underset{(0.0533)}{0.2614}\ \Delta p_{t-2} \\ & + \underset{(0.00675)}{0.0549}\ \Delta pe_t + \textit{constant and 11 dummies} \\ & \text{IV, } T = 111 \text{ (1979(3) - 2007(1))} \\ & \chi^2_s(15) = 19.442[0.1944].\end{aligned} \tag{17.79}$$

The 11 dummies include both seasonals and break dummies, showing that the NKPC equation is not a completely time-invariant structural model. However, abstracting from that problem, equation (17.79) is seen to adhere even closer to the stylized facts of the NKPC than equation (17.78).

As explained above, we want to test the hypothesis that (17.79) encompasses the price equation of the incumbent model. To do that, we first include $ulc_{t-1} - p_{t-1}$ and $ulc_{t-1} - pi_{t-1}$ as explanatory variables to obtain an empirical version of the embedding equation (17.77) above. We then do a general-to-specific search by means of Autometrics in PcGive. The preferred model is reported as equation (17.80):

$$\begin{aligned}\Delta p_t = {} & \underset{(0.069)}{0.003849}\ \Delta p_{t+1} + \underset{(0.0259)}{0.08832}\ \Delta ulc_t + \underset{(0.0577)}{0.1515}\ \Delta p_{t-2} \\ & + \underset{(0.0143)}{0.09883}\ (ulc_{t-1} - p_{t-1}) - \underset{(0.00478)}{0.01943}\ (ulc_{t-1} - pi_{t-1}) + \underset{(0.00518)}{0.05164}\ \Delta pe_t \\ & + \textit{constant and 4 dummies} \\ & \text{IV, } T = 111 \text{ (1979(3) – 2007(1))} \\ & \chi^2_S(16) = 25.434[0.0625].\end{aligned} \tag{17.80}$$

It is seen that the estimated forward coefficient $\hat{\alpha}^f$ is practically zero in (17.80) compared to 0.62 in the NKPC in (17.79). Hence $H^c_0\colon \alpha^f = 0$ cannot be rejected statistically at any meaningful level of significance.

17.4.5 Forecasting for monetary policy

A hallmark of modern and flexible inflation targeting is that the operational target variable is the forecasted rate of inflation (see Svensson, 1997). One argument

for this choice of target is to be ahead of events, rather than to react after actual inflation has deviated from target. In this way one may hope to achieve the target by a minimum of cost to the real economy in terms of, e.g., unwanted output fluctuations or large fluctuations in the exchange rate. However, any inflation forecast is uncertain and might induce the wrong use of policy. Hence, a broad set of issues related to inflation forecasting is of interest for those concerned with the operation and assessment of monetary policy.

A favorable starting point for inflation targeting is when it can be asserted that the central bank's forecasting model is a good approximation to the inflation process in the economy. In this case, forecast uncertainty can be represented by conventional forecast confidence intervals, or by the fan charts used by today's best practice inflation targeters.[16] The point of the probabilistic forecasts is to convey to the public that the forecasted inflation numbers will only coincide with the actual future rate of inflation on average, and that neighbouring inflation rates are almost as probable. By the same token, conditional on the forecasting model's representation of uncertainty, still other inflation rates are seen to be wholly improbable realizations of the future.

However, the idea about model correctness and stationarity of macroeconomic processes is challenged by the high incidence of failures in economic forecasting (see, e.g., Hendry, 2001a). A characteristic of a forecast failure is that forecast errors turn out to be larger, and more systematic, than what is allowed if the model is correct in the first place. In other words, realizations which the forecasts depict as highly unlikely (e.g., outside the confidence interval computed from the uncertainties due to parameter estimation and lack of fit) have a tendency to materialize too frequently. Hence, as a description of real-life forecasting situations, an assumption about model correctness is untenable and represents a fragile foundation for forecast-based interest rate-setting (see Bårdsen *et al.*, 2003).

17.4.5.1 Assumptions about the forecasting situation

In modern monetary policy the forecasted rate of inflation is the intermediate target. It is then of interest to clarify as closely as possible what are the realistic properties of the forecast. Anticipated forecast properties are closely linked to the assumptions we make about the forecasting situation. A useful classification (see Clements and Hendry, 1999, Ch. 1) is:

A The forecasting model coincides with the true inflation process except for stochastic error terms. The parameters of the model are known constants over the forecasting period.
B As in A, but the parameters have to be estimated.
C As in B, but we cannot expect the parameters to remain constant over the forecasting period – structural changes are likely to occur.
D We do not know how well the forecasting model corresponds to the inflation mechanism in the forecast period.

A is an idealized description of the assumptions of macroeconomic forecasting. There is still the incumbency of inherent uncertainty represented by the stochastic

disturbances – even under **A**. Situation **B** represents the situation theoretical expositions of inflation targeting conjure up (see Svensson, 1999). The properties of situation **A** will still hold – even though the inherent uncertainty will increase. If **B** represents the premise for actual inflation targeting, there would be no forecast failures, defined as a significant deterioration of forecast performance relative to in-sample behavior (see, e.g., Hendry, 2006). The within-sample fit of section 17.4.2.2 corresponds to situation **B**, subject to the assumption that NAM is a congruent model of the aggregate Norwegian economy.

The limited relevance of situation **B** for inflation targeting becomes clear once we recognize that, in practice, we do not know what kind of shocks will hit the economy during the forecast period. We generally refer to such changes as *regime shifts,* and their underlying causes include changes in technology and political decisions and, more generally, "the complexity and instability of human behaviour" (Elster, 2007, p. 467). A forecast failure effectively invalidates any claim about a "correct" forecasting mechanism. Upon finding a forecast failure, the issue is therefore whether the misspecification was detectable or not at the time of preparing the forecast. It is quite possible that a model which has been thoroughly tested for misspecification within-sample nevertheless forecasts badly, which may occur in situation **C**.

As discussed by Clements and Hendry (1999), a dominant source of forecast failure is regime shifts in the forecast period, i.e., *after* the preparation of the forecasts. Since there is no way of anticipating them, it is unavoidable that *after-forecast* breaks damage forecasts from time to time. For example, when assessing inflation targeting over a period of years, we anticipate that the forecasters have done markedly worse than they expected at the time of preparing their forecasts, simply because there is no way of anticipating structural breaks before they occur. The task is then to be able to detect the nature of the regime shift as quickly as possible, in order to avoid repeated unnecessary forecast failure.

However, experience tells us that forecast failures are sometimes due to shocks and parameter changes that have taken place prior to the preparation of the forecast, but which have remained undetected by the forecasters. Failing to detect a *before-forecast* structural break might be due to the low power of statistical tests of parameter instability. However, the power is actually quite high for the kind of breaks that are most damaging to model forecasts (see Hendry, 2000). There are also practical circumstances that complicate and delay the detection of regime shifts. For example, there is usually uncertainty about the quality of the provisional data for the period that initialize the forecasts, making it difficult to assess the significance of a structural change or shock.

Hence both after- and before-forecast structural breaks are realistic aspects of real-life forecasting situations that deserve the attention of inflation targeters. In particular, one should seek forecasting models and tools which help cultivate an adaptive forecasting process. The literature on forecasting and model evaluation provide several guidelines (see, e.g., Hendry, 2001a; Granger, 1999).

Situation **D** brings us to the realistic situation, namely one of uncertainty and discord regarding what kind of model approximates reality; in other words, the issues of model specification and model evaluation. In section 17.4.3 we saw that in policy

analysis there is a clear gain from using the congruent model, and avoiding basing policy recommendations on a misspecified model of the supply side. In forecasting, the link between model misspecification and forecast failure is not always as straightforward as one would first believe. The complicating factor is again non-stationarity, regime shifts and structural change. For example, a time series model formulated in terms of the change in the rate of inflation adapts quickly to changes in equilibria – a location shift – and is therefore robust to before-forecast structural breaks of this type, even though it is clearly a misspecified model of the DGP over the historical data period (see Clements and Hendry, 1999, Ch. 5). Therefore a double-differenced device (DDD) can deliver near-unbiased forecasts when a location shift has occurred prior to the preparation of the forecast. Conversely, a forecasting model of the equilibrium correction type is less adaptable. Indeed, following an equilibrium shift, EqCM forecasts tend to move in the opposite direction to the data, thereby causing forecast failure (cf. Hendry, 2006). Eitrheim, Husebø and Nymoen (1999) have shown that this new theory of forecasting has practical relevance for understanding the properties of forecasts from a medium-sized forecasting model of the Norwegian economy, in particular the more adaptive nature of DDD to several historical examples of structural breaks.

The new theory of forecasting that we build on does not deliver carte blanche for using non-congruent models for prediction, though. DDD and near-equivalent forecasting devices are robust for one particular reason: they do not equilibrium correct, and are therefore insulated from changes in the parameters that are most pernicious for forecasting. Replacing a congruent and adequate EqCM with another, less adequate, EqCM model for forecasting is not a good idea. The non-congruent EqCM is also subject to forecast failure on its own premises and, without location shifts, it will forecast worse than the congruent EqCM. In this respect there is a cost to compromising model adequacy also in forecasting. In terms of the contesting supply-side models of section 17.4.3, this is illustrated by Bårdsen, Jansen and Nymoen (2002), who show that, although the PCM is robust to some of the location shifts that can damage forecasting from the ICM, the cost of high forecast variance and bias due to misspecified equilibrium correction dominates.

17.4.5.2 Real-time forecast performance

As mentioned above, NAM is part of the Normetrics system of models which was initiated in 2005. Model-based forecasts for the Norwegian economy are produced in January, March, June and September each year and published on the web. So far, these model-based forecasts have performed relatively well compared to competing forecasts. As an example, Figure 17.9 shows forecasts for core inflation in 2006. Because of inflation targeting, this variable is among the most thoroughly analyzed variables by professional forecasters both in the private and public sectors.

Figure 17.9 shows Norges Bank's projections together with the average of other professional Norwegian forecasters. The third line in the graph shows the sequence of forecasts from the two Normetrics forecasting models, automatized inflation forecasts (AIF) and NAM.[17] The figure shows that Normetrics forecasts were never

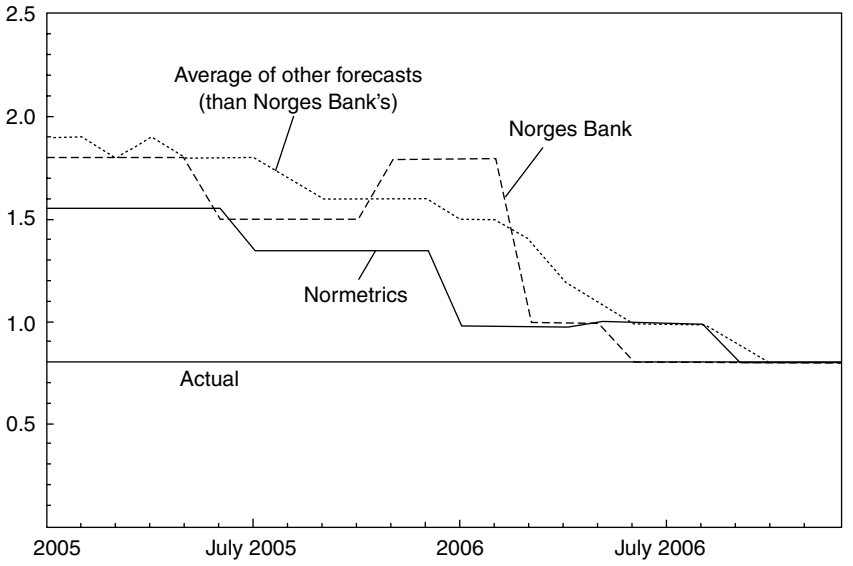


Figure 17.9 Forecasts for annual core inflation in 2006, published at different times. Percent. Monthly figures, January 2005 to December 2006
Source: Norges Bank, *Monetary Policy Report* January 2007, and http://folk.uio.no/rnymoen/Normetrics.

any worse than the average forecasts, and most vintages of Normetrics forecasts are considerably better. All Normetrics forecasts produced in 2005 also improve on Norges Bank's forecasts (June 2005 is the exception). In the period November 2005 to February 2006 the gap between Normetrics and Norges Bank actually widens – as Norges Bank's forecasts were adjusted upward, away from what eventually became the actual inflation rate of 0.8%. However, the forecast in *Monetary Policy Report* February 2006 is accurate, while the Normetrics forecast stays at 1% until September 2006.

Since equilibrium-correction is ubiquitous in macroeconometric models, one can safely assume that some of the "other forecasts" are based on EqCMs. Specifically, Norges Bank's forecast are based on the rational expectations solution of an equilibrium-correction model with leads in variables. Hence, all forecasts in Figure 17.9 may have been damaged by any location shifts that took place in 2006 – they were *after-forecast* structural breaks. In particular, the forecast that was published (early) in 2005 had a large exposure to forecast failure. For the same reason, the forecast errors are reduced as more "2006 information" is conditioned upon. In that perspective we can interpret the figure as evidence that the Normetrics forecasts are more adaptive than the other equilibrium-correction forecasting mechanisms covered by the graph.

The accuracy of the model-based forecasts in Figure 17.9 are less impressive when compared to forecasts from a simple DDD, though. For example, at the start of 2005 a forecast based on the double difference of the log of the CPI-AET (consumer price

index adjusted for electricity and taxes) index would predict a 2006 inflation rate of 0.3%, which is a more accurate forecast than any of the econometric or professional forecasts that were produced in 2005. The DDD forecast is not improved on by Normetrics before January 2006, nor by Norges Bank before the *Monetary Policy Report* January 2006 (from March). However, by this time the DDD forecast could also be have been updated, most simply by replacing 0.3% by the actual rate of inflation in 2005, which was 1.0%. This forecast is practically identical to Normetrics forecasts from before September 2006, and it is not beaten by Norges Bank's forecast before the *Monetary Policy Report* February 2006.

2006 is not the only year for which the Normetrics forecasts compete well with the central bank's forecasts for the monetary policy target variable. Juel, Molnar and Røed (2008) note that, for a relatively long period of time, the automated forecasts from an empirically validated inflation model, AIF above, have been better than Norges Bank's forecasts. Figure 17.10 illustrates the point.

The figure can be used to assess the *ex ante* forecasts based on forecast errors for the period 2004(2)–2007(3). The forecasted variable is the annual rate of CPI-AET inflation. The graphs in the upper panel in the figure show the MFEs. A negative MFE means that the inflation forecasts are on average higher than the actual inflation rates in the period. The biases of the econometric model (AIF) forecasts are

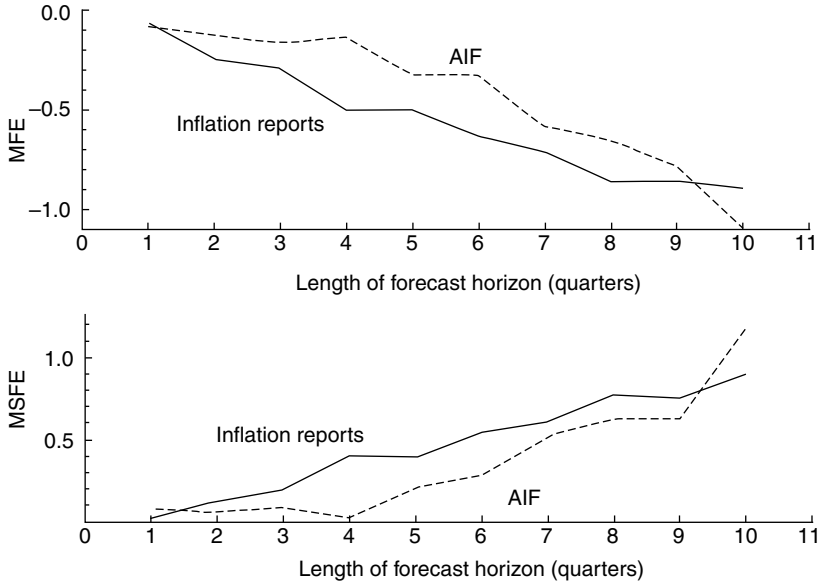


Figure 17.10 Forecasts from an empirically validated inflation model compared to inflation report forecasts

First panel: mean forecast errors (MFEs) for core inflation (annual rate of change in CPI-AET). Inflation report forecasts and AIF forecasts. *Second panel*: mean squared forecast errors (MSFEs).

Source: *Inflation Report*/*Monetary Policy Report* February 2004 to February 2007 and AIFs published at http://folk.uio.no/rnymoen/forecast_ air_index.html.

small for forecast horizons 1, 2, 3 and 4: the MFEs are less than 0.25 percentage points. Norges Bank's inflation forecasts from the *Inflation Report*/*Monetary Policy Reports* are more biased than AIF for horizons 2–8 quarters ahead. The biases of AIF become markedly bigger for forecasts of length 7–10 quarters, and are not much different from the bias of forecasts produced by Norges Bank. The second panel of the figure shows the MSFEs, to which large forecast errors contribute more than small errors. This measure gives more or less the same picture at the MFEs.

17.4.5.3 Ex post *forecast evaluation and robustification*

The discussion of Figure 17.9 illustrated the general point that, although EqCMs forecast well when a process is difference-stationary, they are non-robust if there are non-stationarities due to location shifts in the forecast period. In this particular case, the mean of the rate of inflation seems to have changed, in 2004 or earlier, and the DDD is a more adaptable forecasting mechanism than the EqCMs in the case of the before-forecast structural break. The main benefit of DDDs for model-based forecasting, where one wants to retain the causal information of the model, is that DDDs can help the forecaster to robustify the EqCM forecasts, by intercept corrections. The cost associated with DDDs is, of course, that the forecasts are more "noisy" than EqCM forecasts, hence the forecast-error variances associated with robust forecasts can become large, such as when the first difference of an autoregressive process doubles the one-step forecast variance. In a model with one or two endogenous variables this cost may not be much of an issue, but in a multi-equation forecasting setting there may be a problem of extracting "signal from noise" in practice. In the rest of this section we illustrate these issues by considering system forecasts.

Figure 17.11 shows dynamic NAM forecasts for the period 2003(1)–2007(3).[18] The sample period for the estimation ends in 2002(4). Unlike the inflation forecasts in Figure 17.9 which are real-time *ex ante* forecasts, we now consider *ex post* forecasts, which are conditioned by the true values of the non-modeled variables. Hence the forecasts are not influenced by location shifts in the non-modeled variables (foreign CPI inflation, e.g.), but they are subject to location shifts that are due to changes in the means of the estimated cointegration relationships, or the autonomous growth rates (captured by intercepts after subtraction of equilibrium correction means).

Figure 17.11 serves as the reference case for discussion of the degree of adaptation of NAM forecasts, and for the properties of more robust forecasting devices that we investigate for comparison. There are no less than five possible instances of location shifts in the forecast of these nine variables. First, there is systematic overprediction of the rate of inflation (we use CPI inflation here) over the length of the forecast horizon. In the light of the 70% prediction interval, inflation overprediction is significant (forecast failure) in 2003. Second, unemployment is overpredicted and GDP growth in underpredicted for the three quarters of 2007. Third, the short-term interest rate is significantly overpredicted in 2003(2)–2004(4). Fourth, real credit growth is underpredicted (significantly) from

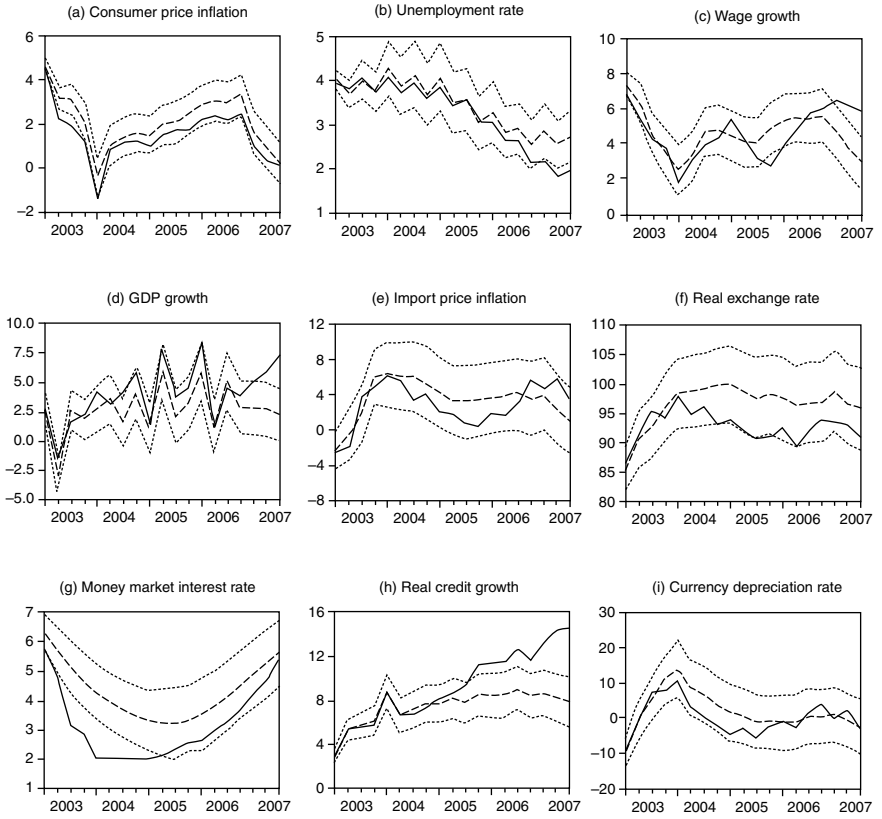


Figure 17.11 Dynamic NAM forecasts 2003(1)–2007(3), with end-of-sample for estimation of parameters in 2002(4)

Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals.

2005(3). Finally, the real exchange rate is systematically overpredicted from 2004(4) and onwards – the real appreciation of the Norwegian krone is not captured by the forecasts.

In the following we show how well the NAM forecasts adapt to these location shifts when we condition on 2004(4) and then 2006(4). In each dynamic forecast the coefficient estimate is updated. Figure 17.12 shows that the 2005(1)–2007(3) forecasts for inflation and, in particular, for the interest rate and the real exchange rate, have improved relative to Figure 17.11. For the other variables there is little change and, if anything, the forecast failures for the rate of unemployment and for GDP growth stand out more clearly than before (also showing that there is a knock-on effect of the high employment forecast on wage inflation). The explanation may be that NAM is unable to adapt, or that the location shifts of the variables are of the after-forecast type.

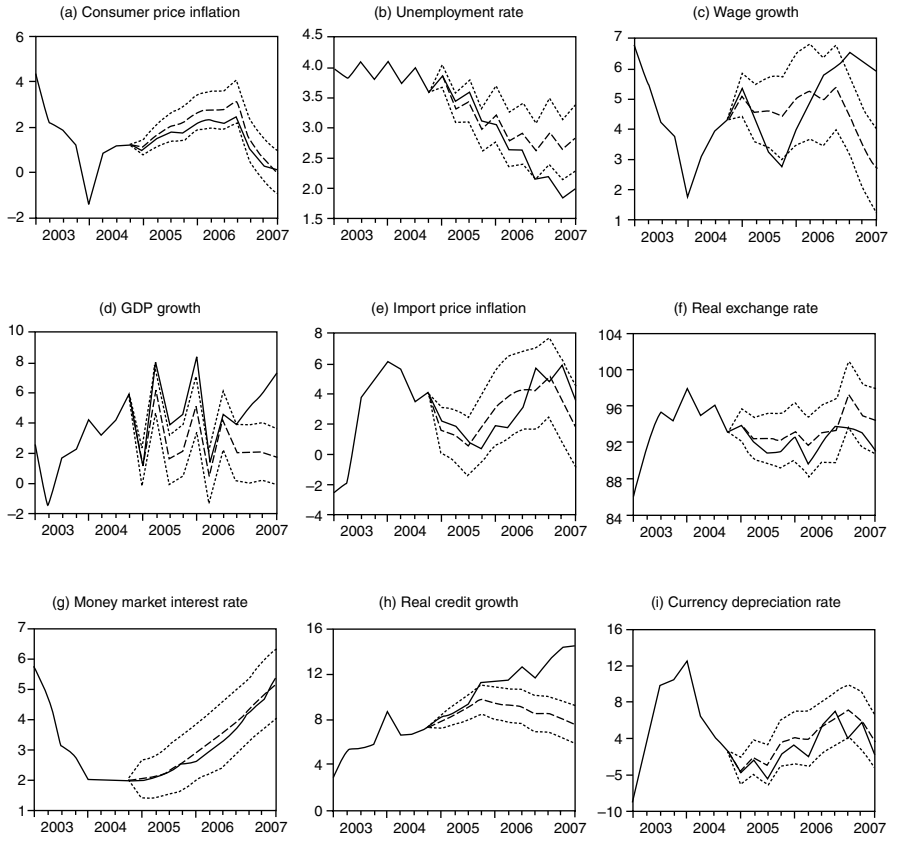


Figure 17.12 Dynamic NAM forecasts 2005(1)–2007(3), with end-of-sample for estimation of parameters in 2004(4)
Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals.

Figure 17.13 focuses on the last three "problem variables" in terms of forecast failure. The rate of unemployment is now much better forecasted in 2007(1)–2007(3), which is a sign of adaptation to a location shift which is of the before-forecast category, where 2006(4) is in the information set. For GDP and real credit growth the forecast failures persist. In the present version of NAM, the high growth rates of 2007 can, in part, be explained by the effects of very high oil prices on demand. In fact, that modeling device was used in Figure 17.6, showing the goodness-of-fit, but clearly would not be known or of any help to a forecaster preparing a forecast for 2007 late in 2006.

As mentioned above, using differencing (DDD) to forecast provides a more robust forecast when non-stationarities are due to location-shifts. As discussed by Hendry (2006), a differenced version of the EqCM may be interpreted as an augmented DDD forecasting rule. We therefore consider forecasts from the differenced NAM

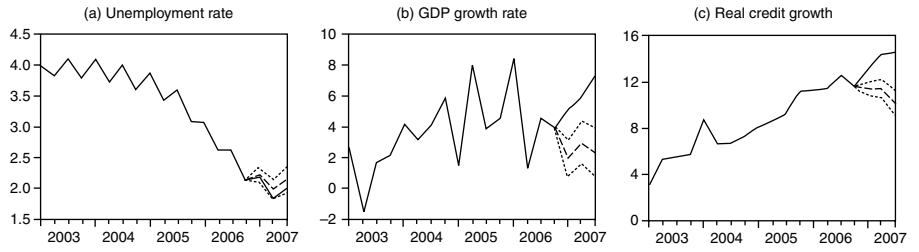


Figure 17.13 Dynamic NAM forecasts 2007(1)–2007(4), with end-of-sample for estimation of parameters in 2006(4)

Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals.

model, and refer to these as dEqCM forecasts. In the differenced forecasting systems, some of the causal information embedded in NAM is retained. The dEqCM has no constants either in the form of means of cointegration relationships or in the form of separate intercept terms (see Hendry, 2006). Hence forecasts from the dEqCM do not equilibrium-correct, thereby reducing the risks attached to EqCM forecasts.

Figure 17.14 shows the dEqCM forecast for the 2003(1)–2007(3) period. Comparison with the NAM forecasts in Figure 17.11 shows that the increase in forecast variance is not a small cost in this case – note the difference in scaling as well. The absolute forecast errors appear to be much worse than the NAM forecast errors as well. For example, unemployment is predicted by the dEqCM to increase over the forecast horizon, and credit growth is underpredicted for the length of the horizon.

Figure 17.15 shows the dEqCM forecasts when we condition on information including 2004(4). Compared to the NAM forecasts that condition on the same information (see Figure 17.12) there is little to be gained in these forecasts. We note that the dEqCM interest rate forecast has adapted, but the same happened with the NAM forecasts. The dEqCM forecasts are still uninformative about the behavior of unemployment and credit growth over the 2006(1)–2007(3) period.

Figure 17.16 indicates that for the three quarters of 2007, the dEqCM forecast for the rate of unemployment is better than the (already quite good) NAM forecasts. However, for GDP growth and credit growth, the dEqCM still does not adjust to the location shift. These results suggest that, in practice, a more discretionary approach may be called for. For example, instead of taking away all the equilibrium correction by differencing, one may concentrate on the sub-set of equations which have failed because of location shifts in recent forecasting rounds, since that will also induce lack of adaptation in the overall forecasting picture. To illustrate the possible benefit from such an approach, Figure 17.17 shows the 2007 forecasts for a dEqCM, where only the equilibrium variables of the aggregate demand and credit equations have been "differenced away." To avoid confusion with the dEqCM used above we refer to this forecasting model as *partial* dEqCM. Figure 17.17 shows the one-step forecast, since any difference in adaptability is then easier to see.

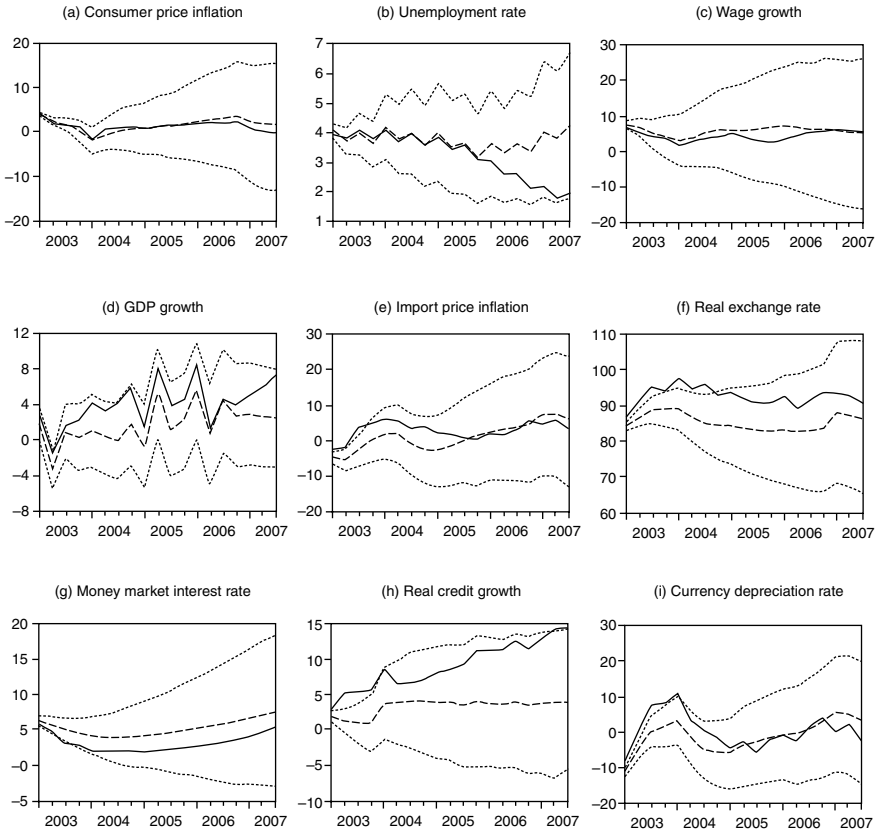


Figure 17.14 Dynamic dEqCM forecasts 2003(1)–2007(3), with end-of-sample for estimation of parameters in 2002(4)

Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals.

Panels (a) and (b) show the NAM forecasts. The forecasts for 2007(1) are the same as in Figure 17.16 for these two variables, but 2007(2) and 2007(3) are different since the NAM forecasts are now conditional on first- and second-quarter information for exogenous and predetermined variables (the coefficients are not updated). The lack of adaptation to the location shifts of the growth rate of the actual series is apparent. Panels (c) and (d) show the corresponding partial dEqCM forecasts. For GDP growth there is less underprediction already in 2007(1), and in 2007(2) the forecast has adapted. Panel (c) shows a marked improvement in adaptation also in the credit growth rate, although not before the second quarter of 2007(2).

17.5 Conclusion

This chapter has presented a case for continuing to use macroeconometric models for policy analysis, including such analyses that rely on instrument use to attain a

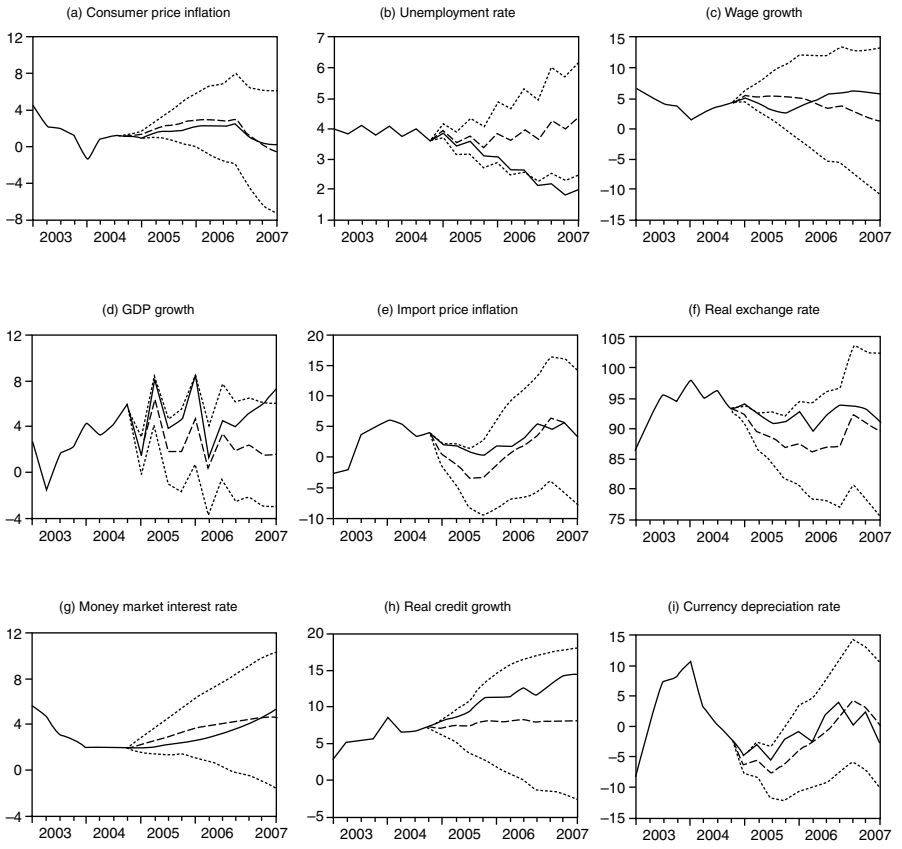


Figure 17.15 Dynamic dEqCM forecasts 2005(1)–2007(3), with end-of-sample for estimation of parameters in 2004(4)
Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals

certain forecast for a target variable like inflation. Paradoxically, perhaps, the main line of argument starts from the recognition that accurate forecasting is a near impossibility in macroeconomics because of the inherent non-stationarity of the economic time series included in the model. Non-stationarity takes different forms, with different implications for macroeconometric modeling and forecasting, and we have distinguished between unit roots and non-stationarity due to structural breaks. We have demonstrated that macroeconometric models can be developed theoretically and empirically in a way that is consistent with a unit root assumption. At the modeling and estimation stage, non-stationarity due to unit roots can in principle be handled by cointegration methods and, given that approach, unit roots are unlikely to be a source of forecast failures. On the contrary, a correct unit root assumption can help in concentrating on predictable functions of variables, like growth rates and linear combinations of target variables.

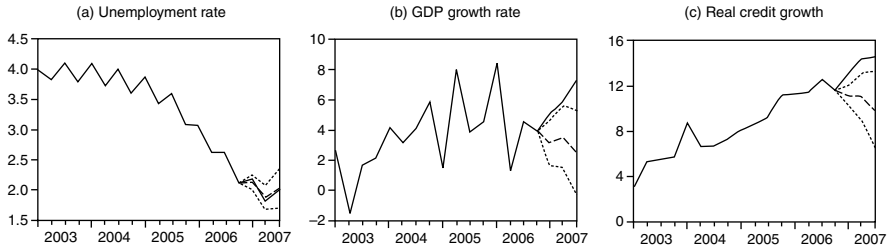


Figure 17.16 Dynamic dEqCM forecasts 2007(1)–2007(4), with end-of-sample for estimation of parameters in 2006(4)
Actual values are shown by solid lines, and forecasts by dashed lines. The distance between the two dotted lines represents 70% prediction intervals.

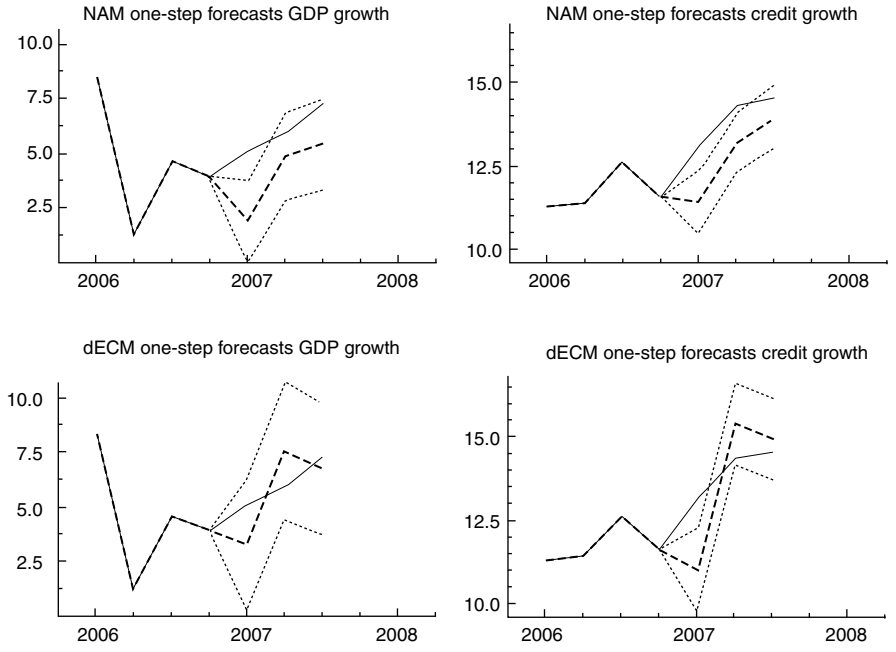


Figure 17.17 One-step forecast from NAM and a partial dEqCM
The forecasts are the dashed lines. The distance between the dotted lines represents 70% prediction intervals.

Non-stationarity due to structural breaks in functional relationships of the economy seem to represent the real challenge to macroeconometric modeling. Structural breaks in the forecast period are particularly harmful since they are untraceable and will make the model forecasts go toward pre-break steady-state relationships. When the sample period is extended, structural breaks represent valuable sample information that provide power to tests of economic hypotheses,

and from that perspective structural breaks are seen to be not entirely negative, since there can be progress in macroeconometric modeling through forecast failure.

It follows from our approach that we agree with those who conclude that guidance from economic theory is important in the forecasting process (cf. Elliott and Timmermann, 2008). But theory and modern econometrics do not provide immunity from new structural breaks. Nevertheless, the case for macroeconometric system of equations models may, to a large extent, depend on their ability to forecast relatively well compared to competing forecasting methods. Making models sufficiently adaptive to a structural break, once the evidence is there that it has occurred, is a necessary step in that project.

It is possible to look to other forecasting disciplines for inspiration. Meteorologists developed their forecasting theory, models and routines between the two world wars. Those forecasts were based on a deterministic model of the dynamics of the atmosphere. Then, in the mid 1960s, the understanding of the dynamics were completely altered with the development and acceptance of chaos theory, with the logical consequence that accurate weather forecasting is impossible. Thirty years after the arrival of the meteorologists' "impossibility theory" for forecasting, the weather forecasts are undeniably more accurate than ever. Hence, the weather is predictable even ten days ahead, despite the chaos represented by the underlying forecasting models. It is perhaps unlikely that we will witness something similar in economics, but there are nevertheless parallels. Meteorologists have precise theory and almost continuous updating of initial conditions. Policy-oriented modeling may have to live with idealized or partial theories, and with variables that are measured at relatively long time intervals, and which are influenced by measurement errors. Yet, as a discipline we have developed methods and strategies that are quite good at making the most of "small data amounts." Hence, while meteorologists can rely on the theoretically specified model of weather dynamics, the interaction between economic theory, econometric theory, good model selection procedures, and diagnostic testing have together greatly improved our capability of modeling the macroeconomy, thus providing models that can aid policy decisions.

Appendix: Data definitions and equation statistics

Variables

The model employs seasonally unadjusted data. Unless another source is given, all data are taken from FPAS, the database of Norges Bank.

The model is developed and estimated with Oxmetrics 5 (http://www.oxmetrics.net) and then re-estimated and simulated with Eviews 6 (http://www.eviews.com).

V Trade-weighted nominal value of the krone based on import shares of trading countries.
G Government sector consumption expenditure, fixed 1991 prices. Millions (Mill): Norwegian krone (NOK).
L Nominal credit volume. Mill. NOK.
R Money market interest rate (three month euro–krone interest rate).

R^* ECU interest rate. For the period 1967(1)–1986(3): effective interest rate on foreign bonds, NOK basket weighted. For the period 1986(4)–1996(4): ECU weighted effective rate on foreign bonds.
R_L Average interest rate on bank loans.
R_B Yield on six year government bond, quarterly average.
R_B^* Yield on long-term foreign bonds. NOK basket weighted.
P Consumer price index (CPI).
P_C "Core" CPI.
P^* Consumer prices abroad in foreign currency.
P_E CPI, electricity, fuel and lubricants.
P_O $US oil price, per barrel Brent–Blend.
P_I Price deflator of total imports.
P_I^* Producer price index, trading partners.
Y Total value added at market prices in the mainland economy (defined as total economy minus North Sea oil and gas production and international shipping). Fixed base year (1991) prices. Mill. NOK.
Z Mainland economy value added per man hour at factor costs.
$T1$ Payroll tax rate, mainland economy.
U Registered rate of unemployment.
W Nominal hourly wage costs in the mainland economy. NOK.

In addition, there is a step dummy, accounting for the introduction of inflation targeting:

$$IT = 0 \text{ until } 2001(1),\ 1 \text{ from } 2001(2).$$

Notation for estimation and misspecification tests

In Table 17.1, the estimation method, which is either OLS or full information maximum likelihood (FIML), is indicated in the first line below each equation, along with the sample size (number of quarterly observations), which is denoted by T, and the residual standard error ($\hat{\sigma}$). For equations estimated with OLS, statistics for residual autocorrelation and ARCH are reported in the second line below the equation. As indicated by the notation, these two statistics are F-distributed under their respective null hypotheses. For example, $F_{AR(1-4)}(4, 44)$ in the exchange rate equation denotes the F-distributed test statistics with 4 and 44 degrees of freedom for the null hypothesis of no autocorrelation against the alternative of fourth-order autocorrelation. In the third line below the estimated OLS equations, we report the chi-square-distributed test of residual normality, and the F-distributed test of heteroscedasticity due to squares of the regressors. For the equations estimated with FIML, systems versions of the misspecification tests are reported and are indicated by the extra subscript *vec*, as $F_{vec,\mathsf{AR}(1-4)}$. The numbers in brackets are p-values for the respective null hypotheses. These, as well as the other standard diagnostics tests, are explained in Doornik and Hendry (2007a) (single-equation diagnostics), and Doornik and Hendry (2007b) (system and simultaneous equations diagnostics).

Acknowledgments

We thank Q. Farooq Akram and Kerry Patterson for helpful comments and discussions.

Notes

1. See, e.g., Hendry (1995a, Ch. 2.3 and 15.3) for a concise definition of structure as the invariant set of attributes of the economic mechanism.
2. This line of thought may lead to the following practical argument against large-scale empirical models: since modeling resources are limited, and some sectors and activities are more difficult to model than others, certain euations of any given model are bound to have less structural content than others, i.e., the model as a whole is no better than its weakest (least structural) equation.
3. See Nymoen (2005) for an analysis of a recent failure in inflation forecasting.
4. Sub-sections 17.2.1–4 draw on Bårdsen, Hurn and Lindsay (2004).
5. Presently, we let the unemployment rate be constant and disregard it for simplicity. We return to the role of the rate of unemployment in section 17.2.6.
6. The distinction between the inferential and model validation facets of modeling is due to Spanos (2008), who conclusively dispels the charge that misspecification testing represents an illegitimate "reuse" of the data already used to estimate the parameters of the statistical model; see also Hendry (1995b, pp. 313–14).
7. It might be noted that the income tax rate $T2$ is omitted from the analysis. This simplification is in accordance with previous studies of aggregate wage formation (see, e.g., Nymoen, 1990, and Nymoen and Rødseth, 2003, where no convincing evidence of important effects from the average income tax rate $T2$ on wage growth could be found).
8. Note that, due to the log-form, $\zeta = is/(1 - is)$, where is is the import share in private consumption.
9. Strictly, we take the expectation through in both equations.
10. NAM is a model project which extends from the early econometric assessment of wage and price inflation in Nymoen (1991), further developed in Bårdsen, Fisher and Nymoen (1998), Bårdsen and Fisher (1999), and the monetary transmission model of Bårdsen and Klovland (2000). Earlier versions of the model are documented in Bårdsen and Nymoen (2001), Bårdsen, Jansen and Nymoen (2003) and Bårdsen *et al.* (2005).

    NAM is used for both research purposes and teaching. The macroeconomic data is from the model databases of Statistics Norway (KVARTS model) and Norges Bank (FPAS database). Specific versions of the model are currently operative for (a) econometric forecasts of the Norwegian macroeconomy (NAM-EF) and (b) model-based analysis of financial stability in Norway (NAM-FS).
11. See http://folk.uio.no/rnymoen/normetrics_index.html.
12. In practice, the policy instrument is the sight deposit rate set by the central bank, but since the sight deposit rate represents (banks') marginal funding cost, changes in the sight rate are transmitted to the money market rate immediately.
13. The size of the depreciation will depend upon the risk premium, and whether expectations counteract or strengthen the initial effect of the interest rate cut, etc. (cf. section 17.3.3).
14. Although the derivations are presented for a single equation with exogenous regressors, for ease of exposition, the techniques are, of course, the same for systems.
15. A full listing of variables is given in the appendix to this chapter.
16. See, e.g., Ericsson (2001) for an accessible discussion of forecast uncertainty, and its presentation in published forecasts.
17. Automatized econometric inflation forecasts have been published twice a year, starting in July 2004. The forecasts are automatized, with a minimum of intervention after the econometric specification of the forecasting mechanism is completed.

18. In this section, we use the model version used for the October 2007 NAM forecasts: see http://folk.uio.no/rnymoen/namforecast_okt07.pdf.

References

Akram, Q. (2007) *Designing Monetary Policy Using Econometric Models*. Oslo: Norges Bank.

Akram, Q.F. and R. Nymoen (2008) Model selection for monetary policy analysis – how important is empirical validity? *Oxford Bulletin of Economics and Statistics* **70**. Forthcoming.

Aldrich, J. (1989) Autonomy. *Oxford Economic Papers* **41**, 15–34.

Andersen, T.M. (1994) *Price Rigidity: Causes and Macroeconomic Implications*. Oxford: Clarendon Press.

Ball, L. (1999) Policy rules for open economies. In J.B. Taylor (ed.), *Monetary Policy Rules*. A National Bureau of Economic Research Conference Report, pp. 127–44. Chicago: University of Chicago Press.

Bårdsen, G., Ø. Eitrheim, E.S. Jansen and R. Nymoen (2005) *The Econometrics of Macroeconomic Modelling*. Oxford: Oxford University Press.

Bårdsen, G. and P.G. Fisher (1993) The importance of being structured. Discussion Paper 02/93, Institute of Economics, Norwegian School of Economics and Business Administration.

Bårdsen, G. and P.G. Fisher (1999) Economic theory and econometric dynamics in modelling wages and prices in the United Kingdom. *Empirical Economics* **24**(3), 483–507.

Bårdsen, G., P.G. Fisher and R. Nymoen (1998) Business cycles: real facts or fallacies? In S. Strøm (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*. No. 32 in Econometric Society Monograph Series, Ch. 16, pp. 499–527. Cambridge: Cambridge University Press.

Bårdsen, G., S. Hurn and K. Lindsay (2004) Linearizations and equilibrium correction models. *Studies in Nonlinear Dynamics and Econometrics* **8**(4), article 5, http://www.bepress.com/snde/vol8/iss4/art5.

Bårdsen, G., E.S. Jansen and R. Nymoen (2002) Model specification and inflation forecast uncertainty. *Annales d'économie et de Statistique* **67**(68), 495–517.

Bårdsen, G., E.S. Jansen and R. Nymoen (2003) Econometric inflation targeting. *Econometrics Journal* **6**(2), 429–60.

Bårdsen, G., E.S. Jansen and R. Nymoen (2004) Econometric evaluation of the New Keynesian Phillips curve. *Oxford Bulletin of Economics and Statistics* **66**(s1), 671–86.

Bårdsen, G. and J.T. Klovland (2000) Shaken or stirred? Financial deregulation and the monetary transmission mechanism in Norway. *Scandinavian Journal of Economics* **102**(4), 563–83.

Bårdsen, G. and R. Nymoen (2001) Rente og inflasjon (Interest rate and inflation). *Norsk Økonomisk Tidsskrift* **115**, 125–48.

Bårdsen, G. and R. Nymoen (2003) Testing steady-state implications for the NAIRU. *Review of Economics and Statistics*, **85**, 1070–75.

Bårdsen, G. and R. Nymoen (2008) U.S. natural rate dynamics reconsidered. In J. Castle and N. Shephard (eds.), *The Methodology and Practise of Econometrics*. Oxford: Oxford University Press. Forthcoming.

Barkbu, B.B., R. Nymoen and K. Røed (2003) Wage coordination and unemployment dynamics in Norway and Sweden. *Journal of Socio-Economics* **32**, 37–58.

Batini, N., B. Jackson and S. Nickell (2005) An open-economy New Keynesian Phillips curve for the U.K. *Journal of Monetary Economics* **52**, 1061–71.

Bjørnstad, R. and R. Nymoen (2008) The New Keynesian Phillips curve testes on OECD panel data. *Economics: The Open Access, Open Assessment E-Jounal* **2**(23), 1–18.

Blanchard, O.J. and L. Katz (1999) Wage dynamics: reconciling theory and evidence. *American Economic Review* **89**(2), 69–74. Papers and Proceedings (May).

Blanchflower, D.G. and A.J. Oswald (1994) *The Wage Curve*. Cambridge: Mass.: MIT Press.
Calmfors, L. and R. Nymoen (1990) Nordic employment. *Economic Policy* **5**(11), 397–448.
Campbell, J. (1994) Inspecting the mechanism: an analytical approach to the stochastic growth model. *Journal of Monetary Economics* **33**(3), 463–506.
Clarida, R., J. Gali and M. Gertler (1999) The science of monetary policy: a New Keynesian perspective. *Journal of Economic Literature* **37**(4), 1661–707.
Clements, M.P. and D.F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.
Doornik, J.A. (2008) Autometrics. In J.L. Castle and N. Shephard (eds.), *The Methodology and Practise of Econometrics*. Oxford: Oxford University Press.
Doornik, J.A. and D.F. Hendry (1997). The implications for econometric modelling of forecast failure. *Scottish Journal of Political Economy* **44**, 437–61.
Doornik, J. and D. H. Hendry (2007a). *Empirical Econometric Modelling. PcGive 12, Volume 1*. London: Timberlake Consultants Ltd.
Doornik, J. and D.H. Hendry (2007b). *Modelling Dynamic Systems. PcGive 12, Volume 2*. London: Timberlake Consultants Ltd.
Eitrheim, Ø., T.A. Husebø and R. Nymoen (1999) Equilibrium-correction versus differencing in macroeconomic forecasting. *Economic Modelling* **16**, 515–44.
Eitrheim, Ø., E.S. Jansen and R. Nymoen (2002) Progress from forecast failure: the Norwegian consumption function. *Econometrics Journal* **5**, 40–64.
Elliott, G. and A. Timmermann (2008) Economic forecasting. *Journal of Economic Literature*, **46**, 3–56.
Elster, J. (2007) *Explaining Social Behaviour. More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
Engle, R.F. and D.F. Hendry (1993) Testing super exogeneity and invariance in regression equations. *Journal of Econometrics*, **56**, 119–39.
Ericsson, N.R. (2001) Forecast uncertainty in economic modeling. In D.F. Hendry and N.R. Ericsson (eds.), *Understanding Economic Forecasts*, Ch. 5, pp. 68–92. Cambridge, Mass.: MIT Press.
Ericsson, N.R. and D. Hendry (1999) Encompassing and rational expectations: how sequential corroboration can imply refutation. *Empirical Economics* **24**(1), 1–21.
Ericsson, N.R. and J.S. Irons (1995) The Lucas critique in practice: theory without measurement. In K.D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, Ch. 8. Dordrecht: Kluwer Academic Publishers.
Fanelli, L. (2008) Testing the New Keynesian Phillips curve through vector autoregressive models: results from the euro area. *Oxford Bulletin of Economics and Statistics* **70**(1), 53–66.
Frisch, R. (1936) On the notion of equilibrium and disequilibrium. *Review of Economic Studies* **3**, 100–5.
Fuhrer, J. (2006) Intrinsic and inherited inflation persistence. *International Journal of Central Banking* **2**, 49–86.
Gali, J. and M. Gertler (1999) Inflation dynamics: a structural econometric analysis. *Journal of Monetary Economics* **44**(2), 233–58.
Gali, J., M. Gertler and D. López-Salido (2001) European inflation dynamics. *European Economic Review* **45**, 1237–70.
Gali, J., M. Gertler and J. Lopez-Salido (2005) Robustness of the estimates of the hybrid New Keynesian Phillips curve. *Journal of Monetary Economics* **52**, 1107–18.
Garratt, A., K. Lee, M.H. Pesaran and Y. Shin (2006) *Global and National Macroeconometric Modelling: A Long-Run Structural Approach*. Oxford: Oxford University Press.
Granger, C.W. (1992) Fellow's opinion: evaluating economic theory. *Journal of Econometrics* **51**, 3–5.
Granger, C.W.J. (1999) *Empirical Modeling in Economics: Specification and Evaluation*. Cambridge: Cambridge University Press.

Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica* **12**, 1–118. Supplement.
Hahn, F. and R. Solow (1997) *A Critical Essay on Modern Macroeconomic Theory*. Cambridge, Mass.: MIT Press.
Hendry, D.F. (1988) The encompassing implications of feedback versus feedforward mechanisms in econometrics. *Oxford Economic Papers* **40**, 132–49.
Hendry, D.F. (1995a) *Dynamic Econometrics*. Oxford: Oxford University Press.
Hendry, D.F. (1995b) Econometrics and business cycle empirics. *Economic Journal* **105**, 1622–36.
Hendry, D.F. (2000) On detectable and non-detectable structural change. *Structural Change and Economic Dynamics* **11**, 45–65.
Hendry, D.F. (2001a) How economists forecast. In D.F. Hendry, *Understanding Economic Forecasts*, Ch. 2, pp. 15–41. Cambridge, Mass.: MIT Press.
Hendry, D.F. (2001b) Modelling UK inflation. *Journal of Applied Econometrics* **16**, 255–75.
Hendry, D. (2004) *Robustifying Forecasts from Equilibrium-Correction Models*. Oxford: Nuffield College, Oxford University.
Hendry, D. (2006) Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics* **135**, 399–426.
Hendry, D.F. and H.M. Krolzig (1999) Improving on "Data mining reconsidered" by K.D. Hoover and S.J. Perez. *Econometrics Journal* **2**, 41–58.
Hendry, D.F. and G.E. Mizon (2000) Reformulating empirical macroeconometric modelling. *Oxford Review of Economic Policy* **16**(4), 138–57.
Hoover, K.D. and S.J. Perez (1999) Data mining reconsidered: encompassing and the general-to-spesific approach to specification search. *Econometrics Journal* **2**, 1–25.
Johansen, L. (1977) *Lectures on Macroeconomic Planning. Volume 1: General Aspects*. Amsterdam: North-Holland.
Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
Johansen, S. (2006) Cointegration: an overview. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics. Volume 1: Econometric Theory*. Basingstoke: Palgrave Macmillan.
Johansen, S. and A.R. Swensen (1999) Testing exact rational expectations in cointegrated vector autoregressive models. *Journal of Econometrics*, **93**, 73–91.
Johansen, S. and A.R. Swensen (2004) More on testing exact rational expectations in cointegrated vector autoregressive models: restricted constant and linear term. *Econometrics Journal* **7**, 389–97.
Juel, S., K. Molnar and K. Røed (2008) *Norges Bank Watch 2008. An Independent Review of Monetary Policymaking in Norway*. Technical Report, Centre for Monetary Economics at the Norwegian School of Management, BI, http://www.bi.no/cmeFiles/NBW2008.
Juselius, K. (2007) *The Cointegrated VAR Model: Methodology and Applications*. Oxford: Oxford University Press.
Kolsrud, D. and R. Nymoen (1998) Unemployment and the open economy wage–price spiral. *Journal of Economic Studies* **25**, 450–67.
Layard, R., S. Nickell and R. Jackman (1991) *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.
Layard, R., S. Nickell and R. Jackman (1994) *The Unemployment Crisis*. Oxford: Oxford University Press.
Layard, R., S. Nickell and R. Jackman (2005) *Unemployment* (second edition). Oxford: Oxford University Press.
Lucas, R.E.J. (1976) Econometric policy evaluation: a critique. In K. Brunner and A.M. Meltzer (eds.), *The Phillips Curve and Labour Markets*, pp. 19–46. Carnegie-Rochester Conference on Public Policy **3**.

Lütkepohl, H. (2006) Vector Autoregressive Models. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics*. *Volume 1*: *Econometric Theory*, pp. 287–325. Basingstoke: Palgrave Macmillan.

Mavroeidis, S. (2006) *Testing the New Keynesian Phillips Curve Without Assuming Identification*. Brown Economic Papers 2006-13, Brown University.

Nymoen, R. (1991) A small linear model of wage- and price-inflation in the Norwegian economy. *Journal of Applied Econometrics* **6**, 255–69.

Nymoen, R. (2005) Evaluating a central bank's recent forecast failure. Memorandum 22/05, Department of Economics, University of Oslo.

Nymoen, R. and A. Rødseth (2003) Explaining unemployment: some lessons from Nordic wage formation. *Labour Economics* **10**, 1–29.

Pagan, A. (2003) Report on modeling and forecasting at the Bank of England. *Bank of England Quarterly Bulletin*, Spring, 60–88.

Patterson, K.D. (1987) Growth coefficients in dynamic time series models. *Oxford Economic Papers* **39**(2), 282–92.

Patterson, K.D. and J. Ryding (1984) Dynamic time series models with growth effects constrained to zero. *Economic Journal* **94**(373), 137–43.

Pesaran, M.H. and R.P. Smith (1985) Evaluation of macroeconomic models. *Economic Modelling* **2**(2), 125–34.

Rudd, J. and K. Whelan (2005) Does labor's share drive inflation? *Journal of Credit and Banking* **37**(2), 297–312.

Rudd, J. and K. Whelan (2007) Modeling inflation dynamics: a critical review of recent work. *Journal of Credit and Banking* **39**(2), 155–70.

Sargan, J.D. (1964) Wages and prices in the United Kingdom: a study of econometric methodology. In P.E. Hart, G. Mills and J.K. Whitaker (eds.), *Econometric Analysis for National Economic Planning*, pp. 25–63. London: Butterworth.

Sargan, J.D. (1980) A model of wage–price inflation. *Review of Economic Studies* **47**, 113–35.

Spanos, A. (2006) Econometrics in retrospect and prospect. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics*. *Volume 1: Econometric Theory*, Ch. 1, pp. 3–58. Basingstoke: Palgrave Macmillan.

Spanos, A. (2008) Sufficiency and ancillarity revisited: testing the validity of a statistical model. In J.L. Castle and N. Shephard (eds.), *The Methodology and Practise of Econometrics*. Oxford: Oxford University Press.

Stanley, T.D. (2000) An empirical critique of the Lucas critique. *Journal of Socio-Economics* **29**, 91–107.

Svensson, L. (1997) Inflation forecast targeting: implementing and monitoring inflation targets. *European Economic Review* **41**, 1111–46.

Svensson, L. (1999) Inflation targeting: some extensions. *Scandinavian Journal of Economics* **101**(3), 337–61.

Uhlig, H. (1999) A toolkit for analysing nonlinear dynamic stochastic models easily. In R. Marimon and A. Scott (eds.), *Computational Methods for the Study of Dynamic Economies*, pp. 30–61. Oxford: Oxford University Press.

18

Monetary Policy, Beliefs, Unemployment and Inflation: Evidence from the UK

S.G.B. Henry

Abstract

Recent applied macroeconomic research has been concerned with the effects of both labor market reforms and the delegation of monetary policy to an inflation-averse central bank as ways of improving inflation and unemployment outcomes. The experiences of the UK following the introduction of changes to the labor market in the 1980s and of inflation targeting and instrument independence for the Bank of England in the 1990s, have often been held up as illustrations of the beneficial effects of regime changes of this sort. Others have contradicted these views, including those who have drawn attention to the weakness in the empirical evidence favoring effects from labor market reforms, and others who argue that a combination of beneficial international events and monetary policy mistakes have played an important part in the UK's recent economic improvement.

We review the case for regime change from either of these sources, labor market and monetary, in an application to the UK using a model that integrates both. The results indicate two things: the importance of allowing for the openness of the UK economy in "behavioral" econometric models of the natural rate, and the importance of allowing for policy "mistakes." Based on our analysis, we conclude that recent changes in UK monetary policy or labor market institutions seem unlikely to have made an important contribution to the improvements in UK economic performance. Effects originating overseas appear to play an important role in unemployment changes in the UK. Policy mistakes have had important effects on inflation over the last two decades, and a proper allowance for these is needed before any firm judgments of the benefits of the delegation of monetary policy can be reached.

18.1 Introduction: inflation and unemployment in the UK

Inflation in the UK over the last 25 years has varied substantially (Figure 18.1). It had a period of seriously high inflation for a decade starting from the early 1970s, with two distinct episodes each coinciding with a huge increase in oil prices.[1] Although this phase was brought under control towards the end of the 1980s, there followed a second – smaller – surge from then until the early 1990s. Although this second phase was less serious than each of the two peaks of the first, it was nevertheless important enough for the UK to join the Exchange Rate Mechanism (ERM) in an effort to control it.

From the date of its departure from the ERM, inflation, growth and unemployment have been unusually good by UK standards although, as Figure 18.2 shows,

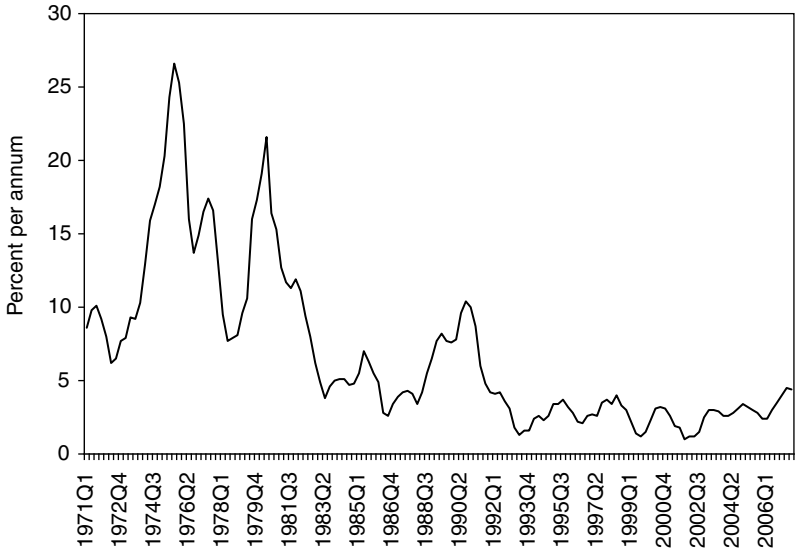


Figure 18.1 RPI inflation in the UK

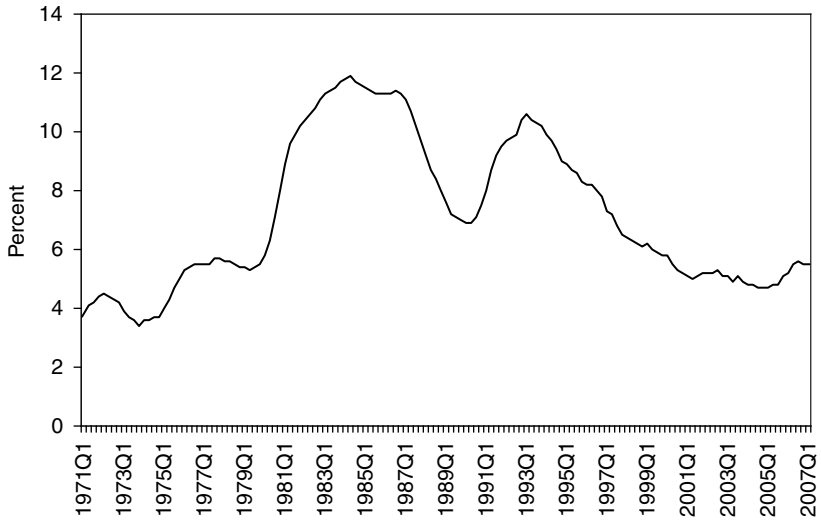


Figure 18.2 UK unemployment (ILO measure)

it was not until 1997 that unemployment fell to the level it reached before the UK joined the ERM.

Starting in 1992, monetary policy has been based on inflation targeting and, from 1997, the Bank of England (BoE) was delegated to set interest rates in pursuit of a preset inflation target (see BoE, 2007, for details). These changes have been heralded as decisive in achieving simultaneously low inflation and unemployment over the last 15 years, both in official circles (Balls and O' Donnell, 2002; BoE, 2007) and elsewhere (Cechetti, 2000). But, in a large US literature focused on testing for structural change in monetary policy, other explanations of improved inflation and growth performance have been suggested. Thus the importance of "good luck" – unusually benevolent world economic developments – has been cited, and yet other US research has emphasized policy mistakes, mainly due to uncertainties about the rate of productive potential and the natural rate (references to these and other parts of the US literature are given in section 18.2). Neither of these issues has received much attention in the UK, where there has been an even longer-standing debate on the possibility that there has been a decline in the UK non-accelerating inflation rate of unemployment (NAIRU) due to labor market reforms in the 1980s. These reforms were largely industrial relations changes to closed-shop arrangements and to procedures for settling industrial disputes, for example, but also included changes to the availability and duration of income out of work. Added to this set of possible alternatives another, which has recently surfaced in the UK, attributes an important and continuing effect on inflation and growth to the UK's membership of the ERM in 1989–92 (see Budd, 2004).

This chapter is directed at assessing some of these alternative explanations for the changes in inflation and unemployment in the UK since the early 1980s. It proceeds

by bringing together what have hitherto been treated as two distinct – and largely separate – possibilities for regime change: that the 1980s labor market reforms had significant effects on unemployment, and the possibility that important effects on inflation and unemployment followed from switching to inflation targeting and central bank independence. The first of these two possible sources of regime change refers to a large "shocks versus institutions" literature (see Blanchard and Wolfers, 2000, amongst others). The other relies on the pioneering research by Sargent (1999) which derives optimal monetary policy in a natural rate model where expectations are formed by recursive learning. As with other research on the policy model, section 18.5 considers to what extent a model of optimal monetary policy with learning conforms to the broad features of UK inflation since 1980. To do this, however, it amends the policy model in the light of econometric results on the determinants of long-run unemployment from section 18.4. These econometric results conclude that the evidence does not support the "wage push" interpretation of changes in long-run unemployment. Instead, it is shown that extensions to the unemployment model that give a major role to international factors, such as real oil prices and measures of international competitiveness, are needed for the model to capture the broad movements in UK unemployment. Once this extension is made to the policy model, it is found that optimal inflation solutions from it conform to the broad pattern of changes to inflation observed over the last 25 years or so without appealing to any regime change. As is evident in this account, the break-point for possible labor market regime changes is the early 1980s and monetary policy regime change is taken to be from the end of 1992 since, as is evident from references listed later, these are the alleged break-points in most of the labor market and monetary policy debates in the UK. Hence, we do not attempt any sort of estimation and inference about likely break-dates, which is an important, but separate, topic to what is presented here.[2]

18.2 A selection of background literature

18.2.1 A baseline New Keynesian policy model

It is helpful, pedagogically, to relate the principal papers referred to later to a "baseline" New Keynesian policy model (NKPM), since they can often be viewed either as a form of complete NKPM or as single equations from it, such as the aggregate supply (AS) or the policy rule for interest rates (the "Taylor rule"). The aggregate demand (AD) equation has received less attention in the literature, and that is the line adopted here.[3] The example below is the closed-economy model found in Henry and Pagan (2004). This baseline NKPM is given by equations (18.1)–(18.3), and microfoundations of these equations are discussed in, *inter alia*, Svensson (2000).

$$\pi_t = \alpha_\pi \pi_{t-1} + (1-\alpha_\pi)\pi^e_{t+1} + \alpha_y(y_{t-1} - y^*_{t-1})] + \xi_t \tag{18.1}$$

$$y_t = \beta_y y^e_{t+1} - \sigma(r_t - \pi^e_{t+1}) + \nu_t \tag{18.2}$$

$$r_t = \bar{r}_t + \beta(\pi^e_{t+1} - \pi^*) + \gamma(y_t - y^*_t). \tag{18.3}$$

Equation (18.1) is the AS (or Phillips curve) equation, where π_t is domestic inflation and $(y_t - y_t^*)$ is the output gap. As written, this is in the so-called "hybrid" form for the New Keynesian Phillips curve (NKPC), which depends on both lagged and future expected inflation rates. Equation (18.2) is the AD equation dependent on expected output and the real interest rate. The last equation (equation (18.3)) is a simple form of policy rule for interest rates, r, which is shown as depending on deviations of expected inflation from target and the output gap.[4] A significant difference in practice is how this equation is treated. First, it may be explicitly derived by optimising a dynamic objective function depending on government macroconomic objectives (inflation and output deviations from their equilibrium levels) subject to the constraints given by the model above, as in Ball (1997) and as in the models in sections 18.3 and 18.5 later, for example. But, most often the policy rule is taken as simply a reasonable description of the authority's behavior and is estimated; however, substantial problems can arise when it is estimated, and some of it is discussed next in a short review of US literature.

18.2.2 Evidence for and against monetary regime change in the US

Under the heading of the "Great Moderation," considerable effort has been directed at finding possible explanations of the marked reduction in the volatility of inflation and output in the US (BoE, 2007, draws attention to similar developments in the UK). It is probably fair to say that the results of this have been inconclusive, with some papers finding evidence for regime change in monetary policy whilst others have reported equally strong findings against. In part, this reflects different modeling approaches, as we illustrate immediately below.

There are now many examples of single equation estimates of both NKPCs and interest rate policy rules, both of which are directed at detecting changes in the effectiveness of monetary policy. They mainly assume that expectations are formed rationally. In the research on the NKPC, a major interest has been whether the degree to which the equation is forward-looking has increased, but this issue is largely unresolved. Thus, for example, using marginal costs rather than output gaps as the driving variable in the equation, Rudd and Whelan (2005) argue for the unimportance of forward-looking terms. In turn, Gali, Gertler and Lopez-Salido (2005) rebut this by observing that Rudd and Whelan use incorrect weights in forming the required forward-looking terms. Turning to the estimated policy reaction function (that is, the estimated version of (18.3) above), this has typically been of the form:

$$r_t = (1-\rho)\alpha + (1-\rho)\beta\pi_{t+n} + (1-\rho)\gamma x_t + \rho r_{t-1} + \varepsilon_t. \tag{18.4}$$

In this equation $\alpha = \bar{r} - \beta\pi^*$, where $\bar{r}$ is the long-run equilibrium nominal interest rate, π^* the target inflation rate and β is the weight on inflation deviations from target in the authority's objective function. The variable x_t is defined as $y_t - y^*$, and in (18.4) allowance is made for interest rate smoothing with a weight ρ. Henry and Pagan (2004) draw attention to a problem of the interpretation of such equations when they are used to infer what central banks' behavior has been. An example of such an interpretation is found in Clarida, Gali and Gertler (2000). Drawing such

inferences from equations like (18.4) is possible only if future expected inflation and output gaps were actually exogenous, which they are not. Complete system adjustments are needed to judge what the interest rate response to changes in either of the right-hand-side variables would be. This interpretative issue is addressed by Dennis (2004), who argues that judgments on the relative weights attached to inflation versus output stabilization require that central bank preferences be estimated, and this requires that all the parameters in the system be estimated.[5] A further econometric issue in these single-equation studies (both the NKPC and the estimated Taylor rules) is the problem of weak instruments. Often, large numbers of instrumental variables are used, and the risk is that the equations may be "overfitted," with predicted values being very close to actual values, with results that are close to ordinary least squares (OLS), as noted by Henry and Pagan (2004).

Moving to complete model estimates, there have been many studies using vector autoregressive models (VARs) and structural vector auto regressive models (SVARs) to evaluate the relative statistical contribution that changes in exogenous shocks versus changes in model structures play in accounting for the changes in output and inflation dynamics. Representative examples used here are Stock and Watson (2002) (henceforth SW) and Boivan and Giannoni (2003) (henceforth BG). SW provide decomposition results for the observed changes in inflation and output volatility in the US, using estimated reduced form VARs of structural models akin to (18.1)–(18.3) above,[6]

$$\tilde{X}_t = \Phi(L)\tilde{X}_{t-1} + u_t, \tag{18.5}$$

where $\tilde{X}_t$ is a four variable vector of gross domestic product (GDP) growth, inflation, the Federal Funds rate and the growth in commodity prices.[7] This latter equation is appended as an *ad hoc* equation, so the set of variables in (18.5), $\tilde{X}_t$, differs from those in the baseline NKPM (18.1)–(18.3) above. SW's VAR estimates are fourth-order VARs for two sub-samples of the period 1960–2001, pre- and post-1984Q1, and the empirical results are noteworthy in that they show that it was changes in the covariance matrix of the unforecastable components of the VARs that account for almost all of the changes in the observed volatility of output. BG also report unrestricted VAR estimates over two sub-samples divided at the end of the 1970s. Broadly speaking, their findings are in line with the results of SW in that BG's unrestricted VAR results show that, if anything, monetary policy effects appear weaker in the second sample.

However, comparative results from the SVARs reported by SW and BG give different conclusions as to the effectiveness of monetary policy over time in the US, although each use a related, but different, version of the NKPM to identify their structural VAR. Thus SW use *a priori* values of the slope of the AD curve, the slope of the AS curve and the weight on forward-looking inflation in the AS curve. When estimated over the two periods, their structurally identified decomposition of output variability implies that most of the reduction in the second period is accounted for by changes in the variability of shocks, not changes in monetary policy coefficients. Although the paper by BG also takes a model based on the closed-economy

version of the NKPM augmented by an *ad hoc* equation for commodity price inflation, their structural model is different. In their "stylized structural model" they use a variant of a dynamic AD function composed of all interest sensitive spending, so amalgamating consumption (the dynamics of which depend on habit persistence) with investment (with dynamics dependent on adjustment costs). Their main finding is that over the later sub-period monetary policy appears more stabilizing, a finding at variance with those reported by SW.

In sum, the evidence from the results of this short, but fairly representative, summary of single-equation and small-complete models of the NKPM sort shows there are findings supporting monetary regime effects in the US, but equally there are findings which reach the opposite conclusion. In spite of this somewhat inconclusive state of play, we take the view that much of the research reviewed here points to the importance of more emphasis on econometric testing. SVAR studies, in particular, have reached diametrically opposite conclusions, in part because different theoretical identifying restrictions are employed by different authors. More, rather than less, empirical testing is one possible way out of this impasse. Also, a lesson from the single-equation research is the probable gains of a more complete treatment of the system as represented by (18.1)–(18.3) above, since there are evident shortcomings in testing for regime change effects with single equations only.

18.2.3 The importance of uncertainty

Much of the research described so far, particularly that in section 18.2.2, has used the rational expectations hypothesis. But there is a burgeoning literature, which again is largely found in applications to the US, which emphasizes uncertainty about the effects of regime change, transmission mechanisms and shocks hitting the economy.

Among the alternative approaches under this heading, the first is the argument, primarily associated with Sargent (1999), emphasizing changes in governments' beliefs as central to inflation and unemployment behavior. In this approach, as in the rest of the literature reviewed here, information is assumed to be symmetric between private and public sectors, with neither side having an informational advantage.[8] In the Sargent model, the crucial assumption is that the authorities learn about the "true" economy over time. (This model is discussed more fully in section 18.3.) An emphasis on more general forms of uncertainty is found in the important work by Orphanides and associates on the effects of uncertain natural rates of unemployment, as in Orphanides (2001, 2002) and Orphanides and Williams (2006). A related development has considered the effects of uncertain rates of technical progress, including work on the effects of technology shocks on monetary policy performance, an example of which is given in Gali, Lopez-Salido and Valles (2003). Much of this line of analysis comes broadly under the "policy mistakes" heading. Lastly, there is the "bad luck" view, which figured in the previous section, and which argues that it was the volatility of exogenous, non-policy, shocks that were primarily responsible for the high volatility of inflation and growth in the 1970s and subsequent falls in these exogenous shocks that were responsible for the improved US performance in later decades. The paper by

Stock and Watson (2002), reviewed above, is an example of this, though it is also emphasized by Sims and Zha (2006), among others.

18.2.4 The effects of openness

It is also the case that the openness of the economy has become a major preoccupation in macroeconomics in general, and this is reflected in recent research on the NKPM, where there has been very considerable debate about the effects of changes in the nominal exchange rate and their "pass-through" into domestic inflation, and this debate continues. At one end of the spectrum comes the so-called "isomorphism" described by Clarida, Gali and Gertler (2001), where closed- and open-economy versions of the "canonical" model with complete pass-through are isomorphic, in the sense that the properties of the model and of its implied monetary policy rules are the same in each version. In recent contributions this characterization is rejected and incomplete pass-through is adopted (examples are Monacelli, 2003, and Engel, 1999, among others). Micro-explanations have been advanced for this incomplete pass-through, such as pricing to market (PTM) or the importance of non-traded goods in consumption, as have macro-explanations such as slow adjustment of prices at the consumer or importer level (see Devereux and Yetman, 2002, and, for UK applications, Balakrishnan and Lopez-Salido, 2002; Kara and Nelson, 2003; Herzberg, Kapetanios and Price, 2003; Batini, Jackson and Nickell, 2005). Unsurprisingly, allowing for non-unitary pass-through fundamentally alters the analysis of monetary policy in an open economy as compared with the closed economy.

In concluding this section, we note the input that this short review provides for the applications reported later in the chapter. Following the account of open-economy issues, an extension of the NKPM to the open economy case is called for; thus section 18.4 calls on the recent empirical research on international determinants of UK price margins contained in some of the work noted immediately above. Section 18.4 also builds on some of the contributions emphasizing model uncertainty, especially those of Sargent (1999) and Orphanides and Williams (2006), when analyzing the effects of monetary policy. The application to the UK reported in section 18.5 uses a model of the aggregate-supply equation that is econometric, following the comments made at the end of section 18.2.2. But, before moving to the application of model uncertainty to the UK case, the next section outlines the basic Beliefs model which underpins it.

18.3 Beliefs and monetary policy

18.3.1 Motivation

A brief outline of Sargent's (1999) model is that it is the Kydland and Prescott (1977) model coupled with the assumption of learning by the authorities, with the results being directed at accounting for variations in US inflation and unemployment. The fundamental assumption is that the authorities have a misspecified model of the Phillips curve, but update the parameters of this in the light of prediction errors. The remarkable finding in Sargent, and confirmed in other studies,

is that, even where the Phillips curve is taken to be static, the model's dynamics due to learning reveal a tendency for the economy to settle in a high-inflation equilibrium regime from which it occasionally "escapes" to occupy a low-inflation one. These "escapes," in turn, depend on an unusual sequence of shocks which move the economy from a sub-optimal high inflation but time-consistent (Nash) strategy, based on the misspecified view of the Phillips curve, to the neighborhood of the low-inflation optimal time-inconsistent strategy based on the "true" Phillips curve. Further analysis identifying the shocks which lead to "escapes" is found in Cho, Williams and Sargent (2002), and other applications from this by now large literature include Tetlow and von zur Muehlen (2001), Sargent, Williams and Zha (2004), McGough (2006) and Ellison and Yates (2007).

18.3.2 A basic learning model of monetary policy

The example of the learning model of monetary policy given below is a closed economy one. Open-economy issues are developed in section 18.4 and are used in the extension to the model in section 18.5. From here on, the approach will be referred to as the "Beliefs" model. The methodology of the Beliefs model is broadly in line with calibration exercises, which posit an AS equation and an assumed government objective function defined over unemployment and inflation, which is then optimized using the systematic part of inflation as the control variable.[9] It is thus a numerical optimal control exercise, the added complication being that the authorities are assumed to have a misspecified Phillips curve, but update their estimate of this according to a recursive updating procedure.

Sargent (1999) and Cho, Williams and Sargent (2002) describe a number of different dynamic versions of their Beliefs model, but it is the one which assumes a static Phillips Curve that will be used to motivate what follows.[10] The basic building block is a government characterized as setting monetary policy dependent upon (their) approximating (that is, misspecified) model of the economy, which is a non-expectational Phillips curve. The true data-generating mechanism, in contrast, is taken to be a vertical expectational Phillips curve in which the natural rate is assumed to be given.[11] The actual or "true" model of the economy used is:

$$u_t = u^* - \theta(\pi_t - \hat{\pi}_t) + \upsilon_{1t} \tag{18.6}$$

$$\pi_t = \hat{\pi}_t + \upsilon_{2t}. \tag{18.7}$$

Equation (18.6) is a natural rate Phillips curve (with u^* being the natural rate), and equation (18.7) shows that actual inflation is then a systematic part ($\hat{\pi}_t$) set by government, together with a random term υ_{2t}.[12] To obtain the policy rule for $\hat{\pi}_t$, the model assumes that the authorities have a perceived (misspecified) Phillips curve of the simple linear form:

$$u_t^p = \gamma_{0t} + \gamma_{1t}\pi_t + \varepsilon_t, \tag{18.8}$$

where u_t^p is perceived unemployment, and it depends on time-varying parameters.

The government's optimal rule for setting the systematic part of inflation is then derived by solving the following control problem:

$$MinE \sum_{t=0}^{\infty} \delta^t (u_t^2 + \pi_t^2), \tag{18.9}$$

where δ^t is the discount rate, using $\hat{\pi}_t$ as the control variable, subject to the authority's misspecified view of the Phillips curve, equation (18.8) above, and equation (18.7). Assuming that, in each period, the government believes its current estimate of the Phillips curve is correct, the optimization problem is also a static one, with the time-varying control rule:

$$\hat{\pi}_t = [-\gamma_{1t}/(1 + \gamma_{1t}^2)]\gamma_{0t}. \tag{18.10}$$

In the cited applications it is assumed that these parameters are updated sequentially using recursive least-squares with constant gain. As noted above, the optimization proceeds by assuming that, in each period. the government treats that period's uncertain parameters as if they were true and optimizes subject to that assumption. Tetlow and von zur Muehlen (2001) review this point and find that the properties of the Beliefs model are robust to wider classes of uncertainty.

Unlike the AS equation (18.1) in the "baseline" NKPM above, this one uses a static Phillips curve, both for the "actual" (18.6) and the "perceived" equation (18.8) where, in each, unemployment is the dependent variable and the "actual" Phillips curve is built on the assumption of a fixed natural rate, from which only inflation surprises produce temporary deviations. Dynamics enter through the distinction between these two Phillips curves, coupled with the crucial assumption of "learning" about the parameters of the perceived Phillips curve using recursive estimation.[13] In this example, these are defined by the equations

$$\gamma_{t+1} = \gamma_t + gP_t^{-1} X_t (u_t - \gamma_{0t} - \gamma_{1t}\pi_t) \tag{18.11}$$

$$P_{t+1} = P_t + g(X_t X_t' - P_t), \tag{18.12}$$

where γ_t is the column vector $(\gamma_{0t}, \gamma_{1t})'$, X_t the vector $(1, \pi_t)'$ and $g = 1 - \vartheta$, where ϑ measures the rate at which past information is discounted. P_t is the 2×2 precision matrix.

An important property of the model is that it depicts the government as pessimistic about the unemployment level needed to reduce inflation, but optimistic about the effect of higher inflation in reducing unemployment; they tend, therefore, to continue to pursue a high inflation policy, which is the basis of the self-confirming equilibrium (SCE) property of the model.[14] However, solutions of the model show that, even when the government is making this assumption, the time path of inflation can undergo abrupt changes, suddenly dropping from high rates to sustained low rates of inflation. Crucially, in this model this happens only because there is a special sequence of shocks which shifts the economy from the SCE of the Nash solution. These dynamics are a highly original way to use this

model, with its underlying assumption of a fixed natural rate and a static Phillips curve to account for the onset of periods of low inflation, and are defined as the most likely path that (government) beliefs will take if they deviate from their mean dynamics in a significant way.[15]

Two recent extensions to Sargent's model are related to the application reported later in section 18.5. Each alters what is taken to be the "true" Phillips curve used above by including further exogenous variables in it. In the first, Ellison and Yates (2007) extend the model to allow for an additional shock which, however, the government is assumed to perceive correctly. The second, by McGough (2006), is closer to the general procedure we follow. He extends the model of the natural rate to allow an effect from real oil prices (OIL), so that equation (18.6) is modified as:

$$u_t = u^* - \theta(\pi_t - \hat{\pi}_t) + \psi OIL_t + \nu_{1t}. \tag{18.13}$$

In this model, the government's approximating model does not have the "true" parameter on this OIL variable and so is:

$$u_t^p = \gamma_{0t} + \gamma_{1t}\pi_t + \varphi_t OIL_t + \psi_t . \tag{18.14}$$

Hence, the assumption is that the government believes that real oil prices may affect the level of unemployment but is unsure of the size of this effect. A related model to this is described and applied to the UK in section 18.5.

18.4 Long-run unemployment: evidence from wage and price equations

In section 18.5, the baseline Beliefs model set out in section 18.3 is extended by embedding in it a long-run unemployment equation estimated, using cointegration, on UK quarterly data. This will be our version of "actual" unemployment (equation (18.6) above) and, like that, uses an aggregate supply equation with unemployment as the dependent variable.[16]

Cointegration analysis is needed to estimate this unemployment equation since the data involved are non-stationary. Ignoring non-stationarities in the variables, and the possibility that cointegrating vectors may exist between them, leads to misspecification, even where the model is estimated in levels, as the cross-equation restrictions due to cointegration will be ignored. It follows that the policy inferences in our later applications are, to this extent, empirically based. In the rest of this section, we will refer only to the long-run results for the unemployment equation, in keeping with the static form of the AS equation of the Beliefs model.

18.4.1 Behavioral models of the labor market

The intention of the rest of the chapter is to integrate tests of the beliefs model, to be given in section 18.5, with the very active program of research by labor economists that uses "behavioral" models of the labor market. The term "behavioral" is used in the sense that the underlying models are derived from flow or stock models of

the labor market, in contrast with more statistical (or time series) models of unemployment, such as Ball and Mankiw (2002) for example. Using such behavioral models, when estimated as reduced-form unemployment equations, many labor economists have concluded that labor supply factors play a crucial role in determining long-term unemployment. For a recent affirmation, see Layard, Nickell and Jackman (2005) and the review of this book by Blanchard (2007). But the empirical claims that effects of labor supply-side or wage-pressure variables, such as unionization, income out of work and the rigor with which rules on this are applied, have been so important is challenged by other economists.[17] Among others, these critics include Madsen (1998), Oswald (1997), Henry and Nixon (2000) and Blanchflower (2007). Empirical reasons for not accepting the claims of the supply-side proponents are discussed in section 18.4.2. The alternative model suggested there is that, based on econometric evidence, the labor supply variables that have characterized much of UK research do not satisfactorily account for the large changes in unemployment since the early 1970s, so the alternative emphasizes other possible determinants of unemployment, based on effects from the external economy.

To make the present study comparable with existing research, the model of wage and price determination used here starts from the model described in Layard, Nickell and Jackman (2005) and Nickell (1998), the basic equations of which are a price equation based on markup pricing and a wage equation derived from union–firm bargaining, which are solved to give an unemployment equation.[18] Thus the price and wage equations (ignoring time subscripts) are:

$$p - w = z_p - \beta_2(p - p^e) \tag{18.15}$$

$$w = \gamma_2 p^e + (1 - \gamma_2)p - \gamma_1 u - \gamma_{11}\Delta u + z_w, \tag{18.16}$$

where p^e is expected prices. Solving (18.15) and (18.16) for u, ignoring the price surprise terms, gives a lagged equation in unemployment and the exogenous wage and price variables (z_p, z_w), that is,

$$u = \frac{\gamma_{11}}{\gamma_1}\Delta u_{-1} + \frac{z_p + z_w}{\gamma_1}. \tag{18.17}$$

This shows that in the long run (where unemployment is not changing, so $\Delta u = 0$), unemployment depends on z_w and z_p, the "push" or driving variables in the underlying wage and price equations respectively. Details on the components of each of these "push" variables are discussed next.

From this point on, the approach is to review the empirical role of each of a large set of wage and price "push" variables, taking those used in recent research on both the wage and the price equation as the starting point. Recent examples of wage "push" variables (where each is an element in the z_w term in (18.17) above) have included the terms of trade (*TT*), a measure of skill shortage (*Skill*), the tax and price wedge (*T*), the replacement ratio (*RR*), a measure of union power (*UP*), an index of industrial turbulence (*IT*), and the real interest rate (*r*) (see, for example, Nickell, 1998, for models of the UK). From the pricing side, recent open-economy

models of UK price formation have included secular trends in the price markup, real import prices and relative competitors prices (the world price of domestic GDP relative to home prices) as determinants of domestic prices (Herzberg, Kapetanios and Price, 2003). Also an empirical (*ad hoc*) effect has been found for real oil prices (Batini, Jackson and Nickell, 2005).[19]

18.4.2 An empirical assessment of the wage and price push variables

Empirical results using sets of the dozen or so contending variables in long-run unemployment equations are reviewed in the rest of this section. In all cases the data are for the UK, and are defined in a short appendix to this chapter. The section starts with a summary of earlier findings by Henry and Nixon (2000) and Henry and Kirby (2007), which both found that the case for using the wage "push" variables listed in the previous section was rejectable in terms of standard statistical criteria. These findings are summarized next and, since this is in the nature of a critique of previous research, uses a sample of quarterly data from 1964Q4 to 1992Q4, which is the sample period used by Nickell (1998), one of the leading proponents of the wage-pressure approach. Following that, the case for using "push" variables from the pricing side is summarized and this uses a sample of quarterly data from 1980Q1 to 2005Q1, as this is the period over which a monetary policy regime change is alleged to have happened, and is the period the application in section 18.5 is concerned with.[20]

The argument of the remainder of this section is that, via a process of testing for parameter stability and weak exogeneity, and using the concept of a minimal set of cointegrating variables, it is possible to arrive at a parsimonious model of long-run unemployment.

18.4.2.1 The wage push variables

Much existing empirical research on long-run unemployment in the UK has emphasized wage "push" variables (z_w in (18.17)) only, in effect treating the driving variables in the price equation as zero. This set of wage "push" variables has used up to seven separate regressors from the wage "push" side to try to account for changes in long-run unemployment. Others have argued that smaller sets of regressors perform better on the grounds of parameter stability of the resulting equation over different sub-samples and the (related) requirement that the regressors in such equations be weakly exogenous. Henry and Nixon (2000), for example, make a case for including only real oil prices, the terms of trade and real interest rates in the long-run unemployment equation.

More recently, Henry and Kirby (2007) reconsider the empirical results for the unemployment equation based on such a large set of regressors according to how the equation is estimated. The equation in question is as follows:

$$\ln u = \beta_1 TT + \beta_2 Skill + \beta_3 T + \beta_4 RR + \beta_5 UP + \beta_6 IT + \beta_7 r. \tag{18.18}$$

Two estimation methods have been used to test the wage "push" thesis: long-run equations that are solutions of autoregressive distributed lag (ARDL) equations,

as in Nickell and Bell (1995) and Nickell (1998); and estimates of a cointegrating equation using Johansen's maximum likelihood (ML) method, as in Nickell and Bell (1995). Henry and Kirby argue that neither satisfactorily explains long-run trends in unemployment in the UK as its proponents claim. Reasons for this are summarised below. First, however, ARDL estimates of the long-run equation are shown in Table 18.1.[21] For this test, the full sample period is 1964Q4 to 1992Q4 as in Nickell (1998). The variables are described in the appendix to this chapter: full definitions are found in Nickell and Bell (1995) and Henry and Kirby (2007).

Table 18.1 ARDL estimates of long-run unemployment equations

Const.	*IT*	*RR**	*TT*	*Skill*	*UP*	*T*	*r*
1964Q4–1992Q4							
−35.5	0.11	0.049	9.99	0.09	1.90	0.035	0.021
**	(0.6)	(1.4)	(2.2)	(3.0)	(2.0)	(2.7)	(2.1)
1964Q4–1992Q4							
−45.5	0.17	2.19	11.2	0.06	1.9	0.039	0.01
(4.0)	(1.2)	(1.5)	(2.89)	(2.1)	(1.82)	(2.7)	(1.5)
1964Q4–1989Q4							
−71.9	0.34	2.5	6.9	0.05	−0.6	0.067	0.01
(4.5)	(2.4)	(1.87)	(1.6)	(2.2)	(0.5)	(3.55)	(1.2)
1964Q4–1984Q4							
−65.1	0.14	0.4	−13.1	0.08	2.0	0.069	0.02
(5.6)	(1.6)	(0.4)	(1.8)	(2.6)	(2.2)	(4.5)	(2.6)

Notes: Each uses a maximum lag of 4 on each variable. *t*-statistics are in (.).
*Nickell and Bell (1995) report estimates of the parameter for this variable of 8.95 and 4.88 for the ARDL and Johansen estimates respectively and appear to use the log of the replacement ratio. The estimate above from Nickell (1998) is apparently its level. We use the former throughout.
**No *t*-statistic is given in the original.

The first equation is taken from Nickell (1998). It is clear that the remaining estimates reveal substantial variation in the estimated parameters and, in crucial cases, a complete turnaround in the significance of the estimates. This feature of parameter instability confirms what a number of critics have pointed out, namely that such wage-pressure variables are unlikely to account for long-run movements in unemployment since they were no worse in the mid 1980s than in the 1960s, yet unemployment still rose.

On a more methodological note, the use of ARDL to estimate long-run equations is not generally acceptable as it requires that the right-hand-side variables do not

themselves cointegrate, though tests in this case show they clearly do (see Henry, Kirby and Riley 2007, for further details). The right-hand-side variables should be weakly exogenous too, which, as discussed later, they are not. However, the problems with the model are not simply due to ARDL estimation, and the difficulty of treating the equation (18.18) as a long-run unemployment equation is not resolved by using an alternative such as the Johansen ML method to estimate a single cointegrating vector, normalized on unemployment, which is then treated as "the" unemployment relationship as in Nickell and Bell (1995). We review problems with the use of Johansen estimation with this dataset next. The purpose of this is not estimation, but to describe what would be required to estimate (18.18) so that it had a behavioral interpretation. This review, incidentally, provides on explanation for the parameter instability noted in Table 18.1.

As the data are mainly non-stationary (see below), the dynamic model underlying equation (18.18) can be written as a vector error correction model (VECM), with eight equations, one for each of the variables in (18.18), as illustrated next:

$$\Delta z_t = \sum_{j=1}^{p-1} \Gamma_j \Delta z_{t-j} + \gamma \alpha' z_{t-1} + \varepsilon_t. \tag{18.19}$$

Here z is a column vector of n variables, n being the eight variables (including the unemployment rate) from equation (18.18). The Γ_j $(j = 1, \ldots, (p-1))$ are a set of (8×8) matrices of parameters on the dynamic terms of the model, where the preset lag-length of the model is p. Attention is focused on the long-run part of the VECM, where γ and α' are the loading weights and cointegrating vectors respectively, and γ is nxr to reflect the reduced rank of the system, where it is implicitly assumed that there are $r < n$ cointegrating vectors in the model, and ε_t is a vector of white-noise error terms, with $\varepsilon_t \sim N(0, \Sigma)$.

Tests of orders of integration of the eight variables reported Table 18.2 reveal that one, *IT, is* $I(0)$ while the others are $I(1)$. In this set it appears there could be up to three cointegrating vectors. Tests for $r \leq 3$ give 32.8 (34.4) for the Johansen eigenvalue test (λ) but 75.2 (75.9) for the Johansen trace test (95% significance levels in brackets).[22] Tests of weak causality show that only the terms of trade (TT) and the tax wedge (T) are weakly exogenous (Table 18.3). In the light of these first-stage results on orders of integration and exogeneity tests, the implied model appears to be a five-equation conditional model plus a two-equation marginal model for TT and T, that is,

$$\Delta y_t = \sum_{j=1}^{p-1} \Gamma_{1j} \Delta z_{t-j} + \gamma \alpha' z_{t-1} + \eta_t \tag{18.20}$$

$$\Delta x_t = \sum_{j=1}^{p-1} \Gamma_{2j} \Delta z_{t-j} + \upsilon_t \tag{18.21}$$

where y_t and x_t are (5×1) and (2×1) column vectors of $I(1)$ endogenous ($\ln u$, *Skill*, *RR*, *UP* and *R*) and $I(1)$ weakly exogenous variables (TT and T), respectively.

Table 18.2 Tests of orders of integration, sample 1964Q4–1992Q4

Variable	*DF*	*ADF(4)*	*DF*	*ADF(4)*
ln u	−0.7	−1.3	−5.4	−4.0
IT	−3.0	−3.4	−8.7	−6.1
T	−2.3	−2.3	−13.5	−4.2
TT	−1.0	−0.3	−13.8	−4.9
Skill	−1.5	−3.3	−4.8	−5.7
RR	−0.2	−2.0	−3.3	−2.7
UP	−1.3	−1.5	−4.0	−3.6
r	−2.1	−2.4	−3.3	−2.7

Notes: The first two columns are for levels, the second two for first differences (ADF followed by ADF(4) in each case). 95% critical value is 2.9.

Table 18.3 Tests for weak exogeneity, sample 1964Q4–1992Q4

Variable	*Wald statistic*
ln *u*	21.9
IT	–
T	2.9
TT	6.4
Skills	22.2
RR	19.7
UP	17.1
r	22.4

Note: The relevant test statistic with three cointegrating vectors is $\chi^2(3)$ with 95% critical value of 7.8.

The stationary variable *IT* is assumed to enter in the Δz_{t-1} vector in a level form.[23] Γ_{1j} are (5×8) and Γ_{2j} are (2×8) matrices of parameters on the dynamic terms of all the variables.

To just identify the cointegrating vectors in the model, r^2 restrictions need to be accepted, with each of the r cointegrating equations having exactly r restrictions successfully applied, as demonstrated in Pesaran and Shin (2002).[24] Further, Wickens and Motto (2001) show that when the cointegrating vector α' in (18.20) is exactly identified, additional overidentication lies in successfully applying restrictions on the loading matrix γ.[25] By applying these procedures, it could, in principle, be possible to derive an equation like (18.18) from the conditional model (18.20), if all the required restrictions on the long-run vectors and the loading matrix needed to get from (18.20) to a single equation of the form of (18.18) were

successfully upheld. Then, estimates of the structural disturbances for the model for which the VECM (18.19) is the reduced form could be obtained, as shown by Wickens and Motto (2001). That is, the responses of unemployment to structural shocks, that is, e_t, defined as $B^{-1}\varepsilon_t$, where B is the matrix of contemporaneous coefficients in the structural model underlying (18.19), could be estimated.[26] In the light of the earlier results on weak exogeneity, such overidentifying restrictions on the loading matrix, in particular, are unlikely to hold.

Hence the status of single-equation estimates of (18.18), such as those given in Nickell (1998), for example, and repeated as the first equation in Table 18.1 above, is then unclear. It is hard to treat it simply as "the" long-run unemployment equation as claimed. Rather, it appears to be part of a fuller dynamic system which involves equations which could be interpreted as determining the real interest rate, movements in skill shortages and the union–non-union wage markup, amongst other things.

The purpose of this last exercise is not to suggest estimation of the full system underlying (18.19) as the way ahead to resolve this issue. Instead, it highlights the dangers of using large sets of potentially $I(1)$, and possibly jointly endogenous, variables if the intention is to estimate a single equation for long-run unemployment. Thus, one important conclusion from this exercise is to emphasize the importance of the approach by Davidson (1998) of determining an irreducible cointegrating (IC) equation. He recommends minimal cointegrating sets of variables as contenders for structural (that is, behavioral) long-run relations.

In what follows, a simpler alternative is proposed which places emphasis on external factors in accounting for the changes in unemployment over the last 25 years.

18.4.2.2 The price push variables

The long-run, or equilibrium, pricing equation which underlies most recent studies on the NKPC is:

$$P_t = \mu_t MC_t, \tag{18.22}$$

where:

$$MC_t = (1/\alpha)(W_t N_t / Y_t).$$

In this equation MC_t is nominal marginal cost, W_t is wages, N_t is employment and Y_t is real output. Assuming a Cobb–Douglas (CD) technology and a constant elasticity demand function, real marginal cost is (in logs):

$$mc_t - p_t = -\ln\alpha + s_{Lt}, \tag{18.23}$$

where s_{Lt} is the labor share. An extension is where technology is not restricted to be CD, and it may be shown that real marginal cost is then affected by the real price of imports (*RPM*) (see Bentolila and Saint-Paul, 1999). In turn, a long-run "equilibrium" price can de defined as P_t^*, where $P_t^* = \mu_t^* MC_t$ and *MC* is nominal (not real) marginal cost, and this equilibrium markup is likely to be time-varying (see Batini, Jackson and Nickell, 2005). Potential determinants of this varying markup

are measures of international price competition (*COM*) and the real price of imports (*RPM*) (see Balakrishnan and Lopez-Salido, 2002; Herzberg, Kapetanios and Price, 2003; Batini, Jackson and Nickell, 2005, amongst others). Other, more *ad hoc* additions which have figured in the empirical literature are also included, such as the real oil price (*OIL*) (see Batini, Jackson and Nickell, 2005; Stock and Watson, 2002; Henry and Nixon, 2000; Boivin and Giannini, 2003). The real exchange rate (*RXR*) has also been used in the wage equation on the grounds of capturing real wage resistance effects (Nickell, 1988), as well as being a way of extending the standard model to allow for internal and external balance (see Layard, Nickell and Jackman, 2005).

Equating the real wage and markup again leads to a long-run unemployment equation, this time of the form:

$$u_t = \beta_1 + \beta_2 COM_t + \beta_3 RPM_t + \beta_4 \tilde{z}_{wt} + \beta_5 RXR_t + \beta_6 OIL_t, \tag{18.24}$$

where $\tilde{z}_{wt}$ is now taken to be the price and tax wedge variable (T).[27]

18.4.2.3 Cointegration results for unemployment

The strategy adopted here is to take the set of possible determinants of unemployment from equation (18.24) and derive a parsimonious version of the long-run unemployment equation using cointegration methods.[28] From earlier findings it appears that there is only one variable (T) which survives as a potential determinant of long-run unemployment from the wage setting side and, in the light of this, the rest of this section considers the empirical case for the use of variables from the price-setting side reviewed above.

The complete set of variables used initially is given in (18.24). In this set there is probably one cointegrating vector. Tests for $r \leq 1$ give 47.4 (40.53) for the Johansen eigenvalue test (λ) and 120.1 (102.6) for the Johansen trace test (95% significance levels in brackets). The tests reject the hypothesis that $r \leq 2$ with $\lambda = 26.3$ (34.4) and *Trace* = 72.7 (76.0). Even so, it remains likely that sub-sets of these six variables also cointegrate, and we build on this idea in following the approach suggested in Davidson (1998) of selecting minimal sets of non-stationary variables as contenders for the long-run equation. This explores the results of dropping each variable (except unemployment) from the cointegrating vector, and the results can be summarized as follows. It is found that a minimal set comprising the two variables *COM* and *OIL*, together with unemployment, form a cointegrating vector (for $r \leq 1$; $\lambda = 40.3$ (22.0) and *Trace* = 51.4 (34.9)). Other contending sub-sets of the six either do not cointegrate or have inappropriate signs and other theoretical shortcomings (further details are given in Henry, Kirby and Riley, 2007).

The minimal cointegrating set selected is then:

$$u_t = 0.76 + 0.85\ OIL_t + 10.7\ COM_t. \tag{18.25}$$

(Tests for orders of integration of these variables are given in Table 18.4.) Although this is simple by design, this long-run relation enters significantly in the error correction model for changes in unemployment based on the just-identified VECM, with a negative loading factor of 0.53 and a t-statistic of 4.6. Tests also show that all

Table 18.4 Tests of orders of integration, sample 1980Q1–2005Q4

Variable	*DF*	*ADF*	*DF*	*ADF*
u	−0.24	−2.1	−2.6	−3.2
OIL	−2.2	−2.14	−8.8	−6.2
COM	−1.7	−2.0	−7.8	−4.1
RPM	−0.27	−0.62	−7.4	−4.7

Notes: See Table 18.2.

Table 18.5 Tests for weak exogeneity, sample 1980Q1–2005Q4

Variable	*Wald statistic*
u	10.9
OIL	0.5
COM	3.2

Note: The relevant test statistic with one cointegrating vector is $\chi^2(1)$ with a 95% critical value of 3.8.

variables except unemployment are weakly exogenous as required (see Table 18.5). It is this empirical model of long-run unemployment that is used in the application of the Beliefs model in the next section.

18.5 Applying the Beliefs model to the UK

This section brings together the positive policy model emphasizing the role of beliefs set out in section 18.3 and the econometric analysis of the UK natural rate from section 18.4. It outlines optimal control solutions from the Beliefs model when the crucial equation for long-run unemployment in the Beliefs model of section 18.3 (equation (18.6)) is replaced with the cointegrating equation (equation (18.25)) derived in the previous section. This econometric equation has important "exogenous" determinants of the natural rate which arguably account for its longer-term variation. In common with other examples of the Beliefs model, it takes a static model of the supply side, and dynamic implications are then due to the assumptions made about the authorities' uncertainty and their processes of learning. This change to the determination of long-run unemployment is profound in its effect on the dynamic properties of the model, which are described later. Uncertainties about the variation in the natural rate, due to changes in the external economy, are advanced as a principal explanation for changes in UK inflation since 1980.

18.5.1 The model

The econometric model of long-run unemployment (equation (18.25)) is taken to represent the "actual" model of the economy but, when solving the model, this is amended to conform more closely to equations used so far in this literature. Thus, the "actual" equation is as shown next:

$$u_t = u^{**} - \theta(\pi_t - \hat{\pi}_t) + d_1 W_{1t} + d_2 W_{2t} + \nu_{1t}, \tag{18.26}$$

where $W_1 = OIL$ and $W_2 = COM$, and the variable u^{**} is the long-run level of unemployment that obtains in the absence of effects from the external economy and inflation surprises. The cointegrating equation derived in section 18.4 (equation (18.25)) is thus taken to be (18.26) when inflation surprises are zero.[29] Inflation surprises are introduced when solving the model below, where the parameter θ in (18.26) is then set at unity, as is standard in the Beliefs literature. Continuing with the rest of the model, actual inflation is again:

$$\pi_t = \hat{\pi}_t + \upsilon_{2t}, \tag{18.27}$$

which gives actual inflation as the systematic part of inflation, $\hat{\pi}_t$, set by the authorities by optimizing (18.9) subject to their "approximating" unemployment model of the economy (18.28):

$$u_t^p = \gamma_{0t} + \gamma_{1t}\pi_t + \delta_{1t} W_{1t} + \delta_{2t} W_{2t} + \eta_t. \tag{18.28}$$

As is evident, it is assumed in the "approximating" model that the authorities know that unemployment is affected by a set of exogenous variables but have a misspecified form of their effect. As written, (18.28) assumes – as in the Sargent model – that the authorities also have a mistaken belief in an exploitable trade-off between inflation and unemployment. As is clear, the time-varying parameters of (18.28) are $(\gamma_{0t}, \gamma_{1t}, \delta_{1t}, \delta_{2t})$, which are again assumed to be updated using standard constant-gain recursive least squares formulae,

$$\xi_{t+1} = \xi_t + g P_t^{-1} X_t (u_t - \gamma_{0t} - \gamma_{1t}\pi_t - \delta_1 W_{1t} - \delta_2 W_{2t}) \tag{18.29}$$

$$= \xi_t + g P_t^{-1} X_t (u^{**} - \theta(\pi_t - \hat{\pi}_t) + \nu_{1t} - \gamma_{0t} - \gamma_{1t}\pi_t + (d_1 - \delta_1) W_{1t} + (d_2 - \delta_2) W_{2t}) \tag{18.30}$$

$$P_{t+1} = P_t + g(X_t X_t' - P_t), \tag{18.31}$$

where $\xi_t = (\gamma_{0t}, \gamma_{1t}, \delta_{1t}, \delta_{2t})'$, $X_t = (1, \pi_t, W_{1t}, W_{2t})$, and P_t is now (4×4).

With this set-up, the solution giving the authorities' optimal setting of the control variable π_t is:

$$\hat{\pi}_t = -\frac{\gamma_{1t}}{1 + \gamma_{1t}^2}(\gamma_{0t} + \delta_{1t} W_{1t} + \delta_{2t} W_{2t}). \tag{18.32}$$

In the remainder of this section, optimal control solutions of the model given by equations (18.9) and (18.26)–(18.32) are described, and possible interpretations of these are advanced as we proceed.

18.5.2 Some model implications

18.5.2.1 Model solutions

Sargent's (1999) model is recursive and is solved given initial values of the natural rate (u^*) (assumed constant), the government's discount rate, the parameter on inflation surprises in the actual Phillips curve (θ) and the variances of the two error processes in the model. These parameters are assigned the values 5% for the natural rate, 0.98 for the discount rate, -1 for θ and the variances of the errors are each taken to be 0.3. The dynamic solutions are calculated over 400 and 1,000 periods, and it is in the latter that the escapes are a prominent feature (see, for example, *ibid.*, Ch. 8, p. 108).

The exercise we report next differs from Sargent's in important ways. It takes the model described in section 18.5.1 and, given the focus on the period since 1980, is solved over 100 quarters using actual data for the two exogenous variables in the unemployment equation starting from 1980Q1.[30] In these solutions, initial values for the parameters of the "perceived" Phillips curve, equation (18.28), $\gamma_{0t}, \gamma_{1t}, \delta_{1t}$ and δ_{2t}, are required and are set at -0.6, -0.9, 2.0 and 10.0, respectively.[31] In each solution, the discount rate is set at unity, θ is minus unity and the gain parameter in the updating equations is fixed at 0.0275 to ensure comparability across the solutions. Stochastic solutions are generated by additive drawings from standard normal distributions.[32]

18.5.2.2 Inflation with a time-varying natural rate

The present exercise is a Beliefs model, but one in which mistakes by the authorities about the evolution of the natural rate as well as a belief in an inflation–unemployment trade-off each can account for inflation dynamics.[33]

In Figure 18.3 the authorities are depicted as remaining uncertain about the effects of external shocks on the natural rate and this, together with their mistaken view about the existence of a trade-off, gives the initial higher rates of inflation. The interactions of their revisions to all the parameters of the perceived Phillips curve are then fairly complex, as we describe below. The effects of these are that they adopt a looser monetary policy to start with in order to reduce unemployment more than is actually required. Subsequently, the parameter estimates are updated and the effects of these changes are that monetary policy is tightened, so bringing down inflation over the first four to five years before it rises modestly and then falls again.

The falls in inflation can be traced to the evolution of the governments' estimates of the parameters on all the variables in the authorities' unemployment equation. As in the other examples of the Beliefs literature cited so far, the authorities' misperception about inflation "surprises" also matter. But, as shown in Figures 18.4(a)–(e), among the parameters in the authorities' misperceived Phillips curve, δ_{1t} and δ_{2t}, each fall, γ_{0t} increases and γ_{1t} falls in absolute value (though it remains negative). These revisions largely account for the changes in inflation shown in Figure 18.3, and they appear to show that the authorities do not learn the full extent of the

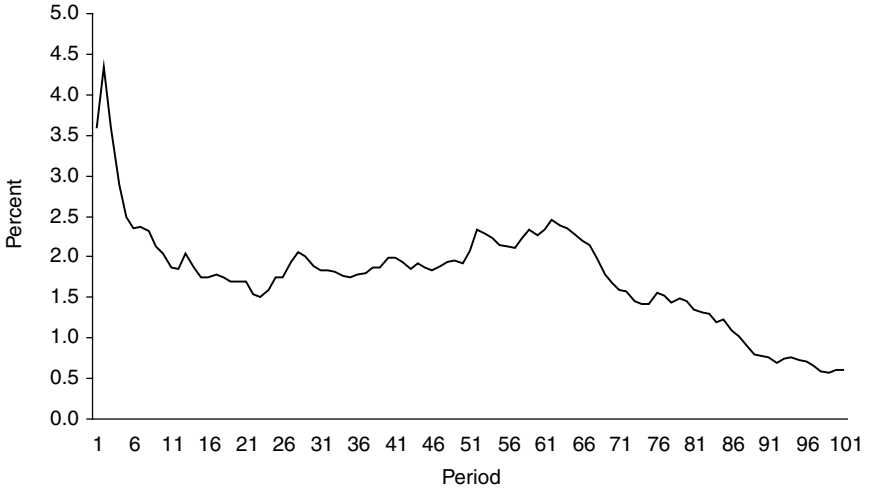


Figure 18.3 The first solution for inflation

dependence on unemployment of external shocks, but instead assume that the unemployment effect on inflation is increasing.

Inflation surprises, as shown in Figure 18.4(e), do not play a part in bringing about the reduction in inflation shown in Figure 18.3, in contrast to the "escapes" model of Sargent (1999), where unusual realizations of the noise processes are the sole explanation for switches between high and low inflation.

In keeping with properties of the Beliefs model, the present one also shows that the authorities' misspecified Phillips curve does not converge on the "true" Phillips curve, and the authorities' estimates of the parameters on the exogenous determinants of unemployment do not converge to their values in the "true" Phillips curve (as estimated separately). This finding is in line with Sargent's depiction of the solution being an SCE.

The next exercise considers what happens to inflation when the authorities are not aware of the effects of the exogenous variables at all, and the parameters on these ($\delta = (\delta_1, \delta_2)$) are set at zero and not updated. The resulting inflation dynamics are shown in Figure 18.5, which shows a significant increase in inflation. The interpretation of this solution advanced here is that it represents the effects of an optimistic view that the natural rate does not worsen in the face of adverse external shocks. Monetary policy is then set as if these adverse shocks had not happened. Below we argue that a parallel case to this is where the authorities believe that the trend rate of productivity growth in the economy has risen when in fact it has not, and this parallel motivates the comments on the economic events at the end of the 1980s made in the following section.

As in the previous case, there are substantial changes in the (γ_0, γ_1) parameters in the authorities' Phillips curve (see Figures 18.6(a) and (b)). This time, however, the authorities' beliefs about the sacrifice ratio (governed by their estimate of γ_1),

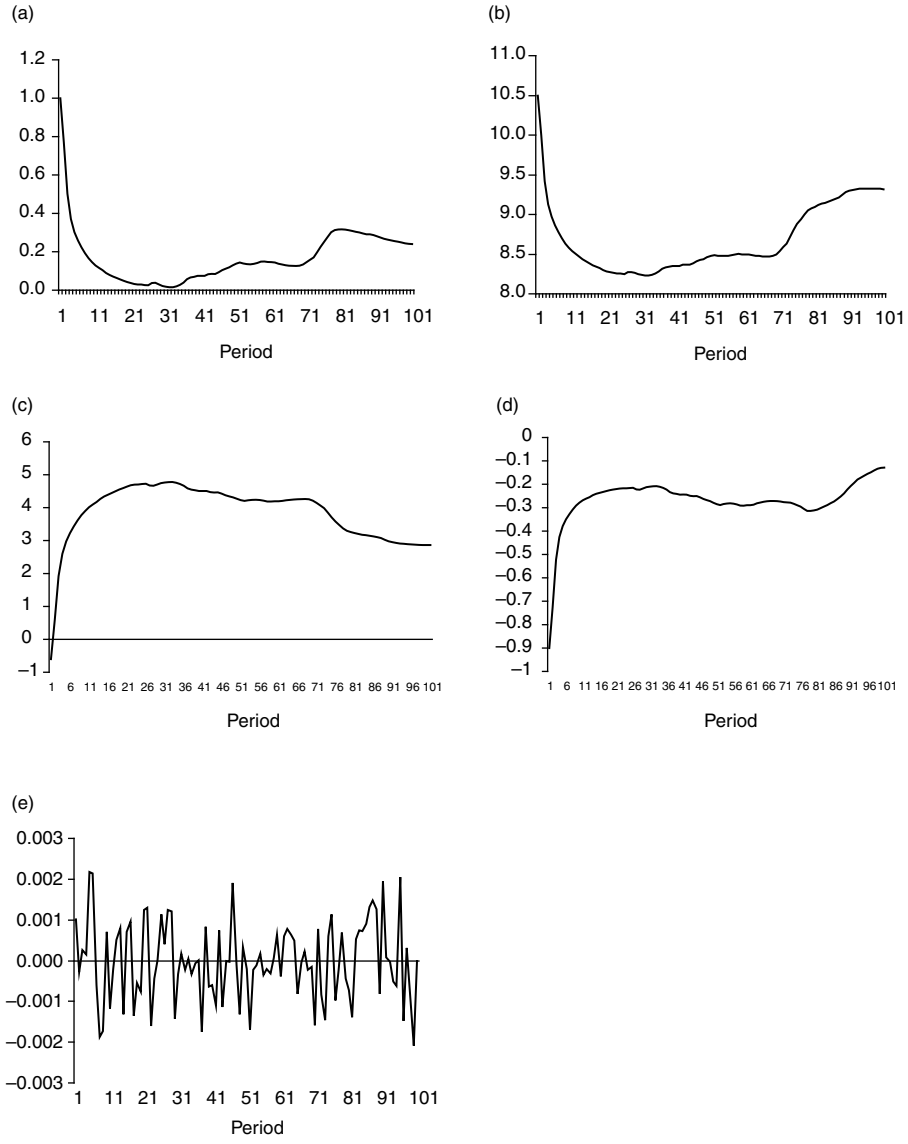


Figure 18.4 (a) δ_{1t}, (b) δ_{2t}, (c) γ_{0t}, (d) γ_{1t}, (e) Inflation surprises (v_{2t})

fluctuate around a fairly constant level, before changing sharply after about 75 quarters; these beliefs becoming more optimistic about the inflation consequences of reduced unemployment.

18.5.2.3 Relating the results to the UK since 1980

In keeping with this branch of the Beliefs literature, the main features of the dynamic solutions for inflation just given are related to events in the UK economy

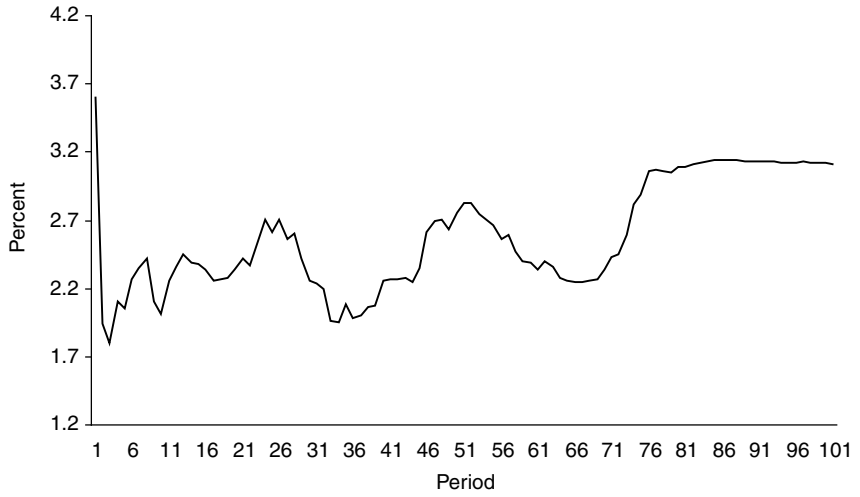


Figure 18.5 The second solution for inflation

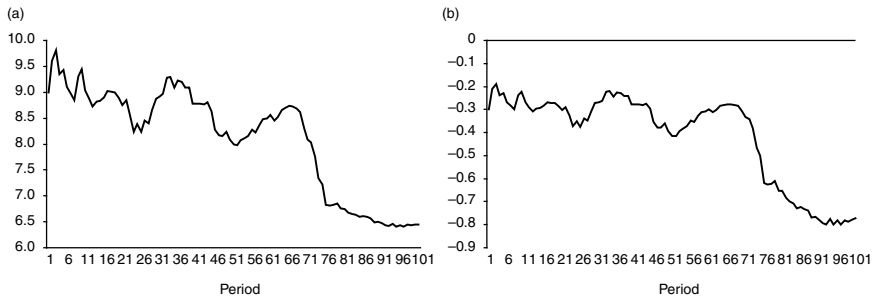


Figure 18.6 (a) γ_{0t}, (b) γ_{1t}

since the 1980s. Thus, starting at the beginning of the 1980s, UK output, unemployment and inflation were dramatically worsened by the 1979 oil price shock, but the effects of this were compounded by a significant switch in the 1980 budget to lower direct tax rates, aiming to offset this with increases in excise tax, petroleum revenue tax and increases in value added tax (VAT). The policy philosophy for control of inflation at this time was governed by the MTFS which was, in turn, predicated on targets for a wide monetary aggregate (£M3) and the (mistaken) belief in a predicable relation between changes in the monetary aggregate and inflation. The interpretation of these events offered here is that, because of its adherence to the MTFS, particularly its assumption that the natural rate was a given, the government expected the effect of the external shock to oil prices on unemployment to be smaller and shorter-lived than it was. In the event, it was not until 1986 that unemployment began to fall significantly. Inflation fell more quickly, and by 1983 was down to about 4%. The numerical solution of the

model shown in Figure 18.3 is broadly consistent with the events of the first half of the 1980s, and suggests that this can be seen as a high inflation outturn due to the authorities' mistaken belief about the natural rate, but revisions to the other parameters in the authorities Phillips curve lead to tighter monetary policy and inflation then falls to a relatively low rate after about three to four years.[34]

However, inflation rose markedly again at the end of the 1980s decade. The argument advanced here is that this further bout of higher inflation was due to yet another mistake by the authorities, this time about the probable trend rate of growth of productive potential in the economy. This mistaken assumption was in keeping with their view that a lower natural rate was to be expected following the "supply side" changes introduced by the government in the first part of the 1980s. The assessment of the evidence summarized in section 18.4 is that these alleged effects from the supply side did not happen. But official estimates of whole economy output per head show that in the period 1985–88 this was estimated to be at a rate of 3.25%, up from the 2% rate of growth for the period 1979–87 (FSBR, 1988, p. 30).[35] Due to this mistaken belief in a faster non-inflationary rate of growth, the Treasury appears to have significantly underpredicted inflation in 1989–90. Barrell, Khoman and Kirby (2007) report that the (retail price index) inflation forecasts made by the Treasury for 1988, 1989 and 1990 were, respectively, 4.5 (6.5), 5.5 (7.6) and 7.25 (10) (inflation outturns in brackets), each of which is evidence of a consistent and marked optimism about inflation. In later FSBR reports, in the face of a slowdown in the rate of output per head, the Treasury placed less emphasis on this trend improvement view. In other words, its belief in an improved productivity performance appears to have been short-lived. We argued above that this case is similar in its effects to the case shown in Figure 18.5, where the authorities are not aware that adverse external shocks on the economy have increased the natural rate and, as result of this mistake, inflation increases. The belief that potential growth is higher than it actually is would have a similar effect. But it is important to note that the solution shown in Figure 18.5 allows for permanent ignorance of the authorities about the effects of external shocks and inflation shoots up as a result. In reality this is too extreme. Thus, to relate this to the parallel case of a mistaken belief in an underlying productivity improvement in the late 1980s in the UK, it needs to be recognized that this belief soon evaporated. Even so, this temporary belief would have produced a spurt in inflation in line with what happened.[36]

18.6 Conclusions and proposals for future work

Evidently, both the empirical and the more theoretical material described in this chapter are limited. Thus, although we have sought to extend the policy model to allow for external (international) effects, in practice these are restricted to having effects only on the evolution of the natural rate. More direct transmissions of international changes in real commodity prices or international structural changes ("globalization") onto UK import price or consumer price inflation, for example, have been ignored and extensions to include some of these developments is an

urgent priority. The quantitative effects of future real oil price increases is another crucial area, with many arguing that future effects will not have the severe consequences that the previous oil price increases had on the international economy.[37] In equal measure, the theoretical model used here is clearly a limited one, and extensions to it, both to enlarge the model of the economy (including further extensions to open economy effects as just noted) and to give a more realistic rendering of monetary policy, are urgently needed. Important as these are, in our judgment, the outstanding item on the agenda is to extend the Phillips curve model used in the Beliefs literature, and in our examples above, to include nominal inertia in the inflation process.

With these important caveats in mind, there are nevertheless some important implications of monetary policy relevance which flow from what has been done here. The theme of the chapter is the central importance of uncertainty in determining the outcomes of monetary policy. The main uncertainty we have concentrated on is uncertainty in the model of the supply side the authorities use in forming their judgments about the appropriate settings for policy. Where it departs from the Sargent approach, which pioneered this research, is most obvious in its treatment of the evolution of the natural rate, where we have sought to combine his insights into the importance of the authorities' "learning" with the ongoing controversy in the UK about the determinants of the long-run movement in unemployment. On this latter point, our emphasis is on exogenous (and largely international) determinants of unemployment, against the prevailing model which places heavy emphasis on domestic labor supply-side factors such as income out of work and union strength. The argument in favor of our alternative is largely evidence-based; the standard model appears to fail conventional statistical tests, mainly because its putative determinants of unemployment would be consistent with little or no change in unemployment in the 1980s as compared with the 1960s when, in fact, actual unemployment rose substantially over this period.

Embedding this empirical model of long-run unemployment in a version of the Sargent model of monetary policy with learning, we suggest, illuminates the sequence of changes in inflation over the last 25 years, a part of which can be accounted for by the evolution of the natural rate, on the one hand, and "misperceptions" of it by the authorities, on the other. It portrays the decline in inflation by the mid 1980s as being a slow recovery from the oil price-induced inflation peak of 1979–80, as the authorities believed the natural rate was not significantly worsened by the shock. Inflation then fell back to reach quite low rates towards the end of the 1980s, so the next interpretive problem is to account for the rise in inflation at the end of the 1980s and early 1990s. In our account, this is portrayed as a further "policy mistake," as the so-called "Lawson boom" of the late 1980s was predicated on a perceived, but mistaken, increase in the economy's productivity trend.[38] We show that such a misperception of a decrease in the natural rate can indeed lead to hikes in the inflation rate. In this light, the subsequent UK membership of the ERM can be seen as an attempt at reducing inflation by importing credibility.[39] Broadly speaking, this interpretation agrees in some measure with the conclusion

reached by Budd (2004) that ERM membership contributed in an important way to the UK's inflation and unemployment record post-membership. According to the analysis here, the major movements in unemployment and inflation in the UK, heralded by some as due to the regime of inflation targeting and central bank independence, can, at best, be only part of the story. The analysis given earlier suggests that a mixture of external shocks and a slow process of recognition of the effects of these by the authorities may also have played an important part in the evolution of UK inflation. Occasionally, too, the economy is diverted by policy mistakes, such as our interpretation of the effects of the "Lawson boom" illustrates, and this conclusion concurs with much of the explanation of US behavior given by Orphanides (2001, 2002) and Primiceri (2005).

18.7 Appendix: Data – definitions and sources

Sample 1964Q4–1992Q4

TT. This is a terms of trade variable defined as $s \ln(P_m/P^*)$, where s is the import share in GDP, P_m is the import price index for the UK, and P^* is the unit value index of manufacturing exports in sterling.
UP. The log of the union/non-union markup, where the markup is a derived series as estimated in Layard, Metcalf and Nickell (1978).
RR. The replacement ratio (percentage) using a weighted average of different family types.
T. This is the tax wedge defined as the sum of the employment tax on firms, the aggregate direct tax rate and an aggregate indirect tax rate.
Skill. This variable is a measure of skill shortages faced by employers, derived from the Confederation of British Industry Industrial Trends Survey. It is the ratio of responses to the questions (i) limits on output due to skill labor shortage, (ii) limits on output due to other labor shortage.
IT. Industrial turbulence, defined as the absolute change in the proportion of employees in production industries as a proportion of total employees in employment.
r. The real interest rate, defined as the Treasury bill rate minus the rate of inflation in the GDP deflator.
ln*u*. Log of unemployment in the UK, males and females.

For sources of all the variables, see Nickell (1998).

Sample 1980Q1–2005Q4

Unemployment (*u*). ILO definition.
Real oil price (*OIL*). Brent spot price less the GDP deflator.
Real international prices (*COM*). Effective export prices for the G7 less GDP deflator.
Real import prices (*RPM*). Implicit price deflator for total imports less the GDP deflator.

Real exchange rate (*RXR*). Effective nominal rates, using the 2000 patterns of trade matrix, deflated by the consumer expenditure deflator.

Source: National Institute database.

Acknowledgments

Thanks are due to Ali Al-Eyd, Joe Pearlman and Malcolm Pemberton for their considerable help with the analysis in section 18.5. I would also like to thank Ray Barrell, Alan Budd, Richard Dennis, Mike Dicks, Stephen Hall, Richard Jackman, Joe Pearlman, Martin Weale and the editors of the Handbook, Terence Mills and Kerry Patterson, for their helpful comments on an earlier draft of the chapter. Simon Kirby provided crucial help at several stages in the preparation of the chapter. All the aforementioned are absolved from responsibility for the contents of this work, which is the sole responsibility of the author.

Notes

1. The inflationary increases of 1979–80 were also exacerbated by the switch from direct to indirect tax in the Budget. Further detail are given in section 18.5.2.4.
2. A recent clear review of this methodology is found in Perron (2006).
3. But see Kara and Nelson (2003) for an example.
4. In what follows, expectations are taken to be conditional on information available at period t.
5. Surico (2006) echoes this view when estimating the NKPC which, he argues, needs to allow for the interest rate reaction function actually in force at the time.
6. This uses the same notation as SW (2002) and follows their convention of ignoring the intercepts in the equations for notational convenience, though these are used in estimation.
7. The standard interpretation is that forward-looking expected variables enter in their lagged representation.
8. The so-called "timing protocol" is largely standard though, whereby the government first sets the systematic part of inflation and then the public form their expectation for inflation.
9. Perhaps the most sophisticated example of this class of studies is that of Svensson (2000).
10. This is largely standard in the cited examples. The present version draws on the Tetlow and von zur Muehlen (2001) account.
11. Note that throughout this chapter, the terms "natural rate" and "NAIRU" are used interchangeably. The switch from one to the other, though possibly confusing to the reader, is made to accord with what is used in the original papers.
12. This version of the AS equation, with unemployment as the dependent variable, is standard in the Beliefs literature. We will refer to it throughout the rest of the chapter as the Phillips curve, hopefully without risk of ambiguity.
13. There are some differences in the precise methods used in the literature. Primiceri (2005), for example, uses Kalman filtering.
14. See Ellison and Yates (2007) for the links between the model's mean dynamics, their stability conditions and the SCE.
15. To establish them a further optimal control problem is used where mean dynamics are perturbed and a weighting function is used that measures the likelihood of the shocks needed to perturb beliefs. For details of this, see Cho, Williams and Sargent (2002).
16. Section 18.5 discusses how the "inflation surprise" term is included in the model.

17. Note that we use the term "supply side" in a narrow sense to refer to the view that the labor market indices noted in the text have the predominant effect on the unemployment trend. This is an influential group in the UK. Elsewhere, economists such as Phelps, who emphasize the supply side, embrace a wider interpretation of supply side effects. See Fitoussi *et al.* (2000).
18. The use of a reduced-form unemployment equation is, in part, a reaction to the well known argument that the standard model of the labor market, such as that found in Layard *et al.* (1991), has a wage equation which is not identified. See Manning (1993) for a clear account of this. Hall and Henry (2006) argue that with non-stationary data this lack of identification is generally not found.
19. This accords with the inclusion of commodity prices in the SVARs estimated in Stock and Watson (2002) and Boivin and Giannoni (2003).
20. Results for longer samples are reported in Henry, Kirby and Riley (2007).
21. All equations in this sub-section use $\ln u$ as the dependent variable. For arguments favoring this, see Nickell and Bell (1995).
22. All tests of cointegrating rank in this chapter use asymptotic tests. For small sample corrections to these tests, see Johansen (2002).
23. See Wickens and Motto (2001) for a discussion of the treatment of $I(0)$ variables in the VECM.
24. These can be nonlinear restrictions.
25. This account ignores the possibility of using restrictions on the short-run dynamics, as our interest is in the long-run model only.
26. The structural model referred to here has unemployment and the remaining seven variables noted in the text as its contemporaneous variables, and not wages and prices. See note 18 for further comment on this point.
27. Full data definitions are given in the appendix to this chapter.
28. Parsimony is based on the concept of a minimal cointegrating vector introduced by Davidson (1998).
29. This treatment follows from the interpretation of equation (18.25) as a long-run unemployment equation.
30. We have also conducted long solutions of the same order as other studies, and these show the same sort of repeated escapes as reported by Sargent (1999) and McGough (2006), for example.
31. Al-Eyd *et al.* (2007) review some issues of the robustness of the findings with respect to these settings.
32. For v_{2t} the variance is scaled by the estimated variance of the acceleration in inflation over the period.
33. Arguably, UK governments at this time did not subscribe to the view that there was a trade-off. The adoption of a Medium Term Financial Strategy (MTFS) by the government in 1979 included a central assumption of a vertical Phillips curve. In spite of this, the model uses the assumption that government believes in a trade-off since, in practice, the government may have continued to act as if there was a trade-off in using the threat of higher unemployment consequences if "excessive" pay demands were accepted.
34. In this account, we are deliberately ignoring other transmission effects of oil price changes on inflation. This is in line with our treatment of these real shocks as impinging on unemployment only. But the partial nature of the account provided should be borne in mind throughout what follows.
35. The 3.25% for this period was partly a forecast.
36. A related issue is the effects of technology shocks on the performance of monetary policy. Recent US research has been directed at this issue too (see, for example, Gali, Lopez-Salido and Valles, 2003).
37. For an alternative view, however, see Nordhaus (2007).

38. It also involved great uncertainty about the measurement of the growth rates of the three measures of GDP at the time, though this plays no part in the analysis we give.
39. The argument at the time was that membership could reduce inflation at reduced unemployment cost due to the credibility gains inherent in "tying one's hands" – essentially the benefits of importing anti-inflation credibility from a then low inflation central bank (the Bundesbank). As unemployment rose to over 10% by 1992 and only reached its pre-entry rate by 1997, it is not clear that, in practice, these benefits actually accrued to the UK.

References

Al-Eyd, A., S.G.B. Henry, J. Pearlman and M. Pemberton (2007) Inflation and institutions: evidence from the UK. Unpublished manuscript, National Institute of Economic and Social Research (NIESR), London.

Balakrishnan, R. and D. Lopez-Salido (2002) Understanding UK inflation: the role of openness. Bank of England Working Paper No. 164.

Balls, E. and G. O'Donnell (eds.) (2002) *Reforming Britain's Economic and Financial Policy: Towards Greater Economic Stability*. London: HM Treasury.

Ball, L. (1997) Efficient rules for monetary policy. NBER Working Paper 5952, March.

Ball, L. and N.G. Mankiw (2002) The NAIRU in theory and practice. *Journal of Economic Perspectives* **16**, 115–36.

Bank of England (BoE) (2007) *The Monetary Policy Committee of the BoE: Ten Years on*. London: Bank of England.

Barrell, R., J. Khoman and S. Kirby (2007) Evaluating forecast uncertainty. *National Institute Review*. London: NIESR.

Batini, N., B. Jackson and S. Nickell (2005) An open-economy New Keynesian Phillips curve for the UK. *Journal of Monetary Economics* **52**, 1061–71.

Bentolila, S. and G. Saint-Paul (1999) Explaining movements in the labor share. CEMFI Working Paper No. 9905.

Blanchard, O. (2007) A review of Layard, Nickell and Jackman's "Unemployment: Macro Performance and the Labor Market." *Journal of Economic Literature* **XLV**, 410–18.

Blanchard, O. and J. Wolfers (2000) The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. *Economic Journal* **110**, C1–33.

Blanchflower, D. (2007) Trends in European labor markets and preferences over unemployment and inflation. Paper presented to the Dresdner Kleinwort Seminar on European Labor Markets and Implications for Inflation and Policy.

Boivin, J. and M. Giannoni (2003) Has monetary policy become more effective? NBER Working Paper No. 9459. Cambridge, Mass.: NBER.

Budd, A. (2004) Black Wednesday – a re-examination of Britain's experience in the Exchange Rate Mechanism. 34th Wincott lecture, Institute of Economic Affairs, London.

Cecchetti, S. (2000) Making monetary policy: objectives and rules. *Oxford Review of Economic Policy* **16**.

Cho, I.-K., N. Williams and T. Sargent (2002) Escaping Nash inflation. *Review of Economic Studies* **69**, 1–40.

Clarida, R., J. Gali and M. Gertler (2000) Monetary policy rules and macroeconomic stability: evidence and some theory. *Quarterly Journal of Economics* **115**, 147–180.

Clarida, R., J. Gali, and M. Gertler (2001) Optimal monetary policy in open versus closed economies: an integrated approach. *American Economic Review*, *Papers and Proceedings*, **91**(2), 248–52.

Davidson, J. (1998) Structural relations, cointegration and identification: some simple results and their application. *Journal of Econometrics* **87**, 87–113.

Dennis, R. (2004) Inferring policy objectives from economic outcomes. In S.G.B. Henry and A. Pagan (eds.), *The Econometrics of the New Keynesian Policy Model*. *Oxford Bulletin of Economics and Statistics* **66**, 735–64.
Devereaux, M.B. and J. Yetman (2002) Price setting and exchange rate pass-through: theory and evidence. Unpublished manuscript, University of British Columbia.
Ellison, M. and T. Yates (2007) Escaping volatile inflation. *Journal of Money, Credit and Banking* **39**, 980–94.
Engel, C (1999) Accounting for US real exchange rate changes. *Journal of Political Economy* **107**, 507–38.
Fitoussi, J., D. Jestaz, E.S. Phelps and G. Zoega (2000) Roots of recent recoveries: labour reforms or private sector forces? *Brookings Papers on Economic Activity* **1**, 237–311.
FSBR (1988) *Financial Statement and Budget Report* 1988–89. London: HMSO.
Gali, J.D. M. Gertler and D. Lopez-Salido (2005) Robustness of estimates of the hybrid New-Keynesian Phillips curve. *Journal of Monetary Economics* **52**(6), 1107–18.
Gali, J., D. Lopez-Salido and J. Valles (2003) Technology shocks and monetary policy performance: assessing the Fed's performance. *Journal of Monetary Economics* **50**, 723–43.
Hall, S.G. and S.G.B. Henry (2006) Identifying the wage equation: an expository note. Unpublished manuscript, NIESR, London.
Hall, S. and M. Wickens (1993) Causality in integrated systems. Discussion Paper No. 27-93. Centre for Economic Forecasting, London Business School.
Hendry, D. (1995) *Dynamic Econometrics*. Oxford: Oxford University Press.
Henry, S.G.B. and S. Kirby (2007) Unemployment in Britain: some more questions. NIESR Discussion Paper No. 293. London, NIESR.
Henry, S.G.B, S. Kirby and R. Riley (2007) A model of UK long-run unemployment. Unpublished manuscript, NIESR, London.
Henry, S.G.B. and J. Nixon (2000) Unemployment dynamics in the UK. *Oxford Economic Papers* **52**, 224–47.
Henry, S.G.B. and A. Pagan (eds.) (2004) *The Econometrics of the New Keynesian Policy Model*. *Oxford Bulletin of Economics and Statistics* **66**, Supplement.
Herzberg, V., G. Kapetanios and S. Price (2003) Import prices and exchange rate pass-through: theory and evidence from the United Kingdom. Working Paper No. 182. Bank of England.
Johansen, S. (2002) A small sample correction for the cointegrating rank in the vector autoregressive model. *Econometrica* **70**, 1929–61.
Kara, A. and E. Nelson (2003) The exchange rate and inflation in the UK. *Scottish Journal of Political Economy* **50**, 585–608.
Kydland, F. and E. Prescott (1977) Rules rather than discretion; the inconsistency of optimal plans. *Journal of Political Economy* **85**, 473–93.
Layard, R., D. Metcalf and S. Nickell (1978) The effects of collective bargaining on relative and absolute wages. *British Journal of Industrial Relations* **16**(3), 287–302.
Layard, R., S. Nickell and R. Jackman (2005) *Unemployment Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.
Madsen, J. (1998) General equilibrium macroeconomic models of unemployment: can they explain the unemployment path in the OECD? *Economic Journal* **108**, 850–67.
Manning, A. (1993) Wage bargaining and the Phillips curve: the identification and specification of aggregate wage equations. *Economic Journal* **103**, 98–118.
McGough, B. (2006) Shocking escapes. *Economic Journal* **116**, 507–28.
Monacelli, T. (2003) Monetary policy in a low pass-through environment. Working Paper No. 227, European Central Bank.
Nickell, S. (1988) The NAIRU: some theory and statistical facts. In R. Cross (ed.), *Unemployment, Hysteresis and the Natural Rate Hypothesis*, pp. 378–85. Oxford: Blackwell.
Nickell, S. (1998) Unemployment: questions and some answers. *Economic Journal* **108**, 802–16.

Nickell, S. and B. Bell (1995) The collapse in demand for the unskilled and unemployment across the OECD. *Oxford Review of Economic Policy* **11**, 40–62.
Nordhaus, W. (2007) Who's afraid of a big oil shock? *Brookings Panel on Economic Activity*, special edition, September.
Orphanides, A. (2001) Monetary rules based on real time data. *American Economic Review* **91**, 964–85.
Orphanides, A. (2002) Monetary policy rules and the great inflation. *American Economic Review*, *Papers and Proceedings*, **92**, 115–20.
Orphanides, A. and J.C. Williams (2006) Inflation targeting under imperfect knowledge. Working Paper 2006-14. Federal Reserve Bank of San Francisco.
Oswald, A.J. (1997) The missing piece of the unemployment puzzle. Inaugural Lecture, Warwick University.
Perron, P. (2006) Dealing with structural breaks. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics*. *Volume 1*: *Econometric Theory.* Basingstoke: Palgrave Macmillan.
Pesaran, M.H. and Y. Shin (2002) Long run structural modelling. *Econometric Reviews* **21**, 49–87.
Primiceri, G.E. (2005) Why inflation rose and fell: policymakers' beliefs and US postwar stabilisation policy. Unpublished manuscript, Department of Economics, Northwestern University, Evanston.
Rudd, J. and K. Whelan (2005) New tests of the New Keynesian Phillips curve. *Journal of Monetary Economics* **52**, 1167–81.
Sargent, T.J. (1999) *The Conquest of American Inflation*. Princeton: Princeton University Press.
Sargent, T.J., N. Williams and T. Zha (2004) Shocks and government beliefs: the rise and fall of American inflation. Working Paper 10764. Cambridge, Mass.: NBER.
Sims, C. and T. Zha (2006) Were there regime switches in US monetary policy? *American Economic Review* **96**, 54–81.
Stock, J.H. and M.W. Watson (2002) Has the business cycle changed and why? In M. Gertler and K. Rogoff (eds.), *NBER Macroeconomics Annual*. Cambridge, Mass.: NBER.
Surico, P. (2006) Monetary Policy Shifts and Inflation Dynamics. In D. Cobham (ed.), *Travails of the Eurozone: Economic Policy and Economic Development*, pp. 42–62. Basingstoke: Palgrave Macmillan.
Svensson, L.O. (2000) Open-economy inflation targeting. *Journal of International Economics* **50**, 155–83.
Tetlow, R. and P. von zur Muehlen (2001) Avoiding Nash inflation: Bayesian and robust responses to model uncertainty. Unpublished manuscript, Federal Reserve System, Washington, DC.
Wickens, M. and R. Motto (2001) Estimating shocks and impulse response functions. *Journal of Applied Econometrics* **16**, pp. 371–87.

Part VII
Applications to Financial Econometrics

This page intentionally left blank

19

Estimation of Continuous-Time Stochastic Volatility Models

George Dotsis, Raphael N. Markellos and Terence C. Mills

Abstract

This chapter reviews some of the key issues involved in estimating continuous-time stochastic volatility models. Such models have become popular recently because they provide a rich variety of alternative specifications which often lead to closed or semi-closed solutions in a variety of asset-pricing applications. An empirical comparison of various stochastic volatility models is also undertaken, along with a discussion of some directions for future research.

19.1 Introduction

It is now widely accepted that volatility in financial markets evolves stochastically over time.[1] The stochastic behavior of volatility has important implications for asset allocation, the pricing and hedging of derivative securities, prudent risk management and the behavior of financial assets in general. There are two central facets to the modeling of time-varying volatility. The first is the estimation of the model's parameters, the second is the filtration of latent volatility given these parameter estimates. The filtration of volatility is particularly important for applications such as option pricing, value-at-risk and portfolio allocation, all of which require volatility estimates. One popular approach that tackles both of

these facets is based on the ARCH/GARCH model introduced by Engle (1982) and Bollerslev (1986).[2] In this approach latent volatility is modeled as a deterministic function of past data available to the econometrician. The significant advantage of the GARCH approach is that empirical estimation can be implemented easily using quasi-maximum likelihood (QML) techniques.

A second class of time-varying volatility models are those termed "stochastic volatility": these are usually specified in continuous time and allow for a separate error process to drive the dynamics of volatility. Continuous-time stochastic volatility models have become fashionable over recent years as they allow a rich variety of alternative specifications. Moreover, stochastic volatility models offer closed or semi-closed solutions in many important asset-pricing applications. Unfortunately, the estimation of stochastic volatility models with discretely sampled data is particularly difficult because the likelihood function is not usually available in a tractable form. This intractability has fueled a significant research effort by financial econometricians. Continuous-time stochastic volatility models originate from the mathematical finance and option-pricing literature.[3] As one of the fathers of continuous-time finance, the late Fisher Black, remarked: "suppose we use the standard deviation of possible future returns on a stock as a measure of its volatility. Is it reasonable to take that volatility as constant over time? I think not" (Black, 1976).

The "official" year of birth of continuous-time stochastic volatility models may be taken to be 1987 as, in that year, Hull and White (1987), Johnson and Shanno (1987), Scott (1987) and Wiggins (1987) all developed option-pricing models with stochastic volatility. These models extended those of Black and Scholes (1973) and Merton (1973) by allowing volatility to follow a separate diffusion process. Scott (1987) and Wiggins (1987) are early attempts in estimating the parameters of the model using a method of moments approach. In this chapter we provide a selective review of some of the other popular methods that have been proposed over the years for estimating continuous-time stochastic volatility models.[4]

The need to estimate stochastic or time-varying volatility stemmed from the desire to explain and reproduce some of the stylized facts that have been observed in financial data:

- *Fat tails*. Since the early studies of Fama (1963, 1965) and Mandelbrot (1963), it has been well documented that asset returns are leptokurtic and violate the assumption of normality. Continuous-time models such as Merton's (1976) jump diffusion can generate non-normality and fat tails.
- *Volatility clustering*. In most financial markets we can observe episodes of high volatility interspersed by episodes of low volatility, so that large returns tend to be followed by large returns and small returns tend to be followed by small returns, irrespective of sign. In fact, one of the reasons for the huge success of GARCH modeling is that it provides a direct link between time-varying volatility, conditional heteroskedasticity and unconditional leptokurtosis. The implied clustering effect is depicted in Figure 19.1, which shows the daily returns of the Standard and Poor (S&P) 500 over the period 1990–2007.

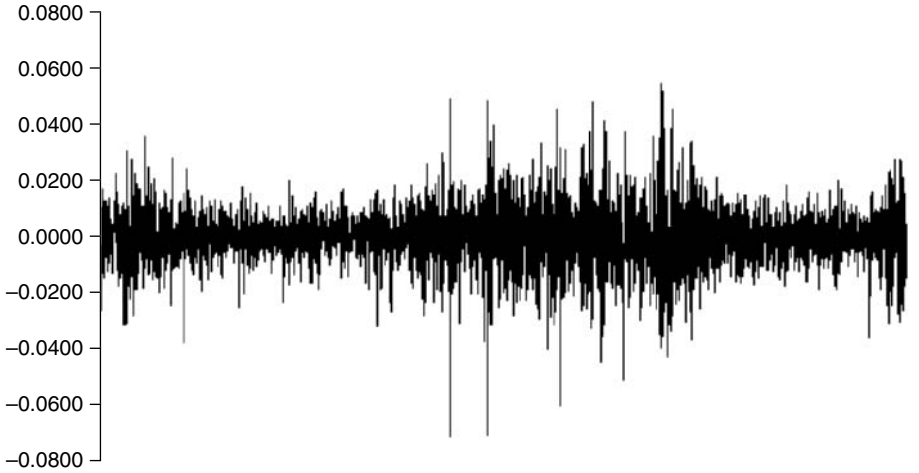


Figure 19.1 Logarithmic returns of the S&P500 over the period January 2, 1990, to December 31, 2007

- *Leverage effect*. Stock returns are negatively correlated with volatility, a phenomenon which Black (1976) coined the "leverage effect." When the stock price of a firm declines the leverage of the firm increases and hence the firm's price becomes more risky and volatile. Modified GARCH processes, such as the threshold-GARCH (TGARCH) model of Glosten, Jagannathan and Runkle (1993) and the exponential-GARCH (EGARCH) model of Nelson (1991), are designed to capture this leverage effect. However, many studies have shown that the asymmetric relationship between asset returns and volatility cannot be explained solely by leverage (for example, see Black, 1976; Christie, 1982; Schwert, 1989).
- *Information arrivals*. Information arrival is non-uniform through time. Clark (1973) linked asset returns to the arrival of information and was one of the first examples of stochastic volatility. The intuition here is that, when information arrival is non-uniform, randomness in business activity can generate randomness in volatility. Easley and O'Hara (1992) developed a market-microstructure model with time deformation that provided, amongst other things, a direct link between market volatility, trading volume and quote arrivals. In continuous-time finance there is a large literature that allows randomness in business time by using time-changed Lévy processes, which can generate stochastic volatility, fat tails and leverage effects (for example, Carr and Wu, 2004).
- *Volatility dynamics*. Stochastic volatility is usually assumed to follow a mean reverting process. Mean reversion in volatility is consistent with the clustering phenomenon and is also consistent with the economic interpretation of volatility as a measure of risk. It implies that volatility oscillates around a long run mean according to the speed with which it reverts to this mean level. The

mean reversion parameter has been found to alter dramatically at different data frequencies, suggesting that volatility may be driven by multiple factors (for example, Chacko and Viceira, 2003). Other studies (such as Bakshi, Ju and Ou-Yang, 2006) suggest that volatility displays nonlinear mean reversion with reversals at both high and low levels of the volatility spectrum. Empirical evidence shows that volatility displays so-called level effects (for example, Jones, 2003) whereby periods of high volatility usually coincide with periods of volatile volatility. Finally, recent studies suggest that volatility and asset returns display correlated jumps during times of market stress (for example, Eraker, Johannes and Polson, 2003).

- *Smiles, skews and implied volatility*. It has long been documented that the Black–Scholes (1973) model is not consistent with observed option prices. Given a set of option prices, one can invert the Black–Scholes formula and backout the implied volatility that sets the observed price equal to the Black–Scholes price. If the Black–Scholes model was correct, then, as a function of strike prices, implied volatility would be a flat line. However, it is well known that implied volatility displays a U-shaped pattern (the implied volatility "smile") in foreign exchange derivatives markets and a downward sloping curve (the implied volatility "skew") in index option markets (see Bates, 1996a). Continuous-time stochastic volatility models have been proposed as an alternative to Black–Scholes in order to explain the empirical patterns of implied volatility curves and smiles. Figure 19.2 depicts the evolution of daily S&P500 prices and the

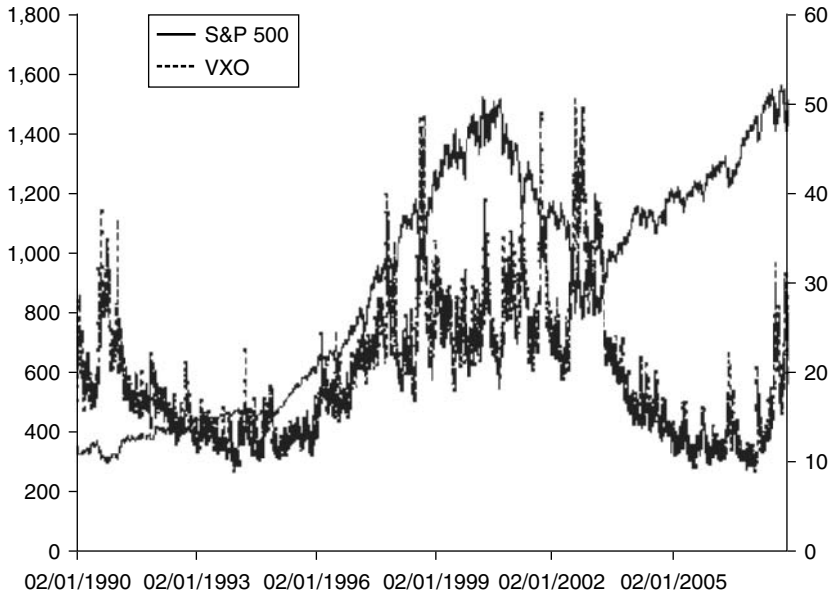


Figure 19.2 S&P500 prices (solid line) and VXO values (dotted line) over the time period January 2, 1990, to December 31, 2007

implied volatility VXO over the period 1990–2007.[5] The relationship appears to be negative, especially during times of market stress. Derivatives markets facilitate empirical estimation by providing an alternative source for backing-out latent volatility (for example, Pan, 2002; Aït-Sahalia and Kimmel, 2006).

The rest of the chapter is structured as follows. In section 19.2 we discuss the properties of some popular stochastic volatility models. Section 19.3 is devoted to the econometric methods used for drawing inferences in stochastic volatility models. First, we review estimation methods when volatility is treated as unobserved, such as efficient method of moments (EMM), Markov chain Monte Carlo (MCMC), and methods based on the empirical characteristic function (ECF). Subsequently, we discuss methods that incorporate information from the derivatives markets into the estimation procedure. Lastly, we review some recent methods that allow inferences to be made in stochastic volatility models using high frequency data. In section 19.4 we conduct an empirical comparison of various stochastic volatility models. Section 19.5 concludes and provides directions for future research.

19.2 Volatility specifications

We assume that the logarithms of an asset, $y_t = \ln(S_t)$, and its latent volatility, V_t, evolve over time according to the following general jump diffusion stochastic volatility model:

$$\begin{aligned} dy_t &= \left(\mu - \frac{1}{2}V_t\right)dt + \sqrt{V_t}dW_t^S + y_t^S dN_t^S \\ dV_t &= a(V_t, t)dt + \sigma(V_t, t)dW_t^V + y_t^V dN_t^V. \end{aligned} \tag{19.1}$$

Here, W_t^S and W_t^V are standard correlated Brownian motions. N_t^S and N_t^V are Poisson processes uncorrelated with these Brownian motions, with constant intensities λ_y and λ_v, and y_t^S and y_t^V are the jump sizes of the asset return and volatility, respectively. When $N_t^S = N_t^V$, jumps in asset returns and volatility occur simultaneously, and when $N_t^S \neq N_t^V$, the jump times are independent. The terms $a(\cdot)$ and $\sigma(\cdot)$ are the drift and diffusion functions of the volatility process, respectively. The parameter vector of (19.1) is denoted as Θ. For simplicity, we assume that the mean return of the asset, μ, is constant, although some models allow the conditional mean return to be a linear function of volatility, as in Merton (1980). In the late 1980s, stochastic volatility models were developed by assuming that the processes driving volatility and asset prices had continuous paths, that is, they followed diffusion processes. By the late 1990s, new types of stochastic volatility models had been introduced which were based on jump diffusion processes for the underlying asset price and, more recently, there have appeared stochastic volatility models that are founded on "double jump" processes, where both the underlying asset price and volatility follow jump-diffusion processes. In the following sub-sections we discuss various specifications of the $a(\cdot)$ and $\sigma(\cdot)$ terms that are nested within the general specification (19.1).[6]

19.2.1 Affine diffusions

The must popular stochastic volatility (SV) models are the so-called affine models. Broadly speaking, affine models are characterized by linearity of the drift and variance functions in (19.1) and provide computational tractability that leads to closed or semi-closed solutions in a variety of applications (see Duffie, Pan and Singleton, 2000). We examine the following affine specifications of the general stochastic volatility model in (19.1) when $N_t^S = N_t^V = 0$.

SV1
$$dV_t = k(\theta - V_t)dt + \sigma dW_t^V$$

SV2
$$dV_t = k(\theta - V_t)dt + \sigma\sqrt{V_t}dW_t^V.$$

In the SV1 model, also called the Ornstein–Uhlenbeck process, volatility follows a Gaussian mean-reverting process. The parameter k captures the speed of mean reversion, θ is the long-run mean of volatility and σ is the volatility of volatility. SV1 was first used by Vasicek (1977) for term structure modeling and has also been used for option pricing by Hull and White (1987), Scott (1987), Stein and Stein (1991), and Brenner, Ou and Zhang (2006), among others. This is the only stochastic volatility model for which the distribution of asset returns in (19.1) can be derived in closed form (see Stein and Stein, 1991). The conditional mean and variance of the SV1 model at time t for a time period $T > t$ is given by:

$$E_t\left(V_T\right) = \theta + \left(V_t - \theta\right)e^{-k(T-t)} \tag{19.2}$$

$$Var_t\left(V_T\right) = \frac{\sigma^2}{2k}\left(1 - e^{-2k(T-t)}\right). \tag{19.3}$$

Unfortunately, the SV1 model is not fully consistent with the empirical properties of volatility. In particular, SV1 implies that volatility can take negative values and that it is homoskedastic, which is evident from the fact that the conditional variance of the process does not depend on the level of volatility. Hence SV1 is not able to capture the positivity of volatility or the level effect. Consequently, despite the analytical tractability offered by this specification, it is no longer in common use.

SV2 was popularized after Heston (1993) and has been extensively used in a variety of applications. It was proposed as an alternative to SV1 that constrained volatility from taking negative values (see, for example, Bates, 1996b, 2000). Under this volatility parameterization the distribution of stock prices in (19.1) is not known in closed form but can be derived from the characteristic function (we discuss this methodology in section 19.3). The conditional mean and variance of the SV2 model are:

$$E_t\left(V_T\right) = \theta + \left(V_t - \theta\right)e^{-k(T-t)} \tag{19.4}$$

$$Var_t\left(V_T\right) = V_t\left(\frac{\sigma^2}{k}\right)\left(e^{-k(T-t)} - e^{-2k(T-t)}\right) + \left(\frac{\sigma^2}{k}\right)\left(1 - e^{-k(T-t)}\right)^2. \tag{19.5}$$

The conditional means of the two models are identical but, from (19.5), it can be seen that the conditional variance of SV2 depends on the level of volatility, thus making the process heteroskedastic.

19.2.2 Affine jump diffusions

Another popular class comprises the affine jump diffusion models, which incorporate a jump component in asset returns and/or volatility. Here we examine the case where jump times in asset returns and volatility occur simultaneously and jump sizes are correlated.

SV3
$$dV = k\left(\theta - V_t\right)dt + \sigma dW_t^V + y_t^V dN_t^V.$$

In SV3, $N_t^S = N_t^V$, the volatility jump size is drawn from an exponential distribution, $f(y^V) = \eta e^{-\eta y^V} 1_{\{y \geq 0\}}$, and the returns jump size follows the conditional distribution $y^S | y^V$: $N(\mu_S + \rho y^V, \sigma_S^2)$. The SV3 model has been used, for example, by Duffie, Pan and Singleton (2000), Eraker, Johannes and Polson (2003), Eraker (2004) and Broadie, Chernov and Johannes (2007). It allows for a rapidly moving and persistent factor to drive asset returns during times of market stress. Under this volatility parameterization the distribution of stock prices in (19.1) can again be derived from the characteristic function. The conditional mean and variance of SV3 are:

$$E_t\left(V_t\right) = V_t e^{-k(T-t)} + \theta\left(1 - e^{-k(T-t)}\right) + \frac{\lambda_v}{k\eta}\left(1 - e^{-k(T-t)}\right) \tag{19.6}$$

$$Var_t\left(V_T\right) = V_t\left(\frac{\sigma^2}{k}\right)\left(e^{-k(T-t)} - e^{-2k(T-t)}\right) + \left(\frac{\sigma^2\theta}{2k}\right)\left(1 - e^{-k(T-t)}\right)^2 + \frac{\lambda_v\sigma^2}{2k^2}\left(1 - e^{-k\tau}\right)^2\frac{1}{\eta} + \frac{\lambda_v}{k}\left(1 - e^{-2k\tau}\right)\frac{1}{\eta^2}. \tag{19.7}$$

19.2.3 Non-affine diffusions

Non-affine models are not particularly popular in the option pricing literature as they do not provide closed form formulae for option pricing. Neither the distribution nor the characteristic function of (19.1) can be obtained in closed form. However, these specifications have been widely used for econometric estimation purposes. Here we assume again that the dynamics in (19.1) do not incorporate a jump component.

SV4
$$dV_t = k\left(\theta - V_t\right)dt + \sigma V_t^{\gamma} dW_t^V$$

SV5
$$d\left(lnV_t\right) = k\left(\theta - \left(lnV_t\right)\right)dt + \sigma dW_t^V$$
$$dV_t = kV_t\left(\theta + \frac{\sigma^2}{2k} - \ln V_t\right)dt + \sigma V_t dW_t^V.$$

SV4 is also called the constant elasticity of variance process and has been used, for example, by Jones (2003) and Aït-Sahalia and Kimmel (2007). The functional form of the diffusion component implies that the model can easily accommodate the level effect. The conditional mean of SV4 generalizes those of SV1 and SV2, which are obtained by setting γ to zero and 0.5, respectively. However, the conditional variance does not have an analytical form and can only be derived using approximations.

SV5 is by far the most popular specification of discrete-time stochastic volatility models as it produces Gaussian log-volatilities (for example, Jacquier, Polson and Rossi, 1994; Kim, Shephard and Chib, 1998). In continuous-time settings, it has been used by Wiggins (1987) for the pricing of equity options, by Melino and Turnbull (1990) for the pricing of currency options, and by Detemple and Osakwe (2000) and Psychoyios, Dotsis and Markellos (2007) for the valuation of volatility options. Andersen, Benzoni and Lund (2002) and Chernov *et al.* (2003) estimated the model empirically. From the volatility levels process we can deduce, first, that the model accounts for the level effect of volatility and, second, that mean reversion depends on the level of V_t, that is, the larger V_t, the larger the mean reversion of the process. The first two conditional central moments are given by:

$$E_t\left(V_t\right) = V_t^{\exp(-k(T-t))} \exp\left(\left(1-e^{-\lambda t}\right)\mu + \frac{\sigma^2}{4\lambda}\left(1-e^{-2\lambda t}\right)\right) \tag{19.8}$$

$$Var_t\left(V_t\right) = V_t^{2\exp(-k(T-t))} \exp\left(2\left(1-e^{-k(T-t)}\right)\theta + \frac{\sigma^2}{2k}\left(1-e^{-k(T-t)}\right)\right) \times \left(\exp\left(\frac{\sigma^2}{2k}\left(1-e^{-2k(T-t)}\right)\right)-1\right). \tag{19.9}$$

19.3 Inference in stochastic volatility models

Although stochastic volatility models are built in continuous time, empirical data are only observed at discrete time intervals. Inference in stochastic volatility models is a difficult task because the likelihood function is typically not available in a tractable form. The Gaussian QML approach of Harvey, Ruiz and Shephard (1994) is appealing because of its simplicity, but many studies have shown that the method fails because volatility models are highly non-Gaussian.

Suppose that $y = \{y_1, \ldots, y_T\}$ is a discrete time series of observed data, $y_t = \log(S_t) - \log(S_{t-1})$, $t = 1, \ldots, T$, and $V = \{V_1, \ldots, V_T\}$ is the vector of latent stochastic volatility. We assume that y is stationary and that $x_t = (y_t, V_t)$ forms a Markov system. The system x_t can be treated as either fully or partially observed. If the former, latent volatility may be extracted from the derivatives markets, whereas, under the latter, volatility is treated as a latent variable that has to be integrated out of the likelihood function.

When $x_t$is fully observed the likelihood function for estimating the parameters is given by:

$$p(x;\theta) = p(x_1;\theta)\prod_{t=1}^{T-1} p(x_{t+1}|x_t;\Theta), \tag{19.10}$$

where $p(x;\Theta)$ is the joint density function and $p(x_1;\Theta)$ is the unconditional density. In many applications (for example, Aït-Sahalia and Kimmel, 2007) the density of the initial observation is omitted since, asymptotically, it does not affect the efficiency of the parameter estimates. Under this approach the difficulty lies in the fact that the conditional densities, $p(x_{t+1}|x_t;\Theta)$, of most stochastic volatility models cannot be derived in closed form. In principle, the transitional densities can be obtained numerically by solving the corresponding Fokker–Planck equation (for example, Lo, 1988). The conditional densities can also be obtained by Fourier inversion of the conditional characteristic function if the model belongs to the affine class or by other approximating methods.

When volatility is not directly observed it has to be integrated out of the likelihood function. In this case, the corresponding marginal likelihood function is given by:

$$p(y;\theta) = \int_{R^T} p(y, V;\Theta)dV = \int_{R^T} p(y|V;\Theta)p(V;\Theta)dV. \tag{19.11}$$

The dimension of the integral depends on the sample size, so that direct evaluation of the likelihood function is difficult. Moreover, when latent volatility is integrated out, the asset price alone is non-Markov due to the presence of correlation and the conditional density of the next period's return depends on the entire set of previous observations, $Y_{t-1} = \{y_{t-1}, \ldots, y_1\}$. The likelihood can be approximated by simulation methods such as MCMC or estimation can be implemented by avoiding the likelihood altogether and resorting to methods of moments estimation such as EMM or generalized method of moments (GMM).

In the next sections we discuss estimation methods when volatility is treated as unobserved and we then extend the analysis to the case where volatility is extracted from option prices.

19.3.1 Simulation-based inference

19.3.1.1 Efficient method of moments

The EMM, developed by Bansal *et al*. (1993, 1995) and Gallant and Tauchen (1996), is an extension of the simulated method of moments (SMM) of Duffie and Singleton (1993). EMM avoids direct computation of the likelihood function and resorts to efficient estimation via GMM and a cautious selection of the moment conditions. The parameter estimates are minimum chi-squared estimators and the optimized chi-squared criterion can be used to evaluate the statistical fitting of the various stochastic volatility models. Hence a significant advantage of the EMM procedure is that it allows empirical comparison between non-nested specifications, such as, for example, SV2 and SV5. However, EMM is computationally demanding and, as with all GMM procedures, it depends on the optimal choice of moment conditions. Estimation of stochastic volatility models with EMM has been undertaken by

Gallant, Hsieh and Tauchen (1997), Andersen and Lund (1997), Andersen, Benzoni and Lund (2002), Chernov and Ghysels (2000) and Chernov *et al.* (2003), while Andersen, Chung and Sørensen (1999) explore the efficiency of EMM estimators in a Monte Carlo study.

The starting point of the EMM procedure is to approximate the conditional density of the data as closely as possible with a discrete-time auxiliary model. This auxiliary model is not related to any particular stochastic-volatility model, its purpose being to capture the probabilistic structure of the data and to provide the moment conditions that must be satisfied by a postulated stochastic-volatility model, which in EMM applications is called the structural model. The auxiliary model usually involves an ARMA-GARCH specification and a semi-nonparametric density based on Hermite polynomials.

Suppose that the conditional density of the auxiliary model is $f_{\mathrm{K}}(y_t|Y_{t-1};.a)$, where α is the parameter vector of the auxiliary model. The parameters can be estimated by QML as follows:

$$\tilde{a} = \arg\max \frac{1}{n} \sum_{t=0}^{n} \log \left[f_{\mathrm{K}}(y_t|Y_{t-1};a \right]. \tag{19.12}$$

The ML estimates $\tilde{a}$ ensure that the quasi-score function satisfies the first-order conditions:

$$\frac{1}{T} \sum_{t=0}^{n} \frac{\partial}{\partial a} \ln f_{\mathrm{K}} \left(y_t|Y_{t-1};\tilde{a} \right) = 0. \tag{19.13}$$

In the second stage, EMM uses the expectation of the score functions of the auxiliary model as the moment conditions that deliver the estimates Θ of the structural model. The expectation is taken under the probability measure, $P\left(Y_t;\Theta\right)$, of the structural model:

$$m\left(\theta,\tilde{a}\right) = E^P \left[\frac{\partial \ln f_{\mathrm{K}}(Y_t;\tilde{a})}{\partial a} \right] = \int \frac{\partial \ln f_{\mathrm{K}}(Y_t;\tilde{a})}{\partial a} dP(Y_t;\Theta). \tag{19.14}$$

Given a set of parameters Θ, the expectation is calculated numerically, using a long simulated series, $\hat{y}_N(\theta)$, from the structural model, as:[7]

$$\tilde{m}\left(\theta,\tilde{a}\right) = \frac{1}{N} \sum_{t=1}^{N} \frac{\partial \ln f_{\mathrm{K}}(\hat{y}_t(\theta)|\hat{Y}_{t-1}(\theta);\tilde{a})}{\partial a}, \tag{19.15}$$

and, as $N \to \infty$, $\tilde{m}\left(\theta,\tilde{a}\right) \to m\left(\theta,\tilde{a}\right)$. The EMM estimator is then obtained by minimizing the quadratic function:

$$\hat{\Theta} = \arg\min_{\Theta} m_N(\Theta,\tilde{a})' W_T m_N(\Theta,\tilde{a}), \tag{19.16}$$

where the weighting matrix W_T is a consistent estimate of the inverse asymptotic covariance matrix of the auxiliary score function. Gallant and Tauchen (1996) show how to derive the weighting matrix and prove that the parameter estimates of the structural model are asymptotically normally distributed. Latent volatility can be filtered out using the reprojection method of Gallant and Tauchen (1998).

19.3.1.2 Markov chain Monte Carlo

The Bayesian Markov chain Monte Carlo method was initially applied by Jacquier, Polson and Rossi (1994) for the estimation of discrete-time stochastic volatility models. However, the method has also become very popular in the estimation of continuous-time models. Johannes and Polson (2006) provide an excellent survey of MCMC applications in a variety of continuous-time asset-pricing models. In the context of stochastic volatility, MCMC has been applied by, for example, Jones (2003), Eraker, Johannes and Polson (2003) and Eraker (2004).

The output of the Bayesian MCMC is the posterior density, $p(V, \theta|y)$, of the parameters and latent variables conditional on the data. The method quantifies parameter uncertainty and model risk, filters out latent volatility, jump times and jump sizes, and avoids optimization routines. However, under the MCMC it is difficult to make comparisons across non-nested models such as SV2 and SV5.

Under the Bayesian approach the posterior density $p(V, \theta|y)$ is proportional to:

$$p(y|V, \Theta)p(V|\Theta)p(\Theta), \tag{19.17}$$

where $p(y|V, \Theta)$ is the full likelihood, $p(V|\Theta)$ is the density of the latent variable, and $p(\Theta)$ is the prior density of the parameters. Sampling from the posterior density is difficult because of the latent variable and, especially, because of the high dimension of the density. The MCMC method samples from the posterior by forming a Markov chain over Θ and V that converges in distribution to $p(V, \theta|y)$. In practice, the posterior is further decomposed into the two conditional densities, $p(V|y, \Theta)$ and $p(\Theta|y, V)$ (see Johannes and Polson, 2006). The two most popular MCMC algorithms for sampling from the two conditionals are the Gibbs sampler and Metropolis–Hastings. If the conditionals can be sampled directly then the MCMC is performed with the Gibbs sampler; otherwise Metropolis–Hastings is applied.

19.3.2 Characteristic function methods

In stochastic volatility models the conditional density is rarely available in closed form. However, for models that belong to the affine class it is feasible to derive the conditional characteristic function. The advantage of the characteristic function methodology is that usually it does not require discretization of the continuous-time process. The characteristic function is a powerful tool that encodes the same information as the conditional density. Though the characteristic function solves the same Kolmogorov forward and backward equations as does the conditional density, the boundary condition for the characteristic function is smoother and this allows the derivation of the characteristic function in a tractable form.[8]

The joint characteristic function conditional on current stock price and volatility is defined as:

$$\phi\left(i\omega_1, i\omega_2, y_t, V_t, \tau; \theta\right) = E\left[e^{i\omega_1 y_{t+\tau} + i\omega_2 V_{t+\tau}} | x_t\right]. \tag{19.18}$$

To keep notation simple, we concentrate on specification (19.1) without jumps. The characteristic function of the process must satisfy the following partial differential

equation (PDE):

$$\frac{1}{2}V_t\frac{\partial^2\phi}{\partial y_t^2}+\rho\sqrt{V_2}\sigma(V_t;\theta)\frac{\partial^2\phi}{\partial y_t\partial V_t}+\frac{1}{2}\sigma(V_t;\theta)^2\frac{\partial^2\phi}{\partial V_t^2}+(\mu-\frac{1}{2}V_t)\frac{\partial\phi}{\partial y_t}+a(V_t;\theta)\frac{\partial\phi}{\partial y_t}-\frac{\partial\phi}{\partial\tau}, \tag{19.19}$$

subject to the boundary condition $\phi\left(i\omega_1, i\omega_2, y_t, V_t, 0, \theta\right) = e^{i\omega_1 y_T + i\omega_2 V_T}$.

When the stochastic volatility model belongs to the affine class, the solution of the joint conditional characteristic function has a simple exponential form, given by:

$$\phi\left(i\omega_1, i\omega_2, y_t, V_t, \tau; \theta\right) = e^{i\omega_1 y_t + A(i\omega_1, i\omega_2, \tau;\theta)V_t + B(i\omega_1, i\omega_2, \tau;\theta).} \tag{19.20}$$

The coefficients $A(i\omega_1, i\omega_2, \tau;\theta)$ and $B(i\omega_1, i\omega_2, \tau;\theta)$ can be derived by solving complex valued Ricatti equations. Duffie, Pan and Singleton (2000) show analytically how to derive the characteristic function for general affine diffusion-jump diffusion stochastic volatility models. Knowledge of the characteristic function also allows the derivation of non-central moments and cross-moments of affine continuous-time stochastic models via simple differentiations.

If volatility is treated as observed then, in principle, the transition density can be derived by Fourier inversion as:

$$p(y_{t+\tau}, V_{t+\tau}|y_t, V_t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(i\omega_1, i\omega_2, y_t, V_t, \tau;\theta\right) \times e^{-i\omega_1 y_{t+\tau} - i\omega_2 V_{t+\tau}} d\omega_1 d\omega_2. \tag{19.21}$$

Estimation with the characteristic function has been applied in latent stochastic volatility models by Singleton (2001), Jiang and Knight (2002), Chacko and Viceira (2003) and Bates (2006). Chacko and Viceira (2003) derive from (19.20) the conditional characteristic function of the asset, $\phi\left(i\omega_1, 0, y_t, V_t, \tau;\theta\right)$, and then integrate out the latent volatility to obtain, in closed-form solution, the characteristic function conditional only on the current asset price, $\phi\left(i\omega_1, 0, y_t, \tau;\theta\right)$. They then estimate various stochastic volatility models using Hansen's (1982) method of moments. However, the estimates are not fully efficient because the only condition is on the current stock price and, as mentioned previously, when volatility is integrated out the stock price on its own is no longer Markov. Jiang and Knight (2002) use iterated expectations and derive the joint unconditional characteristic function up to $t-L$ observations,

$$\phi\left(i\omega_1, i\omega_2, \ldots i\omega_L, y_t, 0.\tau, \theta\right) = E\left[e^{i\omega_1 y_{t-1}, i\omega_2 y_{t-2}, \ldots i\omega_L y_{T-L}}\right].$$

In practice, the joint unconditional characteristic function is derived for a relatively small L. Jiang and Knight (2002) estimate a stochastic volatility model using GMM with non-central and cross-moments. Singleton (2001) and Bates (2006) achieve full efficiency by conditioning on the full history of returns. Singleton proposes a method that combines the simulated method of moments of Duffie and Singleton (1993) with the characteristic function technology. Bates (2006) develops a direct filtration-based maximum likelihood methodology using characteristic functions

and estimates various diffusion/jump diffusion stochastic volatility models. His method does not require any simulations and, in contrast to previous approaches, also allows the filtration of latent volatility.

19.3.3 Derivatives markets

The development of derivatives markets provides an alternative source for filtering latent volatility and estimating the parameters Θ. Some studies use only information from the derivatives markets, whereas others use both option prices and historical returns.

For affine diffusion/jump diffusion stochastic volatility models, European option prices can be obtained in closed form, up to numerical integration, using the characteristic function methodology. For example, the price of a European call option that depends on the parameter vector Θ is given by:

$$C(\cdot;\Theta^*) = SP_1 - Xe^{-rt}P_2. \tag{19.22}$$

Here, P_1 and P_2 are cumulative probability functions that can be calculated by Fourier inversion of the characteristic function (see Duffie *et al.*, 2000) and Θ^* is the set of risk neutral parameters. For a non-affine process, option prices can be calculated either by Monte Carlo simulation (for example, Christoffersen, Jacobs and Mimouni, 2006) or by solving the associated PDE. Many studies (for example, Bakshi, Cao and Chen, 1997; Broadie *et al.*, 2007) calibrate option pricing models with stochastic volatility to observed options prices via least squares:

$$SSE(t) = \min_{\theta} \sum_{i=1}^{N} \left(O_i - \hat{O}_i(\Theta^*)\right)^2, \tag{19.23}$$

where N is the number of options at date t, O_i is the observed option price with strike price K, and $\hat{O}_i(\Theta)$ is the model's implied price given the parameters Θ^*. Here we should note that the set of parameters Θ^* is not the same as the set obtained from historical returns because option prices are calculated under the risk-neutral probability measure. Hence, option prices contain volatility and jump risk premia that are incorporated into certain parameters. For example, under the assumption that the volatility risk premium is linear in volatility (Heston, 1993), the parameters of the SV2 model from calibration would be $k^* = k+\lambda$ and $\theta^* = k\theta/(k+\lambda)$, where λ is the market price of volatility risk. Calibration of the stochastic volatility model to option prices alone does not allow the identification of the risk premium parameter. Another drawback of this approach is that the model has to be recalibrated every day, which will usually produce different parameter estimates.

Another strand of the literature uses options prices and historical returns simultaneously. The advantage of this approach is that the time series V_t can be taken from option prices and it is also feasible to identify risk premia. This approach also imposes consistency between option prices and the time series properties of the underlying returns (see Bates, 1996a). Some well known studies that use option prices and historical returns are Chernov and Ghysels (2000), who use EMM, Pan (2002), using a GMM procedure, and Eraker (2004), employing MCMC. In a recent

breakthrough paper, Aït-Sahalia (2008) derives transition density approximations for multivariate diffusions. In Aït-Sahalia and Kimmel (2007) this method is applied for maximum likelihood estimation of a variety of stochastic volatility models, both affine and non-affine. Aït-Sahalia and Kimmel (2007) provide approximations for the joint density of asset returns and a vector of option prices or the joint density of asset prices and an implied volatility index, which is used as a proxy for latent volatility.

19.3.4 Integrated volatility

Recent advances in high frequency financial econometrics offer alternative methods for making inferences about latent stochastic volatility dynamics. Results from the theory of quadratic variation (e.g., Andersen *et al.*, 2001; Barndorff-Nielsen and Shephard, 2002) show that, when the sampling frequency becomes high, realized integrated volatility converges to the unobserved true integrated volatility. Figure 19.3 shows the evolution of the daily integrated volatility for the S&P500 over the period 1993 to 2004.[9] Integrated volatility over a period $[t, T]$ is computed from summing high frequency log returns and is given by:

$$\lim_{N\to\infty} \sum_{i=1}^{2^N} \left(y_{t+\frac{i}{2^N}(T-t)} - y_{t+\frac{i-1}{2^N}(T-t)} \right)^2 \overset{a.s.}{\to} IV_{t,T} = \frac{1}{T-t} \int_t^T V_s ds. \qquad (19.24)$$

Expression (19.24) contains rich information with respect to the stochastic process followed by the unobserved latent volatility. However, the expression is valid only under the assumption that the sample paths of the asset are continuous. In the presence of jumps in asset returns, integrated volatility also includes a jump component (see Barndorff-Nielsen and Shephard, 2006). For affine models, it is feasible

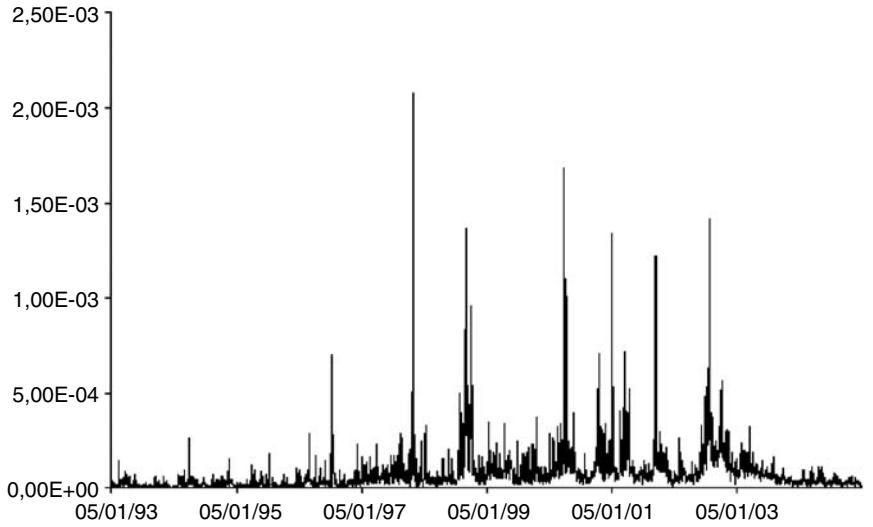


Figure 19.3 Daily integrated volatilities of the S&P500 over the period January 5, 1993, to December 31, 2004

to derive the conditional moments of integrated volatility. For example, the first moment has the general form $E_t\left(IV_{t,T}\right) = a(\tau, \Theta)V_t + b(\tau, \Theta)$. Bollerslev and Zhou (2002) derive the first two conditional moments of integrated volatility and apply a GMM procedure. In order to identify the correlation they also derive in closed form the cross-moment, $E_t\left(y_T IV_{t,T}\right)$. Barndorff-Nielsen and Shephard (2002) develop a QML procedure based on the time series of realized volatility. Chourdakis and Dotsis (2008) estimate non-affine specifications using maximum likelihood and a Markov chain approximation procedure.

19.4 Empirical comparison of volatility processes

In this section we provide an empirical comparison of the models described in section 19.2. A comparison of the econometric methods outlined is section 19.3 is beyond the scope of this chapter. Instead, we estimate the volatility processes autonomously using an implied volatility index.[10] This facilitates estimation but does not allow us to make inferences on the joint dynamics of asset returns and volatility. For reasons explained in Jones (2003) and Bakshi, Ju and Ou-Yang (2006), we consider as a proxy for volatility the implied volatility index VXO. Hence, we set $V_t \equiv VXO_t^2$ over the period 1990–2007, a total of 4,535 daily observations.

The parameters of the various processes are estimated by maximum likelihood, which requires the conditional density function $f[V(t+\tau)|V(t), \Theta]$ $(\tau > 0)$ of the process V_t, where τ denotes the sampling frequency of observations (daily in our application). For a sample $\{V_t\}_{t=1}^{T}$, the log-likelihood function that is maximized is given by:

$$\Im = \max_{\Theta} \sum_{t=1}^{T-\tau} \log\left(f\left(V_{t+\tau}\right)\left|V_t, \Theta\right.\right). \tag{19.25}$$

The standard errors of the estimates are retrieved from the inverse Hessian, evaluated at the estimates. For SV1, SV2 and SV5 the conditional density is known in closed form (see Dotsis, Psychoyios and Skiadopoulos, 2007; Psychoyios, Dotsis and Markellos, 2007). However, for SV3 and SV4, the transition density does not have a closed-form solution. The density of SV4 is obtained by the approximation method of Aït-Sahalia (1999, 2002). The transition density of the SV3 model is obtained by Fourier inversion of the characteristic function (see Singleton, 2001; Dotsis, Psychoyios and Skiadopoulos, 2007). The Fourier inversion of the characteristic function provides the required conditional density function $f[V(t+\tau)|V(t)]$ as:

$$f[V_{t+\tau}|V_t, \Theta] = \frac{1}{\pi}\int_0^{\infty} Re[e^{-isV_{t+\tau}}\phi\left(i\omega, V_t, \tau; \Theta\right)]d\omega. \tag{19.26}$$

Table 19.1 shows the ML estimates for VXO. For each of the processes, the estimated parameters (annualized), the t-statistics (within parentheses), the AIC (Akaike information criterion) and BIC (Bayesian information criterion), and the maximized log-likelihood values (LL) are reported. Likelihood ratio tests were also used to compare nested models: as these supported the ranking obtained from AIC, BIC and LL, they are not reported.

Table 19.1 The parameter estimates (annualized) and the *t*-statistics (reported in parentheses) are based on maximizing the log-likelihood (LL). The data cover the period January 2, 1990, to December 31, 2007, a total of 4,535 observations.

Parameter	*SV1*	*SV2*	*SV3*	*SV4*	*SV5*
k	6.780 (7.71)	5.809 (7.09)	7.999 (11.13)	0.759 (5.49)	3.810 (5.80)
θ	0.044 (19.96)	0.044 (11.52)	0.022 (24.05)	0.105 (−29.85)	−3.368 (−27.78)
σ	0.123 (93.71)	0.447 (93.88)	0.359 (51.34)	3.29 (22.21)	1.96 (94.46)
γ	–	–	–	1.157 (7.89)	–
λ	–	–	31.498 (9.14)	–	–
$1/\eta$	–	–	0.017 (4.56)	–	–
LL	15,459	17,457	17,694	18,446	18,390
AIC	−30,911	−34,908	−35,379	−36,884	−36,773
BIC	−30,892	−34,889	−35,347	−36,850	−36,754

All parameters are significant at the 1% level. According to all model selection criteria, the best fit is provided by SV4, followed by, in order, SV5, SV3, SV2 and SV1. Consistent with previous results in the literature, the Gaussian SV1 model provides a very poor approximation to the data. The square root specification of SV2 improves the fit substantially and the addition of a jump component improves performance further. In the SV3 model, the mean reversion parameter increases so as to pull back the process to its long run mean after a jump event. However, the increase in the speed of mean reversion shows that jumps do not have a persistent effect on volatility. All the statistical criteria suggest that SV5 outperforms the square root specifications. As also reported in Psychoyios, Dotsis and Markellos (2007), this result should not come as a surprise, since SV5 is capable of generating a large increase in volatility at high levels, followed by rapid mean reversion. The best model overall is the constant elasticity of variance specification SV4. The estimate of the exponent in the diffusion, $\hat{\gamma} = 1.16$, suggests that, as volatility increases, its own volatility increases at an even faster rate (see also Jones, 2003). The model also appears capable of capturing strong heteroskedasticity in volatility changes. The empirical results thus point to the conclusion that, at least under this simplified set up, non-affine specifications outperform affine processes.

19.5 Conclusions

The list of methods that we have discussed in this chapter is by no means complete. For example, we have omitted econometric approaches such as the SMM of

Duffie and Singleton (1993), the simulated maximum likelihood (SML) of Santa-Clara (1995) and Pedersen (1995a), the range-based QML estimation technique of Alizadeh, Brandt and Diebold (2002) and the Markov chain approximation of Chourdakis (2002).

We believe there are two avenues for future research. First, more work needs to be done on the comparison of non-affine and affine volatility models. Two questions (at least) suggest themselves – does analytical tractability come at the cost of empirical misspecification?, and what is the impact of non-affine specifications on derivatives valuation? Second, despite this plethora of statistical stochastic volatility models there is still a lack of understanding about the economic underpinnings of volatility. Carr and Wu (2008) show that the market price of volatility risk is large and negative, and Bollerslev and Zhou (2007) find that the volatility risk premium forecasts future excess returns. However, there is still no solid explanation of the economic sources of volatility risk premia and there is also a lack of understanding of the behavior of the pricing kernel as a function of both market returns and return volatilities.

Notes

1. There is a wide variety of interpretations of the term "volatility" within the financial and econometrics literature: for example, variance, (annualized) standard deviation, (total) risk, uncertainty, and so on. In the context of continuous-time models and throughout this chapter, the term is used to describe the latent instantaneous variance.
2. Baillie (2006), for example, provides a recent survey of ARCH/GARCH modeling.
3. See Sundaresan (2000) for a comprehensive review of the development and application of continuous-time methods in finance and Aït-Sahalia (2007) for a survey of estimation methods in continuous-time models. Merton (1990) is the benchmark book in continuous-time finance.
4. Ghysels, Harvey and Renault (1996) provide an extensive review of stochastic volatility models, but they are mainly concerned with models defined in discrete time.
5. VXO is the implied volatility of a synthetic at-the-money option on the S&P500 equity index with a constant time to maturity of 30 calendar days to expiry. In 2003, the Chicago Board Options Exchange (CBOE) introduced a new implied volatility index, coined the VIX. This is calculated in a model-free manner as a weighted sum of out-of-money option prices across all available strikes on the S&P500 index. Carr and Wu (2006) show that the VIX represents the conditional risk-neutral expectation of the return volatility under general market settings. In 2005 CBOE introduced futures on VIX and, in 2006, European calls/puts written on forward VIX.
6. Psychoyios, Skiadopoulos and Alexakis (2003) provide a comprehensive review of alternative volatility processes.
7. See Andersen, Benzoni and Lund (2002) for discussion on the simulation of diffusion/jump stochastic volatility models.
8. The characteristic function methodology has also been used for parameter estimation in discrete time independent and identically distributed (i.i.d) and autoregressive moving average (ARMA) processes by Feuerverger and McDunnough (1981a, 1981b) and Feuerverger (1990).
9. The realized volatilities are based on intraday transactions on the S&P500 index. This dataset has been used by Huang, Liu and Yu (2007) and is available from Professor Yu's web page (http://www.mysmu.edu/faculty/yujun/research.html). The construction

of realized volatility takes into account market microstructure noise using a technique proposed by Zhang, Mykland and Aït-Sahalia (2005).

10. At-the-money implied volatility is sometimes used as a proxy for instantaneous volatility. However, the two are only identical when volatility is uncorrelated with the asset price, the market price of volatility risk is zero, and the time to maturity of the option is short.

References

Aït-Sahalia, Y. (1999) Transition densities for interest rate and other nonlinear diffusions. *Journal of Finance* **54**, 1361–95.

Aït-Sahalia, Y. (2002) Transition densities for interest rate and other nonlinear diffusions: a closed-form approximation approach. *Econometrica* **70**, 223–62.

Aït-Sahalia, Y. (2007) Estimating continuous-time models using discretely sampled data. In R. Blundell, P. Torsten and W.K. Newey (eds.), *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*. Cambridge: Cambridge University Press.

Aït-Sahalia, Y. (2008) Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics* **36**, 906–37.

Aït-Sahalia, Y. and R. Kimmel (2007) Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics* **83**, 413–52.

Aït-Sahalia, Y., P.A. Mykland and L. Zhang (2005) How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* **18**, 351–416.

Alizadeh, S., M.W. Brandt and F.X. Diebold (2002) Range-based estimation of stochastic volatility models. *Journal of Finance* **57**, 1047–91.

Andersen, T.G., L. Benzoni and J. Lund (2002) An empirical investigation of continuous-time equity return models. *Journal of Finance* **57**, 1239–84.

Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys (2001) The distribution of exchange rate volatility. *Journal of the American Statistical Association* **96**, 42–55.

Andersen, T.G., H. Chung and B.E. Sørensen (1999) Efficient method of moments estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Econometrics* **91**, 61–87.

Andersen, T.G. and J. Lund (1997) Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics* **77**, 343–77.

Baillie, R.T. (2006) Modelling volatility. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics. Volume 1: Econometric Theory*, pp. 737–64. Basingstoke: Palgrave Macmillan.

Bakshi, G., C. Cao and Z. Chen (1997) Empirical performance of alternative option pricing models. *Journal of Finance* **52**, 2003–49.

Bakshi, G., N. Ju and H. Ou-Yang (2006) Estimation of continuous-time models with an application to equity volatility dynamics. *Journal of Financial Economics* **82**, 227–49.

Bansal, R., A.R. Gallant, R. Hussey and G.E. Tauchen (1993) Computational aspects of nonparametric simulation estimation. In D.A. Belsley (ed.), *Computational Techniques for Econometrics and Economic Analysis*, pp. 3–22. Boston: Kluwer Academic Publishers.

Bansal, R., A.R. Gallant, R. Hussey and G.E. Tauchen (1995) Nonparametric estimation of structural models for high-frequency currency market data. *Journal of Econometrics* **66**, 251–87.

Barndorff-Nielsen, O.E. and N. Shephard (2002) Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64**, 253–80.

Barndorff-Nielsen, O.E. and N. Shephard (2006) Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* **4**, 1–30.

Bates, D. (1996a) Testing option pricing models. In G.S. Maddala and C.R. Rao (eds.), *Statistical Methods in Finance. Handbook of Statistics, Volume 14*, pp. 567–611. Amsterdam: Elsevier.

Bates, D. (1996b) Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. *Review of Financial Studies* **9**, 69–107.

Bates, D. (2000) Post-87 crash fears in S&P 500 futures options. *Journal of Econometrics* **94**, 181–238.

Bates, D. (2006) Maximum likelihood estimation of latent affine processes. *Review of Financial Studies* **19**, 909–65.

Black, F. (1976) Studies in stock price volatility changes. *Proceedings of the 1976 Business Meeting of the Business and Economic Statistics Section, American Statistical Association*, 177–81.

Black, F., and M. Scholes (1973) The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637–59.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–27.

Bollerslev, T. and H. Zhou (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics* **109**, 33–65.

Bollerslev, T. and H. Zhou (2007) Expected stock returns and variance risk premia. Working Paper, Duke University.

Brenner, M., E.Y. Ou and J.E. Zhang (2006) Hedging volatility risk. *Journal of Banking and Finance* **30**, 811–21.

Broadie, M., M. Chernov and M. Johannes (2007) Model specification and risk premiums: the evidence from the futures options. *Journal of Finance*. Forthcoming.

Carr, P. and L. Wu (2004) Time-changed Lévy processes and option pricing. *Journal of Financial Economics* **71**, 113–41.

Carr, P. and L. Wu (2006) A tale of two indices. *Journal of Derivatives* **13**, 13–29.

Carr, P. and L. Wu (2008) Variance risk premia. *Review of Financial Studies*. Forthcoming.

Chacko, G. and L.M. Viceira (2003) Spectral GMM estimation of continuous-time processes. *Journal of Econometrics* **116**, 259–92.

Chernov, M., A.R. Gallant, E. Ghysels and G. Tauchen (2003) Alternative models for stock price dynamics. *Journal of Econometrics* **116**, 225–57.

Chernov, M. and E. Ghysels (2000) A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of option valuation. *Journal of Financial Economics* **56**, 407–58.

Chourdakis, K. (2002) Continuous-time regime switching models and applications in estimating processes with stochastic volatility and jumps. Working Paper 464, Queen Mary, University of London.

Chourdakis, K. and G. Dotsis (2008) Maximum likelihood estimation and dynamic asset allocation with non-affine volatility processes. Working Paper, University of Essex.

Christie, A. (1982) The stochastic behavior of common stock variances: value, leverage, and interest rate effects. *Journal of Financial Economics* **10**, 407–32.

Christoffersen, P.F., K. Jacobs and K. Mimouni (2006) Models for S&P 500 dynamics: evidence from realized volatility, daily returns, and option prices. Working Paper, McGill University.

Clark, P.K. (1973) A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* **41**, 135–56.

Detemple, J. and C. Osakwe (2000) The valuation of volatility options. *European Finance Review* **4**, 21–50.

Dotsis, G., D. Psychoyios and G. Skiadopoulos (2007) An empirical comparison of continuous time models of implied volatility indices. *Journal of Banking and Finance* **31**, 3584–603.

Duffie, D., J. Pan and K. Singleton (2000) Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**, 1343–76.

Duffie, D. and K. Singleton (1993) Simulated moments estimation of Markov models of asset prices. *Econometrica* **61**, 929–52.
Easley, D. and M. O'Hara (1992) Time and the process of security price adjustment. *Journal of Finance* **47**, 577–605.
Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1006.
Eraker, B. (2004) Do stock prices and volatility jump? Reconciling evidence from spot and option prices. *Journal of Finance* **59**, 1367–403.
Eraker, B., M. Johannes and N. Polson (2003) The impact of jumps in volatility and returns. *Journal of Finance* **53**, 1269–300.
Fama, E.F. (1963) Mandelbrot and the stable Paretian distribution. *Journal of Business* **36**, 420–9.
Fama, E.F. (1965) The behavior of stock market prices. *Journal of Business* **38**, 34–105.
Feuerverger, A. (1990) An efficiency result for the empirical characteristic function in stationary time-series models. *Canadian Journal of Statistics* **18**, 155–61.
Feuerverger, A. and P. McDunnough (1981a) On some Fourier methods for inference. *Journal of the American Statistical Association* **76**, 379–87.
Feuerverger, A. and P. McDunnough (1981b) On the efficiency of empirical characteristic function procedures. *Journal of the Royal Statistics Society, Series B* **43**, 20–7.
Gallant, A.R., D.A. Hsieh and G.E. Tauchen (1997) Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* **81**, 159–92.
Gallant, A.R. and G.E. Tauchen (1996) Which moments to match? *Econometric Theory* **12**, 657–81.
Gallant, A.R. and G. Tauchen (1998) Reprojecting partially observed systems with application to interest rate diffusions. *Journal of the American Statistical Association* **93**, 10–24.
Ghysels, E., A. Harvey and E. Renault (1996) Stochastic volatility. In G.S. Maddala and C.R. Rao (eds.), *Statistical Methods in Finance. Handbook of Statistics, Volume 14*, pp. 119–91. Amsterdam: Elsevier.
Glosten, L.R., R. Jagannathan and D. Runkle (1993) Relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* **48**, 1779–801.
Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.
Harvey, A., E. Ruiz and N. Shephard (1994) Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–64.
Heston, S.L. (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* **6**, 327–43.
Huang, S., Q. Liu and J. Yu (2007) Realized daily variance of S&P 500 cash index: a revaluation of stylized facts. *Annals of Economics and Finance* **8**, 33–56.
Hull, J.C. and A. White (1987) The pricing of options with stochastic volatility. *Journal of Finance* **42**, 281–300.
Jacquier, E., N.G. Polson and P.E. Rossi (1994) Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics* **12**, 371–89.
Jiang, G.J. and J.L. Knight (2002) Estimation of continuous time processes via the empirical characteristic function. *Journal of Business and Economic Statistics* **20**, 198–212.
Johannes, M. and N. Polson (2006) MCMC methods for financial econometrics. In Y. Aït-Sahalia and L. Hansen (eds.), *Handbook of Financial Econometrics*. Forthcoming.
Johnson, H. and D. Shanno (1987) Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis* **22**, 143–51.
Jones, C. (2003) The dynamics of stochastic volatility: evidence from underlying and options markets. *Journal of Econometrics* **116**, 181–224.

Kim, S., N. Shephard and S. Chib (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **65**, 361–93.
Lo, A.W. (1988) Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory* **4**, 231–47.
Mandelbrot, B.B. (1963) The variation of certain speculative prices. *Journal of Business* **36**, 394–416.
Melino, A. and S. Turnbull (1990) Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* **45**, 239–65.
Merton, R.C. (1973) Theory of rational option pricing. *Bell Journal of Economics* **4**, 141–83.
Merton, R.C. (1976) Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* **3**, 125–44.
Merton, R.C. (1980) On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* **8**, 323–63.
Merton, R.C. (1990) *Continuous-Time Finance*. New York: Oxford University Press.
Nelson, D.B. (1991) Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347–70.
Pan, J. (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics* **63**, 3–50.
Pedersen, A. (1995a) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics* **22**, 55–71.
Pedersen, A. (1995b) Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli* **1**, 257–79.
Psychoyios, D., G. Dotsis and R.N. Markellos (2007) A jump diffusion model for VIX options and futures. Working Paper, Athens University of Economics and Business.
Psychoyios, D., G. Skiadopoulos and P. Alexakis (2003) A review of stochastic volatility processes: properties and implications, *Journal of Risk Finance* **4**(3), 43–60.
Santa-Clara, P. (1995) Simulated likelihood estimation of diffusions with an application to the short term interest rate. Ph.D. dissertation, INSEAD.
Schwert, G.W. (1989) Why does market volatility change over time? *Journal of Finance* **44**, 1115–53.
Scott, L. (1987) Option pricing when the variance changes randomly: theory, estimation, and an application. *Journal of Financial and Quantitative Analysis* **22**, 419–38.
Singleton, K. (2001) Estimation of affine asset pricing models using the empirical characteristic function. *Journal of Econometrics* **102**, 111–41.
Stein, E. and J. Stein (1991) Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies* **4**, 727–52.
Sundaresan, S.M. (2000) Continuous time methods in finance: a review and an assessment. *Journal of Finance* **55**, 1569–622.
Vasicek, O. (1977) An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–88.
Wiggins, J. (1987) Option values under stochastic volatility: theory and empirical estimates. *Journal of Financial Economics* **19**, 351–72.
Zhang, L., P.A. Mykland and Y. Aït-Sahalia (2005) Edgeworth expansions for realized volatility and related estimators. Technical Report No. 556, University of Chicago, Department of Statistics.

20
Testing the Martingale Hypothesis

J. Carlos Escanciano and Ignacio N. Lobato

Abstract

This chapter examines testing the Martingale difference hypothesis (MDH) and related statistical inference issues. The earlier literature on testing the MDH was based on linear measures of dependence, such as sample autocorrelations; for example, the classic Box–Pierce portmanteau test and the variance ratio test. In order to account for the existing nonlinearity in economic and financial data, two directions have been entertained. First, to modify these classical approaches by taking into account possible nonlinear dependence. Second, to use more sophisticated statistical tools such as those based on empirical process theory or the use of generalized spectral analysis. This chapter discusses these developments and applies them to exchange rate data.

20.1 Introduction

Martingale testing has received enormous attention in econometrics. One of the main reasons is the efficient market hypothesis and the many ideas related to it. In addition, many economic theories examining dynamic contexts in which expectations are assumed to be rational lead to dependence restrictions of this kind in the underlying economic variables (see, e.g., Hall, 1978; Fama, 1991; LeRoy, 1989; Lo, 1997; Cochrane, 2005). These theories have prompted a great deal of research in macro- and financial economics which has stimulated a huge interest in developing suitable econometric techniques. This econometric research has grown around the theme of the lack of predictability of macro- or financial series, but this

topic has flourished in different branches, emphasized different methodological aspects, and appeared under different subject names.

When looking at asset prices, the idea of lack of predictability has been commonly referred to as the random walk hypothesis. Unfortunately, the term "random walk" has been used in different contexts to mean different statistical objects. For instance, in Campbell, Lo and MacKinlay's (1997) textbook, they distinguish three types of random walks according to the dependence structure of the increment series. Random walk 1 corresponds to independent increments, random walk 2 to mean-independent increments, and random walk 3 to uncorrelated increments. Of these three notions, the two most relevant to financial econometrics are the second and the third. The notion of random walk 1 is clearly rejected in financial data for many reasons, the most important being volatility: the lack of constancy of the variance of current asset returns conditional on lagged asset returns. Within this terminology, this chapter will focus basically on the idea of random walk 2, but we will also discuss some aspects associated with random walk 3. A martingale would correspond to random walk 2, and it plainly means that the best forecast of tomorrow's asset price is today's. The asset returns, which are unpredictable, are then said to form a martingale difference sequence. Since asset prices are not stationary, from a technical point of view it is simpler to handle asset returns, and instead of testing that prices follow a martingale, it is more common to test that returns follow a martingale difference sequence.

Given the huge literature that has developed, it is unavoidable that the present chapter reflects the authors' personal interests. It is important at the outset to stress what this chapter does *not* cover. We do not consider unit root tests, which is a topic covered in many references (see, e.g., Laudrup and Jansson, 2006). We do not address technical analysis, which assumes predictability and focuses on the best ways of constructing a variety of charts to forecast a series. We do not consider out-of-sample prediction tests because they assume particular models under the alternative (see Inoue and Kilian, 2004; Clark and West, 2006). We do not examine chaos tests, which are motivated by deterministic nonlinear models (see references in Barnett and Serletis, 2000; Chan and Tong, 2002). What we address is called conditional mean independence testing in the statistical literature.

The outline of the chapter is as follows. Section 20.2 contains the preliminary definitions and an overview of the data that we will employ to illustrate the different techniques. Section 20.3 studies martingale difference tests based on linear measures of dependence both in the time and frequency domains. Section 20.4 is devoted to tests based on nonlinear measures of dependence. Section 20.5 discusses briefly some hypotheses related to the martingale difference hypothesis and Section 20.6 concludes.

20.2 Preliminaries

The martingale difference hypothesis (MDH) plays a central role in economic models where expectations are assumed to be rational. The underlying statistical object

of interest is the concept of a martingale or, alternatively, the concept of a martingale difference sequence (m.d.s.). Mathematically speaking, we say that X_t forms a martingale, with respect to its natural filtration, when $E\left[X_t \mid X_{t-1}, X_{t-2}, \ldots\right] = X_{t-1}$ almost surely (*a.s.*). As stated in the introduction, from a technical point of view, it is simpler to work with the first differences, $Y_t = X_t - X_{t-1}$, and we say that Y_t follows an mds when $E\left[Y_t \mid Y_{t-1}, Y_{t-2}, \ldots\right] = 0$ *a.s.* More generally, we state that the MDH holds when, for a real-valued stationary time series $\{Y_t\}_{t=-\infty}^{\infty}$, the following conditional moment restriction holds *a.s.*:

$$E\left[Y_t \mid Y_{t-1}, Y_{t-2}, \ldots\right] = \mu, \quad \mu \in \mathbb{R}. \tag{20.1}$$

The MDH slightly generalizes the notion of m.d.s. by allowing the unconditional mean of Y_t to be non-zero and unknown. The MDH states that the best predictor, in the sense of least mean square error, of the future values of a time series, given the past and current information set, is just the unconditional expectation. The MDH is called conditional mean independence in the statistical literature, and it implies that past and current information are of no use for forecasting future values of an m.d.s. In section 20.5 we discuss extensions of this basic version of the MDH.

As noted in the introduction, there is a vast empirical and theoretical literature on the MDH. In order to systematize part of this literature, we start by introducing the following definitions. Let $I_t = \{Y_t, Y_{t-1}, \ldots\}$ be the information set at time t and let $\mathcal{F}_t$ be the σ-field generated by I_t. The following equivalence is then fundamental because it formalizes the characteristic property of an m.d.s.: Y_t is linearly unpredictable given any linear or nonlinear transformation of the past, $w(I_{t-1})$, i.e.,

$$E[Y_t \mid I_{t-1}] = \mu \text{ a.s., } \mu \in \mathbb{R} \iff E[(Y_t - \mu)w(I_{t-1})] = 0, \tag{20.2}$$

for any $\mathcal{F}_{t-1}$-measurable weighting function $w(\cdot)$ (such that the moment exists). Equation (20.2) is fundamental to understanding the motivation and main features behind many tests of the MDH. There are two challenging features present in the definition of an m.d.s.: first, the information set at time t, I_t, will typically include the infinite past of the series; second, the number of functions $w(\cdot)$ is also infinite. We will classify the extant theoretical literature on testing the MDH according to what types of functions $w(\cdot)$ are employed. Section 20.3 analyzes the case where linear $w(\cdot)$ are employed, that is, the use of tests based on linear measures of dependence. Section 20.4 analyzes the case where an infinite number of nonlinear *w*'s are employed, that is, the use of tests based on nonlinear measures of dependence. In both sections, we divide the extensive literature according to whether the tests account for a finite number of lags or not, that is, whether they assume that $w(I_{t-1}) = w(Y_{t-1}, ..., Y_{t-P})$ for some $P \geq 1$ or not.

We shall illustrate some of the available methods for testing the MDH by applying them to exchange rate returns. The martingale properties of exchange rate returns have been studied previously by many authors, leading to mixed conclusions. For instance, Bekaert and Hodrick (1992), Escanciano and Velasco (2006a, 2006b), Fong and Ouliaris (1995), Hong and Lee (2003), Kuan and Lee (2004), LeBaron (1999), Levich and Thomas (1993), Liu and He (1991), McCurdy and Morgan (1988) and

Sweeney (1986) all find evidence against the MDH for nominal or real exchange rates at different frequencies, whereas Diebold and Nason (1990), Fong, Koh and Ouliaris (1997), Hsieh (1988, 1989, 1993), McCurdy and Morgan (1987) and Meese and Rogoff (1983a, 1983b) find little evidence against the MDH. Here we consider data that consists of four daily and weekly exchange rate returns against the US dollar: the euro (Euro), the Canadian dollar (Can), the pound sterling (Pound) and the Japanese yen (Yen). The daily data is taken from January 1, 2004, to August 17, 2007, with a total of 908 observations. For the weekly data, we consider the returns on Wednesdays from January 14, 2000, to August 17, 2007, with a total of 382 observations. The daily noon buying rates in New York City certified by the Federal Reserve Bank of New York for customs and cable transfers purposes are obtained from http://www.federalreserve.gov/Releases/h10/hist. In Figures 20.1 and 20.2 we have plotted the evolution of these four daily and weekly exchange rates, respectively, and, similar to previous analyzes, the two main features of these plots are their unpredictability and their volatility. Table 20.1 provides summary statistics for the most relevant aspects of the marginal distribution of the data. Similar to most financial series, the main feature from Table 20.1 is kurtosis that, in line with previous studies, is larger for daily than for weekly data. Note that skewness is moderate and slightly negative for daily data. As has been observed repeatedly before, the marginal distribution of weekly data is closer to the normal distribution than that of daily data.

20.3 Tests based on linear measures of dependence

Recall the m.d.s. definition in equation (20.2) that should hold for any function $w(\cdot)$. The simplest approach is to consider linear functions, such as $w(I_{t-1}) = Y_{t-j}$,

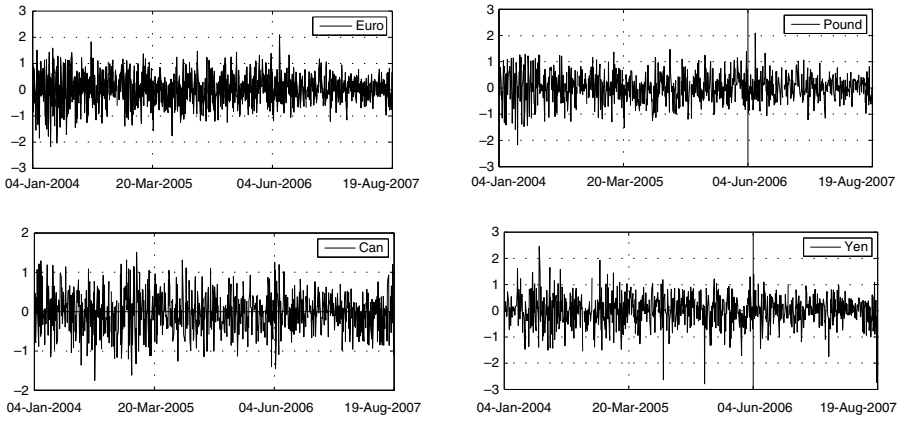


Figure 20.1 Daily returns of the euro, Canadian dollar (Can), pound sterling (Pound) and the Japanese yen (Yen) against the US dollar
Data from January 4, 2004, to August 17, 2007.

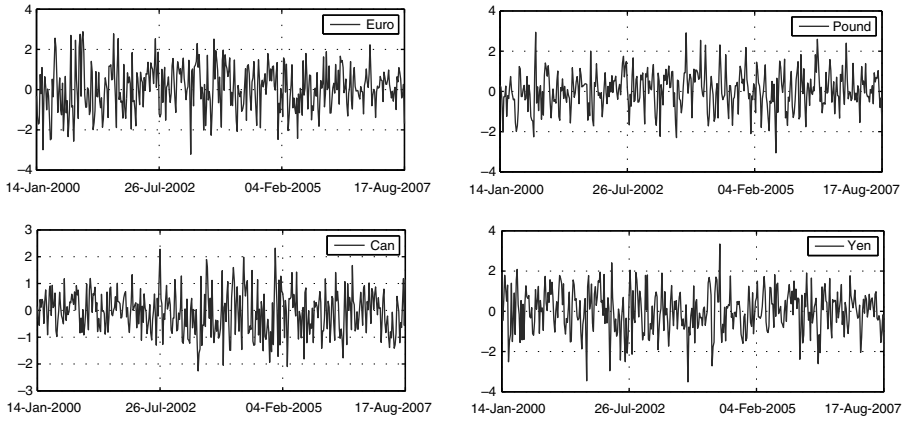


Figure 20.2 Weekly returns of the euro, Canadian dollar (Can), the pound sterling (Pound) and the Japanese yen (Yen) against the US dollar
Data from January 14, 2000, to August 17, 2007.

Table 20.1 Summary statistics of exchange rates returns

	Daily				*Weekly*			
	Euro	*Pound*	*Can*	*Yen*	*Euro*	*Pound*	*Can*	*Yen*
n	908	908	908	908	382	382	382	382
Mean	0.0076	0.0113	−0.0213	0.0068	0.0738	0.0552	−0.0832	0.0352
Median	0.0000	0.0221	−0.0080	0.0279	0.0781	0.0763	−0.0864	0.0141
SD	0.5423	0.5332	0.5036	0.5670	1.3539	1.1407	0.9410	1.2525
Skewness	−0.1263	−0.0976	−0.0196	−0.3763	0.0540	0.0545	0.0846	−0.2945
Kurtosis	3.7602	3.4927	3.1345	5.0746	3.0555	2.9649	2.8875	3.0895
Maximum	1.9358	2.0930	1.5129	2.4519	4.4680	3.4830	2.8128	3.1835
Minimum	−2.0355	−2.1707	−1.7491	−2.7859	−3.1636	−3.2307	−2.7067	−4.3058

for some $j \geq 1$. Hence, a necessary (but not sufficient, in general) condition for the the MDH to hold is that the time series is uncorrelated, i.e.,

$$\gamma_j = Cov(Y_t, Y_{t-j}) = E[(Y_t - \mu)Y_{t-j}] = 0 \qquad \text{for all } j \geq 1, \tag{20.3}$$

where γ_j denotes the autocovariance of order j. In principle, one should test that all autocovariances or autocorrelations are zero. However, the most employed tests just consider that a finite number of autocorrelations are zero. As commented in the introduction, we will address these two cases separately.

Notice that the early literature, which includes some distinguished references such as Yule (1926), Bartlett (1955), Grenander and Rosenblatt (1957) or Durbin and Watson (1950), essentially assumed Gaussianity and, hence, identified three concepts: lack of serial correlation, m.d.s., and independence. In the time series literature the term "white noise" is commonly used to denote an uncorrelated

series that can present some form of dependence. Obviously, a white-noise series is neither necessarily independent nor m.d.s., since dependence can be reflected in other aspects of the joint distribution such as higher-order moments. The distinction between these three concepts has been stressed recently in econometrics. In fact, during the last few years a variety of models designed to reflect nonlinear dependence have been studied in the econometrics literature. For instance, in empirical finance, ARCH and bilinear models have been widely studied (see Bera and Higgins, 1993, 1997, and Weiss, 1986, for a comparison). These models are suitable for reflecting the nonlinear dependence structure found in many financial series.

Tests for white noise have been proposed both in the time domain and in the frequency domain. The time domain has mainly, but not exclusively, focused on a finite number of lags, while the frequency domain has been more suitable for addressing the infinite-dimensional case.

20.3.1 Tests based on a finite-dimensional conditioning set

In the time domain the most popular test (apart from the Durbin–Watson, which is designed to test for lack of first-order serial correlation using regression residuals) has been the Box–Pierce (Box and Pierce, 1970) portmanteau Q_p test. The Q_p test is designed for testing that the first p autocorrelations of a series (possibly residuals) are zero. The number p can be considered to be fixed or to grow with the sample size n. In this section we will assume that p is fixed.

Suppose that we observe raw data $\{Y_t\}_{t=1}^n$. Then γ_j can be consistently estimated by the sample autocovariance:

$$\widehat{\gamma}_j = (n-j)^{-1} \sum_{t=1+j}^{n} (Y_t - \overline{Y})(Y_{t-j} - \overline{Y}),$$

where $\overline{Y}$ is the sample mean, and we also introduce $\widehat{\rho}_j = \widehat{\gamma}_j/\widehat{\gamma}_0$ to denote the jth-order autocorrelation. The Q_p statistic is just:

$$Q_p = n\sum_{j=1}^{p} \widehat{\rho}_j^2,$$

but it is commonly implemented via the Ljung and Box (1978) modification:

$$LB_p = n(n+2)\sum_{j=1}^{p} (n-j)^{-1} \widehat{\rho}_j^2.$$

Note that Q_p (or LB_p) only takes into account the linear dependence up to the lag p. When p is considered fixed, the Q_p test statistic applied to independent data follows a χ_p^2 asymptotic null distribution, since the asymptotic covariance matrix of the first p autocorrelations of an independent series is the identity matrix. Hence, it is useful to write $Q_p = \left(\sqrt{n}\widehat{\rho}\right)' I^{-1} \left(\sqrt{n}\widehat{\rho}\right)$, where $\widehat{\rho} = (\widehat{\rho}_1, ..., \widehat{\rho}_p)'$.

Note, however, that when the series present some kind of nonlinear dependence, such as conditional heteroskedasticity, this asymptotic null covariance matrix is no longer the identity. In fact, denoting $\rho = (\rho_1, ..., \rho_p)'$, for a general time series the asymptotic distribution of $\sqrt{n}(\widehat{\rho} - \rho)$ is $N(0, T)$, where the $p \times p$ matrix T has as (i, j)th element (see, e.g., Romano and Thombs, 1996),

$$\gamma_0^{-2}(c_{ij} - \rho_i c_{0j} - \rho_j c_{0i} + \rho_i \rho_j c_{00}),$$

where, for $i, j = 0, 1, ..., p$,

$$c_{ij} = \sum_{d=-\infty}^{\infty} \Big\{ E\Big[(Y_t - \mu)(Y_{t-i} - \mu)(Y_{t+d} - \mu)(Y_{t+d-j} - \mu)\Big] - E\big[(Y_t - \mu)(Y_{t-i} - \mu)\big] E\Big[(Y_{t+d} - \mu)(Y_{t+d-j} - \mu)\Big] \Big\}.$$

Under alternative assumptions the matrix T can be simplified and this will lead to several modified versions of the Box–Pierce statistic. When this matrix is still diagonal, as happens under m.d.s. and additional moment restrictions, which, for instance, are satisfied by Gaussian GARCH models and many stochastic volatility models, the natural approach is to robustify the Q_p by standardizing it by a consistent estimate of its asymptotic variance, i.e.,

$$Q_p^* = n \sum_{j=1}^{p} \frac{\widehat{\rho}_j^2}{\tau_j},$$

where:

$$\tau_j = \frac{1}{\widehat{\gamma}_0^2} \sum_{t=1+j}^{n} (Y_t - \overline{Y})^2 (Y_{t-j} - \overline{Y})^2.$$

We have followed Lobato, Nankervis and Savin's (2001) notation and denoted the robustified Q_p by Q_p^*. This statistic has appeared in different versions (see, e.g., Diebold, 1986; Lo and MacKinlay, 1989; Robinson, 1991; Cumby and Huizinga, 1992; Bollerslev and Wooldridge, 1992; Bera and Higgins, 1993). The Q_p^* statistic (or its Ljung–Box analog) should be routinely computed for financial data instead of the standard Q_p (or the LB_p). However, this is not typically the case (see Lobato, Nankervis and Savin, 2001, for details).

For the general case, the asymptotic covariance matrix of the first p autocorrelations is not a diagonal matrix. Hence, for this general case both the Q_p and the Q_p^* tests are invalid. However, under m.d.s. the matrix T can be greatly simplified so that its *ij*-th element takes the form $E[(Y_t - \mu)^2 (Y_{t-i} - \mu)(Y_{t-j} - \mu)]$, which can easily be estimated using its sample analog. This is the approach followed by Guo and Phillips (2001). For the general case, which includes m.d.s. and non-m.d.s. processes, the asymptotic covariance matrix of the first p autocorrelations is a complicated non-diagonal matrix. Hence, for this general case, the literature has proposed the following two modifications of the Q_p test. The first is to modify the

Q_p statistic by introducing a consistent estimator of the asymptotic null covariance matrix of the sample autocorrelations, $\widehat{T}$, so that the modified Q_p statistic retains the χ^2_p asymptotic null distribution. Lobato, Nankervis and Savin (2002) define this statistic as $\widetilde{Q}_p = \left(\sqrt{n}\widehat{\rho}\right)' \widehat{T}^{-1} \left(\sqrt{n}\widehat{\rho}\right)$. The main drawback of this approach is that, in order to construct $\widehat{T}$, a bandwidth parameter has to be introduced (see *ibid.*, for details). This approach works for general dependence structures that allow for the asymptotic covariance matrix of the first p autocorrelations to take any form. The second modification has been studied by Horowitz *et al.* (2006), who employ a bootstrap procedure to estimate consistently the asymptotic null distribution of the Q_p test for the general case. They compare two bootstrap approaches, a single and a double blocks-of-blocks bootstrap, and their final recommendation is to employ a double blocks-of-blocks bootstrap after prewhitening the time series. This solution presents a similar problem, though, namely that the researcher has to choose arbitrarily a block length number. The previous papers considered raw data, but Francq, Roy and Zakoïan (2005) have addressed the use of the Q_p statistic with residuals. They propose to estimate the asymptotic null distribution of the Q_p test statistic for the general weak dependent case. However, their approach still requires the selection of p, and of several additional arbitrary numbers necessary to estimate consistently the needed asymptotic critical values.

These references represent an effort to address the problem of testing for m.d.s. using standard linear measures (autocorrelations) but allowing for nonlinear dependence. Lobato (2001) represents an alternative approach with a similar spirit. The target is to avoid the problem of introducing a user-chosen number and the idea is to construct an asymptotically distribution free statistic. Although this approach delivers tests that handle nonlinear dependence and control properly the Type I error in finite samples, its main theoretical drawback is its inefficiency in terms of local power.

A related statistic, which has been commonly employed in the empirical finance literature (see Cochrane, 1988; Lo and MacKinlay, 1989), is the variance ratio, which takes the form:

$$VR_p = 1 + 2\sum_{j=1}^{p-1}(1 - \frac{j}{p})\widehat{\rho}_j.$$

Under independence, $\sqrt{np}(VR_p - 1)$ is asymptotically distributed as $N(0, 2(p-1))$. Although this test can also be robustified and can be powerful on some occasions, it presents the serious theoretical limitation of being inconsistent. For instance, González and Lobato (2003) considered a moving average of order 2 (MA(2)) process $y_t = e_t - 0.4597e_{t-1} + 0.10124e_{t-2}$. For this process $VR_3 = 0$, in spite of the first two autocorrelations being non-zero. The problem with variance ratio statistics resides in the possible existence of compensations between autocorrelations with different signs, and this may affect power severely. Related to variance ratio (VR) tests, Nankervis and Savin (2007) have proposed a robustified version of Andrews and Ploberger's (1996) test that appears to have very good finite sample power with common empirical finance models. In related work, Delgado and Velasco (2007)

have recently considered a large class of directional tests based on linear combinations of autocorrelations. Their tests are shown to be optimal in certain known local alternative directions and are asymptotically equivalent to Lagrange multiplier tests. Finally, we mention Kuan and Lee (2004), who propose a correlation-based test for the MDH that, instead of using lagged values of Y_t as the function $w(\cdot)$, employs some other arbitrary $w(\cdot)$. This test shares with all the tests analyzed in this section the problem of inconsistency, derived from not using a whole family of functions $w(\cdot)$.

20.3.2 Tests based on an infinite-dimensional conditioning set

The approach presented in the previous sub-section lays naturally in the time domain since a finite number of autocorrelations are tested. However, when the infinite past is considered, the natural framework for performing inference is the frequency domain. The advantage of the frequency domain is the existence of one object, namely the spectral density, that contains the information contained in all the autocovariances. Hence, in the frequency domain, the role previously taken by autocorrelations is now carried by the spectral density function. Define the spectral density $f(\lambda)$ implicitly by:

$$\gamma_k = \int_{\Pi} f(\lambda) \exp(ik\lambda) d\lambda \quad k = 0, 1, 2, \ldots,$$

where $\Pi = [-\pi, \pi]$. Define also the periodogram as $I(\lambda) = |w(\lambda)|^2$, where $w(\lambda) = n^{-1/2} \sum_{t=1}^{n} x_t \exp(it\lambda)$. Although the periodogram is an inconsistent estimator of the spectral density, it can be employed as a building block to construct a consistent estimator. The integral of the spectral density is called the spectral distribution, which, under the MDH, is linear in λ.

For this infinite lag case, the MDH implies as the null hypothesis of interest that $\gamma_k = 0$ for all $k \neq 0$, and equivalently, in terms of the spectral density, the null hypothesis states that $f(\lambda) = \gamma_0/2\pi$ for all $\lambda \in \Pi$.

The advantage of the frequency domain is that the problem of selecting p, which was present in the previous sub-section, does not appear because the null hypothesis is stated in terms of *all* autocorrelations, as summarized by the spectral density or distribution. The classical approach in the frequency domain involves the standardized cumulative periodogram, i.e.,

$$Z_n(\lambda) = \sqrt{T} \left(\frac{\sum_{j=1}^{[\lambda T/\pi]} I(\lambda_j)}{\sum_{j=1}^{T} I(\lambda_j)} - \frac{\lambda}{\pi} \right),$$

where $\lambda_j = 2\pi j/n$, $j = 1, 2, ..., n/2$, are called the Fourier frequencies. Based on $Z_n(\lambda)$, the two classical test statistics are the Kolmogorov–Smirnov:

$$\max_{j=1,\ldots,T} \left| Z_n(\lambda_j) \right|,$$

and the Cramér–von Mises:

$$\frac{1}{T}\sum_{j=1,}^{T} Z_n(\lambda_j)^2.$$

These test statistics have been commonly employed (see Bartlett, 1955; Grenander and Rosenblatt, 1957) because, when the series y_t is not only white noise but also independent (or m.d.s. with additional moment restrictions), it can be shown that the process $Z_n(\lambda)$ converges weakly in $D[0,\pi]$ (the space of cadlag functions in $D[0,\pi]$) to the Brownian bridge process (see Dahlhaus, 1985). Hence, asymptotic critical values are readily available for the independent case. In fact, Durlauf (1991) has shown that the independence assumption can be relaxed to conditional homoskedastic m.d.s. For the m.d.s. case with conditional heteroskedasticity (and some moment conditions), Deo (2000) slightly modified this statistic so that the standardized cumulative periodogram retained its convergence to the Brownian bridge. Deo's test can be interpreted as a continuous version of the robustified Box–Pierce statistic, Q_p^*. Notice that, in Deo's set-up, there is no need to introduce any user-chosen number since, under the stated assumptions (see condition A in *ibid.*, p. 293), the autocorrelations are asymptotically independent. As Deo comments, his assumption (vii) is mainly responsible for the diagonality of the asymptotic null covariance matrix of the sample autocorrelations. However, for many common models, such as GARCH models with asymmetric innovations, EGARCH models and bilinear models, Deo's condition (vii) does not hold and the autocorrelations are not asymptotically independent under the null hypothesis. Hence, for the general case, Deo's test is not asymptotically valid. Deo's Cramér–von Mises test statistic can also be written in the time domain as:

$$DEO_n := n\sum_{j=1}^{n-1} \frac{\widehat{\rho}_j^2}{\tau_j}\left(\frac{1}{j\pi}\right)^2.$$

More general weighting schemes for the sample autocovariances $\widehat{\rho}_j$ than the ones considered here are possible. Under the null hypothesis of m.d.s. and some additional assumptions (see Deo, 2000),

$$DEO_n \xrightarrow{d} \int_0^1 B^2(t)dt \text{ as } n \longrightarrow \infty,$$

where $B(t)$ is the standard Brownian bridge on [0,1]. The 10%, 5% and 1% asymptotic critical values can be obtained from Shorack and Wellner (1986, p. 147) and are 0.347, 0.461 and 0.743, respectively. For extensions of this basic approach see also Paparoditis (2000) and Delgado, Hidalgo and Velasco (2005), among others.

Under general weak dependent assumptions (see Dahlhaus, 1985), the asymptotic null distribution of the process $Z_n(\lambda)$ is no longer the Brownian bridge but, in fact, converges weakly in $D[0,\pi]$ to a zero mean Gaussian process with covariance given by:

$$\frac{\pi G(\pi)}{F(\pi)^2}\left\{\frac{G(\lambda\wedge\mu)}{G(\pi)}+\frac{F(\lambda)F(\mu)}{F(\pi)^2}-\frac{F(\lambda)G(\mu)}{F(\pi)G(\pi)}-\frac{F(\mu)G(\lambda)}{F(\pi)G(\pi)}\right.$$
$$\left.+\frac{F_4(\lambda,\mu)}{G(\pi)}+\frac{F_4(\pi,\pi)}{G(\pi)}\frac{F(\lambda)F(\mu)}{F(\pi)^2}-\frac{F_4(\mu,\pi)}{G(\pi)}\frac{F(\lambda)}{F(\pi)}-\frac{F_4(\lambda,\pi)}{G(\pi)}\frac{F(\mu)}{F(\pi)}\right\},$$

where $F(\lambda)$ denotes the spectral distribution function, $F(\lambda)=\int_0^\lambda f(\omega)d\omega$, $G(\lambda)=\int_0^\lambda f(\omega)^2 d\omega$, and $F_4(\lambda,\mu)=\int_0^\lambda\int_0^\mu f_4(\omega,-\omega,-\theta)d\omega d\theta$, where $f_4(\lambda)$, with $\lambda\in\Pi^{q-1}$, denotes the fourth-order cumulant spectral density (see expression (2.6.2) in Brillinger, 1981, p. 25). The important message from the previous complicated covariance is that the asymptotic null distribution depends on the nature of the data generating process of y_t. Therefore, no asymptotic critical values are available. Chen and Romano (1999, p. 628) propose estimating the asymptotic distribution by means of either the block bootstrap or the sub-sampling technique. Unfortunately, these bootstrap procedures require the selection of some arbitrary number and, in a general framework, no theory is available about their optimal selection. Alternative bootstrap procedures which do not require the selection of a user-chosen number, such as resampling the periodogram as in Franke and Hardle (1992) or Dahlhaus and Janas (1996), will not estimate consistently the asymptotic null distribution because of the fourth-order cumulant terms.

Lobato and Velasco (2004) considered the statistic:

$$M_n=\frac{T^{-1}\sum_{j=1}^T I(\lambda_j)^2}{\left(T^{-1}\sum_{j=1}^T I(\lambda_j)\right)^2}-1,$$

under general weak dependence conditions. This statistic was previously considered by Milhøj (1981), who employed M_n as a general goodness-of-fit statistic for time series. Milhøj informally justified the use of this statistic for testing the adequacy of linear time series models but, since he identified white noise with independent and identically distributed (i.i.d.) (see *ibid.*, p. 177), his analysis does not automatically apply in general contexts. Beran (1992) and Deo and Chen (2000) have also employed the M_n statistic as a goodness-of-fit test for Gaussian processes. Statistical inference is especially simple with M_n, since its asymptotic null distribution is normal even after parametric estimation. We note that the continuous version of M_n can be expressed in the time domain as a statistic proportional to $\sum_{j=0}^{n-1}\widehat{\rho}_j^2$, which shows the difficulty of deriving the asymptotic properties in the time domain since the $\widehat{\rho}_j$ may not be asymptotically independent.

In the time domain, Hong (1996) has considered p as growing with n and, hence, has been able to derive a consistent test in the time domain for the case of regression residuals. In this framework p can be interpreted as a bandwidth number that needs to grow with n, so his test can handle the fact that the null hypothesis implies an infinite number of autocovariances. Hong (1996) restricted attention to the independent case while Hong and Lee (2003) have extended Hong's procedure to allow for conditional heteroskedasticity. However, notice that their framework still restricts the sample autocorrelations to be asymptotically independent.

An alternative solution recently explored by Escanciano and Lobato (2007) consists of modifying the Box–Pierce statistic using an adaptive Neyman test that takes the form:

$$AQ_n = Q_{\widetilde{p}}^*,$$

where:

$$\widetilde{p} = \min\{m : 1 \leq m \leq p_n; L_m \geq L_h, h = 1, 2, \ldots, p_n\}, \tag{20.4}$$

and where:

$$L_p = Q_p^* - \pi(p, n, q).$$

p_n is an upper bound that grows slowly to infinity with n, and:

$$\pi(p, n, q) = \begin{cases} p \log n, & \text{if} \quad \max_{1 \leq j \leq p_n} \left| \dfrac{\widehat{\rho}_j^2}{\tau_j} \right| \leq \sqrt{q \log n} \\ 2p, & \text{if} \quad \max_{1 \leq j \leq p_n} \left| \dfrac{\widehat{\rho}_j^2}{\tau_j} \right| > \sqrt{q \log n}, \end{cases}$$

where q is some fixed positive number. We denote this automatic portmanteau test AQ_n. Our choice of q is 2.4 and is motivated by an extensive simulation study in Inglot and Ledwina (2006) and from simulations in Escanciano and Lobato (2007). Small values of q result in the use of Akaike's criterion, while large $q's$ lead to choosing Schwarz's criterion. Moderate values, such as 2.4, provide a "switching effect" which combines the advantages of the two selection rules: when the alternative is of high frequency (that is, when the only significant autocorrelations are at large lags j), Akaike is used whereas, if the alternative is of low-frequency (i.e., if the first autocorrelations are different from zero), Schwarz is chosen. The adaptive test is an improvement with respect to the traditional Box–Pierce and Hong approaches because the AQ_n test is more powerful and less sensitive to the selection of the bandwidth number p_n than these approaches and, more importantly, it avoids the estimation of the complicated variance-covariance matrix T since its asymptotic distribution is χ_1^2 for general m.d.s. processes.

Summarizing, testing the MDH using linear measures of dependence presents two challenging features. The first aspect is that the null hypothesis implies that an infinite number of autocorrelations are zero. This feature has been addressed successfully in the frequency domain under severe restrictions on the dependence structure of the process. The second feature is that the null hypothesis allows the time series to present some form of dependence beyond the second moments. This dependence entails that the asymptotic null covariance matrix of the sample autocorrelations is not diagonal, so that it has n^2 non-zero terms (contrary to Durlauf, 1991, and Deo, 2000, who consider a diagonal matrix, which has only n non-zero elements). This aspect has been handled by introducing some arbitrary user-chosen numbers whose selection complicates statistical inference.

However, all these tests are suitable for testing for lack of serial correlation but not necessarily for the MDH and, in fact, they are not consistent against non-martingale difference sequences with zero autocorrelations. This happens when

Table 20.2 Linear predictability of exchange rates returns

	Daily				*Weekly*			
	Euro	*Pound*	*Can*	*Yen*	*Euro*	*Pound*	*Can*	*Yen*
$\widehat{\rho}_1$	−0.047	0.001	−0.016	−0.020	0.018	0.046	−0.023	0.054
$\widehat{\rho}_2$	0.003	0.007	−0.028	−0.015	−0.002	−0.008	0.031	−0.024
$\widehat{\rho}_3$	−0.046	−0.055	−0.001	−0.016	0.049	−0.031	0.011	0.010
$\widehat{\rho}_4$	−0.002	0.028	−0.060	0.013	0.024	−0.043	0.015	−0.041
$\widehat{\rho}_5$	−0.002	0.003	−0.063	0.039	0.036	−0.024	0.052	−0.095
LB^*_5	4.071	3.586	8.045	2.452	1.795	2.191	1.781	4.900
LB^*_{15}	15.516	13.256	15.181	6.670	9.139	7.451	10.266	18.861
LB^*_{25}	28.552	26.568	19.756	13.155	18.746	32.584	21.786	24.519
LB^*_{50}	61.922	64.803**	49.887	37.428	42.559	59.107	41.140	58.756
DEO_n	0.253	0.055	0.095	0.063	0.043	0.114	0.050	0.167
AQ_n	1.889	0.021	0.253	0.380	0.151	0.827	0.208	1.105

Note: * and ** indicate significantly different from zero at the 5% and 10% level, respectively.

only nonlinear dependence is present, as is commonly the case with financial data, e.g., exchange rates dynamics. These tests are inconsistent because they only employ information contained in the second moments of the process.

To circumvent this problem we could take into account higher-order moments, as in Hinich and Patterson (1992). They proposed using the bispectrum, i.e., the Fourier transform of the third-order cumulants of the process, but again, this test is not consistent against non-martingale difference sequences with zero third-order cumulants.

In Table 20.2 we report the robust (to conditional heteroskedasticity) version of the first five autocorrelations, the Ljung and Box (1978) test, which is a corrected Q^*_p statistic, which we call LB^*_p, Deo's (2000) modification of Durlauf's test statistic, and the Escanciano and Lobato (2007) test based on AQ_n, to check whether or not our exchange rates changes are uncorrelated. (For further evidence of linear independence, see Figures 20.3–20.10 in section 20.4.2.) Table 20.2 is in agreement with previous findings that have shown that exchange rates have no linear dependence (see, e.g., Table 2 in Hsieh, 1989; Bera and Higgins, 1997; Hong and Lee, 2003, and references therein):

20.4 Tests based on nonlinear measures of dependence

Arguably, testing for the MDH is a challenging problem since, in order to verify it, we must check that a very large class of transformations of the past does not help to predict the current value of the series (see (20.2)). An important step, through the development of consistent tests, was made when econometricians realized that it is not necessary to take a very *large* class of functions in (20.2), but just a convenient parametric class of functions, satisfying certain properties. Domínguez and Lobato (2003) called this methodology the global approach and Escanciano (2007a) called

it the integrated approach. Most of this section will be devoted to a careful study of this approach.

Note, however, that there exists an alternative methodology that is based on the direct estimation of the conditional expectation $E[Y_t|\widetilde{Y}_{t,P}]$, where $\widetilde{Y}_{t,P} = (Y_{t-1}, ..., Y_{t-P})'$ for some finite P. This approach can be called the smoothing approach (since smoothing numbers are required for this non-parametric estimation) or a local approach (see Domínguez and Lobato, 2003). Tests within the local approach have been proposed by Wooldridge (1992), Yatchew (1992), Horowitz and Härdle (1994), Zheng (1996), Fan and Huang (2001), Horowitz and Spokoiny (2001) and Guerre and Lavergne (2005), to mention just a few (see Hart, 1997, for a comprehensive review of the local approach when $P = 1$). Among these tests based on local methods, the test recently proposed by Guay and Guerre (2006) seems to be especially convenient for testing the MDH for two reasons. First, it has been justified for time series under conditional heteroskedasticity of unknown form. Second, it is an adaptive data-driven test. Their test combines a chi-square statistic, based on nonparametric Fourier series estimators for $E[Y_t \mid \widetilde{Y}_{t,P}]$, coupled with a data-driven choice for the number of components in the estimator. To construct their test a nonparametric estimator of the unknown conditional variance is needed. Notice that a relevant practical problem of the local approach arises when P is large or even moderate. The problem is motivated by the sparseness of the data in high-dimensional spaces, which leads most test statistics to suffer considerable bias, even for large sample sizes. In the next sub-section, we will consider an approach that helps to alleviate this problem.

This section focuses on integrated tests. We divide the extensive literature within this integrated approach according to whether the tests consider functions of a finite number of lags or not, i.e., whether $w(I_{t-1}) = w(\widetilde{Y}_{t,P})$ for some $P \geq 1$ or not. We stress at the outset that the main advantage of the tests considered in this section is that they are consistent for testing the MDH (at least when the information set has a finite number of variables), contrary to the tests considered in section 20.3. The main disadvantage is that their asymptotic null distributions are, in general, not standard, which means that no critical values are readily available. In this situation, the typical solution is to employ the bootstrap to estimate these distributions.

20.4.1 Tests based on a finite-dimensional conditioning set

The problem of testing over all possible weighting functions can be reduced to testing the orthogonality condition over a parametric family of functions (see, e.g., Stinchcombe and White, 1998). Although the parametric class still has to include an infinite number of elements, the complexity of the class to be considered is substantially simplified and makes it possible to test for the MDH.

The methods that we review in this sub-section use $w(I_{t-1}) = w_0(\widetilde{Y}_{t,P}, x)$ in (20.2), where, as stated above, $\widetilde{Y}_{t,P} = (Y_{t-1}, ..., Y_{t-P})'$ and w_0 is a known function indexed by a parameter x. That is, these methods check for any form of predictability from the lagged P values of the series. The test statistics are based on a "distance" of the sample analogue of $E[(Y_t - \mu)w_0(\widetilde{Y}_{t,P}, x)]$ from zero.

The exponential function $w_0(\widetilde{Y}_{t,P}, x) = \exp(ix'\widetilde{Y}_{t,P})$, $x \in \mathbb{R}$, was first considered in Bierens (1982, 1984, 1990) (see also Bierens and Ploberger, 1997). One version of the Cramér–von Mises (CvM) test of Bierens (1984) leads to the test statistic:

$$CvM_{n,\exp,P} = n^{-1}\widehat{\sigma}^{-2}\sum_{t=1}^{n}\sum_{s=1}^{n}(Y_t - \overline{Y})(Y_s - \overline{Y})\exp\left(-\frac{1}{2}\left|\widetilde{Y}_{t,P} - \widetilde{Y}_{s,P}\right|^2\right),$$

where:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \overline{Y})^2.$$

Indicator functions $w_0(\widetilde{Y}_{t,P}, x) = 1(\widetilde{Y}_{t,P} \leq x)$, $x \in \mathbb{R}$, were used in Stute (1997) and Koul and Stute (1999) for model checks of regressions and autoregressions, respectively, and in Domínguez and Lobato (2003) for the MDH problem.

Domínguez and Lobato (2003), extending to the multivariate case the results of Koul and Stute (1999), considered the CvM and Kolmogorov–Smirnov (KS) statistics, respectively:

$$CvM_{n,P} \quad : \quad = \frac{1}{\widehat{\sigma}^2 n^2}\sum_{j=1}^{n}\left[\sum_{t=1}^{n}(Y_t - \overline{Y})1(\widetilde{Y}_{t,P} \leq \widetilde{Y}_{j,P})\right]^2,$$

$$KS_{n,P} \quad : \quad = \max_{1\leq i\leq n}\left|\frac{1}{\widehat{\sigma}\sqrt{n}}\sum_{t=1}^{n}(Y_t - \overline{Y})1(\widetilde{Y}_{t,P} \leq \widetilde{Y}_{i,P})\right|.$$

As mentioned above, an important problem of the local approach (also shared by other methods) arises in the case where P is large or even moderate. The sparseness of the data in high-dimensional spaces implies severe biases to most test statistics. This is an important practical limitation for most tests considered in the literature because these biases still persist in fairly large samples. Motivated by this problem, Escanciano (2007a) proposed the use of $w_0(\widetilde{Y}_{t,P}, x) = 1(\beta'\widetilde{Y}_{t,P} \leq u)$, where $x = (\beta, u) \in \mathbb{S}^d \times \mathbb{R}$, with $\mathbb{S}^d = \{\beta \in \mathbb{R}^d : |\beta| = 1\}$, and defined CvM tests based on this choice. We denote by $PCVM_{n,P}$ the resulting CvM test in Escanciano. Also recently, Lavergne and Patilea (2007) proposed a dimension-reduction bootstrap consistent test for regression models based on nonparametric kernel estimators of one-dimensional projections. Their proposal falls in the category of local-based methods, though.

As mentioned earlier, the asymptotic null distribution of integrated tests based on $w_0(\widetilde{Y}_{t,P}, x)$ depends on the data-generating process in a complicated way. Therefore, critical values for the test statistics cannot be tabulated for general cases. One possibility, only explored in the literature for the case $P = 1$ by Koul and Stute (1999), consists of applying the so-called Khmaladze's transformation (Khmaladze, 1981) to get asymptotically distribution free tests. Extensions to $P > 1$ are not available yet. Alternatively, we can approximate the asymptotic null distributions by bootstrap methods. The most relevant bootstrap procedure for testing the MDH has been the wild bootstrap introduced in Wu (1986) and Liu (1988). For example, this

Table 20.3 Testing the MDH of exchange rates returns
Empirical *P*-values

	Daily				*Weekly*			
	Euro	*Pound*	*Can*	*Yen*	*Euro*	*Pound*	*Can*	*Yen*
$CvM_{n,\exp,1}$	0.028	0.322	0.744	0.842	0.453	0.086	0.876	0.488
$CvM_{n,\exp,3}$	0.164	0.320	0.898	0.666	0.743	0.250	0.076	0.258
$CvM_{n,1}$	0.020	0.354	0.628	0.822	0.610	0.146	0.863	0.388
$CvM_{n,3}$	0.192	0.424	0.798	0.588	0.916	0.893	0.720	0.500
$KS_{n,1}$	0.016	0.220	0.502	0.740	0.726	0.176	0.836	0.542
$KS_{n,3}$	0.036	0.280	0.734	0.526	0.986	0.810	0.224	0.654
$PCVM_{n,1}$	0.020	0.354	0.626	0.822	0.610	0.146	0.863	0.388
$PCVM_{n,3}$	0.248	0.438	0.790	0.664	0.746	0.443	0.566	0.414

approach has been employed in Domínguez and Lobato (2003) and Escanciano and Velasco (2006a, 2006b) to approximate the asymptotic distribution of integrated MDH tests. The asymptotic distribution is approximated by replacing $(Y_t - \overline{Y})$ with $(Y_t - \overline{Y})(V_t - \overline{V})$, where $\{V_t\}_{t=1}^n$ is a sequence of independent random variables (RVs) with zero mean, unit variance, bounded support and also independent of the sequence $\{Y_t\}_{t=1}^n$. Here, $\overline{V}$ is the sample mean of $\{V_t\}_{t=1}^n$. The bootstrap samples are obtained by resampling from the distribution of V_t. A popular choice for $\{V_t\}$ is a sequence of i.i.d. Bernoulli variates with $P(V_t = 0.5(1-\sqrt{5})) = (1+\sqrt{5})/2\sqrt{5}$, and $P(V_t = 0.5(1+\sqrt{5})) = 1 - (1+\sqrt{5})/2\sqrt{5}$.

We have applied several tests within the integrated methodology to our exchange rate data. In Table 20.3 we report the wild bootstrap empirical values. In our application we have considered the values $P = 1$ and $P = 3$ for the number of lags used in $CvM_{n,\exp,P}$, $CvM_{n,P}$, $KS_{n,P}$ and $PCVM_{n,P}$. Our results favor the MDH for all exchange rates at both frequencies, weekly and daily, with the exception of the daily euro for $P = 1$. Surprisingly enough, we obtain contradictory results for this exchange rate when $P = 3$. These contradictory results have been previously documented in, e.g., Escanciano and Velasco (2006a), and they may be due to a lack of power of the tests for the $P = 3$ case.

Although the consideration of an omnibus test, like those discussed in this section, is naturally the first idea when there is no a priori information about directions in the alternative hypothesis, it is worth noting that omnibus tests present an important limitation: despite their capability to detect deviations from the null in any direction, it is well-known that they only have reasonable nontrivial local power against very few orthogonal directions (see Janssen, 2000, and Escanciano, 2008, for theoretical explanations and bounds for the number of orthogonal directions).

A possible solution for overcoming the "lack" of power of omnibus tests is provided by the so-called Neyman smooth tests. They were first proposed by Neyman (1937) in the context of goodness-of-fit of distributions and, since then, there has been a plethora of research documenting their theoretical and empirical

Table 20.4 Testing the MDH of exchange rates returns
Bootstrap *P*-values. Data-driven tests

	Daily				*Weekly*			
	Euro	*Pound*	*Can*	*Yen*	*Euro*	*Pound*	*Can*	*Yen*
$T_{n,\widetilde{p}}$	0.049	0.847	0.514	0.876	0.622	0.133	0.747	0.299

properties. In the context of MDH testing, a recent data-driven smooth test has been proposed by Escanciano and Mayoral (2007). Their test is based on the principal components of the marked empirical processes resulting from the choice $w_0(\widetilde{Y}_{t,1}, x) = 1(Y_{t-1} \leq x)$ with $x \in \mathbb{R}$. This test is an extension to nonlinear dependence of order one, i.e., for $P = 1$, of the test based on AQ_n. As shown by these authors, this test possesses excellent local power properties and compares favorably to omnibus tests and other competing tests. The test statistic is:

$$T_{n,\widetilde{p}} = \sum_{j=1}^{\widetilde{p}} \widehat{\epsilon}_{j,n}^2, \tag{20.5}$$

with $\widetilde{p}$ as in (20.4), but with $T_{n,p}$, defined by (20.5), replacing Q_p^* there, and where $\widehat{\epsilon}_{j,n}$ are the sample principal components of a certain CvM test (the reader is referred to Escanciano and Mayoral, 2007, for details). The asymptotic null distribution of $T_{n,\widetilde{p}}$ is a χ_1^2.

We have applied the adaptive data-driven test based on $T_{n,\widetilde{p}}$ to our exchange rate data. The results are reported in Table 20.4 and support our previous conclusions. Only the MDH for the daily euro exchange rate is rejected at 1% with $T_{n,\widetilde{p}}$.

20.4.2 Tests based on an infinite-dimensional information set

The aforementioned statistics test the MDH by conditioning on a finite-dimensional information set and, therefore, may miss some dependence structure in the conditional mean at omitted lags. In principle, maximum power could be achieved by using the correct lag-order P of the alternative. However, prior information on the conditional mean structure is usually not available.

There have been some proposals considering infinite-dimensional information sets. First, de Jong (1996) generalized Bierens' test to time series, and although his test had the appealing property of considering an increasing number of lags as the sample size increases, it required numerical integration with dimension equal to the sample size, which makes this test unfeasible in applications where the sample size is usually large, e.g., financial applications. Second, Domínguez and Lobato (2003) suggest constructing a test statistic as a weighted average of all the test statistics established for a fixed number of lags. However, Domínguez and Lobato did not analyze the test any further, nor the selection of the measure to weight the different statistics.

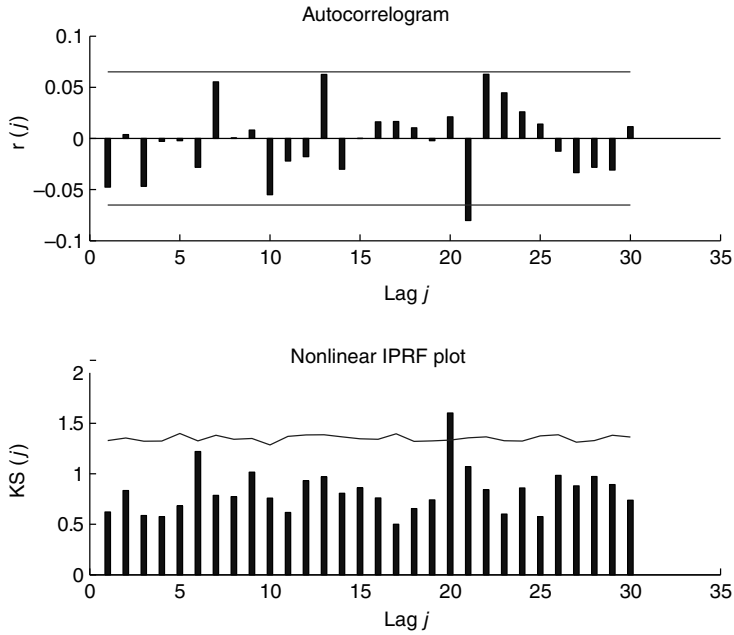


Figure 20.3 IPRF for the daily euro
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

Using a different methodology based on the generalized spectral density approach of Hong (1999), Hong and Lee (2003) proposed an MDH bootstrap test (see also Hong and Lee, 2005). Tests based on the generalized spectral density involve three choices: a kernel, a bandwidth parameter, and an integrating measure; and, in general, statistical inferences are sensitive to these choices. This fact motivated Escanciano and Velasco (2006a, 2006b) to propose testing the MDH by means of a generalized spectral distribution function.

The generalized spectral approach is based on the fact that the MDH implies that:

$$H_0 : \gamma_{j,w}(x) = 0 \quad \forall j \geq 1, \text{ for all } x, \tag{20.6}$$

where $\gamma_{j,w}(x) = E[(Y_t - \mu)w_0(Y_{t-j}, x)]$ and where $w_0(Y_{t-j}, x)$ is any of the parametric functions of the previous section. The generalized spectral approach of Hong is based on the choice $w_0(Y_{t-j}, x) = \exp(ixY_{t-j})$. Escanciano and Velasco (2006a) considered the latter choice, while Escanciano and Velasco (2006b) used $w_0(Y_{t-j}, x) = 1(Y_{t-j} \leq x)$, and called the measure $\gamma_{j,ind}(x) = E[(Y_t - \mu)1(Y_{t-j} \leq x)]$ the integrated pairwise autoregression function (IPAF). The name follows from the fact that:

$$\gamma_{j,ind}(x) = E[(Y_t - \mu)1(Y_{t-j} \leq x)] = \int_{-\infty}^{x} E[Y - \mu \mid Y_{t-j} = z]F(dz),$$

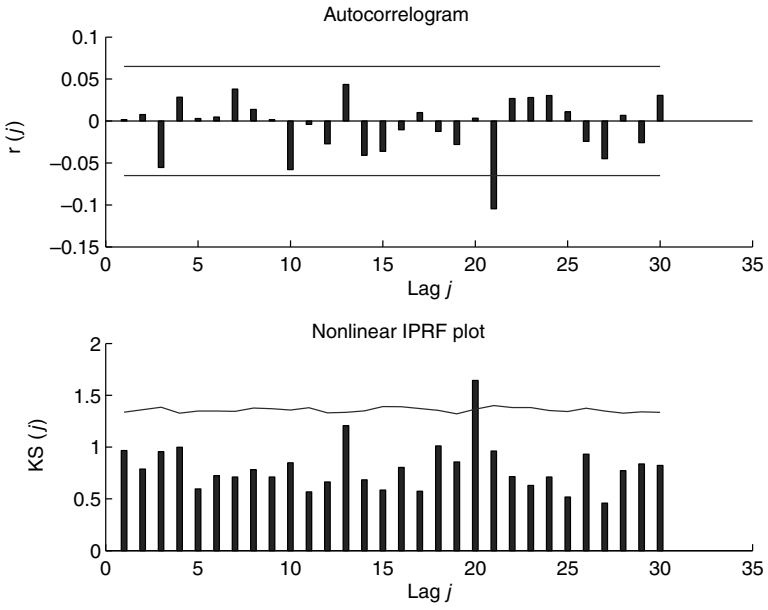


Figure 20.4 IPRF for the daily pound
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

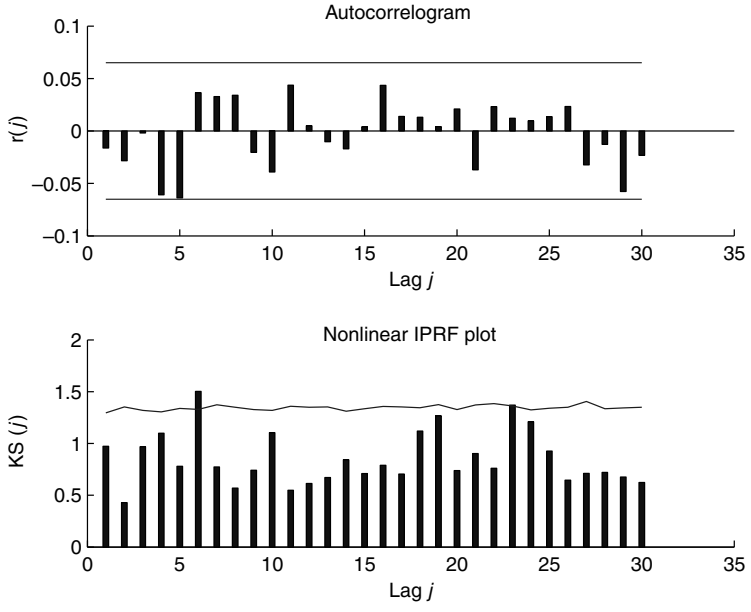


Figure 20.5 IPRF for the daily Can
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

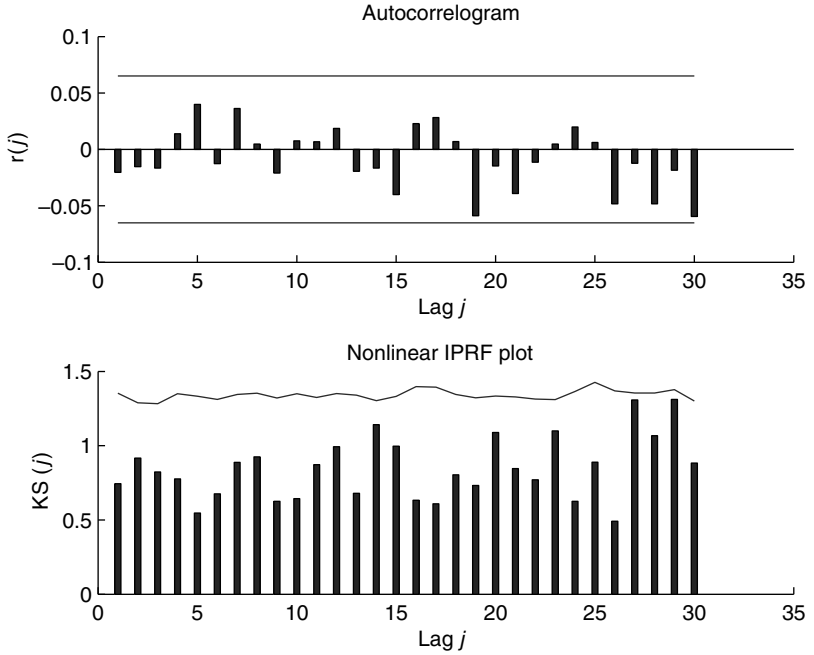


Figure 20.6 IPRF for the daily yen
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

where F is the cumulative distribution function (c.d.f.) of Y_t. The measure $\gamma_{j,w}(x)$ can be viewed as a generalization of the usual autocovariance to measure the conditional mean dependence in a nonlinear time series framework. It can easily be estimated from a sample. For example, the IPAF can be estimated by:

$$\widehat{\gamma}_{j,ind}(x) = \frac{1}{n-j}\sum_{t=1+j}^{n}(Y_t - \overline{Y})1(Y_{t-j} \leq x). \tag{20.7}$$

Moreover, as proposed by Escanciano and Velasco (2006b), nonlinear correlograms can be used to formally assess the nonlinear dependence structure in the conditional mean of the series. These authors define the KS test statistic as:

$$KS(j) := \sup_{x\in[-\infty,\infty]^d}\left|(n-j)^{\frac{1}{2}}\widehat{\gamma}_{j,ind}(x)\right| = \max_{1+j\leq t\leq n}\left|(n-j)^{\frac{1}{2}}\widehat{\gamma}_{j,ind}(Y_{t-j})\right|.$$

The asymptotic quantiles of $KS(j)$ under the MDH can be approximated via a wild bootstrap approach. With these bootstrap critical values we can calculate uniform confidence bands for $\widehat{\gamma}_j(x)$ and the significance of $\gamma_j(x)$ can be tested. The plot of a standardization of $KS(j)$ against the lag parameter $j \geq 1$ can be viewed as a generalization of the usual autocovariance plot in linear dependence to nonlinear conditional mean dependence. Escanciano and Velasco (2006b) call this plot the integrated pairwise regression function (IPRF) plot.

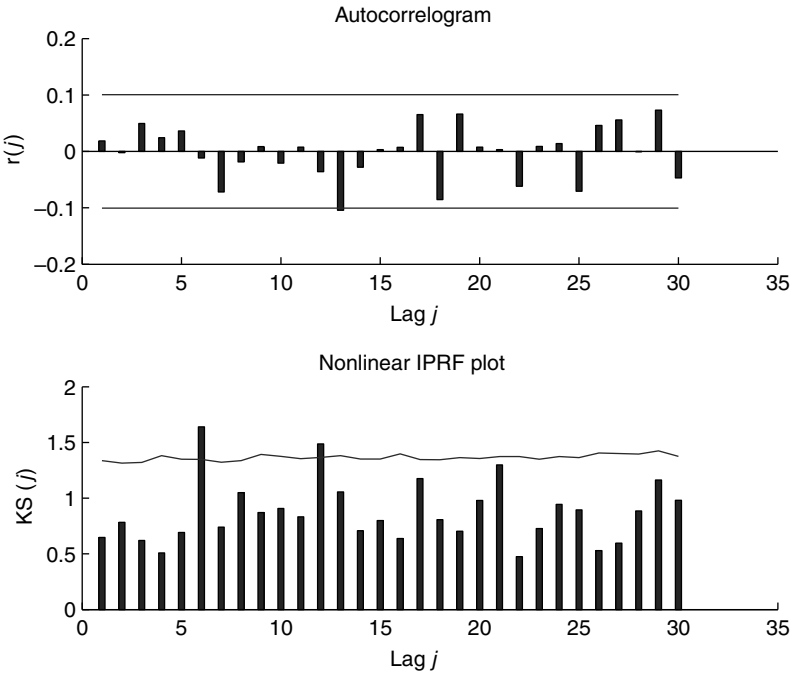


Figure 20.7 IPRF for the weekly euro
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

In Figures 20.3–20.10 we plot the IPRF for our exchange rate returns. The common feature of these graphs is the lack of dependence in the exchange rates, both linear and nonlinear. Only a few isolated statistics seem to be significant, but the evidence is very weak. It seems that the IPRF supports the MDH for these datasets.

We now describe a generalized spectral approach to enable us to consider simultaneously all the nonlinear measures of dependence $\{\gamma_{j,w}(\cdot)\}$. Define $\gamma_{-j,w}(\cdot) = \gamma_{j,w}(\cdot)$ for $j \geq 1$, and consider the Fourier transform of the functions $\gamma_{j,w}(x)$,

$$f_w(\varpi, x) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_{j,w}(x) e^{-ij\varpi} \quad \forall \varpi \in [-\pi, \pi]. \tag{20.8}$$

Note that $f_w(\varpi, x)$ is able to capture pairwise non-martingale difference alternatives with zero autocorrelations. Under the MDH, the condition $f_{0,w}(\varpi, x) = (2\pi)^{-1}\gamma_0(x)$ holds, i.e., the generalized spectral density $f_w(\varpi, x)$ is flat in ϖ. Hong (1999) proposed the estimators:

$$\widehat{f}_w(\varpi, x) = \frac{1}{2\pi} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right)^{\frac{1}{2}} k\left(\frac{j}{p}\right) \widehat{\gamma}_{j,\exp}(x) e^{-ij\varpi},$$

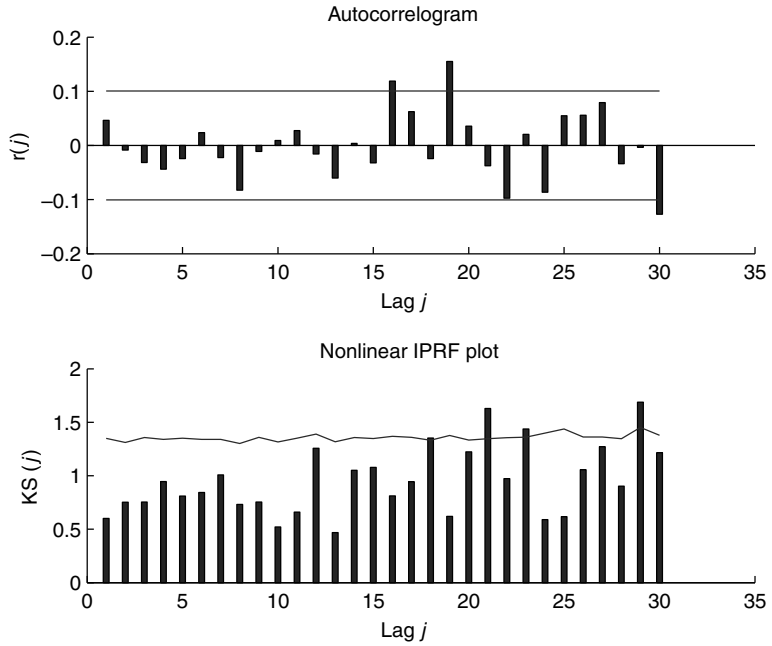


Figure 20.8 IPRF for the weekly pound
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

and:

$$\widehat{f}_{0,w}(\varpi, x) = \frac{1}{2\pi}\widehat{\gamma}_{0,w}(x),$$

to test the MDH, where $k(\cdot)$ is a symmetric kernel and p a bandwidth parameter. He considered a standardization of an L_2-distance using a weighting function $W(\cdot)$:

$$L^2_{2,n}(p) = \frac{\pi}{2}\int_{\mathbb{R}}\int_{-\pi}^{\pi} n\left|\widehat{f}_w(\varpi, x) - \widehat{f}_{0,w}(\varpi, x)\right|^2 W(dx)d\varpi \tag{20.9}$$

$$= \sum_{j=1}^{n-1}(n-j)k^2\left(\frac{j}{p}\right)\int_{\mathbb{R}}\left|\widehat{\gamma}_{j,w}(x)\right|^2 W(dx).$$

Under the null of MDH and some additional assumptions, Hong and Lee (2005) showed that a convenient standardization of $L^2_{2,n}(p)$ converges to a standard normal random variable. The centering and scaling factors in this standardization depend on the higher dependence structure of the series.

Alternatively, the generalized spectral distribution function is:

$$H_w(\lambda, x) = 2\int_0^{\lambda\pi} f_w(\varpi, x)d\varpi \qquad \lambda \in [0, 1],$$

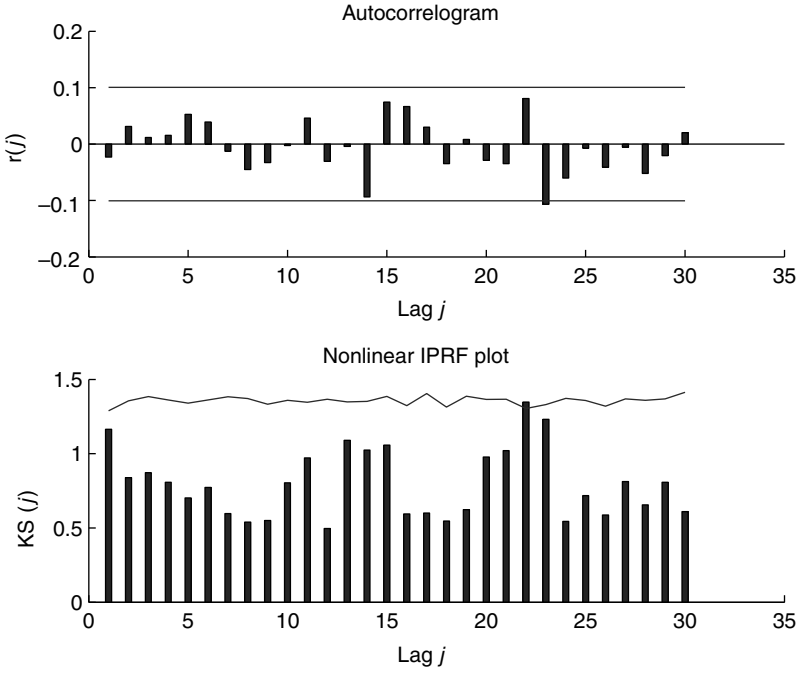


Figure 20.9 IPRF for the weekly Can
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

which, after some algebra, boils down to:

$$H_w(\lambda, x) = \gamma_{0,w}(x)\lambda + 2\sum_{j=1}^{\infty}\gamma_{j,w}(x)\frac{\sin j\pi\lambda}{j\pi}. \tag{20.10}$$

Tests can be based on the sample analogue of (20.10), i.e.:

$$\widehat{H}_w(\lambda, x) = \widehat{\gamma}_{0,w}(x)\lambda + 2\sum_{j=1}^{n-1}(1 - \frac{j}{n})^{\frac{1}{2}}\widehat{\gamma}_{j,w}(x)\frac{\sin j\pi\lambda}{j\pi},$$

where $(1 - \frac{j}{n})^{\frac{1}{2}}$ is a finite sample correction factor. The effect of this correction factor is to put less weight on very large lags, for which we have less sample information. Note that under the MDH, $H_w(\lambda, x) = \gamma_0(x)\lambda$, so that tests for MDH can be constructed based on the discrepancy between $\widehat{H}_w(\lambda, x)$ and $\widehat{H}_{0,w}(\lambda, x) := \widehat{\gamma}_0(x)\lambda$. That is, we can consider the process:

$$S_{n,w}(\lambda, x) = \left(\frac{n}{2}\right)^{\frac{1}{2}}\{\widehat{H}_w(\lambda, x) - \widehat{H}_{0,w}(\lambda, x)\} = \sum_{j=1}^{n-1}(n-j)^{\frac{1}{2}}\widehat{\gamma}_{j,w}(x)\frac{\sqrt{2}\sin j\pi\lambda}{j\pi}, \tag{20.11}$$

to test for the MDH.

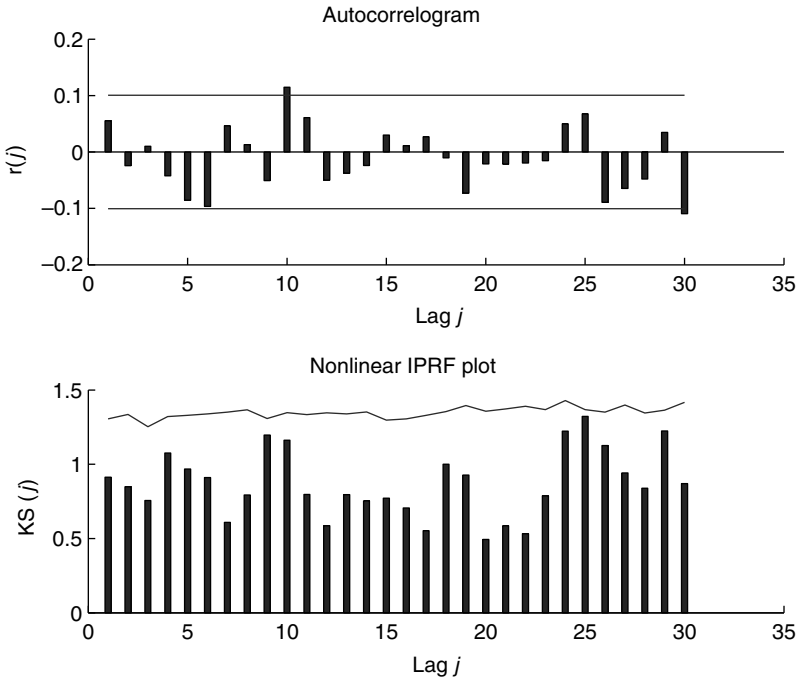


Figure 20.10 IPRF for the weekly yen
Top graph is the heteroskedasticity robust autocorrelation plot. Bottom graph is the IPRF plot.

In order to evaluate the distance of $S_n(\lambda, x)$ from zero, a norm has to be chosen. One norm considered in practice is the Cramér–von Mises norm:

$$D^2_{n,w} = \int_{\mathbb{R}} \int_0^1 \left| S_{n,w}(\lambda, x) \right|^2 W(dx) d\lambda = \sum_{j=1}^{n-1} (n-j) \frac{1}{(j\pi)^2} \int_{\mathbb{R}} \left| \widehat{\gamma}_{j,w}(x) \right|^2 W(dx), \quad (20.12)$$

where $W(\cdot)$ is a weighting function satisfying some mild conditions. $D^2_{n,w}$ has the attractive convenience of being free of choosing any smoothing parameter or kernel, and it has been documented to deliver tests with good power properties (cf. Escanciano and Velasco, 2006a, 2006b).

Among the members of this class of test statistics, the most common choices are:

$$D^2_{n,\exp} = \widehat{\sigma}^{-2} \sum_{j=1}^{n-1} (n-j) \frac{1}{(j\pi)^2} \sum_{t=j+1}^{n} \sum_{s=j+1}^{n} (Y_t - \overline{Y}_{n-j})(Y_s - \overline{Y}_{n-j}) \exp(-0.5(Y_{t-j} - Y_{s-j})^2),$$

and:

$$D^2_{n,ind} = \widehat{\sigma}^{-2} \sum_{j=1}^{n-1} \frac{(n-j)}{n(j\pi)^2} \sum_{t=1}^{n} \widehat{\gamma}^2_{j,ind}(X_t),$$

Table 20.5 Testing the MDH of exchange rates returns
Bootstrap *P*-values. Generalized spectral tests

	Daily				*Weekly*			
	Euro	*Pound*	*Can*	*Yen*	*Euro*	*Pound*	*Can*	*Yen*
$D^2_{n,\exp}$	0.023	0.450	0.680	0.913	0.670	0.123	0.360	0.586
$D^2_{n,ind}$	0.016	0.343	0.640	0.923	0.800	0.253	0.526	0.524

where $\widehat{\gamma}_{j,ind}$ is given in (20.7). The test statistic $D^2_{n,\exp}$ is based on $w_0(Y_{t-j}, x) = \exp(ixY_{t-j})$ and the standard normal c.d.f. as the weighting function $W(\cdot)$, whereas $D^2_{n,ind}$ is based on $w_0(Y_{t-j}, x) = 1(Y_{t-j} \leq x)$ and the empirical c.d.f. as the function W.

We have applied these two generalized spectral distribution based tests to our exchange rate data. The results are reported in Table 20.5 and support our previous conclusions. Only the MDH for the daily euro exchange rate is rejected.

20.5 Related hypotheses

In this chapter we have considered testing the MDH which, in statistical terms, just implies that the mean of an economic time series is independent of its past. The procedures studied in this chapter can be straightforwardly applied to testing the following generalization of the MDH:

$$H_0 : E[Y_t \mid X_{t-1,}X_{t-2,}...] = \mu, \quad \mu \in \mathbb{R},$$

where Y_t is a measurable real-valued transformation of X_t and $\mu = E[Y_t]$. This null hypothesis, which is referred to as the generalized MDH, contains many interesting testing problems as special cases. For instance, when Y_t is a power transformation of X_t, this null hypothesis implies constancy of conditional moments. The leading case in financial applications is where $Y_t = X_t^2$, because when X_t follows an m.d.s., this null hypothesis means that there is no volatility in the series X_t, i.e., X_t is conditionally homoskedastic. The cases $Y_t = X_t^3$ or $Y_t = X_t^4$ would, respectively, test for no dynamic structure in the third moment (conditionally constant skewness) and fourth moment (conditionally constant kurtosis) (see, e.g., Bollerslev, 1987; Engle and González-Rivera, 1991). Another relevant case is when $Y_t = 1(X_t > c)$, $c \in \mathbb{R}^d$. In this case, the null hypothesis represents no directional predictability (see, e.g., Linton and Whang, 2007). Another situation of interest occurs when the null hypothesis is the equality of the regression curves of two random variables, X_{1t} and X_{2t}, say; in this case, $Y_t = X_{1t} - X_{2t}$, $\mu = 0$ (see Ferreira and Stute, 2004, for a recent reference).

Note also that most of the procedures considered in this chapter are also applicable for testing the null hypothesis that a general dynamic nonlinear model is

correctly specified. In this situation, the null hypothesis of interest establishes that:

$$\exists \theta_0 : E[\psi(Y_t, X_t, \theta_0) \mid X_t] = 0,$$

where ψ is a given function, Y_t is a vector of endogenous variables and X_t is a vector of exogenous variables. Test statistics can be constructed along the lines described in this chapter. The main theoretical challenge in this framework is the way of handling the estimation of the parameters. There are basically three alternative approaches. First, to estimate the asymptotic null distribution of the relevant test statistic by estimating its spectral decomposition (e.g., Horowitz, 2006; Carrasco, Florens and Renault, 2007). Second, to use the bootstrap to estimate this distribution (see Wu, 1986; Stute, González-Manteiga and Presedo-Quindmil, 1998). Third, to transform the test statistic via martingalization to yield an asymptotically distribution-free test statistic.

Finally, in this chapter we have considered testing for m.d.s. instead of testing for a martingale. Recall that X_t is a martingale, with respect to its natural filtration, when $E\left[X_t \mid X_{t-1}, X_{t-2}, \ldots\right] = X_{t-1}$ *a.s.* Testing for a martingale presents the additional challenge of handling non-stationary variables. Park and Whang (2005) considered testing that a first-order Markovian process follows a martingale by testing that the first difference of the process, conditionally on the last value, has zero mean, i.e.,

$$E(X_t - X_{t-1} \mid X_{t-1}) = 0 \ a.s. \tag{20.13}$$

Hence they allow for a singular non-stationary conditioning variable. This restrictive Markovian framework has the advantage of leading to test statistics which are asymptotically distribution free and, hence, they do not need to transform their statistics or to use bootstrap procedures to obtain critical values. Similarly, note that many of the procedures described in section 20.4 also lead to asymptotically distribution free tests in this restrictive framework. As shown in Escanciano (2007b), the extension to the multivariate conditioning case in (20.13) leads to non-pivotal tests and some resampling procedure is necessary.

20.6 Conclusions

This chapter has presented a general panoramic of the literature of testing for the MDH. This research started at the beginning of the last century by developing tests for serial correlation and experienced renewed interest recently because of the nonlinear dependence present in economic and, especially, financial series. The initial statistical tools were based on linear dependence measures such as autocorrelations or the spectral density function. These tools were initially motivated by the observation that economic time series follow normal distributions. Since the last 25 years has stressed the non-normal behavior of financial series, the statistics and econometrics literature has followed two alternative approaches. The first's target was to robustify the well-established linear measures to allow for non-linear dependence. This approach has the advantage of simplicity, since it leads typically to standard asymptotic null distributions. However, its main limitation is that it

cannot detect nonlinear dependence. The second approach considered nonlinear measures of dependence. Its advantage is that it is more powerful, its disadvantage is that asymptotic null distributions are non-standard. Nowadays, this feature is hardly a drawback because the increasing availability of computing resources has allowed the implementation of bootstrap procedures, which can estimate the asymptotic null distributions with relative ease.

The definition of a martingale involves the information set of the agent that typically contains the infinite past of the economic series. This feature implies that, in practice, it is practically impossible to construct a test which, although it may be consistent theoretically, has power for any possible violation of the null hypothesis. The *pairwise* approach, which admittedly does not deliver consistent tests, nevertheless leads to tests with reasonable power for common alternatives. Another sensible possibility for reducing this dimensionality problem is to consider alternatives of a single-index structure, i.e., where the conditioning set is given by a univariate, possibly unknown, projection of the infinite-dimensional information set. More research is clearly needed in this direction.

In this chapter we have illustrated the different methodologies with exchange rate data that typically satisfy the MDH, as we have seen. Stock market data are not such a clear-cut case. Rejecting the MDH leads to the challenge of selecting a proper model. In this respect, data-driven adaptive tests are informative, since they provide an alternative model in the case of rejection. Notably, the principal component analysis provided in Escanciano and Mayoral (2007) represents a clear, theoretically well-motivated approach that, coupled with an effective choice for the number of components, can help in this selection process.

Acknowledgments

We are very grateful to Terry Mills for a careful reading of a previous version of this chapter. Escanciano acknowledges financial support from the Spanish Ministerio de Educación y Ciencia, reference numbers SEJ2004-04583/ECON and SEJ2005-07657/ECON. Lobato acknowledges financial support from the Mexican CONACYT, reference number 59028, and from Asociación Mexicana de Cultura.

References

Andrews, D.W.K. and W. Ploberger (1996) Testing for serial correlation against an ARMA (1, 1) process. *Journal of the American Statistical Association* **91**, 1331–42.

Barnett, W. and A. Serletis (2000) Martingales, nonlinearity, and chaos. *Journal of Economics. Dynamics and Control* **24**, 703–24.

Bartlett, M.S. (1955) *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*. London: Cambridge University Press.

Bekaert, G. and R. Hodrick (1992) Characterizing predictable components in excess returns on equity and foreign exchange markets. *Journal of Finance* **47**, 467–509.

Bera, A.K. and M.L. Higgins (1993) ARCH models: properties, estimation and testing. *Journal of Economic Surveys* **7**, 305–62.

Bera, A.K. and M.L. Higgins (1997) ARCH and bilinearity as competing models for nonlinear dependence. *Journal of Business and Economic Statistics* **15**, 43–51.

Beran, J. (1992) A goodness-of-fit test for time series with long range dependence. *Journal of the Royal Statistical Society, Series B* **54**, 749–60.
Bierens, H.J. (1982) Consistent model specification tests. *Journal of Econometrics* **20**, 105–34.
Bierens, H.J. (1984) Model specification testing of time series regressions. *Journal of Econometrics* **26**, 323–53.
Bierens, H.J. (1990) A consistent conditional moment test of functional form. *Econometrica* **58**, 1443–58.
Bierens, H.J. and W. Ploberger (1997) Asymptotic theory of integrated conditional moment tests. *Econometrica* **65**, 1129–51.
Bollerslev, T. (1987) A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics* **69**, 542–47.
Bollerslev, T. and J.M. Wooldridge (1992) Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews* **11**, 143–79.
Box, G.E.P. and D.A. Pierce (1970) Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association* **65**, 1509–26.
Brillinger, D.R. (1981) *Time Series: Data Analysis and Theory*. San Francisco: Holden Day.
Campbell, J.Y., A.W. Lo and A.C. MacKinlay (1997) *The Econometrics of Financial Markets*. Princeton: Princeton University Press.
Carrasco, M., J.P. Florens and E. Renault (2007) Linear inverse problems in structural econometrics. In J.J. Heckman and E.E. Leamer (eds.), *Handbook of Econometrics*, *Volume 6*. Amsterdam: North-Holland.
Chan, K.S. and H. Tong (2001) *Chaos: A Statistical Perspective*. New York: Springer.
Chen, H. and J.P. Romano (1999) Bootstrap-assisted goodness-of-fit tests in the frequency domain. *Journal of Time Series Analysis* **20**, 619–54.
Clark, T.E. and K.D. West (2006) Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics* **135**, 155–86.
Cochrane, J.H. (1988) How big is the random walk in GNP? *Journal of Political Economy* **96**, 893–920.
Cochrane, J.H. (2005) *Asset Pricing*. Princeton: Princeton University Press.
Cumby, R.E. and J. Huizinga (1992) Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variable regressions. *Econometrica* **60**, 185–96.
Dahlhaus, R. (1985) On the asymptotic distribution of Bartlett's U_p-statistic. *Journal of Time Series Analysis* **6**, 213–27.
Dahlhaus, R. and D. Janas (1996) A frequency domain bootstrap for ratio statistics in time series analysis. *Annals of Statistics* **24**, 1934–65.
de Jong, R.M. (1996) The Bierens' tests under data dependence. *Journal of Econometrics* **72**, 1–32.
Delgado, M.A., J. Hidalgo and C. Velasco (2005) Distribution free goodness-of-fit tests for linear models. *Annals of Statistics* **33**, 2568–609.
Delgado, M.A. and C. Velasco (2007) A new class of distribution-free tests for time series models specification. Preprint.
Deo, R.S. (2000) Spectral tests of the martingale hypothesis under conditional heteroscedasticity. *Journal of Econometrics* **99**, 291–315.
Deo, R.S. and W. Chen (2000) On the integral of the squared periodogram. *Stochastic Processes and their Applications* **85**, 159–76.
Diebold, F.X. (1986) Testing for serial correlation in the presence of heteroskedasticity. *Proceedings of the American Statistical Association*, Business and Economics Statistics Section, 323–8.
Diebold, F.X and J.A. Nason (1990) Nonparametric exchange rate predictions? *Journal of International Economics* **28**, 315–32.
Dominguez, M. and I.N. Lobato (2003) Testing the martingale difference hypothesis. *Econometric Reviews* **22**, 351–77.

Durbin, J. and G.S. Watson (1950) Testing for serial correlation in least squares regression: I. *Biometrika* **37**, 409–28.
Durlauf, S.N. (1991) Spectral based testing of the martingale hypothesis. *Journal of Econometrics* **50**, 355–76.
Engle, R.F. and G. González-Rivera (1991) Semiparametric ARCH models. *Journal of Business & Economic Statistics* **9**, 345–59.
Escanciano, J.C. (2007a) Model checks using residual marked empirical processes. *Statistica Sinica* **17**, 115–38.
Escanciano, J.C. (2007b) Weak convergence of non-stationary multivariate marked processes with applications to martingale testing. *Journal of Multivariate Analysis* **98**, 1321–36.
Escanciano, J.C. (2008) On the lack of power of omnibus specification tests. *Econometric Theory*. Forthcoming.
Escanciano, J.C. and I.N. Lobato (2007) Data-driven portmanteau tests for testing for serial correlation. Preprint.
Escanciano, J.C. and S. Mayoral (2007) Data-driven smooth tests for the martingale difference hypothesis. Preprint.
Escanciano, J.C. and C. Velasco (2006a) Generalized spectral tests for the martingale difference hypothesis. *Journal of Econometrics* **134**, 151–85.
Escanciano, J.C. and C. Velasco (2006b) Testing the martingale difference hypothesis using integrated regression functions. *Computational Statistics & Data Analysis* **51**, 2278–94.
Fama, E.F. (1991) Efficient capital markets: a review of theory and empirical work. *Journal of Finance* **25**, 383–417.
Fan, J. and L. Huang (2001) Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association* **96**, 640–52.
Ferreira, E. and W. Stute (2004) Testing for differences between conditional means in a time series context. *Journal of the American Statistical Association* **99**, 169–74.
Fong, W.M., S.K. Koh and S. Ouliaris (1997) Joint variance ratio test of the martingale hypothesis for exchange rates. *Journal of Business and Economics Statistics* **15**, 51–9.
Fong, W.M. and S. Ouliaris (1995) Spectral tests of the martingale hypothesis for exchange rates. *Journal of Applied Econometrics* **10**, 255–71.
Francq, C., R. Roy and J.-M. Zakoïan (2005) Diagnostic checking in ARMA models with uncorrelated errors. *Journal of the American Statistical Association* **13**, 532–44.
Franke, J. and W. Hardle (1992) On bootstrapping kernel spectral estimates. *Annals of Statistics* **20**, 121–45.
González, M. and I.N. Lobato (2003) Contrastes de autocorrelación. *Gaceta de Economía* **14**, 41–57.
Grenander, U. and M. Rosenblatt (1957) *Statistical Analysis of Stationary Time Series*. New York: Chelsea Publishing Company.
Guay, A. and E. Guerre (2006) A Data-driven nonparametric specification test for dynamic regression models. *Econometric Theory* **22**, 543–86.
Guerre, E. and P. Lavergne (2005) Rate-optimal data-driven specification testing for regression models. *Annals of Statistics* **33**, 840–70.
Guo, B. and P.C.B. Phillips (2001) Testing for autocorrelation and unit roots in the presence of conditional heteroskedasticity of unknown form. New Haven: Cowles Foundation for Research in Economics, Yale University.
Hall, R.E. (1978) Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* **86**, 971–87.
Hart, J.D. (1997) *Nonparametric smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.
Hinich, M. and D. Patterson (1992) A new diagnostic test of model inadequacy which uses the martingale difference criterion. *Journal of Time Series Analysis* **13**, 233–52.
Hong, Y. (1996) Consistent testing for serial correlation of unknown form. *Econometrica* **64**, 837–64.

Hong, Y. (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association* **84**, 1201–20.

Hong, Y. and T.H. Lee (2003) Inference on predictability of foreign exchange rate changes via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* **85**, 1048–62.

Hong, Y. and Y.J. Lee (2005) Generalized spectral tests for conditional mean models in time series with conditional heteroskedasticity of unknown form. *Review of Economic Studies* **43**, 499–541.

Horowitz, J. (2006) Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica* **74**, 521–38.

Horowitz, J.L. and W. Härdle (1994) Testing a parametric model against a semiparametric alternative. *Econometric Theory* **10**, 821–48.

Horowitz, J., I.N. Lobato, J. Nankervis and N.E. Savin (2006) Bootstrapping the Box-Pierce Q test: a robust test of uncorrelatedness. *Journal of Econometrics* **133**, 841–62.

Horowitz, J.L. and V.G. Spokoiny (2001) An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69**, 599–632.

Hsieh, D.A. (1988) Statistical properties of daily foreign exchange rates. *Journal of International Economics* **24**, 129–45.

Hsieh, D.A. (1989) Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business* **62**, 339–68.

Hsieh, D.A. (1993) Implication of nonlinear dynamics for financial risk management. *Journal of Finance and Quantitative Analysis* **28**, 41–64.

Inglot, T. and T. Ledwina (2006) Data Driven Score Tests of Fit for Semiparametric Homoscedastic Linear Regression Model. Preprint 665, Institute of Mathematics, Polish Academy of Sciences.

Inoue, A. and L. Kilian (2004) In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Reviews* **23**(4), 371–402.

Janssen, A. (2000) Global power functions of goodness of fit tests. *Annals of Statistics* **28**, 239–53.

Khmaladze, E.V. (1981) Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its Applications* **26**, 240–57.

Koul, H.L. and W. Stute (1999) Nonparametric model checks for time series. *Annals of Statistics* **27**, 204–36.

Kuan, C.-M. and W. Lee (2004) A new test of the martingale difference hypothesis. *Studies in Nonlinear Dynamics & Econometrics* **8**, Article 1.

Laudrup N. and M. Jansson (2006) Improving size and power in unit root testing. In T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics*. *Volume I: Econometric Theory*, Ch. 7, pp. 252–77. Basingstoke: Palgrave Macmillan.

Lavergne, P. and V. Patilea (2007) One for all and all for one: dimension reduction for regression checks. *Journal of Econometrics*. Forthcoming.

LeBaron, B. (1999) Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics* **49**, 125–43.

LeRoy, S.F. (1989) Efficient capital markets and martingales. *Journal of Economic Literature* **27**, 39–44.

Levich, R. and L. Thomas (1993) The significance of technical trading rule profits in the foreign exchange market: a bootstrap approach. *Journal of International Money and Finance* **12**, 451–74.

Linton, O. and Y.-J. Whang (2007) The quantilogram: with an application to evaluating directional predictability. *Journal of Econometrics* **141**, 250–82.

Liu, C.Y. and He, J. (1991) A variance ratio test for random walks in foreign exchange rates. *Journal of Finance* **46**, 773–85.

Liu, R.Y. (1988) Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics* **16**, 1696–708.
Ljung, G.M. and G.E.P. Box (1978) On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.
Lo, A.W. (1997) *Market Efficiency: Stock Market Behaviour in Theory and Practice*, Volumes I and II. Cheltenham: Edward Elgar.
Lo, A.W. and A.C. MacKinlay (1989) The size and power of the variance ratio test in finite samples: a Monte Carlo investigation. *Journal of Econometrics* **40**, 203–38.
Lobato, I.N. (2001) Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association* **96**, 1066–76.
Lobato, I.N., J.C. Nankervis and N.E. Savin (2001) Testing for autocorrelation using a modified Box–Pierce Q test. *International Economic Review* **42**, 187–205.
Lobato, I.N., J.C. Nankervis and N.E. Savin (2002) Testing for zero autocorrelation in the presence of statistical dependence. *Econometric Theory* **18**, 730–43.
Lobato, I.N. and C. Velasco (2004) A general test for white noise. Preprint.
McCurdy, T.H. and I.G. Morgan (1987) Testing of the martingale hypothesis for foreign currency futures with time-varying volatility. *International Journal of Forecasting* **3**, 131–48.
McCurdy, T.H. and I.G. Morgan (1988) Testing of the martingale hypothesis in Deutsche Mark futures with model specifying the form of heteroskedasticity. *Journal of Applied Econometrics* **3**, 187–202.
Meese, R.A. and K. Rogoff (1983a) Empirical exchange rates models of the seventies. *Journal of International Economics* **14**, 3–24.
Meese, R.A. and K. Rogoff (1983b) The out of sample failure of empirical exchange rates models: sampling error or misspecification? In J. Frenkel (ed.), *Exchange Rates and International Economics*. Chicago: University of Chicago Press.
Milhøj, A. (1981) A test of fit in time series models. *Biometrika* **68**, 177–87.
Nankervis, J.C. and N.E. Savin (2007) Testing for serial correlation: generalized Andrews-Ploberger tests. Preprint.
Neyman, J. (1937) Smooth test for goodness of fit. *Scandinavian Aktuarietidskr* **20**, 149–99.
Paparoditis, E. (2000) Spectral density based goodness-of-fit tests for time series models. *Scandinavian Journal of Statistics* **27**, 143–76.
Park, J.Y. and Y.J. Whang (2005) Testing for the martingale hypothesis. *Studies in Nonlinear Dynamics and Econometrics* **2**, Article 2.
Robinson, P.M. (1991) Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* **47**, 67–84.
Romano, J.L. and L.A. Thombs (1996) Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association* **91**, 590–600.
Shorack, G. and J. Wellner (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.
Stinchcombe, M. and H. White (1998) Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* **14**, 295–325.
Stute, W. (1997) Nonparametric model checks for regression. *Annals of Statistics* **25**, 613–41.
Stute, W., W.G. González-Manteiga and M. Presedo-Quindmil (1998) Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **83**, 141–9.
Sweeney, R.J. (1986) Beating the foreign exchange market. *Journal of Finance* **41**, 163–82.
Weiss, A. (1986) ARCH and bilinear time series models: comparison and combination. *Journal of Business and Economic Statistics* **4**, 59–70.
Wooldridge, J. (1992) A test for functional form against nonparametric alternatives. *Econometric Theory* **8**, 452–75.
Wu, C.F.J. (1986) Jacknife, bootstrap and other resampling methods in regression analysis with discussion. *Annals of Statistics* **14**, 1261–350.

Yatchew, A.J. (1992) Nonparametric regression tests based on least squares. *Econometric Theory* **8**, 435–51.

Yule, G.U. (1926) Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Society* **89**, 2–9, 30–41.

Zheng, X. (1996) A consistent test of functional form via nonparametric estimation technique. *Journal of Econometrics* **75**, 263–89.

21
Autoregressive Conditional Duration Models[1]

Ruey S. Tsay

Abstract

This chapter studies the autoregressive conditional duration model. It discusses properties and statistical inference of the model. It also considers some extensions to handle nonlinear durations and interventions. For applications, we apply the model to daily range of the log price of Apple stock and find that adopting the decimal system for the US stock price on January 29, 2001, significantly reduces price volatility.

21.1 Introduction

The autoregressive conditional duration (ACD) model was proposed by Engle and Russell (1998) to model irregularly spaced financial transaction data. It has attracted much interest among researchers and practitioners ever since, and has found many applications outside of modeling transaction data. Duration is commonly defined as the time interval between consecutive events, e.g., the time interval between two transactions of a stock on the New York Stock Exchange or the difference between arrival times of two customers at a service station. The duration between two consecutive transactions in finance is important, for it may signal the arrival

of new information concerning the underlying asset. A cluster of short durations corresponds to active trading and, hence, an indication of the existence of new information.

Since duration is necessarily non-negative, the ACD model has also been used to model time series that consist of positive observations. An example is the daily range of the log price of an asset. The range of an asset price during a trading day can be used to measure its price volatility (e.g., Parkinson, 1980). Therefore, studying range can serve as an alternative approach to volatility modeling. Chou (2005) considers a conditional autoregressive range (CARR) model and shows that his CARR model can improve volatility forecasts for the weekly log returns of the Standard & Poor 500 index over some commonly used volatility models. The CARR model is essentially an ACD model.

In this chapter, we shall introduce the ACD model, discuss its properties, and address issues of statistical inference concerning the model. We then demonstrate its applications via some real examples. We also consider some extensions of the model, including nonlinear duration models and intervention analysis. Using the daily range of the log price of Apple stock, our ACD application shows that adopting the decimal system for US stock prices on January 29, 2001, significantly reduces the volatility of the stock price.

21.2 Duration models

Duration models in finance are concerned with time intervals between trades. For a given asset, longer durations indicate lack of trading activities, which in turn signify a period of no new information. On the other hand, arrival of new information often results in heavy trading and, hence, leads to shorter durations. The dynamic behavior of durations thus contains useful information about market activities. Furthermore, since financial markets typically take a period of time to uncover the effect of new information, active trading is likely to persist for a period of time, resulting in clusters of short durations. Consequently, durations might exhibit characteristics similar to those of asset volatility. Considerations like this lead to the development of duration models. Indeed, to model the durations of intraday trading, Engle and Russell (1998) use an idea similar to that of the generalized autoregressive conditional heteroskedastic (GARCH) models to propose an ACD model and show that the model can successfully describe the evolution of time durations for (heavily traded) stocks. Since intraday transactions of a stock often exhibit certain diurnal patterns, adjusted time durations are used in ACD modeling. We shall discuss methods for adjusting the diurnal pattern later. Here we focus on introducing the ACD model.

Let t_i be the time, measured with respect to some origin, of the ith event of interest with t_0 being the starting time. The ith duration is defined as:

$$x_i = t_i - t_{i-1}, \quad i = 1, 2, \ldots.$$

For simplicity, we ignore, at least for now, the case of zero durations so that $x_i > 0$ for all i. The ACD model postulates that x_i follows the model:

$$x_i = \psi_i \epsilon_i, \tag{21.1}$$

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $E(\epsilon_i) = 1$ and positive support, and ψ_i satisfies:

$$\psi_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{\nu=1}^{q} \beta_\nu \psi_{i-\nu}, \tag{21.2}$$

where p and q are non-negative integers and α_j and β_ν are constant coefficients. Since x_i is positive, it is common to assume that $\alpha_0 > 0$, $\alpha_j \geq 0$ and $\beta_\nu \geq 0$ for $j \in \{1, \ldots, p\}$ and $\nu \in \{1, \ldots, q\}$. Furthermore, the zeros of the polynomial $\alpha(L) = 1 - \sum_{j=1}^{g} (\alpha_j + \beta_j) L^j$ are outside the unit circle, where L denotes the lag operator, $g = \max\{p, q\}$, and $\alpha_j = 0$ for $j > p$ and $\beta_j = 0$ for $j > q$.

Let F_h be the σ-field generated by $\{\epsilon_h, \epsilon_{h-1}, \ldots\}$. It is easy to see that $E(x_i|F_{i-1}) = \psi_i E(\epsilon_i|F_{i-1}) = \psi_i$. Thus, ψ_i is the conditional expected duration of the next transaction given F_{i-1}. Since ϵ_i has a positive support, it may assume the standard exponential distribution. This results in an exponential ACD model. For ease of reference, we shall refer to the model in equations (21.1)–(21.2) as an EACD(p, q) model when ϵ_i follows the standard exponential distribution.

21.2.1 Properties of the EACD model

We start with the simple EACD(1,1) model:

$$x_i = \psi_i \epsilon_i, \quad \psi_i = \alpha_0 + \alpha_1 x_{i-1} + \beta_1 \psi_{i-1}. \tag{21.3}$$

Taking the expectation of the model, we obtain:

$$E(x_i) = E(\psi_i \epsilon_i) = E[\psi_i E(\epsilon_i|F_{i-1})] = E(\psi_i),$$
$$E(\psi_i) = \alpha_0 + \alpha_1 E(x_{i-1}) + \beta_1 E(\psi_{i-1}).$$

Under the weak stationarity assumption, $E(x_i) = E(x_{i-1})$, so that:

$$\mu_x \equiv E(x_i) = E(\psi_i) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}.$$

Consequently, $0 \leq \alpha_1 + \beta_1 < 1$ for a weakly stationary process $\{x_i\}$. Next, making use of the fact that $E(\epsilon_i) = 1$ and $E(\epsilon_i^2) = 2$, we have $E(x_i^2) = 2E(\psi_i^2)$. Again, under weak stationarity:

$$E(\psi_i^2) = \frac{\mu_x^2 [1 - (\alpha_1 + \beta_1)^2]}{1 - 2\alpha_1^2 - \beta_1^2 - 2\alpha_1\beta_1}, \tag{21.4}$$

$$\text{Var}(x_i) = \frac{\mu_x^2 (1 - \beta_1^2 - 2\alpha_1\beta_1)}{1 - 2\alpha_1^2 - \beta_1^2 - 2\alpha_1\beta_1}. \tag{21.5}$$

From these results, for the EACD(1,1) model to have a finite variance, we need $1 > 2\alpha_1^2 + \beta_1^2 + 2\alpha_1\beta_1$. Similar results can be obtained for the general EACD(p, q) model, but the algebra involved becomes tedious.

Forecasts from an EACD model can be obtained using a procedure similar to that of a GARCH model, which in turn is similar to that of a stationary autoregressive moving average (ARMA) model. Again, consider the simple EACD(1,1) model and suppose that the forecast origin is $i = h$. For a one-step-ahead forecast, the model states that $x_{h+1} = \psi_{h+1}\epsilon_{h+1}$ with $\psi_{h+1} = \alpha_0 + \alpha_1 x_h + \beta_1 \psi_h$. Let $x_h(1)$ be the one-step-ahead forecast of x_{h+1} at the origin h. Then:

$$x_h(1) = E(x_{h+1}|F_h) = E(\psi_{h+1}\epsilon_{h+1}) = \psi_{h+1},$$

which is known at the origin $i = h$. The associated forecast error is $e_h(1) = x_{h+1} - x_h(1) = \psi_{h+1}(\epsilon_{h+1} - 1)$. The conditional variance of the forecast error is then ψ^2_{h+1}. For multi-step-ahead forecasts, we use $x_{h+j} = \psi_{h+j}\epsilon_{h+j}$ so that, for $j = 2$,

$$\begin{aligned}\psi_{h+2} &= \alpha_0 + \alpha_1 x_{h+1} + \beta_1 \psi_{h+1} \\ &= \alpha_0 + (\alpha_1 + \beta_1)\psi_{h+1} + \alpha_1 \psi_{h+1}(\epsilon_{h+1} - 1).\end{aligned}$$

Consequently, the two-step-ahead forecast is:

$$x_h(2) = E(\psi_{h+2}\epsilon_{h+2}) = \alpha_0 + (\alpha_1 + \beta_1)\psi_{h+1} = \alpha_0 + (\alpha_1 + \beta_1)x_h(1),$$

and the associated forecast error is:

$$e_h(2) = \alpha_0(\epsilon_{h+2} - 1) + \alpha_1 \psi_{h+1}(\epsilon_{h+2}\epsilon_{h+1} - 1) + \beta_1 \psi_{h+1}(\epsilon_{h+2} - 1).$$

In general, we have:

$$x_h(m) = \alpha_0 + (\alpha_1 + \beta_1)x_h(m-1), \quad m > 1.$$

This is exactly the recursive forecasting formula of an ARMA(1,1) model with autoregressive (AR) polynomial $1 - (\alpha_1 + \beta_1)L$. By repeated substitutions, we can rewrite the forecasting formula as:

$$x_h(m) = \frac{\alpha_0[1 - (\alpha_1 + \beta_1)^{m-1}]}{1 - \alpha_1 - \beta_1} + (\alpha_1 + \beta_1)^{m-1} x_h(1).$$

Since $\alpha_1 + \beta_1 < 1$, we have:

$$x_h(m) \to \frac{\alpha_0}{1 - \alpha_1 - \beta_1}, \quad \text{as} \quad m \to \infty,$$

which says that, as expected, the long-term forecasts of a stationary series converge to its unconditional mean as the forecast horizon increases.

Let $\eta_j = x_j - \psi_j$. It is easy to show that $E(\eta_j) = 0$ and $E(\eta_j \eta_t) = 0$ for $t \neq j$. The variables $\{\eta_j\}$, however, are not identically distributed. Using $\psi_j = x_j - \eta_j$, we can rewrite the EACD(p, q) model in equation (21.2) as:

$$x_i = \alpha_0 + \sum_{j=1}^{g} (\alpha_j + \beta_j) x_{i-j} + \eta_i - \sum_{j=1}^{q} \beta_j \eta_{i-j},$$

where $g = \max\{p, q\}$ and it is understood that $\alpha_j = 0$ for $j > p$ and $\beta_j = 0$ for $j > q$. This is in the form of an ARMA(g, q) model with AR polynomial $1 - \sum_{j=1}^{g}(\alpha_j + \beta_j)L^j$. Consequently, some properties of EACD models can be inferred from those of ARMA models.

21.2.2 Estimation of EACD models

Suppose that $\{x_1, \ldots, x_n\}$ represents a realization of an EACD(p, q) model. The parameter $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)'$ can be estimated by the conditional likelihood method. Again, let $g = \max\{p, q\}$. The likelihood function of the data is:

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}) = f(\boldsymbol{x}_g|\boldsymbol{\theta}) \times \prod_{i=g+1}^{n} f(x_i|\boldsymbol{x}_{i-1}, \boldsymbol{\theta}),$$

where $\boldsymbol{x}_j = (x_1, \ldots, x_j)'$. Since the joint distribution of $\boldsymbol{x}_g$ is complicated and its influence on the overall likelihood function is diminishing as n increases, we adopt the conditional likelihood method by ignoring $f(\boldsymbol{x}_g|\boldsymbol{\theta})$. This results in using the conditional likelihood estimates. Since $f(x_i|F_{i-1}, \boldsymbol{\theta}) = \frac{1}{\psi_i}\exp(-x_i/\psi_i)$, the conditional log-likelihood function of the data then becomes:

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}_n) = -\sum_{i=t_o+1}^{n}\left[\ln(\psi_i) + \frac{x_i}{\psi_i}\right]. \tag{21.6}$$

The usual asymptotics of maximum likelihood estimates apply when the process $\{x_i\}$ is weakly stationary.

21.2.3 Additional ACD models

The EACD model has several nice features. For instance, it is simple in theory and in ease of estimation. But the model also encounters some weaknesses. For example, the use of the exponential distribution implies that the model has a constant hazard function. In the statistical literature, the hazard function (or intensity function) of a random variable X is defined by:

$$h(x) = \frac{f(x)}{S(x)},$$

where $f(x)$ and $S(x)$ are the probability density function and the survival function of X, respectively. The survival function of X is given by:

$$S(x) = P(X > x) = 1 - P(X \leq x) = 1 - \text{CDF}(x), \quad x > 0,$$

which gives the probability that a subject, which follows the distribution of X, survives at the time x. Under the EACD model, the distribution of the innovations is standard exponential so that the hazard function of ϵ_i is 1. As mentioned before, transaction duration in finance is inversely related to trading intensity, which in turn depends on the arrival of new information, making it hard to justify that the hazard function of duration is constant over time.

To overcome this weakness, alternative innovational distributions have been proposed in the literature. Engle and Russell (1998) entertain the Weibull distribution for ϵ_i and Zhang, Russell and Tsay (2001) consider the generalized Gamma distribution. The probability density function of a standardized Weibull random variable X is:

$$f(x|\alpha) = \begin{cases} \alpha \left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^{\alpha} x^{\alpha-1} \exp\left\{-\left[\Gamma\left(1+\frac{1}{\alpha}\right)y\right]^{\alpha}\right\}, & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{21.7}$$

where the α is referred to as the shape parameter and $\Gamma(.)$ is the usual Gamma function. The mean and variance of X are $E(X) = 1$ and $\text{Var}(X) = \Gamma(1+2/\alpha)/[\Gamma(1+1/\alpha)]^2 - 1$. The hazard function of X is:

$$h(x|\alpha) = \alpha \left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^{\alpha} x^{\alpha-1}.$$

Consequently, if $\alpha > 1$, the hazard function is a monotonously increasing function of x. If $0 < \alpha < 1$, then the hazard function is a monotonously decreasing function of x.

The probability density function of a generalized Gamma random variable X with $E(X) = 1$ is:

$$f(x|\alpha,\kappa) = \begin{cases} \frac{\alpha x^{\kappa\alpha-1}}{\lambda^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right], & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{21.8}$$

where $\lambda = \Gamma(\kappa)/\Gamma(\kappa+1/\alpha)$ with $\alpha > 0$ and $\kappa > 0$. Both α and κ are shape parameters so that the hazard function of X becomes more flexible than that of a Weibull distribution.

If ϵ_i of a duration model follows the standardized Weibull distribution with probability density function $f(x|\alpha)$ in equation (21.7), the conditional density function of x_i given F_{i-1} is:

$$f(x,\alpha) = \alpha \left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^{\alpha} \frac{x_i^{\alpha-1}}{\psi_i^{\alpha}} \exp\left\{-\left[\frac{\Gamma\left(1+\frac{1}{\alpha}\right)x_i}{\psi_i}\right]^{\alpha}\right\}, \tag{21.9}$$

which can be used to obtain the conditional log-likelihood function of the data for estimation.

If ϵ_i of a duration model follows the generalized Gamma distribution with $E(\epsilon_i) = 1$ in equation (21.8), the conditional density function of x_i given F_{i-1} is:

$$f(x_i|\alpha,\kappa) = \frac{\alpha x_i^{\kappa\alpha-1}}{(\psi_i\lambda)^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\frac{x_i}{\psi_i\lambda}\right)^{\alpha}\right], \tag{21.10}$$

where, again, $\lambda = \Gamma(\kappa)/\Gamma(\kappa+1/\alpha)$. This density function can be used to perform conditional maximum likelihood estimation of the model.

In what follows, we refer to the duration model in equations (21.1)–(21.2) as the WACD(p,q) or GACD(p,q) model if the innovation ϵ_i follows the standardized Weibull or generalized Gamma distribution, respectively.

21.2.4 Quasi-maximum likelihood estimates

In real applications, the true distribution function of the innovation ϵ_i of a duration model is unknown. One may, for simplicity, employ the conditional likelihood function of an EACD model in equation (21.6) to perform parameter estimation. The resulting estimates are called the quasi-maximum likelihood estimates (QMLE). Engle and Russell (1998) show that, under some regularity conditions, QMLE of a duration model are consistent and asymptotically normal. They are, however, not efficient when the innovations are not exponentially distributed.

21.2.5 Model checking

Let $\hat{\psi}_i$ be the fitted value of the conditional expected duration of an ACD model. We define $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ as the *standardized innovation* or *standardized residual* of the model. If the fitted ACD model is adequate, then $\{\hat{\epsilon}_i\}$ should behave as an i.i.d. sequence of random variables with the assumed distribution. We can use this standardized residual series to perform model checking. In particular, if the fitted model is adequate, both series $\{\hat{\epsilon}_i\}$ and $\{\hat{\epsilon}_i^2\}$ should have no serial correlations. The Ljung–Box statistics can be used to check the serial correlations of these two series. Large values of the Ljung–Box statistics indicate model inadequacy.

In addition, the quantile-to-quantile (QQ) plot of the standardized residuals against the assumed distribution of the innovations can be used to check the validity of the distributional assumption. For instance, under the WACD models, $\hat{\epsilon}_i$ should be close to the standardized Weibull distribution with shape parameter $\hat{\alpha}$. A deviation from the straight line of the QQ-plot suggests that the distributional assumption needs further improvement.

21.3 Some simple examples

In this section, we demonstrate the application of ACD models by considering two real examples.

Example 1 Consider the adjusted transaction durations of IBM stock from November 1 to November 7, 1990. The original durations are time intervals between two consecutive trades measured in seconds. Overnight intervals and zero durations were ignored. The adjustment is made to take care of the diurnal pattern of daily trading activities. The series consists of 3,534 observations and was used in Example 5.4 of Tsay (2005). Figure 21.1(a) shows the adjusted durations and Figure 21.2(a) gives the sample autocorrelation functions of the data. The autocorrelations are not large in magnitude, but they clearly indicate serial dependence in the data.

For illustration, we entertain EACD(1,1), WACD(1,1) and GACD(1,1) models for the IBM transaction durations. The estimated parameters of the three models are given in Table 21.1. The estimates of the ACD equation are rather stable for all three models, consistent with the theory that the estimates based on the exponential likelihood function are QMLE. Figure 21.1(b) shows the standardized innovations and

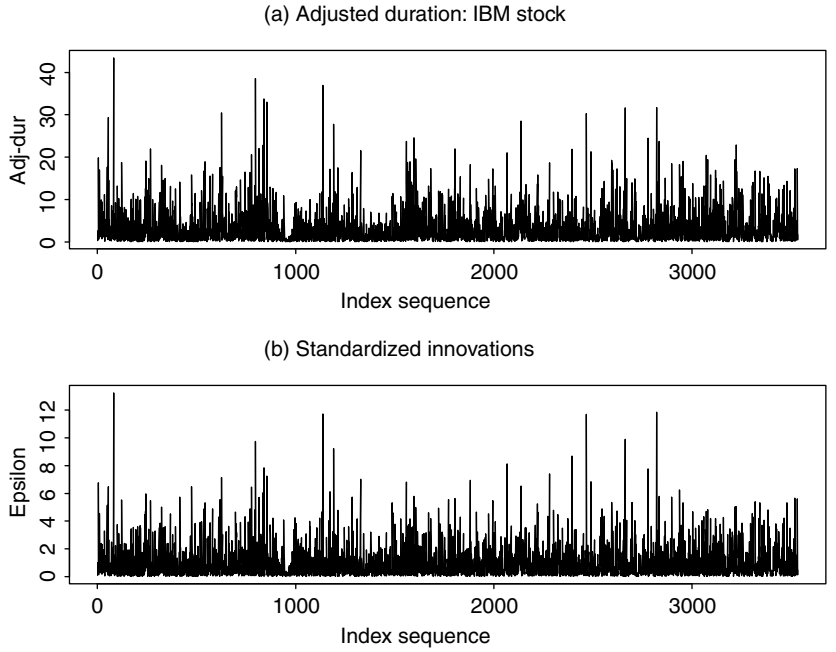


Figure 21.1 Time plots of the IBM transaction durations from November 1 to November 7, 1990: (a) adjusted durations; (b) standardized innovations of a WACD(1,1) model

Figure 21.2(b) gives the sample autocorrelation function (ACF) of the standardized innovations for the fitted WACD(1,1) model. The innovations appear to be random and their ACFs fail to indicate any serial dependence. Indeed, the Ljung–Box statistics for the standardized innovations and the squared innovations are insignificant, so that the fitted models are adequate for describing the dynamic dependence of the adjusted durations.

Figure 21.3 shows the QQ-plot of the standardized residuals versus a Weibull distribution with shape parameter 0.88 and scale parameter 1. The quantiles of the Weibull distribution are generated using a random sample of 30,000 observations. A straight line is imposed on the plot to aid interpretation. From the plot, except for a few large residuals, the assumption of a Weibull distribution seems reasonable. In this particular example, the GACD(1,1) model also fits the data well. We chose the WACD(1,1) model for its simplicity.

Finally, for the WACD(1,1) model, the estimated shape parameter α is less than 1, indicating that the hazard function of the adjusted durations is monotonously decreasing. This seems reasonable for the adjusted durations of the heavily traded IBM stock.

Example 2 In this example, we apply the ACD model to stock volatility modeling. Consider the daily range of the log price of Apple stock from January 4, 1999, to

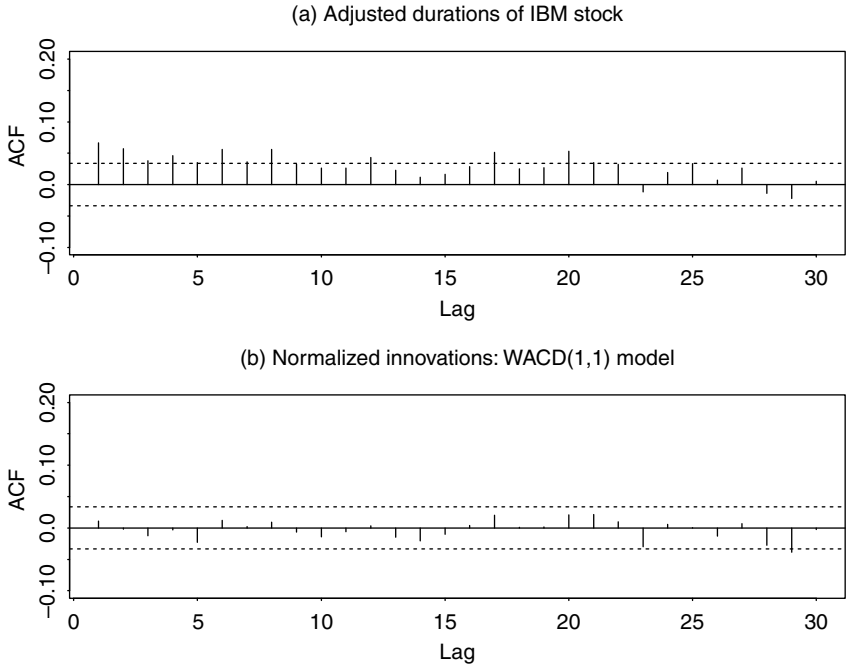


Figure 21.2 The sample autocorrelation function of IBM transaction durations from November 1 to November 7, 1990: (a) ACF of the adjusted durations; (b) ACF of the standardized residual series of a WACD(1,1) model

Table 21.1 Estimation results of EACD(1,1), WACD(1,1) and GACD(1,1) models for the IBM transaction durations of Example 1

Model	*Parameters*					*Checking*	
	α_0	α_1	β_1	α	κ	$Q(10)$	$Q^*(10)$
EACD	0.129	0.056	0.905			4.55	5.48
	(0.037)	(0.009)	(0.018)			(0.92)	(0.86)
WACD	0.125	0.056	0.906	0.880		3.85	5.51
	(0.040)	(0.010)	(0.019)	(0.012)		(0.92)	(0.85)
GACD	0.111	0.056	0.912	0.407	4.016	4.62	5.53
	(0.040)	(0.010)	(0.019)	(0.040)	(0.730)	(0.92)	(0.85)

Notes: The adjusted durations are from November 1 to November 7, 1990, with 3,534 observations. The standard errors of the estimates are in parentheses. The *p*-values of the Ljung–Box statistics are also in parentheses with $Q(10)$ and $Q^*(10)$ for standardized residual series and its squared process, respectively.

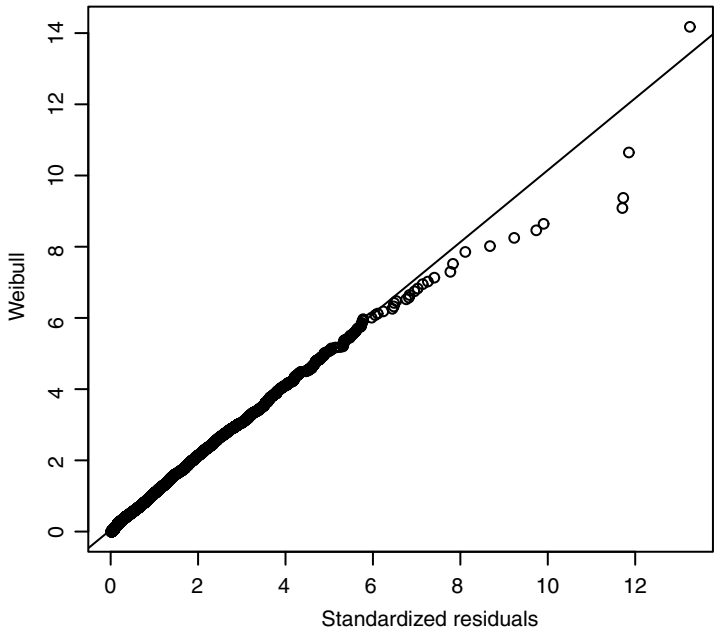


Figure 21.3 QQ-plot of the standardized residuals of the WACD(1,1) model versus a Weibull distribution
The Weibull quantiles are generated from a random sample of 30,000 observations using the shape parameter 0.88 and scale parameter 1.0.

November 20, 2007. The data are obtained from Yahoo Finance and consist of 2,235 observations. The range has been used in the literature as a robust alternative to volatility modeling (see Chou, 2005, and the references therein). Apple stock had two-for-one splits on June 21, 2000, and February 28, 2005, both during the sample period, but for simplicity we make no adjustments for the splits. Also, stock prices in the US markets switched from the tick size 1/16 of a dollar to the decimal system on January 29, 2001. Such a change affected the daily range of stock prices. We shall return to this point later. The sample mean, standard deviation, minimum and maximum of the range of log prices are 0.0407, 0.0218, 0.0068 and 0.1468, respectively. The sample skewness and excess kurtosis are 1.3 and 2.13, respectively. Figure 21.4(a) shows the time plot of the range series. The volatility seems to be increasing from 2000 to 2001, then deceasing to a stable level after 2002. It seems to increase somewhat at the end of the series. Figure 21.5(a) shows the sample ACF of the daily range series. The sample ACFs are highly significant and decay slowly.

Again, we fit the EACD(1,1), WACD(1,1), and GACD(1,1) models to the daily range series. The estimation results, along with the Ljung–Box statistics for the standardized residual series and its squared process, are given in Table 21.2. Again, the parameter estimates for the duration equation are stable for all three models, except for the constant term of the EACD model, which appears to be

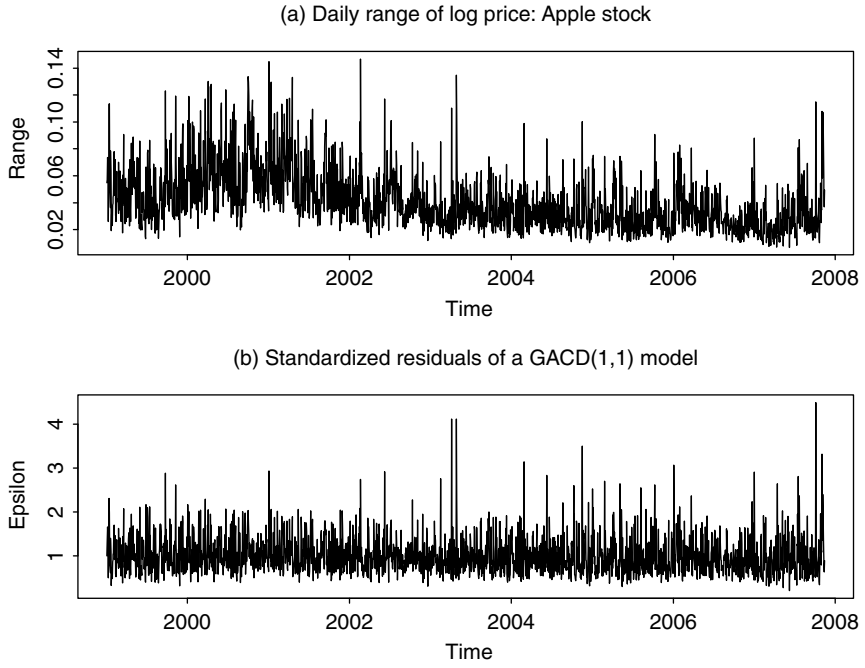


Figure 21.4 Time plots of the daily range of log price of Apple stock from January 4, 1999, to November 20, 2007: (a) observed daily range; (b) standardized residuals of a GACD(1,1) model

statistically insignificant at the usual 5% level. Indeed, in this particular instance, the EACD(1,1) model fares slightly worse than the other two ACD models. Between the WACD(1,1) and GACD(1,1) models, we slightly prefer the GACD(1,1) model, because it fits the data better and is more flexible. Figure 21.6 shows the QQ-plots of the standardized residuals versus the assumed innovation distribution for the GACD(1,1) and WACD(1,1). The plots indicate that further improvement in the distributional assumption is needed for the daily range, but they support the preference of the GACD(1,1) model.

Figure 21.5(b) shows the sample ACFs of the standardized residuals of the fitted GACD(1,1) model. From the plot, the standardized residuals do not have significant serial correlations, even though the lag-1 sample ACF is slightly above its two standard-error limit. We shall return to this point later when we introduce nonlinear ACD models. Figure 21.4(b) shows the time plot of the standardized residuals of the GACD(1,1) model. The residuals do not show any pattern of model inadequacy. The mean, standard deviation, minimum and maximum of the standardized residuals are 0.203, 4.497, 0.999, and 0.436, respectively.

It is interesting to see that the estimates of the shape parameter α are greater than 1 for both WACD(1,1) and GACD(1,1) models, indicating that the hazard function of the daily range is monotonously increasing. This is consistent with the idea

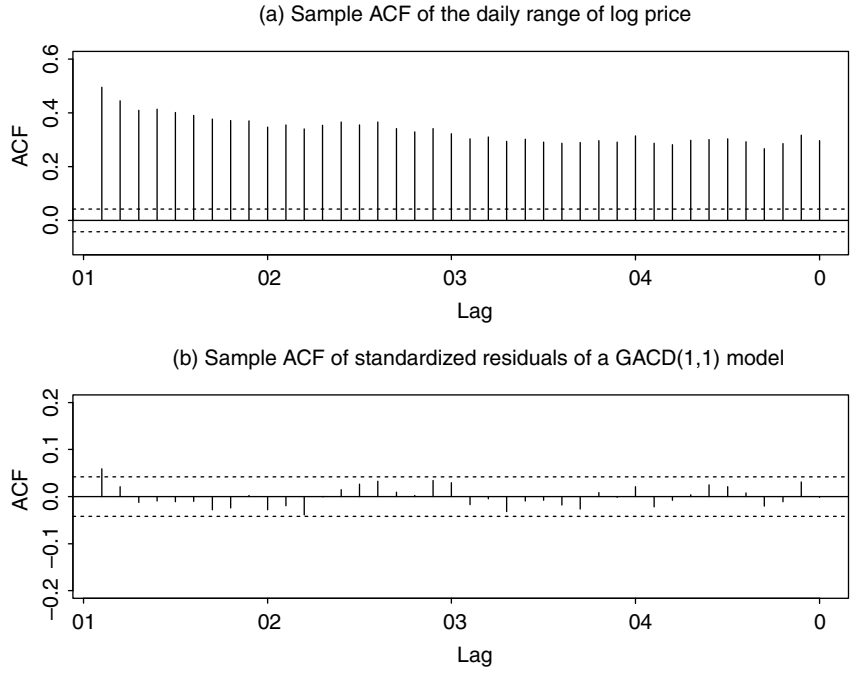


Figure 21.5 The sample autocorrelation function of the daily range of log price of Apple stock from January 4, 1999 to November 20, 2007: (a) ACF of daily range; (b) ACF of the standardized residual series of a GACD(1,1) model

of volatility clustering, for a large volatility tends to be followed by another large volatility. This phenomenon is different from that of the transaction durations in Example 1 for which $\hat{\alpha}$ is less than 1.

21.4 The diurnal pattern

In this section, we discuss a simple method to adjust the diurnal pattern of intradaily trading activities. Figure 21.7(a) shows the trade durations of General Motors (GM) stock from December 1 to December 5, 2003. Again, for simplicity, zero durations are ignored. Figure 21.7(b) shows the time intervals from the market opening (9.30 a.m. Eastern time) to the transaction time. The four vertical drops of the intervals signify the five trading days. From parts (a) and (b) of the figure, the diurnal pattern of trading activities is clearly seen. Specifically, except for a few outliers, the trade durations exhibit a cap-shape pattern within a trading day, namely the durations are in general shorter at the beginning and closing of the market, and longer around the middle of a trading day. One must consider such a diurnal pattern in modeling the transaction durations.

Table 21.2 Estimation results of EACD(1,1), WACD(1,1) and GACD(1,1) models for the daily range of log price of Apple stock from January 4, 1999, to November 20, 2007

Model	*Parameters*					*Checking*	
	α_0	α_1	β_1	α	κ	$Q(10)$	$Q^*(10)$
EACD	0.0007	0.133	0.849			16.65	12.12
	(0.0005)	(0.036)	(0.044)			(0.082)	(0.277)
WACD	0.0013	0.131	0.835	2.377		13.66	9.74
	(0.0003)	(0.015)	(0.021)	(0.031)		(0.189)	(0.464)
GACD	0.0010	0.133	0.843	1.622	2.104	14.62	11.21
	(0.0002)	(0.015)	(0.019)	(0.029)	(0.040)	(0.147)	(0.341)

Notes: The sample size is 2,235. The standard errors of the estimates are in parentheses. The *p*-values of the Ljung–Box statistics are also in parentheses with $Q(10)$ and $Q^*(10)$ for standardized residual series and its squared process, respectively.

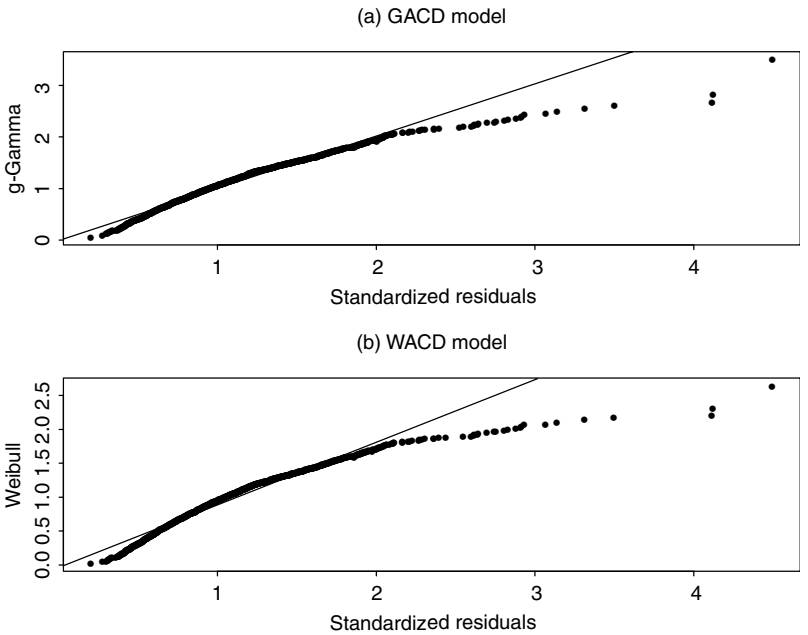


Figure 21.6 QQ-plots for the standardized residuals of ACD models for the daily range of log price of Apple stock from January 4, 1999, to November 20, 2007: (a) GACD(1,1) model; (b) WACD(1,1) model

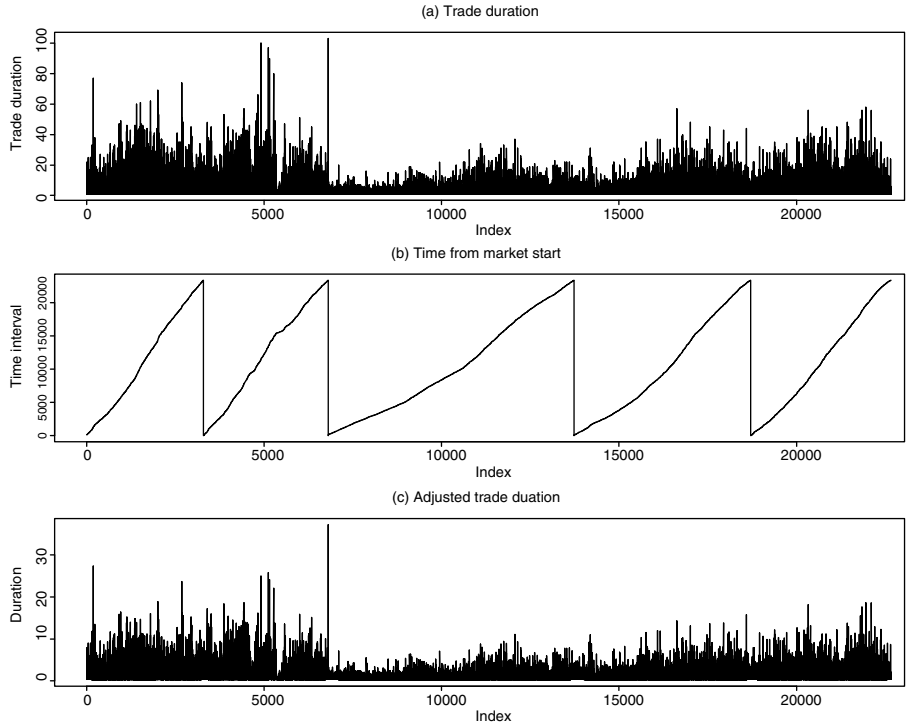


Figure 21.7 Time plots of durations for the General Motors stock from December 1 to December 5, 2003: (a) observed trade durations (positive only); (b) transaction times measured in seconds from midnight; (c) adjusted trade durations

There are many ways to remove the diurnal pattern of transaction durations. Engle and Russell (1998) and Zhang, Russell and Tsay (2001) use some simple exponential functions of time, and Tsay (2005) constructs some deterministic functions of time of the day to adjust the diurnal pattern. Let $f(t_i)$ be the mean value of the diurnal pattern at time t_i, measured from midnight. Then, define:

$$x_i = \frac{z_i}{f(t_i)}, \tag{21.11}$$

to be the adjusted duration, where z_i is the observed duration between the ith and $(i-1)$th transactions. We construct $f(t_i)$ using two simple time functions. Define:

$$O(t_i) = \begin{cases} t_i - 34200 & \text{if } t_i < 43200 \\ 0 & \text{otherwise,} \end{cases} \qquad C(t_i) = \begin{cases} 57600 - t_i & \text{if } t_i \geq 43200 \\ 0 & \text{otherwise,} \end{cases} \tag{21.12}$$

where t_i is the time of the ith transaction measured in seconds from midnight and 34200, 43200, and 57600 denote, respectively, the market opening, noon, and market closing times measured in seconds. Figures 21.8(b) and (c) show the

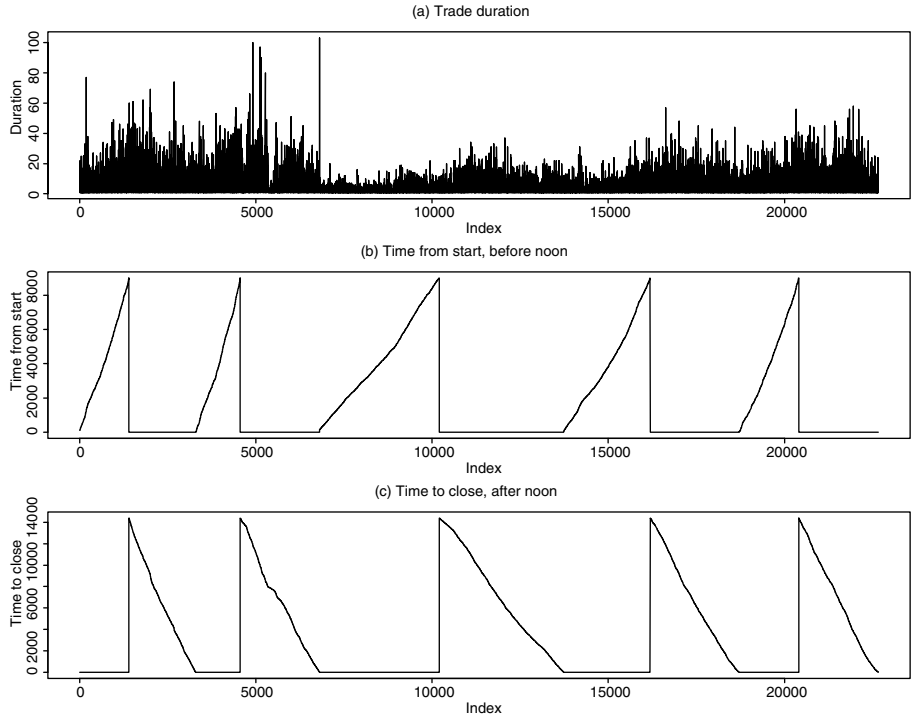


Figure 21.8 Time plots of durations for the G.M. stock from December 1 to December 5, 2003; (a) observed trade durations (positive only); (b) and (c) the time function $O(t_i)$ and time function $C(t_i)$ of equation (21.12)

time plots of $O(t_i)$ and $C(t_i)$ of the GM stock transactions. Figure 21.8(a) shows the observed trade durations as in Figure 21.7(a). From the plots, the use of $O(t_i)$ and $C(t_i)$ is justified.

Consider the multiple linear regression:

$$\ln(z_i) = \beta_0 + \beta_1 o(t_i) + \beta_2 c(t_i) + e_i, \tag{21.13}$$

where $o(t_i) = O(t_i)/10000$ and $c(t_i) = C(t_i)/10000$. Let $\hat{\beta}_i$ be the ordinary least squares estimates of the above linear regression. The residual is then given by:

$$\hat{e}_i = \ln(z_i) - \hat{\beta}_0 - \hat{\beta}_1 o(t_i) - \hat{\beta}_2 c(t_i).$$

The adjusted durations then become:

$$\hat{x}_i = \exp(\hat{e}_i). \tag{21.14}$$

For the GM stock transactions, the estimates of the β_i are 1.015(0.012), 0.133(0.028) and 0.313(0.016), respectively, where the numbers in parentheses denote standard errors. All estimates are statistically significant at the usual 1% level. Note that

the residuals of the regression in equation (21.13) are serially correlated. Thus, the standard errors shown above underestimate the true ones. A more appropriate estimation method of the standard errors is to apply the Newey and West (1987) correction. The adjusted standard errors are 0.018, 0.044 and 0.027, respectively. These standard errors are larger, but all estimates remain statistically significant at the 1% level.

Figure 21.7(c) shows the time plot of the adjusted durations for the GM stock. Compared with part (a), the diurnal pattern of the trade durations is largely removed.

21.5 Nonlinear duration models

The linear duration models discussed in the previous sections are parsimonious in their parameterization and useful in many situations. However, in financial applications, the sample size can be large and the linearity assumption of the model might become an issue. Indeed, our limited experience indicates that some nonlinear characteristics are often observed in transaction durations and daily ranges of log stock prices. For instance, Zhang, Russell and Tsay (2001) showed that simple threshold autoregressive duration models can improve the analysis of stock transaction durations. In this section, we consider some simple nonlinear duration models and demonstrate that they can improve upon the linear ACD models.

21.5.1 The threshold autoregressive duration model

A simple nonlinear duration model is the threshold autoregressive conditional duration (TACD) model. The nonlinear threshold autoregressive (TAR) model was proposed in the time series literature by Tong (1978) and has been widely used ever since (see, e.g., Tong, 1990; Tsay, 1989). A simple two-regime TACD(2;p, q) model for x_i can be written as:

$$x_i = \begin{cases} \psi_i \epsilon_{1i} & \text{if } x_{t-d} \le r, \\ \psi_i \epsilon_{2i} & \text{if } x_{t-d} > r, \end{cases} \tag{21.15}$$

where d is a positive integer, x_{t-d} is the threshold variable, r is a threshold, and:

$$\psi_i = \begin{cases} \alpha_{10} + \sum_{v=1}^{p} \alpha_{1v} x_{i-v} + \sum_{v=1}^{q} \beta_{1v} \psi_{i-v} & \text{if } x_{t-d} \le r, \\ \alpha_{20} + \sum_{v=1}^{p} \alpha_{2v} x_{i-v} + \sum_{v=1}^{q} \beta_{2v} \psi_{i-v} & \text{if } x_{t-d} > r, \end{cases}$$

where $\alpha_{j0} > 0$ and α_{jv} and β_{jv} satisfy the conditions of the ACD model stated in equation (21.2) for $j = 1$ and 2. Here j denotes the regime. The innovations $\{\epsilon_{1i}\}$ and $\{\epsilon_{2i}\}$ are two independent i.i.d. sequences. They can follow the standard exponential, standardized Weibull, or standardized generalized Gamma distribution as before. For simplicity, we shall refer to the resulting models as the TEACD, TWACD, and TGACD model, respectively. The TACD model is a piecewise linear model in

the space of x_{i-d}, and it is nonlinear when some of the parameters in the two regimes are different. The model can be extended to have more than two regimes. In what follows, we assume $p = q = 1$ in our discussion, because ACD(1,1) models fare well in many applications.

The TACD model appears to be simple, and it is indeed easy to use. However, its theoretical properties are very involved. For instance, the stationarity condition stated in equation (21.15) is only sufficient. The necessary condition of stationarity would depend on d and the parameters and deserves further investigation.

A key step in specifying a TACD model for a given time series is the identification of the threshold variable and the threshold, i.e., specifying d and r. The choice of d is relatively simple because $d \in \{1, \ldots, d_0\}$ for some positive integer d_0. For stock transaction durations, $d = 1$ is a reasonable choice as trading activities tend to be highly serially correlated. For the threshold r, a simple approach is to use empirical quantiles. Let $x_{<q>}$ be the qth quantile of the observed durations $\{x_i | i = 1, \ldots, n\}$. We assume that $r \in \{x_{<q>} | q = 60, 65, 70, \ldots, 95\}$. For each candidate $x_{<q>}$, estimate the TACD(2;1,1) model:

$$\psi_i = \begin{cases} \alpha_{10} + \alpha_{11} x_{i-1} + \beta_{11} \psi_{i-1} & \text{if } x_{t-1} \leq x_{<q>}, \\ \alpha_{20} + \alpha_{21} x_{i-1} + \beta_{21} \psi_{i-1} & \text{otherwise,} \end{cases}$$

and evaluate the log-likelihood function of the model at the maximum likelihood estimates. Denote the resulting log-likelihood value by $\ell(x_{<q>})$. The threshold is then selected by:

$$\hat{r} = x_{<q_o>} \quad \text{such that} \quad \ell(x_{<q_o>}) = \max_q \{\ell(x_{<q>}) | q = 60, 65, 70, \ldots, 95\}.$$

21.5.2 Example

In this sub-section, we revisit the series of daily ranges of the log price of Apple stock from January 4, 1999, to November 20, 2007. The standardized innovations of the GACD(1,1) model of section 21.3 have a marginally significant lag-1 autocorrelation. This serial correlation also occurs for the EACD(1,1) and WACD(1,1) models. Here we employ a two-regime threshold WACD(1,1) model to improve the fit. Preliminary analysis of the TWACD models indicates that the major difference in the parameter estimates between the two regimes is the shape parameter of the Weibull distribution. Thus, we focus on a TWACD(2;1,1) model with different shape parameters for the two regimes.

Table 21.3 gives the maximized log-likelihood function of a TWACD(2;1,1) model for $d = 1$ and $r \in \{x_{<q>} | q = 60, 65, \ldots, 95\}$. From the table, the threshold 0.04753 is selected, which is the 70th percentile of the data. The fitted model is:

$$x_i = \psi_i \epsilon_i, \quad \psi_i = 0.0013 + 0.1539 x_{i-1} + 0.8131 \psi_{i-1},$$

where the standard errors of the coefficients are 0.0003, 0.0164 and 0.0215, respectively, and ϵ_i follows the standardized Weibull distribution as:

$$\epsilon_i \sim \begin{cases} W(2.2756) & \text{if } x_{i-1} \leq 0.04753, \\ W(2.7119) & \text{otherwise,} \end{cases}$$

Table 21.3 Selection of the threshold of a TWACD(2;1,1) model for the daily range of the log price of Apple stock from January 4, 1999, to November 20, 2007

Quantile	60	65	70	75	80	85	90	95
$r \times 100$	4.03	4.37	4.75	5.15	5.58	6.16	7.07	8.47
$\ell(r) \times 10^3$	6.073	6.076	6.079	6.076	6.078	6.074	6.072	6.066

Note: The threshold variable is x_{i-1}.

where the standard errors of the two shape parameters are 0.0394 and 0.0717, respectively.

Figure 21.9(a) shows the time plot of the conditional expected duration for the fitted TWACD(2;1,1) model, i.e., $\hat{\psi}_i$, whereas Figure 21.9(b) gives the residual ACFs for the fitted model. All residual ACFs are within the two-standard-error limits. Indeed, we have $Q(1) = 4.01(0.05)$, $Q(10) = 9.84(0.45)$ for the standardized residuals and $Q^*(1) = 0.83(0.36)$ and $Q^*(10) = 9.35(0.50)$ for the squared series of the standardized residuals, where the number in parentheses denotes p-value. Note that the threshold variable x_{i-1} is also selected based on the value of the log-likelihood function. For instance, the log-likelihood function of the TWACD(2;1,1) model assumes the value 6.069×10^3 and 6.070×10^3, respectively, for $d = 2$ and 3 when the threshold is 0.04753. These values are lower than that when $d = 1$.

21.6 The use of explanatory variables

High-frequency financial data are often influenced by external events, e.g., an increase or drop in interest rates by the US Federal Open Market Committee or a jump in the oil price. Applications of ACD models in finance are often faced with the problem of outside interventions. To handle the effects of external events, the intervention analysis of Box and Tiao (1975) can be used. In this section, we consider intervention analysis in ACD modeling. We use the daily range series of Apple stock as an example. Here the intervention is the change in tick size of the US stock markets.

On January 29, 2001, all stock prices on the US markets switched to the decimal system. Before the switch, tick sizes of US stocks went through several transitions, from 1/8 to 1/16 to 1/32 of a dollar. The observed daily range is certainly affected by the tick size.

Let t_o be the time of intervention. For Apple stock, $t_o = 522$, which corresponds to January 26, 2001, the last trading day before the change in tick size. Since more observations in the sample are after the intervention, we define the indicator variable:

$$I_i^{(t_o)} = \begin{cases} 1 & \text{if } i \le t_o, \\ 0 & \text{otherwise,} \end{cases}$$

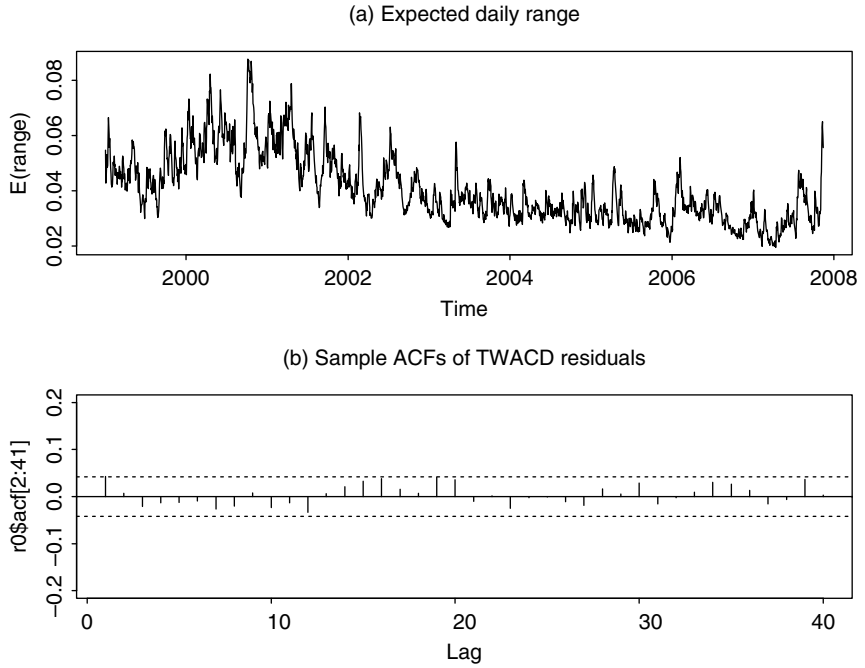


Figure 21.9 Model fitting for the daily range of the log price of Apple stock from January 4, 1999, to November 20, 2007: (a) the conditional expected durations of the fitted TWACD(2;1,1) model; (b) the sample ACF of the standardized residuals

to signify the absence of intervention. Since a larger tick size tends to increase the observed daily price range, it is reasonable to assume that the conditional expected range would be higher before the intervention. A simple intervention model for the daily range of Apple stock is then given by:

$$x_i = \psi_i \begin{cases} \epsilon_{1i} & \text{if } x_{i-1} \le 0.04753, \\ \epsilon_{2i} & \text{otherwise,} \end{cases}$$

where ψ_i follows the model:

$$\psi_i = \alpha_0 + \gamma I_i^{(t_o)} + \alpha_1 x_{i-1} + \beta_1 \psi_{i-1}, \tag{21.16}$$

where γ denotes the decrease in expected duration due to the decimalization of stock prices. In other words, the expected durations before and after the intervention are:

$$\frac{\alpha_0 + \gamma}{1 - \alpha_1 - \beta_1} \quad \text{and} \quad \frac{\alpha_0}{1 - \alpha_1 - \beta_1},$$

respectively. We expect $\gamma > 0$.

The fitted duration equation for the intervention model is:

$$\psi_i = 0.0021 + 0.0011 I_i^{(522)} + 0.1595 x_{i-1} + 0.7828 \psi_{i-1},$$

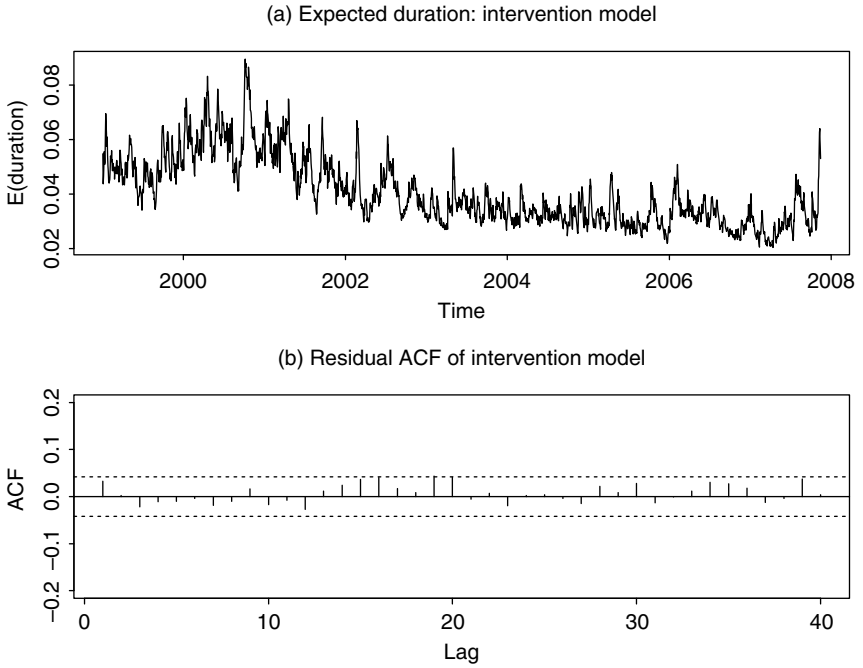


Figure 21.10 Model fitting for the daily range of the log price of Apple stock from January 4, 1999, to November 20, 2007: (a) the conditional expected durations of the fitted TWACD(2;1,1) model with intervention; (b) the sample ACF of the corresponding standardized residuals

where the standard errors of the estimates are 0.0004, 0.0003, 0.0177, and 0.0264, respectively. The estimate $\hat{\gamma}$ is significant at the 1% level. For the innovations, we have:

$$\epsilon_i \sim \begin{cases} W(2.2835) & \text{if } x_{i-1} \leq 0.04753, \\ W(2.7322) & \text{otherwise.} \end{cases}$$

The standard errors of the two estimates of the shape parameter are 0.0413 and 0.0780, respectively. Figure 21.10(a) shows the expected durations of the intervention model and Figure 21.10(b) shows the ACF of the standardized residuals. All residual ACFs are within the two standard error limits. Indeed, for the standardized residuals, we have $Q(1) = 2.37(0.12)$ and $Q(10) = 6.24(0.79)$. For the squared series of the standardized residuals, we have $Q^*(1) = 0.34(0.56)$ and $Q^*(10) = 6.79(0.75)$. As expected, $\hat{\gamma} > 0$ so that the decimalization indeed reduces the expected value of the daily range. This simple analysis shows that, as expected, adopting the decimal system reduces the volatility of Apple stock.

Note that a general intervention model that allows for changes in the dynamic dependence of the expected duration can be used, even though our analysis only allows for a change in the expected duration. Of course, more flexible models are harder to estimate and understand.

21.7 Conclusion

In this chapter, we introduced the autoregressive conditional duration models and discussed their properties and statistical inference. Among many applications, we used the model to study the daily volatility of stock prices and found that, for Apple stock, adopting the decimal system on January 29, 2001, indeed significantly reduces the price volatility.

Note

1. The estimation of all ACD models in this chapter is carried out by the FMINCON function in Matlab.

References

Box, G.E.P. and G.C. Tiao (1975) Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* **70**, 70–9.

Chou, R.Y. (2005) Forecasting financial volatilities with extreme values: the conditional autoregressive range (CARR) model. *Journal of Money, Credit and Banking* **37**, 561–82.

Engle, R.F. and J.R. Russell (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–62.

Newey, W. and K. West (1987) A simple positive semidefinite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 863–98.

Parkinson, M. (1980) The extreme value method for estimating the variance of the rate of return. *Journal of Business* **53**, 61–5.

Tong, H. (1978) On a threshold model. In C.H. Chen (ed.), *Pattern Recognition and Signal Processing*. Netherlands: Sijthoff and Noordhoff.

Tong, H. (1990) *Non-linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.

Tsay, R.S. (1989) Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association* **84**, 231–40.

Tsay, R.S. (2005) *Analysis of Financial Time Series* (second edition). Hoboken, N.J.: John Wiley.

Zhang, M.Y., J.R. Russell and R.S. Tsay (2001) A nonlinear autoregressive conditional duration model with applications to financial transaction data. *Journal of Econometrics* **104**, 179–207.

22

The Econometrics of Exchange Rates

Efthymios G. Pavlidis, Ivan Paya and David A. Peel

Abstract

We provide a selective overview of the econometric methods employed in modeling some of the key relationships which determine the behavior of exchange rates and the efficacy of models employed to forecast them.

22.1 Introduction

The purpose of this chapter is to provide a selective overview of the econometric methods employed in the modeling of exchange rates. Given space constraints, we can only give very brief outlines of the underlying economic theory, which is well covered in, e.g., Taylor (1995), Obstfeld and Rogoff (1996), and Sarno and Taylor (2002).

Whilst always an important focus of applied work, econometric developments, in conjunction with new high-quality datasets and the move to generalized floating exchange rates in 1973, have generated a vast number of empirical papers in the last couple of decades. Perhaps the major change in emphasis over this period has been the application of nonlinear rather than linear methods. These nonlinear models are based on theoretical models that embody factors such as transactions costs, limits to arbitrage and heterogeneity of expectations of market participants (see, e.g., Dumas, 1992; De Grauwe *et al.*, 1993; Shleifer and Vishny, 1997).

An essential building block of many macroeconomic models is that purchasing power parity (PPP) holds in the long run. PPP states that the nominal exchange rate between two currencies should be equal to the ratio of aggregate price levels between the two countries, so that a unit of currency of one country will have the same purchasing power in a foreign country. The first empirical studies employing unit root tests in the late 1980s were consistent in their failure to reject the unit root hypothesis for major real exchange rates (e.g., Taylor, 1988; Mark, 1990). Subsequent research employing longer time series datasets or panel methods suggested that the early non-rejections of the unit root hypothesis was due to low power of the corresponding test (Lothian and Taylor, 1996). However, the implied speeds of adjustment of the real exchange rate in these studies was implausibly slow, typically with half-life in the range of three to five years. Rogoff (1996, p. 647) summarized this position as follows, "How can one reconcile the enormous short-term volatility of real exchange rates with the extremely slow rate at which shocks appear to damp out?"

Perhaps the most important explanation of the Rogoff puzzle is that real exchange rates can be described by a nonlinear data-generating process (DGP) that exhibits a region of unit root or near unit root behavior near the equilibrium real exchange rate. Nonlinear models that capture this type of behavior are the threshold autoregressive model of Tong (1983), and the exponential smooth transition autoregressive model of Ozaki (1978). Econometric testing requires appropriate tests for nonlinearity, where the null can be a stationary linear process or a non-stationary linear process. In addition, the error process can exhibit heteroskedasticity due to changes in regime (e.g., fixed to floating, or different monetary regimes), as well as time-varying volatility. Consequently, the tests have to allow for this feature and critical values have been obtained by either Monte Carlo or bootstrap methods. Because many other empirical tests of aspects of exchange rate behavior employ these tests, we initially consider the econometric tests of PPP in section 22.2.

Average daily global turnover in foreign exchange market transactions was estimated at around $2.7 trillion in 2006. In an efficient speculative market, prices should fully reflect publicly available information so that it should not be possible for a trader to earn systematic abnormal returns. However, empirical estimates show that the spot exchange rate next period moves on average in the opposite direction to that currently predicted by the forward premium, the so-called forward bias puzzle. In section 22.3.1 we consider econometric tests of the covered interest parity condition and, in section 22.3.2, the uncovered interest parity condition and the various explanations of the forward bias puzzle.

If expectations are forward looking then the exchange rate regime in place, as well as the anticipation of the implementation of future exchange rate regimes, will impact on the behavior of the exchange rate. An important example of such policy arrangements are "target zones." Within this framework the authorities intervene to attempt to keep the exchange rate within a band. The recent experience of European currencies between 1979 and 1998 under the European Monetary System (EMS) is one such example.[1] In section 22.4 we consider some of the econometric testing of target zone models.

As with other asset markets, researchers have examined whether exchange rates exhibit rational speculative bubbles.[2] In section 22.5 we briefly discuss some of the more recent developments in the area.

Section 22.6 covers the issue of exchange rate forecasting. In a seminal paper, Meese and Rogoff (1983a) compared the out-of-sample forecasts produced by various exchange rate models with forecasts produced by a random walk (RW) model. On the basis of the root mean square criterion, they concluded that none of the various asset-market exchange rate models they considered outperformed a simple RW. Since then a plethora of papers have been published on this issue. We give an overview of developments. Finally, section 22.7 provides a brief conclusion of the major issues to emerge in this chapter.

22.2 Real exchange rates

The PPP hypothesis states that domestic prices in two countries should be the same when converted to a common currency. Let us define the log real exchange rate as $y_t = s_t - p_t + p_t^*$, where s_t is the logarithm of the spot exchange rate (the domestic price of foreign currency), p_t is the logarithm of the domestic price level and p_t^* the logarithm of the foreign price level.

Initial empirical studies of PPP consisted of fitting a univariate autoregressive model for the real exchange rate, $y_t = \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t$. In the 1980s many empirical studies were unable to reject the unit root null hypothesis for y_t using standard unit root tests such as the augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) (Taylor, 1988; Mark, 1990). In the light of these results, the first econometric issue addressed in the literature was the power of the ADF and PP tests in a linear framework. Lothian and Taylor (1996) undertook the following simulation. They generate a stationary AR(1) process calibrated with empirical estimates on data for 100 years of the dollar–pound real exchange rate as $y_t = 0.87 y_{t-1} + \varepsilon_t$.[3]

The power of both ADF and PP tests was around 45% for a sample size of 100. Caner and Kilian (2001) did a similar experiment but with the null of stationarity. They employ the Kwiatkowski *et al.* (1992) (KPSS) and Leybourne and McCabe (1994) tests. They found that, if real exchange rates were a linear autoregressive process with half-life of about three to five years ($\beta = 0.5^{(1/60)} = 0.9885$ for monthly data, and $\beta = 0.5^{(1/20)} = 0.9659$ for quarterly), they would reject the null of stationarity with about 60% probability for quarterly data and 40% for monthly data. Together, these results pointed out a serious problem with unit root and stationarity tests due to their lack of power and size, respectively.[4]

A number of authors have employed a multivariate cointegration methodology to test for a long-run relationship between exchange rates and foreign and domestic price levels in the recent floating exchange rate period (MacDonald, 1993; Baum *et al.*, 2001). The cointegrating relationship is usually specified as:

$$s_t = \alpha + \beta_1 p_t^* + \beta_2 p_t + u_t. \tag{22.1}$$

The standard empirical findings employing these methods are that cointegration cannot be rejected but the assumption of proportionality and symmetry between the nominal exchange rate and domestic and foreign prices ($\beta_1 = -\beta_2 = 1$) is not supported by the data.[5]

Theoretical analysis of purchasing power parity deviations demonstrate how transactions costs or the sunk costs of international arbitrage induce nonlinear but mean reverting adjustment of the real exchange rate (see, e.g., Dumas, 1992; Sercu *et al.*, 1995; O'Connell and Wei, 1997). Whilst globally mean reverting, these nonlinear processes have the property of exhibiting near unit root behavior for small deviations from PPP, since small deviations are left uncorrected if they are not large enough to cover transactions costs or the sunk costs of international arbitrage. The nonlinearity postulated can be captured by a set of parametric nonlinear autoregressive models. We can classify this set of nonlinear models according to the way the real exchange rate switches between regimes: first, with smooth transition models; second, with threshold autoregressive models.

22.2.1 Smooth transition (STR) models

Consider the following STR model:

$$y_t = \beta' \widetilde{y}_t + \phi' \widetilde{y}_t F(z_{t-d}; \gamma, c) + u_t, \tag{22.2}$$

where $\widetilde{y}_t = (1, y_{t-1}, \ldots, y_{t-p}, w_{t-1}, \ldots, w_{t-q})$, with w_t a vector of exogenous variables, and u_t is a white-noise sequence with mean zero and variance σ_u^2. If $\widetilde{y}_t = (1, y_{t-1}, \ldots, y_{t-p})$, the model is called a smooth transition autoregressive (STAR). We will concentrate on this case hereafter as results can easily be generalized to the STR. The variable z_{t-d} is the transition variable, the one that drives the dependent variable to move between regimes. We will consider the case that $z_{t-d} = y_{t-d}$. The STAR model can then be written as:

$$y_t = \beta_0 + \sum_{j=1}^{p} \beta_j y_{t-j} + \left(\phi_0 + \sum_{j=1}^{q} \phi_j y_{t-j} \right) F(y_{t-d}; \gamma, c) + u_t. \tag{22.3}$$

The variable y_{t-d} (though in general it could be a vector) is assumed to be stationary and ergodic and, for $d \in 1, 2, \ldots, d_{\max}$ with d the delay parameter, must satisfy $1 \leq d \leq r$ for $r = \max(p, q)$. We will only consider the case $p = q$. We are interested in the special case of a unit root in the linear polynomial, $\sum_{j=1}^{p} \beta_j = 1$. The function $F(\cdot)$ is at least fourth-order, continuously differentiable with respect to γ. There are two common forms of the STAR model. The one we will discuss here in detail is the exponential STAR (ESTAR) model of Granger and Teräsvirta (1993), in which transitions between a continuum of regimes are assumed to occur smoothly.[6] The transition function $F(\cdot)$ of the ESTAR model is:

$$F(y_{t-d}; \gamma, c) = [1 - \exp(-\gamma(y_{t-d} - c)^2)]. \tag{22.4}$$

This transition function is symmetric around $(y_{t-d} - c)$ and admits the limits:

$$\begin{aligned} F(\cdot; \gamma) &\rightarrow 1 \text{ as } |y_{t-d} - c| \rightarrow +\infty, \\ F(\cdot; \gamma) &\rightarrow 0 \text{ as } |y_{t-d} - c| \rightarrow 0. \end{aligned}$$

Parameter γ can be seen as the transition speed of the function $F(\cdot)$ towards 1 (or 0) as the deviation grows larger (smaller). The variable y_t moves between an AR of the form $y_t = \beta_0 + \phi_0 + \sum_{j=1}^{p}(\beta_j + \phi_j)y_{t-j} + u_t$ when $F(\cdot)$ is 1 and a unit root, $y_t = \beta_0 + \sum_{j=1}^{p} \beta_j y_{t-j} + u_t$, when the variable is in "equilibrium," $F(\cdot) = 0$. If $\phi_j = -\beta_j$ $\forall j$, the variable y_t would be an AR process that moves between a white noise and a unit root depending on the size of the deviation, $|y_{t-d} - c|$.[7] Unlike in a linear model, the speed of adjustment of these nonlinear models will depend on the size of the deviation from PPP. They can exhibit strong persistence and near unit root behavior.[8] Recent empirical work (e.g., Michael *et al.*, 1997; Taylor *et al.*, 2001; Paya *et al.*, 2003; Paya and Peel, 2006a) has employed monthly real exchange rates for the interwar and post-war float as well as a two-century span of annual rates and showed that the ESTAR model provides a parsimonious fit to the data.

22.2.1.1 Linearity testing against STR

When testing for the existence of the nonlinear part of (22.2) an identification problem arises. The null hypothesis of linearity corresponds to $H_0 : \phi' = 0$. Under H_0, the parameters γ and c could take any value as they are not identified under the null. Alternatively, if the null hypothesis was $H_0 : \gamma = 0$, then parameters ϕ and c would not be identified under the null. In these cases it would not be possible to differentiate between a linear or nonlinear process (see Davies, 1977).[9] This problem is solved by taking a Taylor series approximation of $F(\cdot)$ with respect to γ evaluated at $\gamma = 0$. This method was introduced by Luukkonen *et al.* (1988) and adopted by Teräsvirta (1994). A third-order Taylor expansion of the logistic function would yield:

$$y_t = \beta' \widetilde{y}_t + \frac{1}{4}\gamma \phi' \widetilde{y}_t (y_{t-d} - c) + \frac{1}{48}\gamma^3 \phi' \widetilde{y}_t (y_{t-d} - c)^3. \tag{22.5}$$

In the case of the exponential function a first-order Taylor approximation yields:

$$y_t = \beta' \widetilde{y}_t + \gamma \phi' \widetilde{y}_t (y_{t-d} - c)^2. \tag{22.6}$$

A researcher might not know which STAR model the data follows and a sensible first step would be to have a general linearity test that will include both alternative models. Teräsvirta (1994) proposes a modeling cycle consisting of the following stages:

1. Specification of a linear model.
2. Testing for linearity, $H_{L,0} : \gamma = 0$. Combining (22.5) and (22.6) and recombining in terms of identified parameters, the regression equation becomes:

$$y_t = \beta_1^{'}\tilde{y}_t + \beta_2^{'}\bar{y}_t y_{t-d} + \beta_3^{'}\bar{y}_t y_{t-d}^2 + \beta_4^{'}\bar{y}_t y_{t-d}^3 + u_t, \tag{22.7}$$

   where $\bar{y}_t = (y_{t-1}, \ldots, y_{t-p})$. The linearity test has null hypothesis $H_{L,0} : \beta_2^{'} = \beta_3^{'} = \beta_4^{'} = 0$ and the original hypothesis can be tested by applying the Lagrange multiplier (LM) principle. The appropriate transition variable lag in the STR model can be determined without specifying the form of the transition function. We can compute the F-statistic for $H_{L,0}$ for various values of d (and different z_t variables) and select the one for which the p-value of the test is smallest.
3. Selecting the transition function. The choice between ESTAR and LSTAR models can be based on the following sequence of null hypotheses:

$$\begin{aligned} H_{03} &: \ \beta_4 = 0, \\ H_{02} &: \ \beta_3 = 0 \mid \beta_4 = 0, \\ H_{01} &: \ \beta_2 = 0 \mid \beta_3 = \beta_4 = 0. \end{aligned}$$

   If the p-value for the F-test of H_{02} is smaller than that for H_{01} and H_{03} then we select the ESTAR family, otherwise we choose the LSTAR family.

While Teräsvirta (1994) uses a third-order Taylor expansion of the logistic function and a first-order expansion for the exponential function, Escribano and Jordä (1999) (EJ) augment the regression equation with a second order expansion of the exponential function:[10]

$$y_t = \beta_1^{'}\tilde{y}_t + \beta_2^{'}\bar{y}_t y_{t-d} + \beta_3^{'}\bar{y}_t y_{t-d}^2 + \beta_4^{'}\bar{y}_t y_{t-d}^3 + \beta_5^{'}\bar{y}_t y_{t-d}^4 + v_t. \tag{22.8}$$

EJ claim that this procedure improves the power of both the linearity test and the selection procedure test. The null hypothesis of linearity corresponds to $H_0 : \beta_2^{'} = \beta_3^{'} = \beta_4^{'} = \beta_5^{'} = 0$. Under this null the LM test is asymptotically χ^2. The selection procedure between ESTAR and LSTAR is as follows:

1. Test the null $H_0 : \beta_3^{'} = \beta_5^{'} = 0$ with an F-test (F_E).
2. Test the null $H_0 : \beta_2^{'} = \beta_4^{'} = 0$ with an F-test (F_L).
3. If the p-value of F_E is lower than F_L then select an ESTAR. Choose an LSTAR otherwise.

If the errors display heteroskedasticity, Granger and Teräsvirta (1993) suggest ways of making the testing procedure more robust.[11] However, Lundbergh and

Teräsvirta (1998) conclude that conditional heteroskedasticity may result in severe size distortions and that the robust version of Granger and Teräsvirta (1993) appears to have very low power.[12] Pavlidis *et al.* (2007) investigate the performance of possible alternatives to improve the properties (size and power) of linearity LM tests using heteroskedasticity consistent covariance matrix estimators (HCCME) and the wild bootstrap.[13] They show that in the case of the LM linearity tests, the fixed-design wild bootstrap appears to improve tests both in terms of size and size-adjusted power.

However, besides the functional form of the real exchange rate, a major concern in the PPP literature is that real exchange rates follow a RW. Kapetanios *et al.* (2003a) (KSS) develop a test with a linear unit root null against the alternative of a stationary ESTAR. Their test is also based on a Taylor approximation of the nonlinear autoregressive model. For simplicity, assuming $p = 1$, $d = 1$, $\beta_1 = 1$, $\phi_1 = -\beta_1$, $c = 0$, then (22.3) becomes:

$$y_t = y_{t-1} + \left[1 - \exp\left(-\gamma y_{t-1}^2\right)\right](-y_{t-1}) + u_t. \tag{22.9}$$

Using the first-order Taylor expansion (22.6) in this particular case:

$$y_t = y_{t-1} - \delta(y_{t-1})(y_{t-1})^2 \Rightarrow \Delta y_t = \delta y_{t-1}^3 + u_t. \tag{22.10}$$

Under the null hypothesis of linearity, $H_0 : \gamma = 0$, $\Delta y_t = u_t$. KSS also consider the more general case where model (22.9) includes deterministic components. To ease notation, let $y_t^* = y_t - \hat{c}' x_t$ where $x_t = (1, t)$, and $\hat{c}'$ denotes least squares estimates. Then we can rewrite equation (22.10) as:

$$\Delta y_t^* = \sum_{j=1}^{p} a_j \Delta y_{t-j}^* + \delta y_{t-1}^{*3} + u_t, \tag{22.11}$$

where lags of the dependent variable address the issue of possible error autocorrelation. Testing for $\delta = 0$ against $\delta < 0$ corresponds to testing the null hypothesis. The t-statistic for the null of a linear unit root is given by:

$$t_{NL}(\hat{c}') = \frac{\hat{\delta}}{s.e(\hat{\delta})}, \tag{22.12}$$

where $s.e(\hat{\delta})$ denotes the standard error of $\hat{\delta}$. The asymptotic distribution of (22.12) is not standard but converges weakly to a complicated functional of Brownian motions.[14] Asymptotic critical values for the $t_{NL}(\hat{c}')$ statistics have been tabulated via stochastic simulation. KSS use quarterly data for bilateral real exchange rates for 11 OECD countries against the dollar covering the period 1957–98. While the ADF test was unable to reject the null of a unit root in any of the rates, the KSS test rejected the null in favor of an ESTAR in six cases, thus giving support to PPP.[15]

The linearity tests reviewed so far test the null of linear stationarity or a linear unit root process against a globally stationary nonlinear process in levels. Harvey and Leybourne (2007) (HL) develop a testing procedure for the null of linearity

against nonlinearity. Rejection of the null is therefore indicative of nonlinearity and not that the DGP follows a different linear process.

The HL test consists of two steps. First is the test of linearity. Second, the order of integration of the linear or nonlinear process is determined. Consider the linearity test (22.7) described above for the case of the null of linear $I(0)$. In the case of an $I(1)$ variable, the Taylor expansion would become (considering the transition variable is y_{t-1}):

$$\Delta y_t = \varphi_0 \Delta y_{t-1} + \varphi_1 (\Delta y_{t-1})^2 + \varphi_1 (\Delta y_{t-1})^3 + \varepsilon_t.$$

In order to combine both possibilities, $I(0)$ and $I(1)$, HL propose the following regression model:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1}^2 + \alpha_3 y_{t-1}^3 + \alpha_4 \Delta y_{t-1} + \alpha_5 (\Delta y_{t-1})^2 + \alpha_6 (\Delta y_{t-1})^3 + \varepsilon_t. \tag{22.13}$$

The null hypothesis of linearity would be $H_{0L} : \alpha_2 = \alpha_3 = \alpha_5 = \alpha_6 = 0$. The alternative hypothesis (nonlinearity) would be that at least one of those αs is different from zero. This test is carried out using the Wald statistic:

$$W_T = \frac{RSS_1 - RSS_0}{RSS_0 / T},$$

where the restricted residual sum of squares (RSS_1) comes from an ordinary least squares (OLS) regression of y_t on a constant, y_{t-1}, and Δy_{t-1}. As HL point out, the distribution of W_T under the null differs depending on whether the process followed by y_t is $I(0)$ or $I(1)$. In order to make the limiting distribution of W_T homogeneous under the null they multiply it with a correction that is the exponential of a weighted inverse of the absolute value of the ADF statistic:[16]

$$W_T^* = \exp\left(-b \left| DF_T \right|^{-1}\right) W_T, \tag{22.14}$$

where an expression for the value of b is provided such that, for a given significance level, the critical value of W_T^* coincides with that from a $\chi^2(4)$. They also prove that, under H_1, W_T^* is consistent at the rate $O_p(T)$.

The second step is to test whether the series is an $I(0)$ or an $I(1)$ process. HL use the Harris *et al.* (2003) statistic to distinguish between $I(0)$ or $I(1)$ processes regardless of whether nonlinearity is present or not. HL apply their testing methodology to the monthly real exchange rates of 15 European countries against the dollar and find evidence of PPP (linear $I(0)$ or nonlinear $I(0)$) in only two of them, Finland and the Netherlands. Notwithstanding, from a theoretical point of view, evidence that the real exchange rate follows a nonlinear $I(1)$ process would not be considered evidence in support of PPP.

22.2.1.2 Nonlinear STR estimation

Once the functional form of the real exchange rate has been determined using linearity testing, the next step is to review the nonlinear estimation procedure.[17]

The unrestricted ESTAR model considered here has the following form:

$$y_t^* = \sum_{j=1}^{p} \varphi_j y_{t-j}^* + \left(\sum_{j=1}^{p} \varphi_j^* y_{t-j}^* \right) \left[1 - \exp\left(-\gamma \left(y_{t-d}^* \right)^2 \right) \right] + u_t, \tag{22.15}$$

where y_t^* denotes de-meaned, detrended or in-deviation data.[18] In the estimation of the nonlinear model, γ is estimated by scaling it by the variance of the transition variable. This scaling is suggested for two reasons. One is to avoid problems in the convergence of the algorithm. Second, it makes it easier to compare speeds of adjustment across different studies (see Teräsvirta, 1994).[19]

Since the t-ratio of the estimated coefficient γ in (22.15) does not provide a valid significance test in the usual way, its critical values must be obtained by simulation.[20] The estimation technique can be nonlinear least squares (NLS) or maximum likelihood. Under the assumption that u_t is normally distributed, NLS is equivalent to maximum likelihood, otherwise, NLS estimates can be interpreted as quasi-maximum likelihood estimates. Wooldridge (1994) and Pötscher and Prucha (1997) discuss regularity conditions that allow consistent and asymptotically normal estimators.

The adequacy of the estimated STR model can be evaluated employing the LM-type diagnostic tests for the hypothesis of no error autocorrelation, (the customary portmanteau test has an unknown asymptotic null distribution), nonlinearity and parameter constancy of Eitrheim and Teräsvirta (1996). The last two tests address important issues of misspecification due to neglected nonlinearity and possible parameter instability.

The nonlinear models reported in empirical work have been estimated on data sampled at different levels of aggregation, namely monthly, quarterly and annual (see, e.g., Michael *et al.*, 1997; Taylor *et al.*, 2001; Taylor and Kilian, 2003; Paya *et al.*, 2003).[21] As noted by Taylor (2001), if the true DGP is nonlinear, the temporally aggregated data could exhibit misleading properties regarding the adjustment speeds if a linear model is estimated. Paya and Peel (2006c) complement this work by examining the effects of different levels of temporal aggregation of an ESTAR DGP on aggregate estimates of ESTAR models.[22] They show that ESTAR type nonlinearities are usually preserved under the temporal aggregation schemes examined. However, the dynamic structure of the best fitting models changes and tends to take the form researchers have found to fit well on actual data of the same frequency. Furthermore, comparison of the measured speed of response to shocks with models estimated on the temporally aggregated data and the true DGP shows that the measured speed of adjustment declines the more aggregated the data.

22.2.1.3 Time-varying equilibrium real exchange rate

A variety of theoretical models, such as those of Balassa (1964), Samuelson (1964), Lucas (1982), and Backus and Smith (1993), imply a non-constant equilibrium in the real exchange rate and estimates, including proxies for the equilibrium determinants, appear significant (see, e.g., Lothian and Taylor, 2000; Hegwood and Papell, 2002). Paya and Peel (2004, 2006a), employing various proxies, show the estimated

speeds of mean-reversion to be much faster than when it is assumed constant, with the half-life of shocks to the real exchange rate being less than two years.

However, Paya and Peel (2006a) also highlight the possibility of spurious relationships in nonlinear models if standard critical values are considered as valid when a persistent variable or vector of variables (x_t) are included as proxies for the equilibrium level of the real exchange rate in the ESTAR model:

$$y_t = \alpha + \delta x_t + \exp\left(-\gamma(y_{t-1} - \alpha - \delta x_{t-1})^2\right) \sum_{i=1}^{p} \beta_i (y_{t-i} - \alpha - \delta x_{t-i}) + u_t. \quad (22.16)$$

The bootstrap methodology is used to provide a better finite sample approximation to the distribution of a particular estimator in cases where classical asymptotic theory might not yield a reliable guide.[23] If the true DGP admits heteroskedasticity of unknown form, it cannot be replicated in the bootstrap DGP. The bootstrap method called the wild bootstrap solves this problem by using the following procedure (see, e.g., Wu, 1986; Mammen, 1993; Davidson and Flachaire, 2001).[24]

The null hypothesis is that the coefficients (δ) on the proxy variables for the equilibrium real exchange rate are zero. Accordingly, an "artificial"series for y_t ($\widehat{y}_t^b$) is simulated using previously estimated coefficients of the ESTAR model (22.16) and setting the coefficients of the equilibrium determinants (δ) equal to zero:

$$\widehat{y}_i^b = \widehat{\alpha} + \exp\left(-\widehat{\gamma}(y_{t-1} - \widehat{\alpha})^2\right) \sum_{i=1}^{p} \hat{\beta}_i (y_{t-i} - \widehat{\alpha}) + u_i^b, \quad (22.17)$$

where the $i = 1, \ldots, B$ are replications. The residuals u_i^b are obtained from bootstrapping the estimated residuals ($\widehat{u}_t$) obtained from the ESTAR model (22.16) which includes the equilibrium determinants.[25] In other words, every replication employs the actual residuals from regression (22.16) and creates a new series of residuals (u_i^b) based on $\widehat{u}_t$ as follows:

$$u_i^b = \widehat{u}_t \epsilon_i,$$

where ϵ_i is drawn from the following two-point distribution:

$$\epsilon_i = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } p = \dfrac{(1+\sqrt{5})}{2\sqrt{5}}, \\ (\sqrt{5}+1)/2 & \text{with probability } (1-p). \end{cases}$$

The ϵ_i are mutually independent drawings from a distribution independent of the original data. The distribution has the properties that $E(\epsilon_i) = 0$, $E(\epsilon_i^2) = 1$, and $E(\epsilon_i^3) = 1$.[26] A consequence of these properties is that any heteroskedasticity in the estimated residuals ($\widehat{u}_t$) is preserved in the newly created residuals, u_i^b.[27]

This procedure provides an empirical distribution for $\widehat{\delta}$ and their associated standard errors. The idea in using B replications is to determine the appropriate t-values and F-statistic so we do not reject the null of $\widehat{\delta} = 0$. These critical values can then be used to determine whether the estimates of $\widehat{\delta}$ reject the null or not. Paya and Peel (2006a) find that the hypothesis that the real dollar–sterling rate follows an ESTAR process with time-varying equilibrium proxied by productivity differentials and/or wealth cannot be rejected at the usual significance level.[28]

22.2.2 Threshold autoregressive (TAR) models

If the transition between regimes is assumed abrupt rather than smooth the dynamics of PPP adjustment can be captured by the TAR model of Tong (1983). Empirical studies that use TAR models for deviations from PPP include Obstfeld and Taylor (1997) and Sarno *et al*. (2004a). One of the advantages of this methodology is the direct estimation of the transaction cost band or threshold band. For illustrative purposes we start by describing the estimation of a simple symmetric threshold TAR model of order one employed in some of the empirical work previously mentioned:[29]

$$y_t = \begin{cases} a_0 + a_1 y_{t-1} + u_t & \text{if } z_{t-d} \leq c, \\ b_0 + b_1 y_{t-1} + u_t & \text{if } z_{t-d} > c, \end{cases} \tag{22.18}$$

where z_{t-d} is the transition variable, in our case $z_{t-d} = y_{t-d}$.[30] The integer d is called the delay lag and is typically unknown, so it must be estimated. As we will shortly explain, the least squares principle allows d to be estimated along with the other parameters. Parameter c is the "threshold" that distinguishes two regimes: (i) transition variable z_{t-d} is below c (lower regime); (ii) transition variable z_{t-d} is above c (upper regime). Then, parameter vectors $\boldsymbol{\alpha} = (a_0, a_1)'$ and $\mathbf{b} = (b_0, b_1)'$ determine the response of the real exchange rate to changes in its last period's value.

If the threshold value, c, was known, then to test for threshold behavior all one needs is to test the hypothesis $H_0 : \boldsymbol{\alpha} = \mathbf{b}$. Unfortunately, the threshold value is typically unknown and, under the null hypothesis, parameter c is not identified. The second difficult statistical issue associated with TAR models is the sampling distribution of the threshold estimate. Hansen (1997) provides a bootstrap procedure to test H_0, develops an approximation to the sampling distribution of the threshold estimator free of nuisance parameters, and also develops a statistical technique that allows confidence interval construction for c.[31] In particular, we can write the TAR model (22.18) compactly as:

$$y_t = x_t(c)'\theta + u_t, \tag{22.19}$$

where $x_t(c) = (x_t' \mathbf{1}\{z_{t-d} \leq c\}, x_t' \mathbf{1}\{z_{t-d} > c\})'$ with $x_t = (1, y_{t-1})'$, $\mathbf{1}\{\cdot\}$ the indicator function and $\theta = (\boldsymbol{\alpha}', \mathbf{b}')'$. For a given value of c the least squares (LS) estimate of θ is:

$$\hat{\theta}(c) = \left(\sum x_t(c) x_t(c)' \right)^{-1} \left(\sum x_t(c) y_t \right),$$

with LS residuals $\widehat{u}(c)_t$ and LS residual variance $\sigma_T^2(c) = (1/T) \sum_{t=1}^{T} \hat{u}^2(c)_t$. Then the LS estimate of c is the value:

$$\hat{c} = \arg\min_{c \in C} \sigma_T^2(c), \tag{22.20}$$

where C is an interval (usually trimmed) that covers the sample range of the transition variable. Problem (22.20) can be solved by a direct search over C. The LS

estimate of θ is then $\hat{\theta} = \hat{\theta}(\hat{c})$. Furthermore, the LS principle allows the estimation of the, typically, unknown value d by extending problem (22.20) to a search across the discrete space $[1, \bar{d}]$.

The hypothesis $H_0 : \boldsymbol{\alpha} = \mathbf{b}$ is tested as follows. Let $\{e_t\}_{t=1}^{T}$ be an i.i.d. sequence of $N(0,1)$ draws. Regress e_t on x_t to obtain the residual variance $\hat{\sigma}_T^2$ and on $x_t(c)$ to obtain $\hat{\sigma}_T^2(c)$ and compute $F(c) = T(\hat{\sigma}_T^2 - \hat{\sigma}_T^2(c))/\hat{\sigma}_T^2(c)$. Then compute $F = \sup_{c \in C} F(c)$. Repeat the procedure n times and the asymptotic p-value of the test is given by the percentage of samples for which F exceeds the observed F_T. Hansen (1997) also provides critical values and a method to construct asymptotically valid confidence intervals.[32]

In the last decade the basic TAR model has been extended. New unit root tests against these TAR models have been used as tests for PPP. We detail below some of those tests that include TAR models but differ according to the nature of the transition variable (in levels or in differences), the symmetry or asymmetry of the bands, the autoregressive process within each region (unit root, stationary AR), and the number of regimes.

22.2.2.1 Unit root test versus TAR

Enders and Granger (1998) (EG) modify the Dickey–Fuller (DF) critical values of the F-statistic in order to have the right size to test the unit root null against TAR, or momentum-TAR (M-TAR) models. In particular, they consider the F-statistic of the null unit root hypothesis ($H_0 : \rho_1 = \rho_2 = 0$) in the following TAR model:

$$\Delta y_t = \begin{cases} \rho_1 y_{t-1} + u_t & \text{if } y_{t-1} \leq 0, \\ \rho_2 y_{t-1} + u_t & \text{if } y_{t-1} > 0, \end{cases} \tag{22.21}$$

and also for the M-TAR model:

$$\Delta y_t = \begin{cases} \rho_1 y_{t-1} + u_t & \text{if } \Delta y_{t-1} \leq 0, \\ \rho_2 y_{t-1} + u_t & \text{if } \Delta y_{t-1} > 0. \end{cases} \tag{22.22}$$

EG show that the power of their test improves relative to the standard ADF as the asymmetric adjustment becomes more pronounced.[33]

An alternative threshold unit root test is developed by Basci and Caner (2005) (BC). They consider the following M-TAR model:

$$\Delta y_t = \begin{cases} \theta_1' x_{t-1} + e_t & \text{if } |y_{t-1} - y_{t-m-1}| < \lambda, \\ \theta_2' x_{t-1} + e_t & \text{if } |y_{t-1} - y_{t-m-1}| \geq \lambda, \end{cases} \tag{22.23}$$

where $x_{t-1} = (y_{t-1}, 1, \Delta y_{t-1}, ..., \Delta y_{t-k})$ for $t = 1, 2, ..., T$. e_t is an i.i.d. error term, m represents the delay parameter and $1 \leq m \leq k$. It is possible to rewrite the model above as follows:

$$\Delta y_t = \theta_1' x_{t-1} \mathbf{1}_{\{|y_{t-1} - y_{t-m-1}| < \lambda\}} + \theta_2' x_{t-1} \mathbf{1}_{\{|y_{t-1} - y_{t-m-1}| \geq \lambda\}} + e_t. \tag{22.24}$$

The expression above is estimated using OLS for each $\lambda \in \Lambda$ and the OLS estimate[34] of σ^2 is then $\widehat{\sigma}^2(\lambda) = T^{-1} \sum_{t=1}^{T} \widehat{e}_t(\lambda)^2$. The estimated threshold parameter λ is the one that minimizes the error variance: $\widehat{\lambda} = \arg\min_{\lambda \in \Lambda} \widehat{\sigma}^2(\lambda)$. Testing linearity reduces to $H_0 : \theta_1 = \theta_2$. BC propose the following test:

$$\sup_{\lambda \in \Lambda} W_T(\lambda) = \sup_{\lambda \in \Lambda} T \left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2(\widehat{\lambda})} - 1 \right), \tag{22.25}$$

where $\widehat{\sigma}_0^2$ is the residual variance from simple OLS estimation of the null linear model. Following Caner and Hansen (2001), BC resort to a bootstrap approximation of the distribution of W_T to obtain p-values. They tested the nonlinear behavior of 14 OECD real exchange rates and found that 11 of them display a unit root inside the band and mean reversion outside the band.[35]

This section has examined recent developments in linearity testing and the autoregressive modeling of real exchange rates. The overall conclusion is that nonlinear models provide considerably greater support for the PPP hypothesis and that the PPP puzzle is largely resolved by them.

22.3 International parity conditions

22.3.1 Covered interest parity (CIP)

In the absence of frictions such as transactions costs or limits to arbitrage funds, riskless arbitrage should ensure that the covered interest differential on assets of identical characteristics should be equal to zero. Employing the usual approximations, we have that:

$$i_t - i_t^* = f_t - s_t, \tag{22.27}$$

where i_t, i_t^* are the interest rates on the domestic and foreign assets concerned, f_t is the logarithm of the forward exchange rate (the rate at which the future exchange of currencies is agreed at time t) of the same term to maturity as the assets, and s_t is the spot exchange rate (domestic price of foreign currency).

Whether CIP holds is of interest for at least three reasons. First, absence of CIP, given its riskless nature in principle, would *ceteris paribus* imply that the efficient markets assumption approach to modeling exchange rates (or asset prices in general) had very serious limitations. Second, CIP forms a basis with uncovered interest arbitrage (see below) to determining the properties of the forward rate as a predictor of future movements in the spot rate. Uncovered interest arbitrage could hardly be expected to hold in the absence of CIP. Finally, it is assumed in numerous models that covered and uncovered parity hold. Many empirical tests of the covered interest rate condition have been undertaken.[36] Taylor (1987, 1989), unlike in most previous studies, employs high frequency contemporaneously sampled data for spot and forward dollar–sterling and dollar–Mark exchange rates and corresponding euro-deposit interest rates for a number of maturities and makes allowance for bid-ask spreads and brokerage costs in his calculations for the 1980s and selected

post-war periods. His evidence suggests that there are few profitable violations of CIP, even during periods of market uncertainty and turbulence, which contrasts with the results in earlier studies, thus illustrating the crucial role that appropriately sampled data can play.

A second method for testing CIP widely used in the early literature is to employ regression analysis and test whether $\alpha = 0$ and $\beta = 1$ in the regression:

$$f_t - s_t = \alpha + \beta(i_t - i_t^*) + u_t. \tag{22.28}$$

If CIP holds, on average we should obtain estimates of α and β differing insignificantly from zero and one, respectively. However, as noted by Taylor (1987), $\widehat{\alpha} = 0$ and $\widehat{\beta} = 1$ is a necessary but not a sufficient condition for CIP to hold. These restrictions could be met but the error term might be of such magnitude as to permit substantial arbitrage possibilities.

A more recent approach to modeling deviations from CIP is to employ univariate threshold models (see Balke and Wohar, 1998; Peel and Taylor, 2002).[37] The rationale of the univariate threshold model is, of course, to capture the transactions band that arbitrageurs face in reality. This method of analysis was explained in section 22.2. A complementary method is to model the dynamics of adjustment of each component of CIP by the threshold error correction model set out by Balke and Fomby (1997).

Peel and Taylor (2002) applied this model, as well as the univariate threshold model, to weekly data in the interwar period 1922–25. We outline their method for estimating the threshold error correction model. If we define the vector $X_t = (i_t, i_t^*, f_t - s_t)'$, the deviation from CIP, δ_t, may be viewed as an error correction term relating the three elements of X_t, since $\delta_t = f_t - s_t - (i_t - i_t^*)$. A simple first-order threshold vector error correction model (TVECM) may be written as:

$$\Delta X_t = \begin{cases} E_t & \text{if } |\delta_{t-1}| < \kappa, \\ \theta + \Gamma\delta_{t-1} + E_t & \text{if } \delta_{t-1} \geq \kappa, \\ -\theta + \Gamma\delta_{t-1} + E_t & \text{if } \delta_{t-1} \leq -\kappa, \end{cases} \tag{22.29}$$

where E_t is a (3×1) disturbance vector, and Γ and θ are (3×1) parameter vectors. Within the band, the error correction term has no effect on any of the variables and there is no tendency to adjust toward CIP. However, once outside the band, we expect at least one of the elements in Γ to be non-zero. In that case, one or more of $f_t - s_t$, i_t, and i_t^* adjust toward CIP so that δ_t also adjusts. The statistical significance and relative size of the estimated elements of Γ, the error correction parameters, should give an indication of the relative speeds of adjustment of the components of CIP to large deviations from CIP.

If we define the indicator variables $\mathbf{1}(|\delta_{t-1}| < \kappa)$, $\mathbf{1}(\delta_{t-1} \geq \kappa)$ and $\mathbf{1}(\delta_{t-1} \leq -\kappa)$, each of which takes the value unity when the inequality indicated in parentheses is satisfied, and zero otherwise, the TVECM may be written as a set of dummy variable regressions:

$$\Delta X_t = \mathbf{1}(\delta_{t-1} \geq \kappa)\theta - \mathbf{1}(\delta_{t-1} \leq -\kappa)\theta + [1 - \mathbf{1}(|\delta_{t-1}| < \kappa]\Gamma\delta_{t-1} + E_t. \tag{22.30}$$

Estimation may be carried out by nonlinear least squares through a grid search over κ. However, given the need to form an overall objective function from three sets of residuals, it is easier in the multivariate case to employ maximum likelihood estimation. The concentrated log-likelihood function for this system, for known κ, is:

$$L(\theta, \Gamma, \Sigma, \kappa) = \frac{-T}{2}\left(3\left\{1 + \ln(2\pi)\right\} + \ln\left|\widehat{\Sigma}\right|\right), \tag{22.31}$$

where $\widehat{\Sigma} = (1/T)\Sigma^{T} E_t E_t'$ is the maximum likelihood estimator of the covariance matrix (see Davidson and MacKinnon, 1993, p. 316). This is maximized through a grid search over κ with (22.30) estimated using a full information maximum likelihood (FIML) estimator at each point in the grid. If $\widehat{L}(h)$ is the maximized log-likelihood conditional on a bandwidth parameter h, the resulting estimator of κ may be expressed as:

$$\widehat{\kappa} = \arg\max_{h \in H} \widehat{L}(h),$$

where $H = [0, \max|\delta_t|]$ is the range of the grid search. Hypotheses concerning the parameters can then be tested using an LR statistic defined as $\lambda' = 2(\widehat{L} - \widetilde{L})$, where $\widehat{L}$ denotes the value of the maximized log-likelihood and $\widetilde{L}$ denotes the maximized log-likelihood with the relevant restrictions imposed. Empirical marginal significance levels for this statistic can be calculated using methods set out in Hansen (1997) and explained in section 22.2.2. The results reported by Peel and Taylor were consistent with the conjectures of Keynes (1923) and Einzig (1937). Arbitrage only occurred when significant deviations from CIP occurred and the adjustment back to the implied arbitrage bounds was fairly persistent due to the microstructure faced by arbitragers. Estimation of these threshold (or ESTAR) error correction mechanisms would be of interest in other areas such as PPP.

In the absence of a new study employing high-quality data of the type employed by Taylor in the 1980s, it would appear reasonable to assume that restrictions on arbitrage funds and the ability to make near riskless trades instantaneously implies that CIP holds within a very small transactions band.

However, before we discuss the uncovered condition there is one feature of the nonlinear work applied to CIP that needs comment and has applicability to nonlinear models more generally. Whilst the nonlinear assumption is well motivated, the economic arguments of arbitrage apply to a particular data frequency. For instance, currently we might expect deviations from the arbitrage bound to occur near instantaneously. In the 1920s the appropriate period might have been hours, (possibly days or a week). Unfortunately, if the data frequency available is not that of the DGP then the estimates of a nonlinear model may generate misleading results as to the period of dynamic adjustment or the impact of shocks, as discussed in the context of PPP above. Paya and Peel (2007b) show that systematic sampling from the true DGP, where they employ every *k*th observation from the true DGP, can lead to seriously biased estimates of speeds of adjustment. Estimation of nonlinear models on data sampled at a different frequency to the economic decision would appear to be a serious problem in the evaluation of nonlinear models.[38]

22.3.2 Uncovered interest parity (UIP)

In the absence of frictions an agent should be indifferent between holding domestic or foreign assets of identical type. The return from the foreign asset over a given holding period is its return plus or minus the return from exchange rate changes over the holding period. The latter component is risky so that risk averse agents will require a risk premium. Employing interest rates as the asset's return, the UIP condition is given by:

$$i_t - i_t^* = E_t s_{t+n} - s_t + rp_t, \tag{22.32}$$

where $E_t s_{t+n}$ is the expectation at time t of the exchange rate in n periods time, the time to maturity of the interest rates, conditional on all information available at time t. The uncovered parity condition in conjunction with the covered parity condition imply that:

$$f_t = E_t s_{t+n} + rp_t, \tag{22.33}$$

and assuming rational expectations we obtain:

$$s_{t+n} = f_t - rp_t + \epsilon_{t+n}, \tag{22.34}$$

where ϵ_{t+n} is the rational expectations forecast error, which can follow up to an $n-1$ order moving average error process (see Hansen and Hodrick, 1980). Early tests of the properties of the forward rate were based on testing the hypothesis that $\beta = 1$ in the following OLS regression:

$$s_{t+n} = \alpha + \beta f_t + u_t. \tag{22.35}$$

The estimates of β were typically close to unity. However, the early research, naturally, was unaware that if s_t and f_t are integrated variables, so that (22.35) is a cointegration regression, the t-ratio is not asymptotically standard normal.

A number of different estimation techniques have been employed to estimate (22.35) which remedy this deficiency. For example, Hai *et al.* (1997) estimate α and β with the dynamic OLS (DOLS) and dynamic generalized least squares (DGLS) cointegration vector estimators of Stock and Watson (1993). Moore and Copeland (1995) employ the fully modified maximum likelihood procedure (FM-OLS) of Phillips and Hansen (1990), which treats equation errors in a general semi-parametric way to estimate α and β. Phillips *et al.* (1996), building on Phillips (1995), estimate the coefficients employing the fully modified least absolute deviations (FM-LAD) estimator.[39]

A comparison by Phillips of the FM-LAD estimates of α and β with those obtained from FM-OLS and OLS on interwar data suggests it can make a major difference to the magnitude and significance of the estimated coefficients. The properties of the FM-LAD estimator appear to us to make it a prime candidate for use in many areas of finance, e.g., the relationship between asset prices and fundamentals and tests of bubbles (see section 22.5). It appears to us a neglected contribution.

Overall, the estimates of β do not appear to differ significantly from unity.[40] The estimates of α do appear to differ from zero but this, of course, could be the influence of a stationary, non-zero mean risk-premium. However, an estimate of

β that does not differ significantly from unity does not imply that markets are efficient or expectations are formed rationally. It is also consistent with a variety of non-rational expectations processes.

Orthogonality tests are a standard method for testing the rationality of expectations. Consider the following specifications:

$$s_{t+n} - s_t = \lambda + \theta(f_t - s_t) + v_{t+n}, \tag{22.36}$$

$$s_{t+n} - f_t = \lambda + (\theta - 1)(f_t - s_t) + v_{t+n}, \tag{22.37}$$

where v_{t+n} is the error term. Given rational expectations and risk-neutrality, (22.34) implies that $\lambda = 0$ and $\theta = 1$ and the error term exhibits up to an $n-1$ order moving average error. In fact, a vast amount of empirical work has reported estimates of θ that are not only significantly different from unity but also significantly negative (see, e.g., Fama, 1984; Hodrick, 1987; Backus *et al.*, 1993). The negative value of θ implies that the more the foreign currency is at a premium in the forward market, the less the home currency is predicted to depreciate. In that case the spot exchange rate next period moves, on average, in the opposite direction to that currently predicted by the forward premium. This implication has become the forward bias puzzle in the literature. Of course, one explanation could be the absence of rational or informed expectations and numerous evidence on the properties of survey evidence on expectation formation support this view (Frankel and Froot, 1987). However, the systematic nature of the pattern in empirical estimates of θ over both the interwar and post-war period rather suggests *a priori* that expectation formation cannot play a major role in the explanation even if one attaches reasonable weight to the quality of survey data.

Numerous reasons have been set out to explain the puzzle and we now consider some of these. Fama (1984) considers the implications of a time-varying risk-premium for estimates of (22.36) and (22.37). Assuming rational expectations, we have as a property that:

$$s_{t+n} - s_t = Es_{t+n} - s_t + \epsilon_{t+n}. \tag{22.38}$$

Also, from rearrangement of (22.33):

$$f_t - s_t = E_t s_{t+n} - s_t + rp_t, \tag{22.39}$$

and from (22.34):

$$s_{t+n} - f_t = -rp_t + \epsilon_{t+n}. \tag{22.40}$$

Now the OLS estimate of θ in (22.36) must satisfy asymptotically:

$$\begin{aligned} plim(\widehat{\theta}) = \theta &= \frac{Cov(Es_{t+n} - s_t + \epsilon_{t+n}, E_t s_{t+n} - s_t + rp_t)}{Var(f_t - s_t)} \\ &= \frac{Var(E_t s_{t+n} - s_t) + Cov(rp_t, Es_{t+n} - s_t)}{Var(f_t - s_t)}; \end{aligned} \tag{22.41}$$

also, from (22.37):

$$plim(\widehat{\theta} - 1) = \theta - 1 = \frac{Cov(-rp_t + \epsilon_{t+n}, E_t s_{t+n} - s_t + rp_t)}{Var(f_t - s_t)}$$
$$= \frac{-Var(rp_t) - Cov(rp_t, Es_{t+n} - s_t)}{Var(f_t - s_t)}. \tag{22.42}$$

We note from (22.41) that a negative estimate of θ implies that $Cov(rp_t, Es_{t+n} - s_t) < 0$. This is the important point made by Fama (1984), that negativity of estimates of θ require a negative covariation between the risk premium and the expected rate of depreciation. In addition, this covariation has greater absolute magnitude than $Var(E_t s_{t+n} - s_t)$. From (22.42), a negative estimated coefficient implies that $Var(rp_t)$ has greater absolute magnitude than $Cov(rp_t, Es_{t+n} - s_t)$.

A time-varying risk premium is well motivated. For example, in a consumption capital asset pricing model (CAPM) framework, assuming logarithmic utility and that all variables are jointly lognormally distributed, we can derive that:

$$f_t - E_t s_{t+1} = 0.5 Var_t(s_{t+1}/s_t) - cov_t([s_{t+1}/s_t].p_{t+1}/p_t)$$
$$- \delta cov_t([s_{t+1}/s_t].\log(c_{t+1}/c_t)), \tag{22.43}$$

where δ is the coefficient of relative risk-aversion.[41] In general, optimizing models built on microtheoretic underpinnings will imply that the risk-premium depends on the variance of the exchange rate. As with other asset prices, there is considerable evidence of time-varying volatility in spot exchange rates at high frequencies. Various authors have employed extensions of the work of Engle (1982) and estimated multivariate GARCH models and included the own conditional variance in the mean equation (see, e.g., Baillie and Bollerslev, 1990; Bekaert and Hodrick, 1993). For instance, the basic idea (without the multivariate generalization) is to estimate:

$$s_{t+1} - s_t = \alpha_0 + \alpha_1(f_t - s_t) + \alpha_2 h_{t+1} + \epsilon_{t+1}, \tag{22.44}$$

$$h_{t+1} = \delta_0 + \delta_1 h_t + \delta_2 \epsilon_t^2 + \delta_3 \left| f_t - s_t \right|, \tag{22.45}$$

where variables are defined as above and h_t is the conditional variance of the error term. The absolute difference $|f_t - s_t|$ is included in the variance equation based on empirical observation by Hodrick (1989).

GARCH effects disappear with aggregation (Drost and Nijman, 1993) and so are not usually statistically significant in low-frequency data such as monthly or quarterly. Also consumption and price data are not available at weekly or daily levels. Consequently, some of the terms in the risk-premium have to be assumed constant in empirical work. Nevertheless, it would appear that a time-varying risk-premium does not rationalize the forward premium anomaly.[42] However, it is interesting in this context to note the results reported by Flood and Rose (1996). They found that estimates of θ were positive in credible periods in the EMS target zone when risk-premia might, *a priori*, be expected to be smaller in magnitude.[43]

Another important possibility is inference problems arising from the different statistical properties of changes in the the spot rate and forward premium. Baillie and Bollerslev (2000) consider a sample of monthly observations on the DM/$ spot and one-month forward rates from January 1974 to December 1991, realizing a total of 215 observations. They report that the monthly sample standard deviation of percentage changes in the spot rate is 2.75 and the corresponding figure for the monthly forward premium as 0.217. These figures are typical. Changes in the spot rate usually exhibit a standard deviation at least 100 times bigger than the forward premium. In addition, the forward premia are generally very persistent whilst changes in the spot rate are not. In fact, Baillie and Bollerslev (1994), Byers and Peel (1996), and Maynard and Phillips (2001) have argued that the temporal dependencies in the forward premium can be parsimoniously described by a fractionally integrated, or I(d), process.

Mathematically, the autoregressive fractional integrated moving average (ARFIMA) model for a time series process y_t can be written as:

$$\phi(L)(1-L)^d(y_t-\mu)=\theta(L)\epsilon_t, \tag{22.46}$$

where $\phi(z)=0$ and $\theta(z)=0$ have all roots lying in the unit circle and $\{\epsilon_t\}$ is a martingale difference sequence. The differencing operator, $(1-L)^d$, is defined as follows:

$$(1-L)^d=\sum_{j=o}^{\infty}\frac{\Gamma(j-d)L^j}{\Gamma(-d)\Gamma(j+1)},$$

where Γ is the gamma or generalized factorial function. A fractionally integrated process is one that exhibits long memory, with persistent local trends, but which nonetheless eventually "reverts to the mean." The degree of persistence is measured by the real-valued parameter d, lying on the unit interval.[44] Smallwood (2005) has an interesting discussion of the properties of the various estimators of the fractional parameter d in the context of PPP. One important property of fractional processes is the self-similarity property. This implies that the estimate of d should remain invariant over different temporal aggregates of the process. Ohanissian *et al.* (2007) have developed a statistical test based on this property. Their simulations show that the test has good size and power properties against some alternatives such as Markov-switching. The potential use of this test in exchange rate econometrics seems great. We also consider that the power of this test against alternatives such as ESTAR would be of interest.

From a purely statistical perspective the different properties of changes in the spot rate and the forward premium suggest that the orders of integration of the dependent and explanatory variables are not the same. One method of dealing with the long-memory characteristics of the forward premium, as suggested by Baillie and Bollerslev, is to regress the spot return on the fractionally differenced forward premium (see Maynard and Phillips, 2001; Abadir and Talmain, 2006):

$$s_{t+1}-s_t=\lambda+\theta(1-L)^d(f_t-s_t)+\epsilon_{t+1}. \tag{22.47}$$

Such a regression appears to imply rejection of efficiency. Granger and Joyeux (1980) illustrate how long memory can arise via aggregation. Alternatively, Granger and Hyung (2004) and Diebold and Inoue (2001) show that structural breaks or regime switching can generate spurious long-memory behavior in an observed time series. Granger and Teräsvirta (1999) provide an abstract example of a nonlinear model that can generate data with the misleading linear property of long memory. They suggest that other nonlinear models with this property are worth searching for. Byers and Peel (2003) show that data generated from an ESTAR can exhibit the long-memory property whether in raw or temporally aggregated form. That this might be the case was an early conjecture of Acosta and Granger (1995). In this respect, recent applied work which has tried to explain the anomaly by nonlinearities induced by transactions costs and other frictions seems promising (see, e.g., Coakley and Fuertes, 2001; Leon *et al.*, 2003; Sarno *et al.*, 2006).[45]

We know from analysis of the empirical work on PPP discussed in section 22.2 that fractional processes can appear to parsimoniously explain PPP deviations but are not as theoretically well motivated as the nonlinear models that also appear to parsimoniously explain the data. Leon *et al.* (2003) and Sarno *et al.* (2006) provide a nice rationale for nonlinearity based on arguments by Lyons (2001) as to the limits to speculation (see Shleifer and Vishny, 1997). The idea is that financial institutions will only engage in uncovered arbitrage – a currency trading strategy – if the strategy yields a Sharpe ratio at least equal to an alternative investment strategy, such as a buy-and-hold equity strategy. The Sharpe ratio is defined as $(E[R(s)] - R(f))/\sigma_s$, where $E[R(s)]$ is the expected return on the strategy, $R(f)$ is the risk-free interest rate, and σ_s is the standard deviation of the returns to the strategy. In the foreign exchange market the excess return equals $E[R(s)] - R(f) = E[s_{t+1} - s_t - (f_t - s_t)]$ and the denominator is determined by the exchange rate variances. In the case of multiple-exchange rate strategies, the covariances among the exchange rates considered in the currency strategy would also be included in the denominator.

The Sharpe ratio can be interpreted as the expected excess return from speculation per unit of risk. Sarno *et al.* (2006) point out that the Sharpe ratio for a buy-and-hold equity strategy has averaged about 0.4 on an annual basis for the US over the last 50 years or so. Only when θ departs from unity does the numerator of the Sharpe ratio becomes positive. In fact, perhaps surprisingly, only when $\theta \leq -1$ or $\theta \geq 3$ is the Sharpe ratio for currency strategies about the same magnitude as the average from a buy-and-hold equity strategy, i.e., 0.4 (see Lyons, 2001, p.210). Consequently, there is a band of inaction such that, if $-1 < \theta < 3$, financial institutions would have no incentive to take up a currency strategy. Deviations from uncovered interest parity are too small to attract speculative funds so that the spot exchange rate and forward exchange rate need not move together.

As with PPP, these types of considerations suggest either a threshold or an ESTAR type of adjustment mechanism. The latter is justifiable by an appeal to heterogeneous agents, who face different levels of position limits and the like. Consider the following ESTAR adjustment mechanism estimated by Sarno *et al.* (2006):[46]

$$s_{t+1} - s_t = \lambda_1 + \theta_1(f_t - s_t) + [\lambda_2 + \theta_2(f_t - s_t)] \times \Psi\left(s_t - f_{t-1}; \gamma\right) + \epsilon_{t+1}, \tag{22.48}$$

where $\Psi\left(s_t - f_{t-1}; \gamma\right) = 1 - \exp\left(-\gamma(s_t - f_{t-1})^2\right)$. Equivalently:

$$s_{t+1} - f_t = \lambda_1 + (\theta_1 - 1)(f_t - s_t) + [\lambda_2 + \theta_2(f_t - s_t)]\Psi\left((s_t - f_{t-1}); \gamma\right) + \epsilon_{t+1}. \tag{22.49}$$

When $s_t - f_{t-1}$ is small we obtain from (22.48):

$$s_{t+1} - s_t = \lambda_1 + \theta_1(f_t - s_t), \tag{22.50}$$

and when $s_t - f_{t-1}$ is large:

$$s_{t+1} - s_t = \lambda_1 + \lambda_2 + (\theta_1 + \theta_2)(f_t - s_t). \tag{22.51}$$

Sarno *et al.* (2006) report that the restriction $\theta_1 + \theta_2 = 1$ cannot be rejected for the currencies they examine and also that $\theta_1 < 0$. Simulated data from this model generate negative estimates of θ in the standard regressions (22.36) and (22.37). In fact, if the constraint that $\theta_1 + \theta_2 = 1$ is imposed, the model collapses to the form employing expected excess returns as:

$$s_{t+1} - f_t = -\theta_2(f_t - s_t)\exp\left(-\gamma\left[E_t(s_{t+1} - f_t)\right]^2\right) + \epsilon_{t+1}, \tag{22.52}$$

when $\lambda_1 + \lambda_2 = 0$, where $E_t(s_{t+1} - f_t)$ is the expected excess return formed on information available in period t. In this form the model allows expectations to be formed rationally, as in the arbitrage consistent STAR (ARBSTAR) model proposed by Peel and Venetis (2005), which remedies some economic difficulties when TAR, ESTAR, or LSTAR models are employed to model arbitrage.[47]

The estimates of Sarno *et al.* (2006)[48] suggest that the relationship between the excess return and the forward premium is nonlinear. However, it is not clear that the efficiency proposition is tested, or indeed that there is a well defined null, except under the assumption of risk neutrality. Time-varying risk premia will imply the lack of a unit relationship in the outer regimes. It would be of interest to estimate the Sarno *et al.* (2006) model when risk premia are *a priori* small – e.g., in credible periods in the Exchange Rate Mechanism (ERM), as in the analysis of Flood and Rose (1996) mentioned above.

An extension of the nonlinear STAR model has recently been employed to model forward premia and PPP deviations (see Smallwood, 2005; Baillie and Kapetanios, 2005, 2006).[49] The idea is that the fractional difference of a series is an ESTAR or LSTAR model, and the model is called FI-STAR. In particular, the FI-STAR model for a time series y_t is defined as:

$$(1-L)^d y_t = \left(\alpha_{1,0} + \sum_{j=1}^{p} \alpha_{1,j}(1-L)^d y_{t-j}\right) + \left(\alpha_{2,0} + \sum_{j=1}^{p} \alpha_{2,j}(1-L)^d y_{t-j}\right) G(y_{t-k}; \gamma, c) + \varepsilon_t, \tag{22.53}$$

where $G(\cdot)$ is the transition function of an LSTAR or ESTAR model and k is the delay parameter. Under the null hypothesis, $\gamma = 0$, the time series process is distributed as a long-memory ARFIMA(p,d,0). We outline Smallwood's test for joint fractional and ESTAR nonlinearity, which seems more appropriate in the context of exchange rate econometrics where issues of transactions bands and limits to arbitrage suggest ESTAR (or threshold) as the parametric form of nonlinearity.[50] The Smallwood test procedure is similar to that outlined in section 22.2.1 and consists of a first-order Taylor series expansion of model (22.53):

$$(1-L)^d y_t = \left(\alpha_{1,0} + \sum_{j=1}^{p} \alpha_{1,j}(1-L)^d y_{t-j}\right) + \left(\sum_{j=1}^{p} \alpha_{2,j}(1-L)^d y_{t-j} y_{t-k}\right) + \left(\sum_{j=1}^{p} \alpha_{3,j}(1-L)^d y_{t-j} y_{t-k}^2\right) + e_t. \tag{22.54}$$

Assuming the error term is Gaussian, the null hypothesis of a linear fractional process is given by $H_0 : \alpha_{2,j} = \alpha_{3,j} = 0$, $j = 1, \ldots, p$. Smallwood illustrates that the existence of the fractional differencing parameter complicates the construction of the LM-type test statistic based on (22.54). However, he shows that a χ^2 and F version can be calculated as follows. One first estimates an ARFIMA(p,d,0) model and obtains the estimate of d ($\widehat{d}$), and the set of residuals $\widehat{\varepsilon}_t$. The sum of squared errors, denoted SSR_R, is then constructed from the residuals $\widehat{\varepsilon}_t$. Second, a regression of $\widehat{\varepsilon}_t$ is run on $\sum_{j=1}^{t-1} \widehat{\varepsilon}_{t-j}/j$, 1, $(1-L)^{\widehat{d}} y_{t-1}, \ldots, (1-L)^{\widehat{d}} y_{t-p}$, $(1-L)^{\widehat{d}} y_{t-1} y_{t-k}, \ldots, (1-L)^{\widehat{d}} y_{t-p} y_{t-k}$, and $(1-L)^{\widehat{d}} y_{t-1} y_{t-k}^2, \ldots, (1-L)^{\widehat{d}} y_{t-p} y_{t-k}^2$. The unrestricted sum of squared residuals, SSR_{UR}, is formed from this regression. The χ^2 version of the LM test statistic is calculated as $\mathrm{LM}_{\chi^2} = T(SSR_R - SSR_{UR})/SSR_R$, and is distributed as a $\chi^2(2p)$. The F version of the LM test statistic is calculated as $\mathrm{LM}_F = \left[(SSR_R - SSR_{UR})(2p)^{-1}\right]\left[SSR_{UR}(T-3p-1)^{-1}\right]^{-1}$, and is distributed as an $F(2p, T-3p-1)$.

In practice, of course, the long-memory parameter is generally unknown. Different methods have been employed to obtain a consistent estimate in the first step of the test. Smallwood prefers the estimator of Beran (1995) based on the conditional likelihood function of the time series process. Baillie and Kapetanios (2006) employ the local Whittle semiparametric estimator.[51] We would also suggest that the test of Ohanissian *et al.* (2007) mentioned above would be worth exploring in this context. Its size properties given data generated from a FI-STAR model would be of interest and, similarly, its power properties for data generated from an ESTAR.[52]

22.4 Target zone models

There have been a large number of papers that have examined the behavior of exchange rates in target zones. The basic theory is due to Flood and Garber (1983), with the particular application to target zones by Krugman (1991). The model of

Krugman, cast in continuous time, is based on the log-linear monetary model with instantaneous purchasing power parity so that the reduced form for the logarithm of the exchange rate $s(t)$ is given by:

$$s(t) = f(t) + \alpha \frac{E[ds(t)]}{dt}, \tag{22.56}$$

where $f(t)$ is the logarithm of the fundamental and α represents the interest rate semi-elasticity of money demand. The fundamental is assumed equal to:

$$f(t) = m(t) + v(t), \tag{22.57}$$

where $m(t)$ represents the policy instrument and $v(t)$, which is assumed to follow a Brownian motion without drift, contains all the other determinants of the exchange rate which impact through the term $f(t)$. Krugman assumes a credible target zone exists so that $s^l \leq s \leq s^u$, where s^l and s^u are the lower and upper bounds respectively. The authorities are assumed to intervene by movements in m when the exchange rate reaches either boundary value s^l or s^u. The formal solution of the model given by Krugman (1991) or Taylor (1995) has the form:

$$s(t) = m(t) + v(t) + A\left[\exp\left(\theta(m+v)\right) - \exp\left(-\theta(m+v)\right)\right], \tag{22.58}$$

where $\theta = \sqrt{2/(\alpha\sigma^2)}$. A is uniquely determined by the "smooth-pasting" conditions (see, e.g., Taylor, 1995). The formal solution to the basic (symmetric) target zone model is an S-shaped function. The S shape illustrates the "honeymoon effect" and the "smooth-pasting conditions." If s is close to s^u then the probability that the exchange rate will fall is higher than that it will rise as the authorities intervene to stop the exchange rate breaching the upper band. Consequently, the exchange rate will be lower than if the exchange rate was freely floating. Similar considerations apply near the lower bound. This behavior implies that variation in the exchange rate will be smaller, for any variation in the fundamental, than under a freely floating regime. This is called the "honeymoon effect." The so-called "smooth-pasting" conditions ensure the absence of riskless speculative gains.

There have been numerous empirical tests of the target zone model with various refinements to the basic model.[53] We will consider the models of Iannizzotto and Taylor (1999), Taylor and Iannizzotto (2001) and Lundbergh and Teräsvirta (2006), which are empirical tests assuming that the zone is credible. Suitable data points are therefore chosen by them for the analysis given this assumption.

22.4.1 Method of simulated moments (MSM)

Iannizzotto and Taylor (1999) and Taylor and Iannizzotto (2001) employ the MSM. This method is based on work by Lee and Ingram (1991) and Duffie and Singleton (1993). The essential idea is to simulate data from the chosen target zone model for a range of parameter values and compare the statistical moments of the simulated data with the statistical moments of the real data. A loss function which penalizes the deviation between the actual and simulated moments is minimized over the

various parameter values. The simulated moments estimator is found by minimizing the following loss function given weak regularity conditions and a symmetric weighting matrix W_z:

$$L = \{H_z(k) - H_N[\gamma(\beta)]\}' W_z\{H_z(k) - H_N[\gamma(\beta)]\}, \tag{22.59}$$

where $H_z(k) = (1/Z)\sum_{z=1}^{Z} h(k_z)$, $H_n[\gamma(\beta)] = (1/N)\sum_{j=1}^{N} h[y_j(\beta)]$ are the sample moments of, respectively, the observed data for Z observations $(k_z, z = 1, 2, \ldots, Z)$; and the simulated data for N observations of the Krugman model conditional on a vector of parameters $\beta, (y_j(\beta), j = 1, 2, \ldots, N)$. Taylor and Iannizzotto note that Hansen (1982) shows that:

$$W_z^* = \left(1 + \frac{1}{n}\right)^{-1} \Omega^{-1},$$

is an optimal choice for the weighting matrix in that it yields the smallest asymptotic covariance matrix for the estimator. $\Omega = \sum_{i=-\infty}^{\infty} R_x(i)$, where $R_x(i)$ is the ith autocovariance matrix of the population moments of the observed process, and $n = N/Z$ is the ratio of the length of the simulated series to the length of the observed series. Given W_z^*, the MSM estimator converges in distribution to the normal:

$$\sqrt{Z}(\widehat{\beta}_{zN} - \beta_0) \xrightarrow{D} N\left(0, \left[B'\left(1 + \frac{1}{n}\right)^{-1} \Omega^{-1} B\right]^{-1}\right) \text{ as } Z, N \to \infty,$$

where $B \equiv E[\partial h(y_j(\beta))/\partial\beta]$. The moment restrictions are tested by Taylor and Iannizzotto by exploiting a result of Hansen (1982) that the minimized value of the loss function converges asymptotically to a χ^2 distribution given the null hypothesis of no errors in specification:

$$Z\{H_z(k) - H_N[\gamma(\widehat{\beta}_{zN})]\}' W_z^*\{H_z(k) - H_N[\gamma(\widehat{\beta}_{zN})]\} \xrightarrow{D} \chi^2(\varpi - k),$$

where ϖ is the number of moment conditions and k the number of parameters being estimated. Taylor and Iannizzotto's papers improve on previous literature in a number of respects. First, they employ daily data, making appropriate allowance for holidays and weekends, so that the frequency is more consistent with the underlying theoretical model. Second, they use data only from periods when the target zone was *a priori* credible – again trying to map the data to the underlying theoretical assumptions of the model. They report statistically significant parameters of plausible magnitudes and the inability to reject the model using specification tests. However, the degree of nonlinearity implied by their parameter estimates is very small, so that the estimated honeymoon effect is small. Another interesting finding, though perhaps not too surprising given their results on the magnitude of the honeymoon effect, is that standard unit root tests have low power against data generated from a credible target zone. They found, in fact, that the chances of non-rejection of the unit root hypothesis may exceed 90% even when the exchange rate conforms to a fully credible Krugman target zone.[54]

22.4.2 Smooth transition autoregressive target zone

Lundbergh and Teräsvirta (2006) propose a flexible parametric target zone model that nests the model of Krugman but also allows estimation of an implicit target zone if it exists. A feature of their model is that it allows joint modeling of both the conditional mean and the conditional variance. Their model builds on an earlier contribution of Bekaert and Gray (1998). They call their model the Smooth Transition Autoregressive Target Zone (STARTZ) model. The STARTZ model is a parameterization of the first and second moments of s_t, the deviation of the exchange rate from the central parity. The STARTZ model is given by the following equation:

$$s_t = \lambda_t + \epsilon_t, \tag{22.60}$$

with $\epsilon_t = \sqrt{z_t h_t}$, where $\{z_t\} \sim$ i.i.d.$(0,1)$ and h_t is the conditional variance of ϵ_t. The conditional mean λ_t is defined by:

$$\lambda_t = \varphi' x_t + (\mu s^l - \varphi' x_t) G^l(s_{t-1}; \gamma_a, \theta_a, \mu s^l) + (\mu s^u - \varphi' x_t) G^u(s_{t-1}; \gamma_a, \theta_a, \mu s^u),$$

where $x_t = (1, s_{t-1}, \ldots, s_{t-n})'$ and $\varphi = (\varphi_0, \varphi_{1,\ldots,} \varphi_n)'$ is the corresponding parameter vector. The two transition functions G^l and G^u have the form:

$$G^l(s_{t-1}; \gamma, \theta, c) = \left[1 + \exp\left(-\gamma(c - s_{t-1})\right)\right]^{-\theta},\ \gamma > 0, \theta > 0,$$
$$G^u(s_{t-1}; \gamma, \theta, c) = \left[1 + \exp\left(-\gamma(s_{t-1} - c)\right)\right]^{-\theta},\ \gamma > 0, \theta > 0,$$

where s_{t-1} is the transition variable, and γ, c and θ are the slope, location and asymmetry parameters, respectively. The lower and upper bounds of the zone are defined by s^l and s^u, so that $c = \mu s^l$ and $c = \mu s^u$ are location parameters.[55]

The target zone literature implies that the conditional variance of the exchange rate must be very small near the edges of credible, implicit or explicit, bands. Lundbergh and Teräsvirta (2006) parameterize this feature by assuming that the conditional variance has a parametric specification similar to the conditional mean and given by:

$$h_t = \eta' w_t + (\delta - \eta' w_t) G^l(s_{t-1}; \gamma_b, \theta_b, \mu s^l) + (\delta - \eta' w_t) G^u(s_{t-1}; \gamma_b, \theta_b, \mu s^u),$$

where $\eta = (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)'$, $w_t = (1, \varepsilon^2_{t-1}, \ldots, \varepsilon^2_{t-q}, h_{t-1}, \ldots, h_{t-p})'$. It is assumed that $\delta > 0$ which, together with the restrictions $\alpha_0 > 0, \alpha_i \geq 0$, $i = 1, \ldots, q$, $\beta_j \geq 0$, $j = 1, \ldots, p$, is sufficient to ensure the conditional variance is positive. Because $\epsilon_t = s_t - \lambda_t$, such that φ is assumed not to depend on η, the conditional variance is a nonlinear function of the elements of w_t.[56]

Lundbergh and Teräsvirta obtain their parameter estimates by maximizing the log-likelihood under the assumption that $\{z_t\}$ is a sequence of independent standard normal errors. Under this assumption the (quasi) log-likelihood function equals:

$$l_t = const - 0.5 \ln h_t - 0.5 \frac{\epsilon_t^2}{h_t}.$$

They set out a battery of diagnostic tests for the model, some of which we discussed in section 22.2. The empirical results reported from employing daily data for the Swedish and Norwegian krone suggest the STARTZ model provides a parsimonious representation of the behavior of the Swedish krone between 1985 and 1991, and for the Norwegian krone between 1989 and 1990. The estimates accord with the theoretical models. We suggest that it would be useful to apply the STARTZ model to the daily datasets examined by Taylor and Iannizzotto. It would also be interesting to find out the properties of standard unit root or fractional tests from a simulated STARTZ model, where one expects that such tests will exhibit low power.

22.5 Speculative bubbles

22.5.1 Theory

Consider the discrete time stochastic differential equation that occurs in asset market exchange rate models (see Engel and West, 2005):

$$s_t = (1-b)a_1' x_t + b a_2' x_t + b E_t s_{t+1}, \quad b \in (0,1). \tag{22.61}$$

The above equation states that the exchange rate depends upon the current level of economic fundamentals x_t plus the discounted expected spot rate next period, where b is the discount factor. In the absence of rational bubbles, the forward solution to the above equation is:

$$s_t = (1-b)E_t\left(\sum_{j=0}^{\infty} b^j a_1 x_{t+j}\right) + bE_t\left(\sum_{j=0}^{\infty} b^j a_2 x_{t+j}\right). \tag{22.62}$$

The logarithm of the exchange rate can be written as the discounted sum of current and expected future fundamentals, such as interest rates, prices, money supplies and income. This is a general form of several exchange rate determination models based on macroeconomic fundamentals that can provide several insights concerning the empirical findings of studies on exchange rate forecasting discussed in section 22.6. By assuming Cagan-style money demand functions for the home and foreign countries with common parameters, we obtain:

$$s_t = m_t - m_t^* + y_t - \gamma(w_t - w_t^*) + \lambda(i_t - i_t^*), \tag{22.63}$$

where m_t is the log of the money supply, w_t is the log of real income, i_t is the short-term interest rate, γ and λ denote the income elasticity and interest rate semi-elasticity of money demand, and y_t is the real exchange rate. An asterisk denotes foreign quantities. The deviations from UIP, $rp_t = i_t - i_t^* - E_t\left(\Delta s_{t+1}\right)$, can be thought of as unobserved fundamentals. Substituting into (22.63) yields:

$$s_t = m_t - m_t^* + y_t - \gamma(w_t - w_t^*) + \lambda rp_t + \lambda E_t\left(\Delta s_{t+1}\right). \tag{22.64}$$

By iterating forward we obtain:

$$s_t = \frac{1}{1+\lambda} E_t \left(\sum_{j=0}^{\infty} \left(\frac{\lambda}{1+\lambda} \right)^j (m_{t+j} - m^*_{t+j} + y_{t+j} - \gamma(w_{t+j} - w^*_{t+j})) \right) + \frac{\lambda}{1+\lambda} E_t \left(\sum_{j=0}^{\infty} \left(\frac{\lambda}{1+\lambda} \right)^j rp_{t+j} \right). \tag{22.65}$$

In this case, $b = \lambda/(1+\lambda)$ and $a_2' x_t = rp_t$. If we assume, further, that PPP and UIP hold, then the exchange rate model simplifies to:

$$s_t = \frac{1}{1+\lambda} E_t \left(\sum_{j=0}^{\infty} \left(\frac{\lambda}{1+\lambda} \right)^j f_{t+j} \right), \tag{22.66}$$

where $f_t = a_1' x_t = \left(m_t - m^*_t\right) - \left(w_t - w^*_t\right)$.[57] However, it is well known that the above equation is a single solution from a potentially infinite set. Letting $s_{f,t}$ denote the fundamental solution, the rational expectations solutions to (22.64) are given by:

$$s_t = s_{f,t} + B_t, \tag{22.67}$$

where:

$$E_t B_{t+1} = \frac{(1+b)}{b} B_t. \tag{22.68}$$

The term B_t is the speculative bubble, which has to follow the form given by (22.68).[58] We can also write the solution for the bubble as (Salge, 1997):

$$B_t = \frac{M_t}{\lambda^t}. \tag{22.69}$$

From (22.69) and (22.68):

$$E_t B_{t+1} = \frac{E_t M_{t+1}}{\lambda^{t+1}} = \frac{1}{\lambda} \frac{M_t}{\lambda^t}, \tag{22.70}$$

so that:

$$E_t M_{t+1} = M_t, \tag{22.71}$$

implying that M_t is a martingale process. A more general form of a stochastic martingale process is given by:

$$M_t = \rho_t M_{t-1} + u_t, \tag{22.72}$$

where $E_t \rho_{t+1} = 1$, $E_{t-j} u_t = 0$, $j = 1, \ldots, n$, $E_t \rho_t M_t = 0$, and $E_t u_t M_t = 0$. In general, a bubble can depend on its lagged values, called a Markovian bubble, on an extraneous process "M," which follows a martingale process, or on fundamentals, called an intrinsic bubble (Froot and Obstfeld, 1991).[59]

Bubbles, if they exist, can of course pop and a variety of forms of rational bubbles that exhibit this property and are consistent with (22.68) have been proposed (see,

e.g., Blanchard and Watson, 1982; Evans, 1991). The bubble proposed by Evans takes the form:

$$B_{t+1} = \begin{cases} \lambda^{-1} B_t \epsilon_{t+1} & \text{if } B_t \leq k, \\ \left[\delta + \pi^{-1}\lambda^{-1}\phi_{t+1}\left(B_t - \lambda\delta\right)\right]\epsilon_{t+1} & \text{if } B_t > k, \end{cases} \tag{22.73}$$

where $E_t\epsilon_{t+1} = 1$, and ϕ_{t+1} takes the value one or zero with probabilities π or $1-\pi$. Taking expectations of the second equation in (22.73) we have that:

$$E_t B_{t+1} = \pi\left[\delta + \pi^{-1}\lambda^{-1}1\left(B_t - \lambda\delta\right)\right] + (1-\pi)\,\delta = \frac{B_t}{\lambda}. \tag{22.74}$$

In this case, when the Evans Markovian bubble exceeds the value of k then it grows at a faster rate till it pops to a value of δ. In expectation the bubble is explosive, which implies that s_t in (22.67) is explosive regardless of the integration properties of the fundamental. There appears to be no reason why an intrinsic bubble that pops could also not be specified – though as yet no research has investigated such bubbles.

22.5.2 Testing and evidence

Though the majority of applications are based on stock price data, Evans (1986) and Meese (1986) are two applications to exchange rates.[60] The empirical tests to date on exchange rates are inconclusive as to the existence of bubbles. However, in an important new development in testing for bubbles, Phillips *et al.* (2006) set out a unit root testing procedure that has good power characteristics in finite samples and enables dating the origination and the collapse of bubbles.[61] Their study is based on the presumption that bubbles can be identified by way of manifestation of explosive characteristics in the data. This can be achieved by estimating the regression equation:

$$s_t = \mu + \phi s_{t-1} + \sum_{j=1}^{J} \xi_j \Delta s_{t-j} + \varepsilon_{s,t}, \quad \varepsilon_{s,t} \sim NID(0, \sigma_s^2), \tag{22.75}$$

and testing the null hypothesis of a unit root, $H_0 : \phi = 1$ against the alternative $H_1 : \phi > 1$. Phillips *et al.* (2006) propose two tests, a right-side ADF test and a sup test, based on the recursive estimation of (22.75). Recursive estimation is implemented by fitting (22.75) for a fraction of the sample, say r_0, and sequentially increasing this fraction by including successive observations. Under the null the corresponding test statistics, denoted by ADF_r and $\sup_{r\in[r_0,1]} ADF_r$, are:

$$ADF_r \Rightarrow \frac{\int_0^r W dW}{\int_0^r W^2}, \tag{22.76}$$

$$\sup_{r\in[r_0,1]} ADF_r \Rightarrow \sup_{r\in[r_0,1]} \frac{\int_0^r W dW}{\int_0^r W^2}, \tag{22.77}$$

where W denotes Brownian motion and $r \in [r_0, 1]$. If the null hypothesis is rejected then confidence intervals for the parameter ϕ can be constructed on the basis of the work of Phillips and Magdalinos (2007) regarding the asymptotic distribution theory for mildly explosive processes.

Application of these methods to exchange rates and their fundamental determinants as well as forward premia would seem to be of interest. It might also be interesting to examine the power of their tests against certain nonlinear processes, given that the forward premium should embody any rational bubble and that it has been modeled by such processes as discussed in section 22.3.

Another important property of Markovian bubbles was proved initially by Lux and Sornette (2002). The standard empirical finding is that the distribution of asset returns belongs to the class of so-called fat-tailed distributions with hyperbolic decline of probability mass in the tails. They derived the implications of rational bubbles of the Blanchard–Watson type for the unconditional distribution of prices, price changes and returns. They proved that the Blanchard–Watson (1982) bubble exhibited a tail index of less than unity (see, e.g., Koedijk *et al.*, 1990; Loretan and Phillips, 1994; Huisman *et al.*, 2001; Wagner and Marsh, 2005).[62] Yoon (2005) proved the same result for the Evans (1991) bubble and this property was transferred to asset returns. In fact, the empirical results for exchange rates typically generate tail estimates of around 2–6, suggesting the absence of bubbles. However, the results of Phillips *et al.* (2006) are relevant here. For instance, the standard ADF test suggested the absence of bubbles when applied to the full sample of Nasdaq price data that they considered – February 1973 to June 2005. However, their new test procedure detects the presence of a bubble in June 1995 continuing until July 2001.

This suggests that one could, in principle, employ the Phillips procedure to indicate the potential presence of bubbles and use the indicated bubble samples to obtain tail estimates. It would also seem of interest to examine the properties of the cointegrating residuals using the Phillips *et al.* (2006) tests as well as their tail properties, perhaps estimating the cointegrating vector by the Phillips *et al.* (1996) FM-LAD estimator, the properties of which seem ideal in this context.

22.6 Exchange rates, economic fundamentals and forecasting

In their landmark paper, Meese and Rogoff (1983a) employed rolling regressions in order to generate forecasts for the level of the spot rate based on a comprehensive range of exchange rate determination models: the flexible-price monetary model (Frenkel–Bilson), the sticky-price monetary model (Dornbusch–Frankel), and the sticky-price asset model (Hooper–Morton). At the time there was widespread optimism about the potential of monetary models to explain the fluctuations of floating exchange rates.

A general "prediction" equation that nests all models under examination is:

$$s_{t+k} = a_{0,k} + a_{1,k}(m_t - m_t^*) + a_{2,k}(w_t - w_t^*) + a_{3,k}(i_t - i_t^*) + a_{4,k}(\pi_t^e - \pi_t^{e*}) + a_{5,k}tb_t + a_{6,k}tb_t^* + u_t, \quad (22.78)$$

where s_{t+k} denotes the log exchange rate (domestic price of foreign currency), m_t is the log of the money supply, w_t is the log of real income, i_t is the short-term interest rate, π_t^e denotes expected inflation, tb_t is the cumulated trade balance, u_t is a possibly serially correlated error term, and the a_ks are parameters corresponding to the kth forecast horizon, with $k = 1, 3, 6, 12$ months. An asterisk denotes foreign quantities. Meese and Rogoff (1983a) estimated the above regression by OLS, as well as GLS and instrumental variables (IVs) (Fair, 1970) so as to deal with the presence of serial correlation in the residuals and simultaneous equation bias due to the endogeneity of the variables. Surprisingly, on the basis of root mean squared error (RMSE), none of these models outperformed the naive RW model for horizons up to a year, even though realized values of the forcing variables were used.

The poor forecasting performance of the considered models can be attributed to a number of factors. First, the fact that the variables are highly persistent may lead to biased estimates of the coefficients (Rossi, 2005). If the error term is non-stationary (i.e., the exchange rate is not cointegrated with fundamentals), the coefficients will be inconsistent and forecasting will be meaningless. Second, equation dynamics (MacDonald and Taylor, 1994) are omitted from the regression equation and the case of a nonlinear DGP is not considered (see Taylor and Peel, 2000; Meese and Rose, 1990). Further, parameter instability may characterize empirical exchange rate models (Wolff, 1987; Rossi, 2006) and simultaneous equation bias may not have been properly accounted for. As far as the latter explanation is concerned, Meese and Rogoff (1983b) impose coefficient restrictions based on the theoretical and empirical literature on money demand and the rate at which shocks to the real exchange rate appear to damp out. This enables the examination of longer forecasting periods. Their findings indicate that, for horizons greater than a year, there are cases where the RMSE of the RW is larger than that of structural models, suggesting the presence of simultaneous equation bias and that the performance of structural models may improve at longer horizons (Meese and Rogoff, 1983b; Rogoff, 1999). These findings support the estimation of long-horizon regressions and the application of advanced inference procedures (see, e.g., Mark, 1995; Chinn and Meese, 1995).

22.6.1 Long-horizon regressions

Consider the monetary model (22.66) and let the fundamental term, f_t, follow a driftless RW. Then equation (22.66) reduces to $s_t = f_t$. The relationship between the exchange rate and the monetary fundamentals motivates Mark (1995) to examine whether current deviations of the exchange rate from its fundamental value, $z_t \equiv f_t - s_t$, contain predictive power for future movements in the exchange rate, as well as the horizon at which the predictive power of the deviations becomes apparent. Obviously, this contrasts with the view that exchange rates are best characterized

by a no-change model:

$$s_{t+k} - s_t = \varepsilon_{t+k}, \quad k = 1, 4, 8, 12, 16 \text{ quarters}, \tag{22.79}$$

and motivates the use of long-horizon regressions:

$$s_{t+k} - s_t = a_k + \beta_k z_t + \varepsilon_{t+k}, \quad k = 1, 4, 8, 12, 16 \text{ quarters}. \tag{22.80}$$

The exchange rate is expected to rise (fall) when it is below (above) its fundamental value. Thus the slope coefficient β_k should be positive and statistically significant. Formally, the null hypothesis to be tested is $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k > 0$, or the joint hypothesis $H_0 : \beta_k = 0 \, \forall k$ versus $H_1 : \beta_k > 0$ for some k. The evaluation of the out-of-sample performance of the model involves generating recursive or rolling forecasts based on equations (22.66) and (22.79) and estimating a forecast evaluation statistic such as Theil's U-statistic, or the DM-statistic of Diebold and Mariano (1995).

In evaluating the statistical significance of the results Mark confronted a number of econometric problems. First, a highly persistent explanatory variable implies biased OLS estimates of the slope coefficients in finite samples (see Neely and Sarno, 2002, and the references therein). Second, the fact that the k-period change in s_t is used as the regressand induces serial correlation in the disturbances of order $(k - 1)$, for $k > 1$. In order to correct for serial correlation Mark used the Newey–West covariance matrix estimator based on either a fixed truncation lag of 20 or a truncation lag specified by Andrews' (1991) procedure. Finally, although the DM-statistic follows the standard normal distribution for non-nested models, long-horizon regressions nest the RW model and, therefore, the distribution of the DM-statistic is not known in general (McCracken, 1999). To this end, Mark proposed a bootstrap procedure:

1. Estimate the following vector autoregression (VAR), where the null of no predictability has been imposed:

$$\Delta s_t = a_0 + u_{s,t}, \tag{22.81}$$

$$z_t = \mu + \sum_{j=1}^{p} b_j z_{t-j} + u_{z,t}. \tag{22.82}$$

2. Use the estimates of the fitted model and draw from the bivariate normal distribution with mean 0 and covariance matrix equal to the covariance matrix of the estimated residuals, in order to recursively generate pseudo observations for Δs_t and Δz_t. Alternatively, if the error term is not normal, resample from the observed residuals.
3. In turn, estimate the long-horizon regression so as to obtain the slope coefficient, the t-statistics and R^2 for the simulated series and generate forecasts based on the monetary and the driftless RW model and compute Theil's U-statistic and the DM-statistic.

4. Repeat steps 2 and 3 2,000 times so as to obtain bootstrap distributions and the corresponding *p*-values.

Mark (1995) examines quarterly US dollar exchange rates for the Canadian dollar, Deutsche Mark, Japanese yen and the Swiss franc and the corresponding fundamentals from 1973:2 to 1991:4. The last 40 quarters are used for the out-of-sample forecasting exercise. His findings indicate that the hypothesis of no in-sample predictability, $H_0 : \beta_k = 0 \ \forall k$, can be rejected at the 5% significance level for Switzerland and Germany. The slope coefficients, the R^2s and the test statistics tend to increase with the forecast horizon for all countries. Furthermore, in several cases the monetary model forecasts are superior to the RW model, especially for long horizons. Mark (1995) concludes that, given the small size of the dataset, these results support the view that exchange rates are predictable.

22.6.1.1 Turning on the microscope

A number of subsequent studies questioned the validity of Mark's methodology and the resultant conclusions (see Neely and Sarno, 2002). An important assumption in Mark's study was that the process of the deviations of the exchange rate from its fundamental value is stationary. By the Granger representation theorem the exchange rate and the fundamentals must posses a vector error correction model (VECM) representation with cointegrating vector $(1, -1)$. The VECM model is expressed:

$$\Delta s_t = \mu_s + \lambda_s z_{t-1} + \sum_{i=1}^{p-1} \phi_i^s \Delta s_{t-i} + \sum_{i=1}^{p-1} \psi_i^s \Delta f_{t-i} + u_{s,t}, \tag{22.83}$$

$$\Delta f_t = \mu_f + \lambda_f z_{t-1} + \sum_{i=1}^{p-1} \phi_i^f \Delta s_{t-i} + \sum_{i=1}^{p-1} \psi_i^f \Delta f_{t-i} + u_{f,t}, \tag{22.84}$$

where λ_s and λ_f determine the speed of adjustment, μ_s and μ_f are intercepts, ϕ_i and ψ_i are parameters, and the disturbance terms $u_{s,t}$ and $u_{f,t}$ are i.i.d. Stationarity requires one of the speed of adjustment terms to be different from zero and $\lambda_f - \lambda_s < 0$.

Berkowitz and Giorgianni (2001) use the VECM representation and derive the following expression for the slope coefficient in the long-horizon regression (22.80):

$$\beta_k = \beta_1 \frac{1 - \theta^k}{1 - \theta}, \tag{22.85}$$

where $\theta = 1 + \lambda_f - \lambda_s < 0$ and $\beta_1 = \lambda_s$. The above equation provides several insights concerning long-horizon regressions under the assumption of a linear DGP. The rejection of the null hypothesis that the slope coefficients in the long-horizon regressions are zero implies that the exchange rate is weakly exogenous. In this case, although the fundamentals do not contain predictive power, the existence of a long-run relationship between fundamentals and the exchange rate is not ruled out. It follows that, in the context of a linear model, the fundamentals cannot

contain predictive power at long horizons without the presence of short-run predictability, given that $\beta_1 = 0$ implies $\beta_k = 0 \forall k$. Thus, either Mark's results are spurious, or the DGP is misspecified.[63]

Spurious inference may arise due to the fact that Mark's procedure is based on the assumption that the deviation process is stationary. Conditioning on cointegration when cointegration fails may result in false inferences. The LS estimates of the long-horizon regression will be inconsistent without having any economic interpretation and conventional statistical tests will not be valid, especially for long-horizons (see Berben and van Dijk, 1998). Let s_t follow an RW: it follows that, for large k, the dependent variable $s_{t+k} - s_t$ will also approximate an RW. Given that $z_t \sim I(1)$, the long-horizon regression involves regressing an $I(1)$ variable on another $I(1)$ variable. This gives rise to a near-spurious regression problem (Granger and Newbold, 1974).[64] To this end, Berkowitz and Giorgianni (2001) apply the Horvath–Watson test (Horvath and Watson, 1995) without being able to reject the null of no cointegration for all countries except Switzerland. It is noted, however, that if the deviation process is, in fact, nonlinear then linear cointegration techniques may suffer from low power (Paya and Peel, 2007a).

Kilian (1999) emphasizes that the stability of the bootstrap DGP is not ensured in the procedure proposed by Mark. To this end, he suggests feasible generalized least squares (FGLS) estimation subject to a stability constraint. Furthermore, he notes that Mark's approach is inconsistent and may result in spurious inference due to the fact that a drift is included in the bootstrap exchange rate series but forecasts are based on the no-change model. Thus, the superior forecast performance of the long-horizon regression may be due to either the contribution of the fundamentals or just the inclusion of the drift term. The VECM bootstrap proposed consists of the following steps:

1. Define the VECM model as the DGP:

$$\Delta s_t = \mu_s + \lambda_s z_{t-1} + \sum_{i=1}^{p-1} \phi_i^s \Delta s_{t-i} + \sum_{i=1}^{p-1} \psi_i^s \Delta f_{t-i} + u_{s,t},$$

$$\Delta f_t = \mu_f + \lambda_f z_{t-1} + \sum_{i=1}^{p-1} \phi_i^f \Delta s_{t-i} + \sum_{i=1}^{p-1} \psi_i^f \Delta f_{t-i} + u_{f,t}.$$

2. Impose the null that all the coefficients in the first equation except for the intercept are zero and specify the lag order by using an information criterion such as the AIC. Estimate the model by FGLS subject to the stability constraint.
3. Use the coefficient estimates and draw with replacement from the observed recentered residuals to recursively generate pseudo-observations for Δs_t and Δf_t.
4. Estimate the long-horizon regressions for the pseudo-series and construct the test statistics under examination.
5. Repeat steps 3 and 4 2,000 times so as to obtain bootstrap distributions of the test statistics and the corresponding p-values.

Kilian extends the sample period from 1991:4 to 1997:4 and applies both Mark's bootstrap methodology and the VECM bootstrap. The former methodology indicates that for the extended data set the *p*-values of the various statistics are stable or increasing and there is overall predictability only for Switzerland.[65] Kilian also shows that the effect of small sample bias, together with the fact that Mark's bootstrap is inconsistent, has a substantial impact on inference. When allowing for a drift in the forecasts of the RW, both bootstrap methodologies detect overall predictability for Switzerland and Canada.

22.6.1.2 Forecast evaluation measures

In the late 1990s, the scientific consensus was that the Meese and Rogoff (1983a, 1983b) results still stood (Rogoff, 1999). However, the findings of Clark and McCracken (2001, 2003, 2005) raise concerns regarding the power of commonly used *t*-type tests. This motivates McCracken and Sapp (2005) to investigate the out-of-sample performance of exchange rate determination models using new test statistics regarding the comparison of nested models.[66]

Let $y_{t+k} = s_{t+k} - s_t$ denote the variable to be predicted and $x_{2,t} = (x'_{1,t}, x'_{22,t})'$ be a vector of predictors. The number of in-sample and out-of-sample observations is R and P, respectively, so that $T = R + P$. Forecasts are generated by estimating two linear models of the form $x_{1,t}\beta_1$ and $x'_{22,t}\beta_2$ recursively by OLS. The forecast errors are $\hat{u}_{1,t+k} = y_{t+k} - x'_{1,t}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+k} = y_{t+k} - x'_{22,t}\hat{\beta}_{2,t}$. Under the null hypothesis the first model is nested within the second and the two forecast errors are identical. In the present context, the first model is the RW with a drift and the second the long-horizon regression model, which implies that $x_{1,t} = 1$ and $x_{22,t} = z_t$.

The first two test statistics examined are used to test for forecast accuracy. The former is a *t*-type test and the latter is its *F*-type counterpart. The null hypothesis is that the two mean squared errors (MSEs) are equal against the alternative that the MSE for the second model is smaller. Let $\widehat{d}_{t+k} = \widehat{u}^2_{1,t+k} - \widehat{u}^2_{2,t+k}$, $\bar{d} = (P - k + 1)^{-1} \sum_{t=R}^{T-k} \widehat{d}_{t+k} = \text{MSE}_1 - \text{MSE}_2$, $\widehat{\Gamma}_{dd}(j) = (P - k + 1)^{-1} \sum_{t=R+j}^{T-k} \widehat{d}_{t+k}\widehat{d}_{t+k-j}$ for $j \geqslant 0$ and $\widehat{\Gamma}_{dd}(j) = \widehat{\Gamma}_{dd}(-j)$, and let $\widehat{S}_{dd} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\widehat{\Gamma}_{dd}(j)$ be the long-run covariance of d_{t+k} estimated using a kernel-based estimator with function $K(\cdot)$, bandwith parameter M and maximum number of lags $\bar{j}$. The tests for forecast accuracy are:

$$\text{MSE} - t = (P - k + 1)^{1/2} \frac{\bar{d}}{\widehat{S}_{dd}^{1/2}}, \tag{22.86}$$

$$\text{MSE} - F = (P - k + 1)^{1/2} \frac{\bar{d}}{\text{MSE}_2}. \tag{22.87}$$

The next two tests concern forecast encompassing and are built upon the statistic used by Harvey *et al.* (1998) for non-nested models. In this case, the null hypothesis is that the forecast of the RW model encompasses that of the structural exchange rate model and, therefore, the covariance between the forecast

errors of the RW and the difference of the forecasts errors of the two models will be equal to or smaller than zero. Under the alternative, the deviations from fundamentals contain valuable information, implying that the covariance is positive. Let $\widehat{c}_{t+k} = \widehat{u}_{1,t+k}(\widehat{u}_{1,t+k} - \widehat{u}_{2,t+k})$. The forecast-encompassing test statistics are:

$$\text{ENC} - t = (P - k + 1)^{1/2} \frac{\bar{c}}{\hat{S}_{cc}^{1/2}}, \tag{22.88}$$

$$\text{ENC} - F = (P - k + 1)^{1/2} \frac{\bar{c}}{MSE_2}. \tag{22.89}$$

The limiting distributions of the test statistics considered so far are non-standard and depend upon nuisance parameters. Thus, critical values for the test statistics should be generated by bootstrap procedures, such as the VECM bootstrap of Kilian (1999).

The final statistic, derived by Chao *et al.* (2001), is also used to test for forecast encompassing and follows a χ^2 distribution. Let $\widehat{h}_{t+k} = \widehat{u}_{1,t+k}x_{1,t}$, $\widehat{b}_{t+k} = \widehat{u}_{1,t+k} + x_{22,t}$, $\bar{b} = (P - k + 1)^{-1} \sum_{t=R}^{T-k} \widehat{b}_{t+k}$, $\widehat{F} = -(P - k + 1)^{-1} \sum_{t=R}^{T-k} x_{22,t}x'_{1,t}$, and $\widehat{B} = [(P - k + 1)^{-1} \sum_{t=R}^{T-k} x_{1,t}x'_{1,t}]^{-1}$. Furthermore, let $\hat{\Gamma}_{bb}(j) = (P - k + 1)^{-1} \sum_{t=R+j}^{T-k} \hat{b}_{t+k}\hat{b}'_{t+k-j}$, $\hat{\Gamma}_{hh}(j) = (P - k + 1)^{-1} \sum_{t=R+j}^{T-k} \hat{h}_{t+k}\hat{h}'_{t+k-j}$, $\hat{\Gamma}_{bh}(j) = (P - k + 1)^{-1} \sum_{t=R+j}^{T-k} \hat{b}_{t+k}\hat{h}'_{t+k-j}$, for $j \geqslant 0$, with $\hat{\Gamma}_{bb}(j) = \hat{\Gamma}_{bb}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}_{hh}(-j)$, $\hat{\Gamma}_{bh}(j) = \hat{\Gamma}_{bh}(-j)$. Finally, let $\hat{S}_{bb} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{bb}(j)$, $\hat{S}_{hh} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{hh}(j)$, $\hat{S}_{bh} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{bh}(j)$. The test statistic is written as:

$$\text{CCS} = (P - k + 1)\bar{b}'\Omega\bar{b}, \tag{22.90}$$

where $\Omega = \hat{S}_b b + \hat{\lambda}_{bh}(\hat{F}\hat{B}\hat{S}'_{bh} + \hat{S}'_{bh}\hat{B}'\hat{F}') + \hat{\lambda}_{bb}\hat{F}\hat{B}\hat{S}'_{hh}\hat{B}'\hat{F}'$, $\hat{\pi} = (P - k + 1)R^{-1}$, $\hat{\lambda}_{bh} = 1 - \hat{\pi}^{-1}ln(1 + \hat{\pi})$, $\hat{\lambda}_{bh} = 2[1 - \hat{\pi}^{-1}ln(1 + \hat{\pi})]$. The null hypothesis of no predictability based on macroeconomic fundamentals requires the covariance between $u_{1,t+k}$ and $x_{22,t}$ to be zero. If fundamentals contain predictive power then the covariance should deviate from zero.

McCracken and Sapp (2005) employ the long-horizon regression (22.80), but follow Meese and Rogoff (1983a, 1983b) and determine the fundamental value of the exchange rate according to various structural models. This approach results in a vast number of tests, which raises concerns about the reliability of inference.[67] In order to mitigate the multiple testing problem, McCracken and Sapp (2005) follow recent developments in the statistical genetics literature and calculate *q*-values along with the *p*-values.[68] Using both *p*-values and *q*-values, they find evidence of predictability for many cases. The encouraging results can be attributed to the fact that the new *F*-type tests are more powerful than the *t*-type tests. Despite the fact that RMSEs are similar to those reported by Kilian (1999), the *F*-type tests are able to detect the superiority of the structural models over the RW with a drift. As far as the monetary model is concerned, *F*-type tests of equal forecast accuracy indicate

more short horizon predictability for Germany and more long-horizon predictability for Canada and Switzerland. Moreover, tests of forecast encompassing appear to be superior in detecting predictability of exchange rates compared to tests of forecast accuracy.

22.6.2 Nonlinear models

Taylor and Kilian (2003) investigate the forecasting performance of long-horizon regressions in the presence of ESTAR dynamics in the real exchange rate. They employ the long-horizon regression of Mark (1995) and draw inferences from bootstrap distributions of the test statistics generated under the null hypothesis that the nominal exchange rate s_t follows an RW and the real exchange rate z_t is an ESTAR process:[69]

$$s_t = \mu_s + u_{s,t}, \tag{22.91}$$

$$z_t = \mu_z + \left[\phi_1(z_{t-1} - \mu_z) + (1 - \phi_1)(z_{t-2} - \mu_z)\right] \times \exp\left(-\gamma \sum_{d=1}^{5} (z_{t-d} - \mu_z)^2\right) + u_{z,t}. \tag{22.92}$$

Taylor and Kilian (2003) use quarterly data for seven OECD countries for the period 1973:1 to 1998:4. Long-horizon regressions appear to be significantly more accurate than the naive RW model in several cases, especially when the Newey–West standard errors are used with the truncation lag specified by Andrews' procedure. However, the out-of-sample results are not as encouraging.[70] The *DM*-statistics indicate that the long-horizon regression model is capable of beating the RW model only for the UK and Switzerland at the three-year horizon. The authors conclude that incorporating nonlinearities increases the predictability of models based on macroeconomic fundamentals. However, it is difficult to detect the improvement in the forecast accuracy due to the small time span and the rarity of large deviations from the fundamentals.

Another group of recent studies focuses on the Markov-switching (MS) model, which allows exchange rate dynamics to alternate between regimes.[71] Clarida *et al.* (2003) argue that the forward rate has predictive content regarding the spot rate. The authors build upon the results of Clarida and Taylor (1997) and apply an MS-VECM model, which allows for shifts in the intercept and the error variance. Their findings indicate that the MS-VECM outperforms both the linear VECM and the naive RW model, especially at long horizons.

Sarno *et al.* (2004b) employ a long span of data for the US dollar exchange rate and show that fundamentals are useful in explaining the behavior of numerous exchange rates under different monetary regimes by estimating MS-VECM models. Frömmel *et al.* (2005), motivated by the market microstructure literature (Lyons, 2001) and questionnaire surveys (e.g., Cheung and Chinn, 2001) showing that market participants regard the importance of fundamentals as time-varying, establish that there are significant regime changes in real interest differential (RID) variants of the monetary model (Frankel, 1979) for three major US dollar exchange rates.[72]

The MS model employed can be written as (see Hamilton, 1994):

$$r_t = \alpha(S_t) + \beta(S_t)\text{fund}_t + \varepsilon_t, \quad S_t = 1, \ldots, M, \tag{22.93}$$

where r_t is the 12-month percentage change of the exchange rate and fund_t denotes the vector of RID fundamentals, which covers relative changes in money supply, industrial production, money market interest rates and the government bond yield. The vectors of coefficients, α and β, are governed by the unobservable state variable, S_t. In MS models the regime-generating process is an ergodic Markov chain with a finite number of states, M, defined by the transition probabilities $p_{ij} = Pr(S_{t+1} = j | S_t = i)$ with $\sum_{j=1}^{M} p_{ij} = 1 \; \forall i, j \in 1, \ldots, M$. Frömmel *et al.* (2005) set the number of states equal to two and assume that the error term, ε_t, is a white-noise process with constant variance. The estimation of the model is implemented by using an expectation maximization (EM) algorithm (Kim and Nelson, 1999).

Wald tests indicate that the null hypothesis of constant parameters can be rejected for all exchange rates. For each exchange rate the coefficients are in line with the RID model for one of the regimes. Furthermore, the MS-RID model produces a lower RMSE than the RW model. In contrast to the in-sample results, the out-of-sample performance of the MS-RID model is not as encouraging. Frömmel *et al.* (2005) use a rolling sample of ten years and calculate the conditional expectations of the percentage change of the exchange rate. Although the MS-RID model is superior to the standard RID model, it cannot beat the naive RW on the basis of the *DM*-statistic. This finding is not surprising, since non-linear models in general, and MS models in particular, may produce superior in-sample fits compared to linear models but not necessarily superior forecasts (see Dacco and Satchell, 1999).

22.6.3 Real-time forecasting and market expectations

It is common practice in studies of exchange rate forecasting to employ the most recent datasets on macroeconomic fundamentals. However, these datasets are subject to extensive revisions and are not available to real-time forecasters. Provided that market participants' expectations depend on currently available data on fundamentals, real-time data may lead to a better approximation of market expectations than revised data (Neely and Sarno, 2002). According to the present value model (22.61), exchange rates are influenced by market expectations of future fundamentals and, therefore, real time data may improve upon the predictability of exchange rates.

Faust *et al.* (2003) investigate the impact of using real time data for the currencies examined by Mark (1995). Their findings indicate that long-horizon predictability is present only in less than a two-year window around the vintage used by Mark. In order to isolate the effect of data revisions from the sample period, the authors fix the estimation and the forecast periods to be the same as Mark's and use all vintages of data from 1992 onward. Overall, the evidence of predictability weakens. As far as real-time forecasting is concerned, the out-of-sample performance of the monetary model is poor. The RMSEs are generally greater than those of the RW model and increase with the horizon. However, real-time forecasts produce significantly lower

Theil U-statistics than forecasts based on the revised data, indicating the superiority of real-time data.[73]

Engel and West (2005) provide a thorough analysis of the role of market expectations and the value of the discount factor. According to the rational expectations present value model (22.61), the importance of expected future fundamentals relative to current fundamentals increases with the discount factor, b. For large b and non-stationary fundamentals, the movement in the exchange rate at time t will be almost uncorrelated with information known at time $(t-1)$, since the exchange rate will be largely driven by the expected future path of the fundamentals. By the Engel and West (2005) theorem, if $a_1' x_t \sim I(1)$ and $a_2 = 0$ or $a_2' x_t \sim I(1)$ then, as $b \to 1$, the exchange rate exhibits near random walk behavior.[74] The theorem highlights the fact that movements in the exchange rate reflect changes in expectations. If expectations contain valuable information about future fundamentals then changes in the exchange rate should be useful in forecasting fundamentals. This provides an alternative approach concerning the evaluation of the performance of the various models under examination.

A different perspective for the implications of the Engel and West theorem is provided by Evans and Lyons (2005) in the context of market microstructure. Lyons show that micro-based models can establish a link between expectational surprises and specific types of non-public information. The key idea is that the trades of private agents reveal new information to the market makers about future fundamentals. In this framework prices are determined by the market makers' expectations, E_t^m, about the future values of fundamentals. The market makers construct their information sets and revise their forecasts on the basis of the order flow, $x_{o,t}$, that is signed transaction flow. Suppose that innovations in order flow are correlated with the innovations in fundamentals growth:

$$\Delta x_{o,t} = \lambda \Delta x_{o,t-1} + \eta_t, \tag{22.94}$$

$$\Delta f_t = \phi \Delta f_{t-1} + u_t + \delta \eta_t, \tag{22.95}$$

and that market makers observe order flow innovations and, hence, the current state of the economy, with a time delay $f_t - E_t^m f_t = \delta \eta_t$. Evans and Lyons (2005) show that, under these assumptions, the present value model (22.61) implies that changes in the exchange rate depend on lagged order flow:

$$\Delta s_{t+1} = \frac{1-b}{b}(s_t - E_t^m f_t) + \frac{1}{1-b\phi} u_{t+1} + \frac{[1+\phi(1-b)]\,\delta}{1-b\phi}(x_{o,t} - \lambda x_{o,t-1}). \tag{22.96}$$

It follows from the above equation that $b \to 1$ does not rule out forecastability. In order to test whether order flow contains predictive power for the movements in the exchange rate, they employ the following regression equations:

$$s_{t+1} - s_t = \alpha_0 + \alpha x_{o,t}^{agg} + \varepsilon_{t+1}, \tag{22.97}$$

where $x_{o,t}^{agg}$ denotes aggregate order flow, and:

$$s_{t+1} - s_t = \alpha_0 + \sum_{j=1}^{6} \alpha_j x_{o,t}^{j} + \varepsilon_{t+1}, \tag{22.98}$$

where $x^{j}_{o,t}$ denotes order flow from end-user segment j. Six end-user segments are considered: trades implemented in the US and non-US markets for non-financial firms, investors and leveraged trades. The forecasting performance of the micro-based model is compared to that of an RW and a structural model using data for the largest spot market, the US dollar/euro, for the period from 1993 to 1999. In general, the findings indicate that, for horizons between 10 and 20 days, the micro-based models are able to beat the RW benchmark and the structural model, irrespective of the type of order flow used.

Large values of the discount factor and highly persistent fundamentals provide an explanation for the failure of structural models in out-of-sample forecasting. However, forecastability of the changes in the exchange rate is still possible in the presence of stationary terms. Consider the present value model (22.61) and let $f_t = a_1' x_t$ and $a_2' x_t = rp_t$. If the deviations from the UIP, rp_t, follow a stationary AR(1) process:

$$rp_t = \theta rp_{t-1} + e_t, \tag{22.99}$$

where $e_t \sim$ i.i.d. and $\sigma_e^2 = var(e_t)$, and f_t is an RW process, with innovation η_t and $\sigma_\eta^2 = var(\eta_t)$, then the forward solution for the exchange rate is:

$$s_t = f_t + \frac{b}{1 - b\theta} rp_t. \tag{22.100}$$

The k-period change in the exchange rate is:

$$\begin{aligned} s_{t+k} - s_t &= f_{t+k} - f_t + \frac{b}{1 - b\theta}(rp_{t+k} - rp_t) \\ &= \sum_{j=1}^{k} \eta_{t+k} + \frac{b}{1 - b\theta}(\theta^k - 1) rp_t + \frac{b}{1 - b\theta} \sum_{j=1}^{k} \theta^{k-j} e_{t+k} \\ &= \sum_{j=1}^{k} \eta_{t+k} + (1 - \theta^k) z_t + \frac{b}{1 - b\theta} \sum_{j=1}^{k} \theta^{k-j} e_{t+k}, \end{aligned} \tag{22.101}$$

where z_t are the deviations from the observed fundamentals, and the corresponding R_k^2 is:

$$\begin{aligned} R_k^2 &= \frac{(\theta^k - 1)^2 \text{var}(f_t)}{\text{var}(s_{t+k} - s_t)} \\ &= \frac{(1 - \theta^k)^2 \sigma_e^2}{(1 - \theta^k)^2 \sigma_e^2 + (1 - \theta^{2k})^2 \sigma_e^2 + k(1 - \theta^2)(1 - b\theta)^2 \sigma_\eta^2 / b^2}. \end{aligned} \tag{22.102}$$

Engel *et al.* (2007) set $b = 0.9$, $\theta = 0.95$ and $\sigma_\eta / \sigma_e = 3$ and calibrate the model. They find that predictability, in terms of R_k^2, exhibits a hump shaped pattern with respect to k. At short horizons there is not much evidence of predictability, e.g., $R_1^2 = 0.02$,

but, as the horizon increases, the value of R^2 becomes larger and reaches a maximum at a horizon of 44 quarters, where $R^2_{44} = 0.38$. The authors argue that, in this framework, panel estimation techniques may be useful in detecting predictability by picking up a common element to the risk-premium across exchange rates.

22.6.4 Panels

Empirical studies based on country-by-country estimations are confronted with the problem of low power and poor parameter estimates. The possibility of common elements in the DGPs motivates the use of pooled time series estimation in order to increase the power of predictability tests.[75] A number of recent studies using panel datasets provide evidence in favor of a long-run relationship of the deviations of the exchange rate from its fundamental value.[76]

Mark and Sul (2001) investigate if panel estimation techniques are useful in forecasting exchanges rates by exploiting interdependencies of exchange rates with the same numeraire, namely the US dollar, the Swiss franc and the Japanese yen. The dataset includes 19 OECD countries and spans the period from 1973:1 through 1997:1. Their study centers on the panel version of the equation employed in Mark (1995):

$$s_{i,t+1} - s_{i,t} = \beta z_{i,t} + e_{i,t+1}, \tag{22.103}$$

$$e_{i,t+1} = \gamma_i + \theta_{t+1} + u_{i,t+1}, \tag{22.104}$$

where γ_i is a country-specific effect, θ_{t+1} is a time-specific error and $u_{i,t+1}$ is an idiosyncratic error. Before examining the out-of-sample performance of the multi-country monetary model, Mark and Sul (2001) follow Berkowitz and Giorgianni (2001) and test for cointegration. However, due to the fact that the standard least squares dummy variable (LSDV) estimator suffers from second-order bias, causing the t-statistic to diverge asymptotically, the panel dynamic OLS estimator is adopted. The system of equations for the changes in the exchange rates is:

$$s_{i,t+1} - s_{i,t} = \gamma_i + \theta_t + \beta z_{i,t-1} + \sum_{j=-p_i}^{p_i} \delta_{ij} \Delta x_{i,t-j-1} + u_{i,t}. \tag{22.105}$$

Although the corresponding t-ratio is asymptotically normally distributed, Mark and Sul (2001) also use a bootstrap procedure to account for possible finite sample bias. The null DGP for the bootstrap is a restricted VAR:

$$\Delta s_{i,t} = \mu^i_s + \varepsilon^i_{s,t}, \tag{22.106}$$

$$\Delta z_{i,t} = \mu^i_z + \sum_{j=1}^{q_i} \phi^i_{1,j} \Delta s_{i,t-j} + \sum_{j=1}^{q_i} \phi^i_{2,j} \Delta z_{i,t-j} + \varepsilon^i_{z,t}. \tag{22.107}$$

The equations for $\Delta z_{i,t}$ are fitted by iterated seemingly unrelated regression (SUR).[77] For all three numeraire currencies, both the asymptotic and the

bootstrap p-values indicate that the null hypothesis of no cointegration is rejected at the 5% level. In turn, Mark and Sul (2001) test for short-horizon in-sample predictability in the presence of cointegration.[78] The results provide strong evidence in favor of predictability based on both monetary and PPP fundamentals. Finally, the authors examine whether macroeconomic fundamentals contain power in forecasting exchange rates at horizons $k = 1, 16$. They use Theil's U-statistic and construct bootstrap critical values under the assumption of cointegration. Overall, forecasts based on monetary fundamentals dominate forecasts based on the PPP fundamentals and the RW with a drift for the majority of countries when the US dollar or the Swiss franc is the numeraire currency, but not in the case of the Japanese yen.[79]

In this section we have shown how, more than 30 years after the breakdown of Bretton Woods, the difficulty of forecasting exchanges rates using economic fundamentals has become a stylized fact in international finance. Although the availability of longer datasets on modern floating rates and the application of recent sophisticated econometric techniques regarding panel data, nonlinear models, as well as forecast evaluation measures, are promising, researchers and practitioners are still faced with the problem of deriving models with robust behavior in terms of out-of-sample forecasting across exchange rates and time periods.

22.7 Conclusions

We have provided a selective overview of a few of the key relationships which will play critical roles in determining the behavior of the exchange rate and an evaluation of the efficacy of forecasting methods. The major change in their analysis over the last decade has been the application of more sophisticated time series techniques motivated by theoretical considerations such as the limits to arbitrage and the microstructure of the exchange market. Nonlinear models seem able to provide some explanation of the PPP puzzle and the forward bias problem. However, issues of how data sampled at a different frequency to the economic decision impacts on the nonlinear models needs further investigation. Also, the recent finding of joint long memory and nonlinearity in these relationships is a new puzzle. The possibility that the exchange market can exhibit bubbles is of long-standing interest. The new analysis of Phillips *et al.* (2006) would appear to be an exciting development and might provide new insights on this issue over the next few years.

The nonlinear methods appear to offer scope for improved forecasts than the previous linear models. However, researchers and practitioners are still faced with the problem of deriving models with robust behavior in terms of out-of-sample forecasting across exchange rates and time periods, even when the model employed for estimation or forecasting is appropriate for the policy regime in operation (or anticipated).

Notes

1. In fact, the first example of such arrangements was apparently Austria-Hungary between 1896 and 1914 (Flandreau and Komlos, 2003). Other important examples where the

anticipation of a change in policy regime has been an important focus of the analysis of exchange rates are: (i) the possible anticipation of the return to gold of sterling prior to April 1925 (e.g., Flood and Garber, 1983; Michael *et al.*, 1997), (ii) the anticipation of a fixed exchange rate between the East and West German Mark following German monetary union (Burda and Gerlach, 1993).

2. Similarly, chaotic behavior, which we do not have space to consider (see, e.g., De Grauwe *et al.*, 1993).
3. This would imply a half-life of around five years. Assuming $y_t \sim$ AR(1), the half-life (h) of a unit shock would be $0.5 = \beta^h$, or taking logs, $\ln 0.5 = h \ln \beta$, $h = \frac{\ln 0.5}{\ln \beta}$.
4. Linear univariate autoregressive time series models for the real exchange rate have not been restricted to integer orders of integration. Explosive as well as fractional processes have been used to model real exchange rates. Bleaney *et al.* (1999) fitted a stochastic unit root process to the real exchange rate of high inflation countries (Argentina, Brazil, Chile, and Israel). The real exchange rate process is as follows:

$$y_t = (1 + \delta_t) y_{t-1} + \nu_t,$$

where $\nu_t \sim$ i.i.d.$(0, \sigma_\nu^2)$ $\delta_t \sim$ i.i.d.$(0, \sigma_\delta^2)$. Leybourne *et al.* (1996) derive a test for the null hypothesis $H_0 : \sigma_\delta^2 = 0$ against $H_1 : \sigma_\delta^2 > 0$. It is worth mentioning that, even under the null, PPP is assumed non-stationary in this model. A different order of integration for PPP deviations is proposed by the fractional literature (see section 22.3.2: see, e.g., Diebold *et al.*, 1991; Pippenger and Goering, 1993; Taylor *et al.*, 2001). These papers also show that standard unit root tests may exhibit low power against the fractional alternative. However, neither the stochastic unit root nor fractional process have clear theoretical underpinnings.
5. Paya and Peel (2007a) also analyze the asymptotically efficient estimator for cointegration regression introduced by Saikkonen (1991). Cointegration is then tested using the Shin (1994) statistic, which is a residual-based test where the null hypothesis is that of cointegration or stationary residuals in the Saikkonen regression. Using simulated data, Paya and Peel (2007a) find that the Saikkonen estimator produces estimates of the cointegrating weights which are much closer on average to their true values, and with much smaller standard errors than the Johansen method.
6. The smooth adjustment process is suggested in the analysis of Dumas (1992), noted by Teräsvirta (1994) and demonstrated by Berka (2002).
7. The other common form of STAR is the logistic STAR (LSTAR), where the transition function $F(\cdot)$ is logistic:

$$F(z_{t-d}; \gamma, c) = [1 + \exp(-\gamma(z_{t-d} - c))]^{-1}.$$

The LSTAR transition function is asymmetric about $(y_{t-d} - c)$ and admits the limits:

$$F(\cdot; \gamma) \to 1 \text{ as } (z_{t-d} - c) \to +\infty,$$
$$F(\cdot; \gamma) \to 0 \text{ as } (z_{t-d} - c) \to -\infty.$$

LSTAR models have also been fitted to real exchange rates (see Sarantis, 1999; Copeland and Heravi, 2007). Even though the theoretical argument is not as strongly supported as with ESTAR, there are some attempts to rationalize the asymmetric adjustment in the real exchange rate (see Campa and Goldberg, 2002).
8. The effect of nonlinearities on traditional cointegration techniques have been examined. Paya and Peel (2007a) assume that the DGP is given by the ESTAR process. Using simulated data from such a model, in which proportionality, $(1, -1)$, is imposed, they examine the empirical results obtained when the Johansen method is employed to determine whether the spot exchange rate is cointegrated with domestic and foreign prices and whether proportionality can be rejected. Empirical results show that the Johansen method produces

poor estimates, on average, of the cointegrating vector, with a range of values that include those reported in the literature.

9. Conventional maximum likelihood theory is therefore not applicable. If one were to use a maximum likelihood estimator, the testing procedure would be analogous to the one described below but using the partial derivatives of the likelihood function evaluated under the null (see Granger and Teräsvirta, 1993, Ch. 6).
10. That is, $y_t = \beta' \widetilde{y}_t + \gamma \phi' \widetilde{y}_t (y_{t-d} - c)^2 + \gamma \phi' \widetilde{y}_t (y_{t-d} - c)^4$. Note that even powers of the Taylor approximation of the logistic function are all zero while odd powers of the Taylor approximation of the exponential are all zero. The logistic function has one inflection point while the exponential possesses two, which is the point of using a second-order Taylor expansion.
11. In particular, they propose to use the White heteroskedasticity consistent covariance matrix in the LM test. Their analysis is based on MacKinnon and White (1985). See also Wooldridge (1990).
12. See Lundbergh and Teräsvirta (1998) for the specification, estimation and evaluation of models with nonlinear behavior in the mean (STAR) and in the conditional variance (STGARCH), the STAR-STGARCH model.
13. See below for a description and references of this methodology.
14. However, its distribution is the same when the transition variable y^*_{t-1} is substituted by y^*_{t-d} or moving averages of y^*_{t-1} (see Venetis *et al.*, 2005). What changes is the form of the auxiliary regressions, which generalize to:

$$\Delta y^*_t = \delta y^*_{t-1} y^{*2}_{t-d} + error,$$

$$\Delta y^*_t = \sum_{j=1}^{p} a_j \Delta y^*_{t-j} + \delta y^*_{t-1} y^{*2}_{t-d} + error.$$

The corollary results in Venetis *et al.* are particularly important since it is possible to generalize KSS tests against wider alternatives that assume longer adjustment periods.

15. The linear unit root hypothesis against an ESTAR has also been tested using a different methodology to that of a Taylor approximation. Kiliç (2003) overcomes the identification problems for γ and c in his test by using a grid search over the space of values for the parameters γ and c to obtain the largest possible t-value for ϕ in the following regression:

$$\Delta y^*_t = \phi y^{*3}_{t-1} \left[1 - \exp(-\gamma (z_t - c)^2) \right] + error,$$

where z_t is the transition variable, which in our framework would be Δy^*_{t-1}. He tests the null hypothesis of $H_0 : \phi = 0$ (unit root case) against $H_1 : \phi < 0$. Kiliç (2003) claims that the advantages of his procedure over KSS is twofold. First, it computes the test statistic even when the threshold parameter needs to be estimated in addition to the transition parameter. Second, it claims to have higher power. Using quarterly data for 17 real exchange rates of developed countries against the dollar for the floating period, Kiliç finds strong evidence of nonlinear ESTAR behavior.

16. This approach is suggested by Vogelsang (1998).
17. PPP has also been tested using nonlinear cointegration. Kapetanios *et al.* (2003b) (KSSb) propose a testing procedure to detect the presence of a cointegrating relationship that follows a STAR process. Venetis *et al.* (2005) and Paya and Peel (2006a) find evidence of nonlinear cointegration between the real exchange rate and productivity proxies. Further support for nonlinear cointegration has been found with a cointegration method that uses a transformation of the variables. Breitung (2001) suggests a rank test procedure based on the difference between the sequences of ranks of the variables involved in the cointegrating relationship. Haug and Basher (2005) also apply the Breitung test for the dollar and DM based real exchange rates of the G10 countries and only found evidence of nonlinear long-run PPP in two of them, the pound/dollar and Belgian franc/DM.

18. The order of autoregression is chosen through inspection of the PACF function of y_t^*.
19. In the ESTAR model (22.15), if the error variance is not reported as standardized but has a standard error of *se*, it is necessary to multiply the estimated speed of adjustment parameter γ by (se^2) for comparison purposes.
20. Under the null hypothesis $H_0 : \gamma = 0$, y_t is generated as an RW with initial values $\{y_i\}_{i=-\max\{d,p\}}^{0} = 0$ and sample size set equal to the observed sample size. The underlying noise process u_t is $NID(0, s^2)$. The value of s is chosen to be equal to the standard error of each estimated model. Regression (22.15) is estimated using the respective d values and the t-ratios are stored for the estimates of γ. This is repeated 10,000 times and the critical values are obtained from the upper empirical quantiles since the empirical distributions are not symmetric and interest rests on the one-tail alternative $\gamma > 0$. See Paya and Peel (2006b) for a discussion of possible bias in the estimated γ for small samples.
21. We know from the transition function $F(\cdot)$ in ESTAR models that adjustment is time-varying and depends on the size of the deviation. However, for comparison purposes, we need a value of speed in time periods. The generalized impulse response function (GIRF) introduced by Koop *et al.* (1996) successfully confronts the challenges that arise in defining impulse responses for nonlinear models and is defined as:

$$\widehat{GIRF}_h(h, \delta, \omega_{t-1}) = E(y_{t+h}|u_t = \delta, \omega_{t-1}) - E(y_{t+h}|u_t = 0, \omega_{t-1}),$$

where $h = 1, 2, \ldots$ denotes horizon, u_t is a random shock of size δ occurring at time t, ω_{t-1} is defined by all sets $\{y_{1+i}^*, \ldots, y_{d+i}^*\}_{i=0}^{T-d-1}$, and Ω_{t-1} is a random variable defining possible history sets. Since analytic expressions for the conditional expectations involved in the expression above are not available for $h > 1$, Gallant *et al.* (1993) and Koop *et al.* (1996) used stochastic simulation to approximate it (see Venetis *et al.*, 2007, for an analytical expression for the "naive" impulse response function of an ESTAR model). Taylor and Peel (2000) conduct GIRF analysis on the deviations of real exchange rates from monetary fundamentals and Taylor *et al.* (2001) use impulse response functions to gauge how long shocks survive in real exchange rate nonlinear models. The half-life of shocks is dramatically shorter than that obtained or implied by linear models.
22. Analytic results are available on the impact of temporal aggregation on a linear series (e.g., Rossana and Seater, 1995) but not for nonlinear series.
23. For instance, MacKinnon and White (1985) showed that in finite samples the White HCCME can be seriously biased.
24. In Paya and Peel (2006a) the real dollar–sterling exchange rate series spans several exchange rate regimes, and within this context, a parametric form may not adequately capture the conditional heteroskedasticity in the data (see Gonçalves and Kilian, 2003).
25. Wild resampling typically underestimates the variance of the parent distribution. This can be remedied by replacing the observed residuals with "leveraged" residuals $\frac{\widehat{u}_t}{1-h_t}$, where h_t is the leverage for the ith residual estimated from the parametric model (see Davidson and Hinkley, 1997).
26. An alternative wild bootstrap has the following distribution: $\varepsilon_i = 1$ with $p = 0.5$; and $\varepsilon_i = -1$ with $p = 0.5$ (see Davidson and Flachaire, 2001).
27. The wild bootstrap matches the moments of the observed error distribution around the estimated regression function at each design point $(\widehat{y}^b)$. Liu (1988) and Mammen (1993) show that the asymptotic distribution of the wild bootstrap statistics are the same as the statistics they try to mimic.
28. For the cases where the time-varying equilibrium is defined by deterministic components, such as dummies and time trends, see Paya and Peel (2003, 2004).
29. Kapetanios (1999) considers the properties of model selection using information criteria in the context of nonlinear threshold models: in particular, Akaike information criterion (AIC), Schwarz criterion (SC), Hannan–Quinn (HQ) criterion, the generalized information criterion (GIC) and the informational complexity criterion (ICOMP). In a Monte Carlo

exercise he shows that standard information criteria (AIC, SC, HQ) have an important role to play in model selection for nonlinear threshold models. Others (GIC, ICOMP) are less reliable.

30. This particular type of TAR model is called a self-exciting threshold autoregressive (SETAR) model.
31. Peel and Taylor (2002) adopted a similar strategy in a threshold model under the null of RW. See Duarte *et al.* (2005, Appendix B) for the case where the errors display autocorrelation.
32. Estimate the model using the actual data for a set of values of c in the range C and in each case calculate the likelihood ratio (LR) statistic $LR(c)$ for that value of c against the value of the likelihood obtained by unrestricted LS, i.e., $LR(c) = T(\hat{\sigma}^2_T(c) - \hat{\sigma}^2_T(\hat{c}))/\hat{\sigma}^2_T(\hat{c})$. Notice that for $c = \hat{c}$ we get $LR(c) = 0$. Then plot $LR(c)$ against c and draw a flat line that corresponds to the β-level critical value $c^*(\beta)$ given in Hansen (1997, table 1, p. 5). For $\beta = 5\%$, $c^*(\beta) = 7.35$. The confidence interval LR^c is given by $LR^c = \{c : LR(c) \leq c^*(\beta)\}$.
33. This is generalized for the case where the attractor might be different than 0, say a_0, and also for the case of a linear trend, $a_0 + a_1 t$.
34. The threshold value λ should be between the 15th (λ_1) and 85th percentile (λ_2) of y_t. $\lambda \in \Lambda = [\lambda_1, \lambda_2]$.
35. Further support for the TAR behavior of real exchange rates is found in De Jong *et al.* (2007). They consider an extension of the simple TAR model above: an asymmetric TAR model where the adjustment in the upper part of the "action band" might differ from the one in the lower part of the "action band." The asymmetric TAR can be written as:

$$\Delta y_t = \begin{cases} u_t & \text{if } c_1 \leq y_{t-1} \leq c_2, \\ \rho_1(y_{t-1} - c_1) + u_t & \text{if } y_{t-1} < c_1, \\ \rho_2(y_{t-1} - c_2) + u_t & \text{if } y_{t-1} > c_2, \end{cases} \tag{22.26}$$

where $\rho_1, \rho_2 \in (-2, 0)$. De Jong *et al.* (2007) propose a three-regime threshold unit root (TUR) test that is robust to errors that are not i.i.d. They propose an asymptotically pivotal test statistic that optimizes over the parameter that is unidentified under the null and allows for weakly dependent errors. De Jong *et al.* (2007) apply their statistic to monthly real exchange rates of six developed countries against the dollar. While ADF and PP could not reject the null of a unit root, their test statistic found evidence against unit root real exchange rates.

36. See Taylor (1995) and Sarno and Taylor (2002) for numerous references and further discussion.
37. Canjels *et al.* (2004) employ threshold autoregression to analyze the efficiency of international arbitrage under the gold standard from 1879 to 1913. A theoretical model is developed and estimated employing continuous daily data.
38. For example, Monoyios and Sarno (2002) employ an ESTAR model to describe the mispricing between spot and futures. Employing daily data, they report a lag structure of five days. The impulse response functions reported by them imply an arbitrage process taking weeks for small shocks and days for large shocks. However, other researchers using minute by minute data for the same market, and ESTAR or TAR adjustment mechanisms, report adjustment mechanisms taking several minutes (Dwyer *et al.*, 1996; Taylor *et al.*, 2000).
39. This estimator retains the properties of the FM-OLS estimator in that it can deal with non-stationary data and serial dependence in the equation errors. However, the FM-LAD estimator is based on a "fully modified" extension of the LAD regression estimator, which is a robust procedure well known in the treatment of non-normality of error terms when the regressors are stationary (Bassett and Koenker, 1978; Phillips, 1991). The FM-LAD estimator developed by Phillips has all the features of the LAD estimator but is applicable in models where the regressors are non-stationary, there is endogeneity in the regressors and serial dependence in the errors. In addition, the FM-LAD estimator is valid even when the data do not have finite variances.

40. When s_t and f_t are integrated variables, a regression of s_{t+n} on f_t will give the same coefficient as one of s_t on f_t, at least asymptotically. Maynard and Phillips (2001) show that in small samples the overlapping errors induced by employing s_{t+n} in the regressions can give rise to different estimates.
41. See Minford and Peel (2002) for a textbook derivation.
42. However, we note that the properties of regressions with $f_t - s_t$ as the dependent variable and the conditional variance as a regressor would be of interest in this context. They would determine whether the estimates of the conditional variance have predictive content for the forward premium.
43. An alternative economic explanation of the anomaly that might be relevant for particular periods is the one advanced by McCallum (1994). He suggests that the negative coefficient may be the result of simultaneity induced by the existence of a policy reaction function where the interest rate differential is set in order to stabilize current exchange rate movements. Another important possibility is set out by Evans and Lewis (1995) and Spagnola *et al.* (2005). They show that if the "long" swings in exchange rate regimes between depreciating and appreciating periods have *ex ante* predictability, then in small samples a peso problem occurs. They assume that the exchange rate regimes can be captured by a Markov-switching process (see, e.g., Hamilton, 1990). Again, whilst the peso problem is undoubtedly important to some policy periods, the systematic nature of the forward anomaly over different time periods and different numeraire currencies suggests there are other factors at work.
44. At the one extreme, $d = 0$ represents the short memory case. If $d > 0.5$, the process is not wide-sense stationary, having infinite variance. And at the other extreme, $d = 1$ corresponds to the ordinary integrated process, familiarly known as an RW, which is well known not to revert to the mean, but eventually to wander arbitrarily far from the starting point. The autocorrelations of a fractional process dissipate but at a slow hyperbolic rate rather than the geometric rate of a standard ARMA process (see Granger, 1980; Granger and Joyeux, 1980). (see Granger , 1980; Granger and Joyeux, 1980).
45. Peel and Davidson (1998) fit a nonlinear error correction mechanism to the spot-forward relationship based on the idea of nonlinearities in the unobserved risk-premium which might be captured by a bilinear process.
46. Baillie and Kiliç (2006) prefer the logistic function as the transition function. This implies asymmetric behavior of the deviations as to whether they are positive or negative. Given this, their results have the same qualitative implications as those of Sarno *et al.* (2006), even though the symmetric transition function appears to be better economically motivated.
47. The standard ESTAR model exhibits a maximum expected deviation from equilibrium. Note that if the dataset does not actually include a maximum or minimum then the value of γ estimated has to be such, i.e., smaller, to ensure that a max or min occurs outside of the range of estimated data. An economic rationale for the maxima or minima is unclear. This line of reasoning suggests a simple modification of the ESTAR model, which is to assume that the variable forcing the process towards its equilibrium value is not observed from past deviations but from the expected deviation at time t.
48. They employ survey data to measure anticipated excess returns.
49. See also Van Dijk *et al.* (2002) for an earlier application to unemployment rates, and Tsay and Härdle (2007), who set out a general class of Markov-switching ARFIMA processes.
50. Baillie and Kapetanios (2005) consider a test based on neural networks.
51. The FI-STAR model is clearly of interest. If it is the case that forward premia (or PPP) are parsimoniously described by such a process then it seems to pose a challenge for theoretical work.
52. We note that numerous studies have been conducted to determine the empirical relationship between real interest rates and the real exchange rate. From the uncovered parity conditions, ignoring any time-varying risk-premia for simplicity:

$$i_t - E_t p_{t+n} + p_t = i_t^* - E_t p_{t+n}^* + p_t^* + E_t y_{t+n} - y_t, \tag{22.55}$$

where i and i^* are the domestic and foreign nominal interest rates, p and p^* are logs of the domestic and foreign price level, and y is the log of the real exchange rate. We note immediately that empirical work that examines the relationship between real interest rates and excludes the real exchange rate will, in general, be misspecified unless the real exchange rate follows an RW as suggested by Roll (1979). However, this does not appear to be either theoretically or empirically justified (see Taylor and Sarno (2004); Minford and Peel (2007), and the references suggesting real exchange rates are nonlinear mean reverting processes above).

The empirical studies on the relationship between real interest rates and real exchange rates are problematic and inconsistent. For example, a number of studies have examined the relationship in a cointegration framework, though there are no theoretical grounds for expecting the real exchange rate to be cointegrated with the real interest rate differential, as noted by Baxter (1994).

The implications of nonlinear real exchange rate adjustment have also not been integrated into the empirical literature. As discussed above, temporal aggregation of the "monthly" process changes the form of the ESTAR process (Paya and Peel, 2006b). In particular, the number of autoregressive terms increases. One procedure, if the correct DGP is an ESTAR process, is to derive multi-period forecasts from it using Monte Carlo methods. These forecast changes should be employed as regressors in empirical work. Our exercise shows why, in empirical studies, the reported results and their implications are likely to change as the horizon of expectations and the temporal nature of the data changes. These implications appear worthy of investigation in further work.

53. Bertola and Svensson (1993) consider imperfect credibility, Miller and Weller (1991) consider price-stickiness. Bauer *et al.* (2007) develop a non-rational model based on chartists and fundamentalists.
54. Chung and Tauchen (2001), using the efficient method of moments proposed by Gallant and Tauchen (1996, 2000), allow for intermarginal intervention. They report evidence that this model parsimoniously describes the dynamics of the French franc/Deutsche Mark exchange rate from January 1987 until July 1993. We note that they employ weekly data, and further tests with daily data appear warranted.
55. Lundbergh and Teräsvirta (2006) point out that the asymmetry parameter is essential to ensure that movements of the exchange rate are restricted by the bounds. The parameter restrictions $\gamma > 0$, $\theta > 0$ are identifying restrictions, and $\mu < 1$ identifies an implicit bound. Non-symmetry around the lower or upper bound – explicit or implicit – can be allowed for by different values of γ, θ.
56. The STARTZ model can capture the dynamics of behavior implied by many theoretical target zone models. When the exchange rate is near the centre of the band, such that G^l and G^u are close to zero, then the exchange rate will depend on its own lags, $\varphi' x_t$. Given previous research we would anticipate that the exchange rate would appear to be a unit root process in this vicinity. As the exchange rate approaches the edges of the bands, so that G^l and G^u are close to unity, then the exchange rate process will be described by white noise like behavior around μs^l or μs^u. Also, as the deviation from the central parity increases, so that G^l and G^u approach unity, there is smooth transition from the standard GARCH model represented by $\eta' w_t$ towards a constant $\delta > 0$ that is expected to be close to zero. The assumption that δ is non-zero means there is a positive probability that the exchange rate could leave the band even though no realignment takes place. This feature implies that a credible zone can be one in which occasional breaches occur.
57. Following Mark (1995), the money demand income elasticity is set equal to one.
58. The mathematical form of the bubble is the same as that obtained in the stock market. If the exchange model exhibited sticky prices the process followed by the speculative bubble would be different although the features are qualitatively the same. For example, with a simple sticky-price adjustment of the form $p_t - p_{t-1} = \theta(s_t - p_{t-1})$, $\theta \in [0, 1]$, the bubble

takes the form:

$$B_t = \frac{b}{b+\theta} E_t B_{t+1} - \frac{(1-\theta)b}{b+\theta} E_{t-1} B_t + \frac{(1-\theta)b}{b+\theta} B_{t-1}.$$

For $\theta = 1$ we obtain the bubble solution as in (22.68) above. If $E_{t+1}B_t = \delta B_t$ then the analysis can proceed as above. The magnitude of $\delta > 1$ could be important as to whether the bubble is asymptotically stationary (see Yoon, 2005).

59. In the latter case, the form of the bubble depends on the time series process followed by the fundamentals. For the geometric process:

$$B_t = c s_{f,t}^{\beta} \ln s_{f,t} - \ln s_{f,t-1} = \mu + \omega_t,$$

where $\omega_t \sim N(0, \sigma_h^2)$ and μ is a drift term, β is the solution of $\beta\mu + 0.5\beta^2\sigma_h^2 + \ln\lambda = 0$ and c is an arbitrary constant. Minford and Peel (2002) provide a textbook treatment, while Bidarkota and Dupoyet (2007) extend the analysis.

60. Diba and Grossman (1988) tested for bubbles by applying unit root tests to the asset price, real stock prices and dividends. If the fundamental is an integrated process, say $I(1)$, then from (22.67) the bubble will imply rejection of cointegration. The important insight of Evans was to show via simulation that standard unit root and cointegration tests have little power to detect his periodically collapsing bubble. Yoon (2005) demonstrates that Evans bubbles tail indices are less than one, a property we comment on below. Gurkaynak (2005) has a useful survey of many of the empirical tests for bubbles.
61. Phillips *et al.* (2006) demonstrate that Evans bubbles with a π as low as 0.25 may be detectable.
62. Note that, for comparison, the tail index for the student t-distribution is its degrees of freedom.
63. The latter justification is in line with recent empirical evidence which suggests that the relationship between the exchange rate and the fundamental value is characterized by significant nonlinearities (Taylor and Peel, 2000; Taylor *et al.*, 2001).
64. However, spurious regression problems may rise even at short horizons. Ferson *et al.* (2003) decompose asset returns into expected returns and an unpredictable noise component. In this setting, although the dependent variable is not a persistent process, the possibility of persistent expected returns and explanatory variables may lead to spurious results.
65. Hence, Mark's conjecture that the mixed evidence on long-horizon predictability is due to the small sample size is not supported by the data.
66. Abhyankar *et al.* (2005) adopt a different approach which utilizes the realized end-of-period wealth so as to evaluate forecasts based on monetary fundamentals. Their findings suggest that there is evidence of economic value to exchange rate predictability, especially at long horizons.
67. Four hundred tests are implemented since four models, four bilateral exchange rates, five forecast horizons and five test statistics are considered.
68. A q-value is defined as the minimum possible false discovery rate for which the null hypothesis is rejected (Storey, 2003). The false discovery rate is the ratio of the number of tests for which we reject the null, F, over the total number of tests, S, given that the null is true, $E(F/S)$. A detailed description of the computation of q-values and their properties is provided in McCracken and Sapp (2005) and the references therein.
69. The particular ESTAR specification remains the same under the null and the alternative, and is determined in the preliminary analysis on the basis of the goodness-of-fit, the significance of the coefficients and residual diagnostics (Eitrheim and Teräsvirta, 1996). The bootstrap procedure is similar to the VECM bootstrap of Kilian (1999) and was described in section 22.6.1.1.
70. Taylor and Kilian (2003) conduct Monte Carlo experiments to investigate the power and size properties of the tests. Their findings indicate that in-sample tests have substantially higher power than out-of-sample tests.

71. A regime switching behavior of the exchange rate can be attributed to factors such as the heterogeneity of opinions among agents, the presence of transaction costs, the interaction of chartists and fundamentalists, the peso problem, different monetary and fiscal policies between countries, as well as the implications of the dirty floating exchange rate regime (see Engel and Hamilton, 1990; De Grauwe and Vansteenkiste, 2001; Lee and Chen, 2006).
72. The use of MS is also motivated by parameter instability in empirical exchange rate models (see Rossi, 2006), which may be attributed to "swings" in expectations about future values of the exchange rate (Frankel, 1996), as well as by rational expectations models of exchange rate determination in which the weight attached to fundamentals by practitioners changes over time (Bacchetta and van Wincoop, 2004).
73. In a related study, Sarno and Valente (2005) examine the evolution of the relationship between fundamentals and the exchange rate by employing the recursive procedure of Pesaran and Timmermann (1995) to real-time data. The authors use a broad set of fundamentals for five major US dollar exchange rates over the post-Bretton Woods era. In the preliminary analysis, a "virtual search" is conducted over all possible models and the optimal combination of fundamentals is determined period by period. In the cases where the best model outperforms the RW, a real-time forecasting exercise is implemented. The main implication of the experiment is that fundamentals contain predictive power for the movements of the exchange rate. However, the importance of each of the fundamental variables changes over time. Furthermore, conventional model selection criteria cannot identify the "correct" model to beat the RW in real time.
74. Engel and West (2005) note that estimates of the interest semi-elasticity of money demand λ typically range from 29 to 60, which implies that in the monetary model b is between 0.97 and 0.98.
75. However, Sarno and Taylor (1998) note that panel unit root tests tend to reject the null of a unit root even if a single series is stationary. Moreover, Rapach and Wohar (2004) emphasize the importance of the homogeneity assumption. If such an assumption is not empirically supported then pooling data across countries may result in false inference.
76. For example, Frankel and Rose (1996) and Lothian (1997), among others, show that real exchange rates mean revert in the long run by using pooled data; Groen (2000) and Groen and Kleibergen (2003) find evidence of cointegration between monetary fundamentals and exchange rates.
77. Bootstrap samples are generated according to the fitted VAR by sampling from the estimated residuals with replacement. Next, the slope coefficient is estimated by panel dynamic OLS and the corresponding t-ratio is computed. The above procedure is repeated 2,000 times so as to obtain bootstrap distributions and p-values.
78. The bootstrap procedure is the same as before, with the exception that the second equation of the previous null DGP becomes:

$$\Delta z_{i,t} = \mu_z^i + \gamma_i z_{i,t-1} + \sum_{j=1}^{q_i} \phi_{1,j}^i \Delta s_{i,t-j} + \sum_{j=1}^{q_i} \phi_{2,j}^i \Delta z_{i,t-j} + \varepsilon_{z,t}^i.$$

    Cointegration requires $-2 < \gamma_i < 0$.
79. Groen (2005) emphasizes the importance of expectations on the validity of the monetary model. He examines the European Union exchanges rates of Canada, the US and Japan for the period from 1975 to 2000. The cointegration framework adopted is an extension of the Johansen method for a panel of VECM models which allows heterogeneous short-run dynamics (see Groen and Kleibergen, 2003). Overall, exchange rates appear to be predictable at medium- to long-term horizons, i.e., one to four years. However, the results are sensitive to the cointegrating parameters restriction.

References

Abadir, K. and G. Talmain (2006) Distilling co-movements from persistent macro and financial series. Mimeo, University of York.

Abhyankar, A., L. Sarno and G. Valente (2005) Exchange rates and fundamentals: evidence on the economic value of predictability. *Journal of International Economics* **66**, 325–48.

Acosta, F. and C. Granger (1995) A linearity test for near for near-unit root time series. Discussion Paper No. 95-12, Department of Economics, University of California, San Diego.

Andrews, D. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–58.

Bacchetta, P. and E. van Wincoop (2004) A scapegoat model of exchange-rate fluctuations. *American Economic Review* **94**, 114–18.

Backus, D., A. Gregory and C. Telmer (1993) Accounting for forward rates in matrkets for foreign currency. *Journal of Finance* **48**, 1887–908.

Backus, D. and G. Smith (1993) Consumption and real exchange rates in dynamic economies with non-traded goods. *Journal of International Economics* **35**, 297–316.

Baillie, R. and T. Bollerslev (1990) A multivariate generalized ARCH approach to modelling risk premia in forward foreign exchange markets. *Journal of International Money and Finance* **9**, 309–24.

Baillie, R. and T. Bollerslev (1994) The long-memory of the forward premium. *Journal of International Money and Finance* **13**, 309–24.

Baillie, R. and T. Bollerslev (2000) The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance* **19**, 471–88.

Baillie, R. and G. Kapetanios (2005) Testing for neglected nonlinearity in long memory models. Working Paper 528, Queen Mary, University of London.

Baillie, R. and G. Kapetanios (2006) Non-linear models with strongly dependent processes and applications to forward premia and real exchange rates. Working Paper 570, Queen Mary, University of London.

Baillie, R. and R. Kiliç (2006) Do asymmetric and nonlinear adjustments explain the forward premium anomaly? *Journal of International Money and Finance* **25**, 22–47.

Balassa, B. (1964) The purchasing power parity doctrine: a reappraisal. *Journal of Political Economy* **72**, 584–96.

Balke, N. and T. Fomby (1997) Threshold cointegration. *International Economic Review* **38**, 627–46.

Balke, N. and M. Wohar (1998) Nonlinear dynamics and covered interest parity. *Empirical Economics* **23**, 535–59.

Basci, E. and M. Caner (2005) Are real exchanges rates nonlinear or nonstationary? Evidence from a new threshold unit root test. *Studies in Nonlinear Dynamics and Econometrics* **9**, Article 2.

Bassett, G. and R. Koenker (1978) Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* **363**, 618–22.

Bauer, C., P. De Grauwe and S. Reitz (2007) Exchange rate dynamics in a target zone – a heterogeneous expectations approach. Mimeo, Deutsche Bundesbank, No. 11.

Baum, C., J. Barkoulas and M. Caglayan (2001) Nonlinear adjustment to purchasing power parity in the post-Bretton Woods era. *Journal of International Money and Finance* **20**, 379–99.

Baxter, M. (1994) Real exchange rates and real interest rate differentials: have we missed the business cycle relationship? *Journal of Monetary Economics* **33**, 5–37.

Bekaert, G. and S. Gray (1998) Target zones and exchange rates: an empirical investigation. *Journal of International Economics* **45**, 1–35.

Bekaert, G. and R. Hodrick (1993) On biases in the measurement of foreign exchange risk premiums. *Journal of International Money and Finance* **12**, 115–38.

Beran, J. (1995) Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models. *Journal of the Royal Statistical Society, Series B* **57**, 659–72.
Berben, R.-P. and D. van Dijk (1998) Does the absence of cointegration explain the typical findings in long horizon regressions? Econometric Institute Report 145, Erasmus University, Rotterdam.
Berka, M. (2002) General equilibrium model of arbitrage trade and real exchange rate persistence. Mimeo, University of British Columbia.
Berkowitz, J. and L. Giorgianni (2001) Long-horizon exchange rate predictability. *Review of Economics and Statistics* **83**, 81–91.
Bertola, G. and L. Svensson (1993) Stochastic devaluation risk and the empirical fit of target-zone models. *Review of Economic Studies* **60**, 689–712.
Bidarkota, P. and B. Dupoyet (2007) Intrinsic bubbles and fat tails in stock prices. A note. *Macroeconomic Dynamics* **3**, 405–22.
Blanchard, O. and M. Watson (1982) Bubbles, rational expectations, and financial markets. In P. Wachter (ed.), *Crises in the Economic and Financial Structure*, pp. 295–315. Lexington, Mass.: Lexington Books,
Bleaney, M., S. Leybourne and P. Mizen (1999) Mean reversion of real exchange rates in high-inflation countries. *Southern Economic Journal* **65**, 839–54.
Breitung, J. (2001) Rank tests for nonlinear cointegration. *Journal of Business and Economic Statistics* **19**, 331–40.
Burda, M. and S. Gerlach (1993) Exchange rate dynamics and currency unification: the Ostmark–DM rate. *Empirical Economics* **18**, 417–29.
Byers, J. and D. Peel (1996) Long-memory risk premia in exchange rates. *Manchester School* **64**, 421–38.
Campa, J. and L. Goldberg (2002) Exchange rates pass-through into import prices: a macro or micro-phenomenon? NBER Working Paper 8934.
Caner, M. and B. Hansen (2001) Threshold autoregression with a unit root. *Econometrica* **69**, 1555–96.
Caner, M. and L. Kilian (2001) Size distortions of tests of the null hypothesis of stationarity: evidence and implications for the PPP debate. *Journal of International Money and Finance* **20**, 639–57.
Canjels, E., G. Prakash-Canjels and A. Taylor (2004) Measuring market integration, foreign exchange arbitrage and the gold standard, 1879–1913. NBER Working Paper 10583.
Chao, J., V. Corradi and N. Swanson (2001) Out-of-sample tests for Granger causality. *Macroeconomic Dynamics* **5**, 598–620.
Cheung, Y.-W. and M. Chinn (2001) Currency traders and exchange rate dynamics: a survey of the US market. *Journal of International Money and Finance* **20**, 439–71.
Chinn, M. and R. Meese (1995) Banking on currency forecasts: how predictable is change in money? *Journal of International Economics* **38**, 161–78.
Chung, C.-S. and G. Tauchen (2001) Testing target-zone models using efficient method of moments. *Journal of Business and Economic Statistics* **19**, 255–69.
Clarida, R., L. Sarno, M. Taylor and G. Valente (2003) The out-of-sample success of term structure models as exchange rate predictors: a step beyond. *Journal of International Economics* **60**, 61–83.
Clarida, R. and M. Taylor (1997) The term structure of forward exchange premiums and the forecastability of spot exchange rates: correcting the errors. *Review of Economics and Statistics* **79**, 353–61.
Clark, T. and M. McCracken (2001) Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105**, 85–110.
Clark, T. and M. McCracken (2003) Evaluating long horizon forecasts. Manuscript, Federal Reserve Bank of Kansas City and University of Missouri – Columbia.

Clark, T. and M. McCracken (2005) The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics* **124**, 1–31.
Coakley, J. and A. Fuertes (2001) Rethinking the forward premium puzzle in a nonlinear framework. Mimeo, University of Warwick.
Copeland, L. and S. Heravi (2006) Structural breaks in the real exchange rate adjustment mechanism. Cardiff Economics Working Papers E2006/21.
Dacco, R. and S. Satchell (1999) Why do regime-switching models forecast so badly? *Journal of Forecasting* **18**, 1–16.
Davidson, A. and D. Hinkley (1997) *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.
Davidson, R. and E. Flachaire (2001) The wild bootstrap, tamed at last. Queen's Institute for Economic Research Working Paper No. 1000.
Davidson, R. and J. MacKinnon (1993) *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Davies, R. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 179–90.
De Grauwe, P., H. Dewachter and M. Embrechts (1993) *Exchange Rate Theory Chaotic Models of Foreign Exchange Markets*. Oxford: Blackwell.
De Grauwe, F. and I. Vansteenkiste (2001) Exchange rates and fundamentals. CESifo Discussion Papers No. 577, Munich.
De Jong, R., C.-H. Wang and Y. Bae (2007) Correlation robust threshold unit root tests. Working Paper, Department of Economics, Ohio State University.
Diba, B. and H. Grossman (1988) Explosive rational bubbles in stock prices. *American Economic Review* **78**, 520–30.
Diebold, F., S. Husted and M. Rush (1991) Real exchange rates under the gold standard. *Journal of Political Economy* **99**, 1252–71.
Diebold, F. and A. Inoue (2001) Long memory and regime switching. *Journal of Econometrics* **105**, 131–59.
Diebold, F. and R. Mariano (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253–63.
Drost, F. and T. Nijman (1993) Temporal aggregation of GARCH processes. *Econometrica* **4**, 909–27.
Duarte, A., I. Venetis and I. Paya (2005) Predicting real growth and the probability of recession in the Euro-area using the yield spread. *International Journal of Forecasting* **21**, 261–77.
Duffie, D. and K. Singleton (1993) Simulated moments estimation of Markov models of asset prices. *Econometrica* **61**, 929–52.
Dumas, B. (1992) Dynamic equilibrium and the real exchange rate in spatially separated world. *Review of Financial Studies* **5**, 153–80.
Dwyer, G., P. Locke and W. Yu (1996) Index arbitrage and nonlinear dynamics between the S&P 500 futures and cash. *Review of Financial Studies* **9**, 301–32.
Einzig, P. (1937) *The Theory of Forward Exchange*. London: Macmillan.
Eitrheim, Ø. and T. Teräsvirta (1996) Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics* **74**, 59–75.
Enders, W. and C. Granger (1998) Unit root tests and asymmetric adjustment with an example using the term structure of interest rates. *Journal of Business and Economic Statistics* **16**, 304–11.
Engel, C. and J. Hamilton (1990) Long swings in the dollar: are they in the data and do markets know it? *American Economic Review* **80**, 689–713.
Engel, C., N. Mark and K. West (2007) Exchange rate models are not as bad as you think. NBER Working Paper 13318.
Engel, C. and K. West (2005) Exchange rates and fundamentals. *Journal of Political Economy* **133**, 485–517.

Engle, R. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.
Escribano, A. and O. Jordá (1999) Improved testing and specification of smooth transition regression models. In P. Rothman (ed.), *Nonlinear Time Series Analysis of Economic and Financial Data*. Dordrecht: Kluwer Academic Publishers.
Evans, G. (1986) A test for speculative bubbles in the sterling–dollar exchange rate: 1981–84. *American Economic Review* **76**, 621–36.
Evans, G. (1991) Pitfalls in testing for explosive bubbles in asset prices. *American Economic Review* **81**, 922–30.
Evans, M. and K. Lewis (1995) Do long-term swings in the dollar affect estimates of the risk premia? *Review of Financial Studies* **3**, 709–42.
Evans, M. and R. Lyons (2005) Meese–Rogoff redux: micro-based exchange rate forecasting. *American Economic Review* **95**, 405–14.
Fair, R. (1970) The estimation of simultaneous equation models with lagged endogenous variables and first order serially correlated errors. *Econometrica* **38**, 507–16.
Fama, E.F. (1984) Forward and spot exchange rates. *Journal of Monetary Economics* **14**, 319–38.
Faust, J., J. Rogers and J. Wright (2003) Exchange rate forecasting: the errors we've really made. *Journal of International Economics* **60**, 35–59.
Ferson, W., S. Sarkissian and T. Simin (2003) Spurious regressions in financial economics? *Journal of Finance* **58**, 1393–414.
Flandreau, M. and T. Komlos (2003) Target zones in history and theory: lessons from an Austro Hungarian experiment (1896–1914). Mimeo, University of Munich.
Flood, R. and P. Garber (1983) A model of stochastic process switching. *Econometrica* **3**, 537–51.
Floor, R. and A. Rose (1996) Fixes: of the forward discount puzzle. *Review of Economics and Statistics* **4**, 748–52.
Frankel, J. (1979) On the mark: a theory of floating exchange rate based on real interest differentials. *American Economic Review* **69**, 610–22.
Frankel, J. (1996) How well do foreign exchange markets work: might a tobin tax help? In M. ul Haq, I. Kaul and I. Grunberg (eds.), *The Tobin Tax: Coping with Financial Volatility*. New York and Oxford: Oxford University Press.
Frankel, J. and K. Froot (1987) Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* **1**, 133–53.
Frankel, J. and A. Rose (1996) A panel project on purchasing power parity: mean reversion within and between countries. *Journal of International Economics* **40**, 209–24.
Frömmel, M., R. MacDonald and L. Menkhoff (2005) Markov switching regimes in a monetary exchange rate model. *Economic Modelling* **22**, 485–502.
Froot, K. and M. Obstfeld (1991) Intrinsic bubbles: the case of stock prices. *American Economic Review* **81**, 1189–214.
Gallant, A., P. Rossi and G. Tauchen (1993) Nonlinear dynamic structures. *Econometrica* **61**, 871–908.
Gallant, A. and G. Tauchen (1996) Which moments to match? *Econometric Theory* **12**, 65–81.
Gallant, A. and G. Tauchen (2000) EMM: a program for effcient method of moments estimation. Technical Report, Duke University.
Gonçalves, S. and L. Kilian (2003) Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. CIRANO Working Paper 2003s-17.
Granger, C. (1980) Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* **14**, 227–38.
Granger, C. and N. Hyung (2004) Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* **11**, 399–421.

Granger, C. and R. Joyeux (1980) An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.
Granger, C. and P. Newbold (1974) Spurious regressions in econometrics. *Journal of Econometrics* **2**, 111–20.
Granger, C. and T. Teräsvirta (1993) *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
Granger, C. and T. Teräsvirta (1999) A simple nonlinear time series model with misleading linear properties. *Economics Letters* **62**, 161–5.
Groen, J. (2000) The monetary exchange rate model as a long-run phenomenon. *Journal of International Economics* **52**, 299–319.
Groen, J. (2005) Exchange rate predictability and monetary fundamentals in a small multi-country panel. *Journal of Money, Credit and Banking* **37**, 495–516.
Groen, J. and F. Kleibergen (2003) Likelihood-based cointegration analysis in panels of vector error correction models. *Journal of Business and Economic Statistics* **21**, 295–318.
Gurkaynak, R. (2005) Econometric tests of asset price bubbles: taking stock. Finance and Economics Discussion Series. Division of Monetary Affairs Board of Governors of the Federal Reserve System.
Hai, W., N. Mark and Y. Wu (1997) Understanding spot and forward exchange rate regressions. *Journal of Applied Econometrics* **12**, 715–36.
Hamilton, J. (1990) Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**, 39–70.
Hamilton, J. (1994) *Time Series Analysis*. Princeton: Princeton University Press.
Hansen, B. (1997) Inference in tar models. *Studies in Nonlinear Dynamics and Econometrics* **2**, 1–14.
Hansen, L. (1982) Large sample properties of generalised method of moments estimators. *Econometrica* **50**, 1029–54.
Hansen, L. and R. Hodrick (1980) Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. *Journal of Political Economy* **5**, 829–53.
Harris, D., B. McCabe and S. Leybourne (2003) Some limit theory for autocovariances whose order depend on sample size. *Econometric Theory* **19**, 829–64.
Harvey, D. and S. Leybourne (2007) Testing for time series linearity. *Econometrics Journal* **10**, 149–65.
Harvey, D., S. Leybourne and P. Newbold (1998) Tests for forecast encompassing. *Journal of Business and Economic Statistics* **16**, 254–59.
Haug, A. and S. Basher (2005) Unit roots, nonlinear cointegration and purchasing power parity. Econometrics 0401006, EconWPA.
Hegwood, N. and D. Papell (2002) Purchasing power parity under the gold standard. *Southern Economic Journal* **69**, 72–91.
Hodrick, R. (1987) *The Empirical Evidence on the Efficiency of the Forward and Futures Foreign Exchange Market*. London: Harwood.
Hodrick, R. (1989) Risk, uncertainty, and exchange rates. *Journal of Monetary Economics* **23**, 433–59.
Horvath, M. and M. Watson (1995) Testing for cointegration when some of the cointegrating vectors are prespecified. *Econometric Theory* **11**, 984–1014.
Huisman, R., K. Koedijk, J. Clemens and F. Kool (2001) Tail-index estimates in small samples. *Journal of Business and Economic Statistics* **19**, 208–16.
Iannizzotto, M. and M. Taylor (1999) The target zone model, non-linearity and mean reversion: is the honeymoon really over? *Economic Journal* **109**, C96–110.
Kapetanios, G. (1999) Model selection in threshold models. *Journal of Time Series Analysis* **22**, 733–54.
Kapetanios, G., Y. Shin and A. Snell (2003a) Testing for a unit root in the nonlinear STAR framework. *Journal of Econometrics* **112**, 359–79.

Kapetanios, G., Y. Shin and A. Snell (2003b) Testing for cointegration in nonlinear STAR error correction models. Working Paper No. 497, Queen Mary, University of London.
Keynes, J. (1923) *A Tract on Monetary Reform*. London: Macmillan.
Kilian, L. (1999) Exchange rates and monetary fundamentals: what do we learn from long-horizon regressions? *Journal of Applied Econometrics* **14**, 491–510.
Kiliç, R. (2003) A testing procedure for a unit root in the STAR model. Working Paper, School of Economics, Georgia Institute of Technology.
Kim, C.-J. and C. Nelson (1999) *State-Space Models with Regime Switching: Classical and Gibbs Sampling Approaches with Applications*. Cambridge, Mass.: MIT Press.
Koedijk, K., M. Schafgans and C. De Vries (1990) The tail index of exchange rate returns. *Journal of International Economics* **29**, 93–108.
Koop, G., H. Pesaran and S. Potter (1996) Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* **74**, 119–47.
Krugman, P. (1991) Target zones and exchange rate dynamics. *Quarterly Journal of Economics* **106**, 669–82.
Kwiatkowski, D., P. Phillips, P. Schmidt and Y. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* **54**, 159–78.
Lee, B. and B. Ingram (1991) Simulation estimation of time-series models. *Journal of Econometrics* **47**, 197–205.
Lee, H.-Y. and S.-L. Chen (2006) Why use Markov-switching models in exchange rate prediction? *Economic Modelling* **23**, 662–8.
Leon, H., L. Sarno and G. Valente (2003) Limits to speculation and nonlinearity in deviations from uncovered interest parity: empirical evidence and implications for the forward bias puzzle. Mimeo, IMF.
Leybourne, S. and B. McCabe (1994) A consistent test for a unit root. *Journal of Business and Economic Statistics* **12**, 157–66.
Leybourne, S., B. McCabe and A. Tremayne (1996) Can economic time series be differenced to stationarity. *Journal of Business and Economic Statistics* **14**, 435–46.
Liu, R. (1988) Bootstrap procedure under some non i.i.d. models. *Annals of Statistics* **16**, 1696–708.
Loretan, M. and P. Phillips (1994) Testing the covariance structure of heavy-tailed time series. *Journal of Empirical Finance* **1**, 211–48.
Lothian, J. (1997) Multi-country evidence on the behavior of purchasing power parity under the current float. *Journal of International Money and Finance* **16**, 19–35.
Lothian, J. and M. Taylor (1996) Real exchange rate behaviour: the recent float from the perspective of the past two centuries. *Journal of Political Economy* **104**, 488–509.
Lothian, J. and M. Taylor (2000) Purchasing power parity over two centuries: strengthening the case for real exchange rate stability: a reply to Cuddington and Liang. *Journal of International Money and Finance* **19**, 759–64.
Lucas, R. (1982) Interest rates and currency price in a two-country world. *Journal of Monetary Economics* **10**, 335–59.
Lundbergh, S. and T. Teräsvirta (1998) Modelling economic high frequency time series with STAR–STGARCH models. Stockholm School of Economics, Working Paper Series in Economics and Finance No. 291.
Lundbergh, S. and T. Teräsvirta (2006) A time series model for an exchange rate in a target zone with applications. *Journal of Econometrics* **131**, 579–609.
Luukkonen, R., P. Saikkonen and T. Teräsvirta (1988) Testing linearity against smooth transition autoregressive model. *Biometrika* **75**, 491–99.
Lux, T. and D. Sornette (2002) On rational bubbles and fat tails. *Journal of Money, Credit and Banking* **34**, 589–610.
Lyons, R. (2001) *The Microstructure Approach to Exchange Rates*. Cambridge, Mass.: MIT Press.

MacDonald, R. (1993) Long-run purchasing power parity: is it for real? *Review of Economics and Statistics* **75**, 690–5.
MacDonald, R. and M. Taylor (1994) The monetary model of the exchange rate: long-run relationships, short-run dynamics and how to beat a random walk. *Journal of International Money and Finance* **13**, 276–90.
MacKinnon, J. and H. White (1985) Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–25.
Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics* **21**, 255–85.
Mark, N. (1990) Real and nominal exchange rates in the long run. *Journal of International Economics* **28**, 115–36.
Mark, N. (1995) Exchange rates and fundamentals: evidence on long-horizon predictability. *American Economic Review* **85**, 201–18.
Mark, N. and D. Sul (2001) Nominal exchange rates and monetary fundamentals: evidence from a small post-Bretton Woods panel. *Journal of International Economics* **53**, 29–52.
Maynard, A. and P. Phillips (2001) Rethinking an old empirical puzzle: econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics* **6**, 671–708.
McCallum, B. (1994) A reconsideration of the uncovered interest parity relationship. *Journal of Monetary Economics* **33**, 105–32.
McCracken, M. (1999) Asymptotics for out-of-sample tests of causality. Unpublished manuscript, Department of Economics, Louisiana State University.
McCracken, M. and S. Sapp (2005) Evaluating the predictability of exchange rates using long-horizon regressions: mind your p's and q's! *Journal of Money, Credit and Banking* **37**, 473–94.
Meese, R. (1986) Testing for bubbles in exchange markets: a case of sparkling rates? *Journal of Political Economy* **94**, 345–73.
Meese, R. and K. Rogoff (1983a) Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics* **14**, 3–24.
Meese, R. and K. Rogoff (1983b) The out-of-sample failure of empirical exchange rate models: sampling error or misspecification? In J. Frenkel (ed.), *Exchange Rates and International Macroeconomics*. Chicago: University of Chicago Press.
Meese, R. and A.K. Rose (1990) Nonlinear, nonparametric, nonessential exchange rate estimation. *American Economic Review* **80**, 192–96.
Michael, P., A. Nobay and D. Peel (1997) Transactions costs and nonlinear adjustment in real exchange rates: an empirical investigation. *Journal of Political Economy* **105**, 862–79.
Miller, M. and P. Weller (1991) Exchange rate bands with price inertia. *Economic Journal* **101**, 1380–99.
Minford, A. and D. Peel (2002) *Advanced Macroeconomics: A Primer*. Cheltenham: Edward Elgar.
Minford, A. and D. Peel (2007) On the equality of real interest rates across borders in integrated capital markets. *Open Economies Review* **18**, 119–25.
Monoyios, M. and L. Sarno (2002) Mean reversion in stock index futures markets: a nonlinear analysis. *Journal of Futures Markets* **22**, 285–314.
Moore, M. and L. Copeland (1995) A comparison of Johansen and Phillips–Hansen cointegration tests of forward market efficiency, Baillie and Bollerslev revisited. *Economics Letters* **47**, 131–5.
Neely, C. and L. Sarno (2002) How well do monetary fundamentals forecast exchanges rates? *Federal Reserve Bank of St. Louis Review* **84**, 51–74.
Obstfeld, M. and K. Rogoff (1996) *Foundations of International Macroeconomics*. Cambridge, Mass.: MIT Press.

Obstfeld, M. and A. Taylor (1997) Nonlinear aspects of goods-market arbitrage and adjustment: Hecksher's commodity points revisited. *Journal of the Japanese and International Economies* **11**, 441–79.

O'Connell, P. and S. Wei (1997) The bigger they are, the harder they fall: how price differences between US cities are arbitraged. Discussion Paper, Department of Economics, Harvard University.

Ohanissian, A., J. Russell and R. Tsay (2007) True or spurious long memory? A new test. *Journal of Business and Economic Statistics*. Forthcoming.

Ozaki, T. (1978) Non-linear models for non-linear random vibrations. Technical Report, Department of Mathematics, UMIST.

Pavlidis, E., I. Paya and D. Peel (2007) Linearity testing in the presence of heteroskedasticity. In *ESRC Seminar Series: Nonlinear Economics and Finance Research Community*, Brunel.

Paya, I. and D. Peel (2003) Purchasing power parity adjustment speeds in high frequency data when the equilibrium real exchange rate is proxied by a deterministic trend. *Manchester School* **71**, 39–53.

Paya, I. and D. Peel (2004) Real exchange rates under the gold standard: nonlinear adjustments. *Southern Economic Journal* **71**, 302–13.

Paya, I. and D. Peel (2006a) A new analysis of the determinants of the real dollar–sterling exchange rate: 1871–1994. *Journal of Money, Credit and Banking* **38**, 1971–90.

Paya, I. and D. Peel (2006b) On the speed of adjustment in ESTAR models when allowance is made for bias in estimation. *Economics Letters* **90**, 272–7.

Paya, I. and D. Peel (2006c) Temporal aggregation of an ESTAR process: some implications for purchasing power parity adjustment. *Journal of Applied Econometrics* **21**, 655–68.

Paya, I. and D. Peel (2007a) On the relationship between nominal exchange rates and domestic and foreign prices. *Applied Financial Economics* **17**, 105–17.

Paya, I. and D. Peel (2007b) Systematic sampling of a nonlinear model. Mimeo, Lancaster University.

Paya, I., I. Venetis and D. Peel (2003) Further evidence on PPP adjusment speeds: the case of effective real exchange rates and the EMS. *Oxford Bulletin of Economics and Statistics* **65**, 421–38.

Peel, D. and J. Davidson (1998) A non-linear error correction mechanism based on the bilinear model. *Economics Letters* **2**, 165–70.

Peel, D. and M. Taylor (2002) Covered interest arbitrage in the interwar period and the Keynes–Einzig conjecture. *Journal of Money, Credit and Banking* **34**, 51–75.

Peel, D. and I. Venetis (2005) Smooth transition models and arbitrage consistency. *Economica* **72**, 413–30.

Pesaran, H. and A. Timmermann (1995) Predictability of stock returns: robustness and economic significance. *Journal of Finance* **50**, 1201–28.

Phillips, P. (1991) Shortcut to LAD estimator asymptotics. *Econometric Theory* **4**, 450–63.

Phillips, P. (1995) Robust nonstationary regression. *Econometric Theory* **5**, 912–51.

Phillips, P. and B. Hansen (1990) Statistical inference in instrumental variables regression with $I(1)$ processes. *Review of Economic Studies* **57**, 99–125.

Phillips, P. and T. Magdalinos (2007) Limit theory for moderate deviations from unity under weak dependence. In G. Phillips and E. Tzavalis (eds.), *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*. Cambridge: Cambridge University Press.

Phillips, P., J. McFarland and P. McMahon (1996) Robust tests of forward exchange market efficiency with empirical evidence from the 1920s. *Journal of Applied Econometrics* **1**, 1–22.

Phillips, P., Y. Wu and J. Yu (2006) Explosive behavior and the Nasdaq bubble in the 1990s: when did irrational exuberance escalate asset values? Mimeo, Yale University.

Pippenger, M. and G. Goering (1993) A note on the empirical power of unit root tests under threshold processes. *Oxford Bulletin of Economics and Statistics* **55**, 473–81.
Pötscher, B. and I. Prucha (1997) *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. New York: Springer-Verlag.
Rapach, D. and M. Wohar (2004) Testing the monetary model of exchange rate determination: a closer look at panels. *Journal of International Money and Finance* **23**, 867–95.
Rogoff, K. (1996) The purchasing power parity puzzle. *Journal of Economic Literature* **34**, 647–68.
Rogoff, K. (1999) Monetary models of dollar/yen/euro nominal exchange rates: dead or undead? *Economic Journal* **109**, F655–9.
Roll, R. (1979) Violations of purchasing power parity and their implications for efficient international commodity markets. In G.S.S. Marshall (ed.), *International Finance and Trade, Volume 1*, pp. 133–76. Cambridge, Mass.: Ballinger.
Rossana, R. and J. Seater (1995) Temporal aggregation and economic time series. *Journal of Business and Economic Statistics* **13**, 441–51.
Rossi, B. (2005) Testing long-horizon predictive ability with high persistence, and the Meese–Rogoff puzzle. *International Economic Review* **46**, 61–92.
Rossi, B. (2006) Are exchange rates really random walks? Some evidence robust to parameter instability. *Macroeconomic Dynamics* **10**, 20–38.
Saikkonen, P. (1991) Asymptotically efficient estimation of cointegration regressions. *Econometric Theory* **7**, 1–21.
Salge, M. (1997) *Rational Bubbles*. Berlin: Springer-Verlag.
Samuelson, P. (1964) Theoretical notes on trade problems. *Review of Economics and Statistics* **46**, 145–54.
Sarantis, N. (1999) Modelling nonlinearities in real effective exchange rates. *Journal of International Money and Finance* **18**, 27–45.
Sarno, L. and M. Taylor (1998) Real exchange rates under the recent float: unequivocal evidence of mean reversion. *Economics Letters* **60**, 131–7.
Sarno, L. and M. Taylor (2002) *The Economics of Exchange Rates*. Cambridge and New York: Cambridge University Press.
Sarno, L., M. Taylor and I. Chowdhury (2004a). Nonlinear dynamics in deviations from the law of one price. *Journal of International Money and Finance* **23**, 1–25.
Sarno, L. and G. Valente (2005) Exchange rates and fundamentals: footloose or evolving relationship? In *Exchange Rate Determinants and Economic Impacts*. Frankfurt: Joint Workshop of the European Central Bank and the Bank of Canada.
Sarno, L., G. Valente and H. Leon (2006) Nonlinearity in deviations from uncovered interest parity: an explanation of the forward bias puzzle. *Review of Finance* **10**, 443–82.
Sarno, L., G. Valente and M. Wohar (2004b) Monetary fundamentals and exchange rate dynamics under different nominal regimes. *Economic Inquiry* **42**, 179–93.
Sercu, P., R. Uppal and C. Van Hulle (1995) The exchange rate in the presence of transaction costs: implications for tests of purchasing power parity. *Journal of Finance* **50**, 1309–19.
Shin, Y. (1994) A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory* **10**, 91–115.
Shleifer, A. and R. Vishny (1997) The limits of arbitrage. *Journal of Finance* **52**, 35–55.
Smallwood, A. (2005) Joint tests for non-linearity and long memory: the case of purchasing power parity. *Studies in Nonlinear Dynamics and Econometrics* **9**, 1–28.
Spagnola, F., Z. Psaradakis and M. Sola (2005) Testing the unbiased forward exchange rate hypothesis using a Markov switching model and instrumental variables. *Journal of Applied Econometrics* **20**, 423–37.
Stock, J. and M. Watson (1993) A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* **4**, 783–820.

Storey, J. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–35.
Taylor, A. (2001) Potential pitfalls for the purchasing-power parity puzzle? Sampling and specification biases in mean-reversion tests of the law of one price. *Econometrica* **69**, 473–98.
Taylor, M. (1987) Covered interest parity: a high-frequency, high-quality data study. *Economica* **54**, 429–38.
Taylor, M. (1988) An empirical examination of long run purchasing power parity using cointegration techniques. *Applied Economics* **20**, 1369–81.
Taylor, M. (1989) Covered interest arbitrage and market turbulence. *Economic Journal* **99**, 376–91.
Taylor, M. (1995) The economics of exchange rates. *Journal of Economic Literature* **33**, 13–47.
Taylor, M. and M. Iannizzotto (2001) On the mean-reverting properties of target zone exchange rates: a cautionary note. *Economics Letters* **71**, 117–29.
Taylor, M. and L. Kilian (2003) Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics* **60**, 85–107.
Taylor, M. and D. Peel (2000) Nonlinear adjustment, long-run equilibrium and exchange rate fundamentals. *Journal of International Money and Finance* **19**, 33–53.
Taylor, M., D. Peel and L. Sarno (2001) Nonlinear mean-reversion in real exchange rates: toward a solution to the purchasing power parity puzzles. *International Economic Review* **42**, 1015–42.
Taylor, M., D. Peel and L. Sarno (2004) International real interest rate differentials, purchasing power parity and the behaviour of real exchange rates: the resolution of a conundrum. *International Journal of Finance* **9**, 15–23.
Taylor, N., D. Van Dijk, P. Franses and A. Lucas (2000) SETS, arbitrage activity, and stock price dynamics. *Journal of Banking and Finance* **24**, 1289–306.
Teräsvirta, T. (1994) Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**, 208–18.
Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag.
Tsay, W. and W. Härdle (2007) A generalized ARFIMA process with Markov-switching fractional differencing parameter. SFB 649, Discussion Paper 2007-022.
Van Dijk, D., T. Teräsvirta and P. Frances (2002) Smooth transition autoregressive models – a survey of recent developments. *Econometrics Reviews* **21**, 1–47.
Venetis, I., I. Paya and D. Peel (2005) Do real exchange rates "mean revert" to productivity? A nonlinear approach. Mimeo, Lancaster University.
Venetis, I., I. Paya and D. Peel (2007) Deterministic impulse response in a nonlinear model An analytic expression. *Economics Letters* **95**, 315–19.
Vogelsang, T. (1998) Trend function hypothesis testing in the presence of serial correlation. *Econometrica* **66**, 123–48.
Wagner, N. and Marsh, T.A. (2005) Measuring tail thickness under GARCH and an application to extreme exchange rate changes. *Journal of Empirical Finance* **12**, 165–85.
Wolff, C. (1987) Time-varying parameters and the out-of-sample forecasting performance of structural exchange rate models. *Journal of Business and Economic Statistics* **5**, 87–97.
Wooldridge, J. (1990) A unified approach to robust, regression-based specification tests. *Econometric Theory* **6**, 17–43.
Wooldridge, J. (1994) On the limits of GLM for specification testing: comment. *Econometric Theory* **10**, 409–18.
Wu, C. (1986) Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statististics* **14**, 1261–350.
Yoon, G. (2005) Some properties of periodically collapsing bubbles. Mimeo, Pusan National University and the University of York.

This page intentionally left blank

Part VIII
Growth/Development Econometrics

This page intentionally left blank

23
The Econometrics of Convergence

Steven N. Durlauf, Paul A. Johnson and Jonathan R.W. Temple

Abstract

The presence or absence of convergence between rich and poor countries represents one of the most important questions in the new growth economics. New growth theories have been explicitly designed to explain forms of divergence which do not appear in their neoclassical counterparts. Despite substantial empirical work on convergence, there is no consensus as to whether it is present in cross-country data. This chapter surveys the econometrics of convergence as well as the range of empirical claims that have appeared. Particular attention is given to the relationship between statistical versus economic notions of convergence. We argue that the disparities in claims across empirical studies can to some extent be understood as reflecting inadequate attention to the relationship between the statistical and economic notions.

23.1 Introduction

The question of whether nations converge or diverge has long been of interest to historians and social scientists. One classic passage of Edward Gibbon's *The Decline and Fall of the Roman Empire*[1] expresses the hope that all societies will exhibit progress in the long run:

> If, in the neighbourhood of the commercial and literary town of Glasgow, a race of cannibals has really existed, we may contemplate, in the period of Scottish

> history, the opposite extremes of savage and civilised life. Such reflections tend to enlarge the circle of our ideas; and to encourage the pleasing hope, that New Zealand may produce, in some future age, the Hume of the Southern Hemisphere. (book II, ch. XXV)

But another famous passage notes the permanent effects of particular events, arguing that, had the Franks not defeated the Arabs at the Battle of Poitiers in 732, "Perhaps the interpretation of the Koran would now be taught in the school of Oxford, and her pulpits might demonstrate to a circumcised people the sanctity and truth of the revelation of Mahomet" (book V, ch. LIII). Modern social science is neither as literary nor as broad in its sweep, yet Gibbon's remarks resonate with the broad questions involved in the study of convergence. Over long epochs, will initially different societies or economies evolve towards a common form, or will their initial conditions play a role in determining their long-run outcomes?

The nature of interest in convergence has varied over time, referring to different countries and different socioeconomic forms. From the vantage point of the Cold War, one debate concerned the extent of the differences between the West and the Soviet bloc. It was not uncommon to argue that capitalist and socialist economies would converge over time, as market institutions began to shape socialist economies, while capitalism might increasingly be accompanied by extensive government regulation and intervention, and a range of activist social welfare policies. More recently, a new convergence debate has focused on issues related to the persistence or transience of differences between rich and poor countries, with parallel interest in the income differences of regions, states, and districts. Convergence here is typically conceived in terms of narrowing differences in per capita income, rather than broader socioeconomic institutions. It is this literature that we will review in this chapter. The possibility of income convergence has been studied more intensively than any other hypothesis in growth economics, even if the effort to identify growth determinants, both proximate and fundamental, has become the main area of current empirical research.

Contemporary interest in the convergence hypothesis stems from at least three factors. The first is the enormously high levels of international inequality, and attendant levels of human suffering in poorer societies. This inequality means that the extent to which economies are converging or diverging, and at what rate, forms the background for almost any discussion of globalization and the work of international institutions and aid agencies. In popular commentary, it is often claimed that the gap between rich countries and poor countries is widening, and the study of convergence helps to evaluate such claims in a rigorous way.

Second, endogenous growth models, at least when based on increasing returns to scale, often imply long-run divergence of per capita incomes. Hence the investigation of convergence came to be seen – perhaps mistakenly, as we will discuss below – as the best way to discriminate between the new growth theories and their neoclassical predecessors. Finally, the greater availability of internationally comparable data for a broad cross-section of countries, primarily due to the work

of Summers and Heston (1988, 1991), has meant that it is now possible to study convergence for a wide range of countries.[2]

In this chapter, we first describe various statistical notions of convergence that have arisen as part of the modern economic growth literature. Our goal is not only to characterize the range of convergence notions that have been used in empirical work, but also to give some sense of their links to substantive economic claims.

23.2 β-convergence

The first statistical convergence concept used in the modern growth literature is based on the relationship between initial income and subsequent growth. Intuitively, this convergence notion is simple: two countries exhibit convergence if the one with lower initial income grows faster than the other and so tends to "catch up" with the higher-income country. This is the concept used in Abramovitz (1986), Baumol (1986) and Marris (1982), and which also plays a central role in Barro (1991), Barro and Sala-i-Martin (1992) and Mankiw, Romer and Weil (1992).[3] One posited driving force behind catching-up is that a position well below the technological frontier creates the potential for rapid advancement, through the installation of capital embodying the current frontier technology, for example. Another convergence mechanism, which is usually associated with the neoclassical growth model and which has played a greater role in the literature, emphasizes the role of diminishing returns. It predicts that countries which begin with a relatively low level of income will grow relatively rapidly, but this growth will slow down as the economy approaches its balanced growth path and the marginal product of capital declines towards its steady-state level.[4]

23.2.1 Convergence and the neoclassical growth model

Convergence as a property of the neoclassical growth model may be understood in terms of the behavior of output around the model's unique and stable steady-state. Let $Y_{i,t}$ denote output, $L_{i,t}$ the labor force, and $A_{i,t}$ the level of (labor-augmenting) efficiency in economy i at time t. From these, following the standard logic by which steady-states are constructed in the neoclassical growth model, define $y^E_{i,t} = Y_{i,t}/(A_{i,t}L_{i,t})$ as output per efficiency unit of labor input at any time t and $y^E_{i,\infty} = \lim_{t\to\infty} y^E_{i,t}$ as its associated stable steady-state value. Assuming that $y^E_{i,0} > 0$, a log-linear approximation around the stable steady-state implies the law of motion:

$$\log y^E_{i,t} = (1 - e^{-\lambda_i t}) \log y^E_{i,\infty} + e^{-\lambda_i t} \log y^E_{i,0}. \tag{23.1}$$

The parameter λ_i, which may be shown to be positive, depends on the other parameters of the model and characterizes the speed with which $y^E_{i,t}$ adjusts towards its steady-state value.[5]

Given this general law of motion for output per efficiency unit of labor, it is straightforward to describe the behavior of the observable output per unit of labor input, $y_{i,t} = Y_{i,t}/L_{i,t}$. Letting g_i be the (constant) rate of (labor-augmenting)

technological progress, so that $A_{i,t} = A_{i,0}e^{g_i t}$, equation (23.1) may be rewritten as:

$$\log y_{i,t} - g_i t - \log A_{i,0} = (1 - e^{-\lambda_i t}) \log y^E_{i,\infty} + e^{-\lambda_i t}(\log y_{i,0} - \log A_{i,0}) \qquad (23.2)$$

so that:

$$\log y_{i,t} = g_i t + (1 - e^{-\lambda_i t}) \log y^E_{i,\infty} + (1 - e^{-\lambda_i t}) \log A_{i,0} + e^{-\lambda_i t} \log y_{i,0}.^6 \qquad (23.3)$$

Letting $\gamma_i = t^{-1}\left(\log y_{i,t} - \log y_{i,0}\right)$ denote the growth rate of $y_{i,t}$ between 0 and t and subtracting $\log y_{i,0}$ from both sides of (23.3), division by t yields:

$$\gamma_i = g_i + \beta_i(\log y_{i,0} - \log y^E_{i,\infty} - \log A_{i,0}), \qquad (23.4)$$

where $\beta_i = -t^{-1}(1 - e^{-\lambda_i t})$.

Equation (23.4) decomposes the growth rate of $y_{i,t}$ into two parts. The first, g_i, is growth due to technological progress, and the second, $\beta_i(\log y_{i,0} - \log y^E_{i,\infty} - \log A_{i,0})$ is growth due to the closing of the initial gap between output per worker and the steady-state value – "catching up." Because $\beta_i \to 0$ as $t \to \infty$, the importance of this term, and hence the role of initial conditions in determining contemporaneous output, diminishes to zero. The long-run rate of growth of the economy is g_i, the rate of technological progress.

While equation (23.4) provides a characterization of the sources of economic growth, it is not yet in a form that can be estimated. One reason is that the parameters λ_i and g_i are country-specific. The empirical growth literature often assumes that these parameters are identical across countries, so that we have $\lambda_i = \lambda$ and $g_i = g$, $\forall i$. Under these assumptions, (23.4) simplifies to:

$$\gamma_i = g - \beta \log y^E_{i,\infty} - \beta \log A_{i,0} + \beta \log y_{i,0}. \qquad (23.5)$$

Equation (23.5) implies, *ceteris paribus*, a negative relationship between average rates of growth and initial levels of output per capita, over any time period, when estimated for a cross-section of countries. Those countries with low income are further below their balanced growth path and will grow relatively quickly: their low income implies that the capital-output ratio is lower, and the marginal product of capital higher, than in countries starting with a higher level of income. This mechanism leads to a period of relatively fast growth, so that the countries initially behind will catch up with other countries that have the same levels of steady-state output per effective worker and initial efficiency. Similarly, countries that begin above their balanced growth path, perhaps because some determinants of steady-state income have deteriorated over time, must grow relatively slowly. In these economies, the capital-output ratio is high and the marginal product of capital relatively low, leading to a period of growth at below the rate of technical progress. This movement towards a balanced growth path is the economic notion of convergence implied by the neoclassical model.

23.2.2 Cross-country regressions and β-convergence

In order to translate this economic notion of convergence into a statistical model, it is necessary to address the unobservable variables $y^E_{i,\infty}$ and $A_{i,0}$. One possibility is to assume that (i) all countries have a common steady-state so that $y^E_{i,\infty}$ is a constant, y^E_{∞}; and (ii) $\log A_{i,0} = \log A + \varepsilon_i$ where, ε_i is a country-specific shock.[7] Subsuming y^E_{∞}, $\log A$ and g into the constant term, unconditional β-convergence holds if $\beta < 0$ in the regression:

$$\gamma_i = \alpha + \beta y_{i,0} + \varepsilon_i. \tag{23.6}$$

When using a sample of countries, one rarely finds unconditional β-convergence, unless the sample is restricted to similar entities such as the Organization for Economic Cooperation and Development (OECD) member countries. There is rarely a strong correlation between initial income and subsequent growth in large, heterogeneous samples, such as those found in the Penn World Table.[8]

Relative to equation (23.6), heterogeneity in countries naturally suggests that $y^E_{i,\infty}$ is not, in fact, constant. In the augmented Solow model estimated by Mankiw *et al.* (1992), for example, $y^E_{i,\infty}$ is shown to depend on the rates of physical and human capital accumulation and the rate of population growth. More generally, letting Z_i be the set of variables that determine a country's steady-state income level, conditional β-convergence is said to hold if $\beta < 0$ in the regression:

$$\gamma_i = \alpha + \beta y_{i,0} + Z_i\Gamma + \varepsilon_i. \tag{23.7}$$

While many differences exist in the choice of controls, most studies include several determinants inspired by some version of the Solow growth model. Unlike unconditional β-convergence, evidence of conditional β-convergence has been found in many contexts. For the cross-country case, the finding is generally attributed to Barro (1991), Barro and Sala-i-Martin (1992) and Mankiw *et al.* (1992).

The Mankiw *et al.* (1992) analysis is of particular interest, as it is based on a regression suggested by the dynamics of the Solow growth model. Their findings have been widely interpreted as evidence in favor of diminishing returns to capital (the source of $\beta < 0$ in the Solow model) and as evidence against some endogenous growth models. The analysis especially calls into question models which emphasize increasing returns in capital accumulation (either human or physical) as a source of perpetual growth. However, some endogenous growth models are consistent with β-convergence, and therefore some caution is needed in drawing inferences about the nature of the growth process from the results of β-convergence tests.[9]

Given an estimate of β, an estimate of λ, the implied rate of convergence, can be obtained from $\beta = -t^{-1}(1 - e^{-\lambda t})$. In cross-section studies, this typically yields an estimated convergence rate of about 2% per year; in other words, over the course of a year, countries will close just 2% of the gap between their current position and the balanced growth path.[10] This result is found using datasets on a wide variety of economic entities and from time periods more than 100 years apart.[11] While

some have (perhaps with a degree of irony) accorded the 2% value a status akin to that of a universal constant in physics, others have been rather more skeptical of its generality. There is nothing in the logic of the neoclassical growth model to suggest that this parameter should be invariant across environments, and this in itself might lead to skepticism about the supposed universality of the 2% figure. On the other hand, there are reasons to believe that the time series structure of per capita output data could produce a "universal" convergence estimate because of inadequate attention to temporal dependence in the data. Quah (1996b), for example, suggests that the 2% finding may be a statistical artifact that arises for reasons unrelated to convergence *per se*. Specifically, he argues that, if each per capita output series in a cross-section regression contains a unit root, this can produce a β estimate such that a 2% convergence rate is produced, even if the series are independent. Quah's argument is important in motivating the importance of time series approaches to evaluating convergence, which are discussed later.

23.2.3 Critiques

The evidence of convergence obtained from cross-country growth regressions has been subjected to a number of criticisms. Here we focus on those that seem most important (see Durlauf, Johnson and Temple, 2005, for a more complete treatment).

The first criticism relates to the choice of control variables, and whether claims about convergence are robust to alternative choices. Any claims about conditional convergence, in particular, necessarily depend on a specific choice for the set of control variables Z_i. This is a serious concern, given the lack of consensus in growth economics about which growth determinants are empirically important. This lack of consensus is reflected in the "growth regression industry" that has arisen, as researchers have added a range of controls, of varying degrees of plausibility, to those of the basic Solow model.[12]

The most conceptually satisfying response to this problem has been the use of model averaging methods, as in Fernandez, Ley and Steel (2001) and Sala-i-Martin, Doppelhofer and Miller (2004). The model averaging approach constructs estimates (or posterior means, depending on whether one is evaluating convergence along frequentist or Bayesian lines) based upon a model space of candidate growth regressions. Information on convergence in each model is aggregated with weights corresponding to posterior model probabilities. These studies show that the cross-country finding of conditional β-convergence is robust to the choice of controls. Both studies conclude that the posterior probability that initial income is part of the linear growth model is high. The Sala-i-Martin *et al.* study reports a posterior expected value for the β regression parameter of -0.0085, implying an estimated convergence rate of 1% per year.

A second critique, which is perhaps better called a class of critiques, is that there are many good reasons to believe that the model errors in growth regressions are correlated with the associated regressors, leading to inconsistent estimates. One can understand a number of developments in the econometrics of β-convergence as efforts to address this problem.

A first source of correlation derives from country-specific heterogeneity. Cross-section growth regressions assume that the error terms ε_i are uncorrelated with the $\log y_{i,0}$ terms, but this is unlikely to hold if there is country-specific unobserved heterogeneity in output levels. If such effects were present, they would typically imply a link between ε_i and $y_{i,0}$. For this reason, a number of researchers have investigated convergence using panel data. This leads to models of the form:

$$g_{i,t} = c_i + y_{i,t-1}\beta + Z_{i,t}\gamma + \varepsilon_{i,t} \tag{23.8}$$

where growth is now measured between $t-1$ and t. This approach means that individual ("fixed") effects can be used to control for (time-invariant) unobserved heterogeneity. In practice, this has often been supplemented with the use of instrumental variables to address the endogeneity of variables like investment rates. Panel analyses have been conducted by Bond, Hoeffler and Temple (2001), Caselli, Esquivel and Lefort (1996), Islam (1995, 1998) and Lee, Pesaran and Smith (1997, 1998), among others. These studies generally find that conditional convergence takes place at higher rates than estimated in the cross-section studies. For example, Caselli *et al.* (1996) report annual convergence rate estimates of 10%.

As discussed in Durlauf and Quah (1999) and Durlauf *et al.* (2005), panel data approaches to convergence suffer from the problem that, once country-specific effects are allowed, it becomes harder to interpret the results in terms of specific economic explanations. One problem is that, once the growth model includes individual effects, then the question of convergence is changed, at least if the goal is to understand whether initial conditions matter; simply put, the country-specific effects may partly reflect the effects of varying initial conditions. When studies such as Lee *et al.* (1997, 1998) allow for rich forms of parameter heterogeneity across countries, β-convergence becomes the equivalent of the proposition that there is some mean reversion in a country's output process. The rate of mean reversion could be informative about the extent of diminishing returns, but not about whether certain types of contemporaneous inequalities are increasing or decreasing. This does not lessen the interest of these studies as statistical analyses, but means their economic import can be unclear.

A second source of correlation between the error term and explanatory variables is that some variables are endogenously determined. Variables such as investment rates and initial income[13] are themselves equilibrium outcomes, in the same way as growth rates. This has led some authors to propose instrumental variables approaches to estimating β. Barro and Lee (1994) analyze growth data in the periods 1965–75 and 1975–85 using five-year lagged explanatory variables as instruments, but find that this makes little difference to the coefficient estimates. Although motivated by the possibility of measurement error, Romer (1990) finds that estimating a growth regression using instrumental variables (IVs) eliminates the negative and significant coefficient on initial income typically found when the equation is estimated by ordinary least squares (OLS). As noted above, Caselli *et al.* (1996) find estimates of β of the order of 10% – much larger than the typical 2% of cross-section studies – using a generalized method of moments

(GMM) estimator to analyze a panel variant of the standard cross-country growth regression. The variation in the outcomes of these responses to endogeneity suggests that the effects of endogeneity on β-convergence tests remain an open question.

A distinct third source of correlation between errors and regressors is measurement error. This is a particular concern in growth contexts since, despite the best efforts of those compiling the data, it is inevitable that output per capita will be mismeasured, particularly in developing economies. Measurement errors in initial levels of per capita income will tend to bias estimates of β in favor of the β-convergence hypothesis, through the effects of the standard attenuation bias. One way to see this is to consider convergence in terms of the association (or correlation) between final output and initial conditions. When convergence is rapid, final output will be only weakly associated with initial conditions; when convergence is slow, the two will be strongly associated. The standard attenuation bias implies that when initial output is measured with temporary error, this weakens the partial correlation between initial and final output, and so biases the regression finding towards a rate of convergence that is too fast. Exactly the same logic carries over when the dependent variable is the growth rate, since a model relating growth to initial log income can alternatively be reparameterized as a model that relates the log of final output to the log of initial output.[14]

However, as Temple (1998) notes, in more general settings the actual direction of the bias will depend on the stochastic properties of the measurement error itself, as well as the possibility of measurement errors in several of the Z_i control variables. He investigates the effects of allowing for measurement error in the models estimated by Mankiw *et al.* (1992), using the measurement error diagnostics developed by Klepper and Leamer (1984) and Klepper (1988), together with classical method of moments adjustments. The possibility of small errors in the measurement of initial income implies a lower bound on the estimated rate of convergence that, while positive, is too close to zero to give conditional convergence the status of a stylized fact.

Mindful of the possible effects of measurement error, Romer (1990) estimates a growth equation by both OLS and IV using the number of radios per 1,000 inhabitants and (the log of) per capita newsprint consumption as instruments for initial income and the literacy rate. In the OLS case, he finds a negative and significant coefficient on initial income, but in the IV case the coefficient is insignificant, perhaps suggesting that the significance in the OLS case is attributable to measurement error. Using lagged income as instruments for initial income, Barro (1991) and Barro and Sala-i-Martin (2004) find little change in the estimated convergence rates compared to OLS, and conclude that measurement error is not an important factor behind their findings supporting β-convergence.

Another important criticism of β-convergence regressions concerns the power of the test against non-convergent alternatives, such as models with endogenous growth or poverty traps. As shown above, $\beta < 0$ is an implication of the neoclassical growth model, but $\beta < 0$ is also potentially consistent with economically interesting alternatives. To see this, assume that there is no technical change or

population growth, that each country has a common set of control variables Z_i, and that convergence does not occur because of the presence of a threshold externality, as in Azariadis and Drazen (1990). In the Azariadis–Drazen model, such an externality can produce multiple steady-states, with the long-run outcome for an economy depending on whether its initial capital stock is above or below a threshold. Those starting below the threshold will converge to one steady-state, while those starting above will converge to another.

Relative to the economic idea of convergence as manifested in the neoclassical model, the Azariadis–Drazen model does not exhibit convergence, and different initial conditions lead to different steady-states. Yet the data generated by the Azariadis–Drazen model will not necessarily lead to the finding that $\beta \geq 0$, as shown in Bernard and Durlauf (1996). To see why, consider the growth process:

$$\gamma_i = k + \beta(\log y_{i,0} - \log y^*_{l(i)}) + \varepsilon_i, \tag{23.9}$$

where country i has steady-state $l(i)$ with associated output per capita $y^*_{l(i)}$, a value common across all countries with that steady-state. Suppose that the (now misspecified) cross-country growth regression (23.6) is employed to test for convergence. The regression coefficient will, in the probability limit, equal:

$$\beta_m = \beta\left(1 - \frac{\text{cov}(\log y^*_{l(i)}, \log y_{i,0})}{\text{var}(\log y_{i,0})}\right). \tag{23.10}$$

The sign of β_m cannot be determined *a priori*, as it depends on $\text{cov}(\log y^*_{l(i)}, \log y_{i,0})$, which is determined by the relationship between initial incomes and steady-states. It is clearly possible for β_m to be negative, implying statistical convergence, defined as $\beta < 0$, despite the absence of economic convergence.

This problem is more than a theoretical possibility, as shown in Durlauf and Johnson (1995). They estimate a model with multiple growth regimes motivated by the Azariadis and Drazen (1990) framework and find that it fits cross-country data better than does the linear Solow model. The issue is also highlighted in work by Liu and Stengos (1999) and Durlauf, Kourtellos and Minkin (2001), who find that β appears to depend nonlinearly on initial conditions and may be equal to zero for some countries. These and related findings will be discussed in the next section on nonlinearities and multiple regimes.

Similar results may be derived in linear environments in which the distinction between neoclassical and endogenous growth theories depends on returns to scale in the aggregate production function as embodied in a particular parameter value. Kocherlakota and Yi (1995) analyze a representative agent model in which $y_t = A_t k_{t-1}^{\alpha}$, so that $\log y_t = \log A_t + \alpha \log k_{t-1}$. For this production function, the difference between the neoclassical and endogenous growth models concerns the value of the parameter α. The case of $\alpha \geq 1$ represents a version of endogenous growth: the model yields perpetual growth regardless of whether there is a deterministic drift in technology.

Specifically, the Kocherlakota–Yi analysis provides conditions under which:

$$E(\log y_t - \log y_{t-1} \,|\log A_1, \log k_1) \text{ is decreasing in } \log A_1, \tag{23.11}$$

even if $\alpha \geq 1$. This means that a finding of β-convergence may occur when endogenous growth is the relevant case. The key assumption for this case is that $\log A_t = g + \rho \log A_{t-1} + \varepsilon_t$ with $0 < \rho < 1 - \delta$, where δ is the discount rate. In this case, the presence of an endogenous growth component is swamped by transitional dynamics, as the effects of shocks die out. These findings parallel the result in Bernard and Durlauf (1996) but using a different mechanism.

Kocherlakota and Yi (1995) also give conditions under which:

$$E(\log y_t - \log y_{t-1} \,|\log A_1, \log k_1) \text{ is increasing in } \log A_1, \tag{23.12}$$

even if $\alpha < 1$. This means that a positive β can occur in a cross-country growth regression, even when returns to capital are diminishing. The key to the result is a unit root in the level of technology; formally, they assume $\log A_t = g + \log A_{t-1} + \varepsilon_t$. Intuitively, intertemporal optimization, at least for their choice of utility function, means that higher initial income leads to higher investment, and this can yield a positive correlation between growth and initial income.

These findings show the difficulties that can arise when convergence findings are used to discriminate between growth models. The above analyses may be understood as arguing that tests for β-convergence fail to distinguish between behavior along a transition path to a steady-state and behavior in the steady-state, in the way needed to allow reliable discrimination between neoclassical growth models and newer alternatives.

23.3 Nonlinearities and multiple growth regimes

Clearly, tests for β-convergence may have low power against the alternative hypothesis of multiple steady-states. With this in mind, some studies have explicitly searched for statistical evidence of multiple steady-states. Durlauf and Johnson (1995) use classification and regression tree (CART) methods to search for nonlinearities in the growth process implied by the existence of multiple steady-states.[15] This procedure identifies sub-groups of countries that obey a common linear growth model based on the Solow variables, and enables a test of the null hypothesis of a common growth regime against the alternative hypothesis of multiple regimes. In their study, allowing for multiple regimes means that economies with similar initial conditions (such as literacy rates) are allowed to converge or behave in similar ways, without imposing any requirement that steady-states are unique. Using the Mankiw *et al.* (1992) data, Durlauf and Johnson (1995) reject the single regime model required for global convergence. Instead, they conclude that there is a role for initial conditions in explaining variation in cross-country growth behavior, even after controlling for the structural heterogeneity implied by the

augmented Solow model estimated by Mankiw *et al.* (1992).[16] These findings are extended in Tan (2008), who employs GUIDE (generalized, unbiased interaction detection and estimation) and finds strong evidence that measures of institutional quality and ethnic fractionalization define sub-groups of countries which obey common growth models.[17]

Other research has produced additional evidence consistent with multiple regimes using alternative statistical methods. Some of these study the behavior of the entire cross-country distribution of per capita income, as we discuss below. Here we highlight the work of three other researchers. Desdoigts (1999) uses projection pursuit methods and finds several interesting clusters, which can be described as the OECD member countries, Africa, Southeast Asia, and Latin America.[18] He argues that the first of these is identified by variables that proxy for the effects of initial conditions on subsequent growth behavior. Kourtellos (2003) uses projection pursuit to construct models of the growth process and finds evidence of two steady-states. Canova (2004) utilizes a procedure with Bayesian origins that both estimates the number of groups and assigns countries to groups. The researcher orders the countries by various criteria (for example, output per capita in the pre-sample period) and the estimation procedure chooses cluster boundaries and memberships to maximize the predictive ability of the overall model. He finds that ordering the data by initial income divides the regions of Europe into four clusters, with statistically and economically significant differences in long-run income levels and little across-cluster mobility, consistent with the existence of multiple steady-states. He also finds two clusters among the OECD countries with initial per capita income again being the preferred ordering variable. There is little mobility between the clusters, and the implied long-run difference in the average incomes is "economically large."

As discussed in Durlauf and Johnson (1995) and Durlauf *et al.* (2005), studies of nonlinearity also suffer from identification problems with respect to questions of convergence. One problem is that a given dataset cannot fully uncover the nature of growth nonlinearities without strong additional assumptions. As a result, it becomes difficult to extrapolate those relationships between predetermined variables and growth to infer steady-state behavior. Durlauf and Johnson (1995) give an example of a data pattern that is compatible with both a single steady-state and multiple steady-states. A second problem concerns the interpretation of the conditioning variables in these exercises. Suppose one finds, as in Durlauf and Johnson, that high- and low-literacy economies are associated with different aggregate production functions. One interpretation of this finding is that the literacy rate proxies for unobserved fixed factors, for example, culture, implying that these two sets of economies will never obey a common production function, and so will never exhibit convergence. Alternatively, the aggregate production function could structurally depend on the literacy rate, so that as literacy increases, the aggregate production functions of economies with low current literacy will converge to those of the high literacy ones. Data analyses of the type that have appeared to date cannot easily distinguish between these possibilities.

23.4 σ-convergence

An alternative statistical convergence concept focuses on the cross-section dispersion of log per capita output across countries, and whether it is increasing or shrinking. This has a natural connection to debates on whether inequality across countries is widening or diminishing. Discussion of this form of convergence often emphasizes the cross-section variance of log incomes, but the variance is not a sufficient statistic for the overall dispersion, and so can mask interesting forms of cross-section inequality. Although most empirical studies use the log variance, other measures of inequality can easily be used, such as the Gini coefficient or the Atkinson (1970) class of measures. These will sometimes be preferable to the log variance on axiomatic grounds.

A reduction in the dispersion of log income is interpreted as convergence because, as with β-convergence, it suggests that contemporary income differences are transitory. Letting $\sigma^2_{\log y,t}$ denote the variance across i of $\log y_{i,t}$, σ-convergence occurs between t and $t+T$ if:

$$\sigma^2_{\log y,t} - \sigma^2_{\log y,t+T} > 0. \tag{23.13}$$

Barro and Sala-i-Martin (2004, Ch. 11) report declines in the variance of the logarithm of income for the US states between 1880 and 2000, for the Japanese prefectures between 1930 and 1990, and for regions within five European countries between 1950 and 1990. In contrast, the variance of log income per capita in the world as a whole increased between 1960 and 2000.

These results are consistent with the outcomes of unconditional β-convergence tests, but there is no necessary relationship between β- and σ-convergence. It is easy to see how σ-divergence can occur even when, by an economic measure, convergence is present. For example, if all countries start from the same position but shocks are present, then divergence will occur, regardless of the speed of mean reversion. A more subtle point is that, if output changes for all countries obey $y_{i,t} - y_{i,t-1} = \beta y_{i,t-1} + \varepsilon_{i,t}$, then $\beta < 0$ is compatible with a constant cross-sectional variance, which in this example will equal the variance of $y_{i,t}$. The alternative, but mistaken, idea that mean reversion in a time series must imply a falling variance is known as Galton's fallacy. The mistake here is to ignore the role of ongoing shocks in sustaining the variance. The relevance of these types of arguments to understanding the relationship between convergence concepts in the growth literature was identified by Friedman (1992) and Quah (1993a).

Several authors have recently proposed tests for σ-convergence that employ regression specifications. Following Friedman (1992), Cannon and Duck (2000) argue that a possible test for σ-convergence could use regressions of the form:

$$\gamma_i = T^{-1}(\log y_{i,t+T} - \log y_{i,t}) = \alpha + \pi \log y_{i,t+T} + \varepsilon_i. \tag{23.14}$$

As the probability limit of the OLS estimator of π is $T^{-1}\left(1 - \sigma_{\log y_{i,t},\log y_{i,t+T}} / \sigma^2_{\log y_{i,t+T}}\right)$, a negative estimate of π implies $\sigma_{\log y_{i,t},\log y_{i,t+T}} > \sigma^2_{\log y_{i,t+T}}$. This

inequality in turn implies $\sigma^2_{\log y,t} > \sigma^2_{\log y,t+T}$, as otherwise $(\sigma_{\log y_{i,t},\log y_{i,t+T}})^2 < \sigma^2_{\log y_{i,t+T}}\sigma^2_{\log y_{i,t}}$, the condition for positive definiteness of the variance-covariance matrix of $\log y_{i,t}$ and $\log y_{i,t+T}$, would be violated. Hence a test that rejects the null hypothesis $\pi = 0$ in favor of the alternative $\pi < 0$ is evidence of σ-convergence. Their application of the test finds convergence among the US states and the European regions but not among the countries of the world.

Egger and Pfaffermayr (2007) suggest a test of conditional σ-convergence that, in the growth context, would involve a test of the hypothesis that $\pi_T = 1-\sigma^2_{u_T}/\sigma^2_{\log y_{i,t}}$ against the alternative that $\pi_T < 1 - \sigma^2_{u_T}/\sigma^2_{\log y_{i,t}}$, where π_T is the coefficient on $\log y_{i,t}$ in the cross-section regression $\log y_{i,t+T} = \pi_T \log y_{i,t} + Z_i\Gamma + u_{iT}$, where, as above, Z_i is a vector of country-i specific, time-invariant control variables. This test generalizes the unconditional σ-convergence test of Carree and Klomp (1997), which those authors use to provide evidence of σ-convergence within the OECD countries. Egger and Pfaffermayr (2007) apply their test to a large dataset on the size of European manufacturing firms; its power for the samples typically used in growth applications is not yet clear.

Bliss (1999a, 1999b) points out that the interpretation of tests of σ-convergence can be problematic in the presence of non-stationarities; an evolving distribution for the data makes it difficult to think about the distributions of test statistics under the null hypothesis. Further difficulties arise when unit roots are present.

23.5 Convergence and the cross-country distribution of per capita income

Both β- and σ-convergence are directly motivated by the law of motion of the neoclassical growth model. A distinct approach to convergence was pioneered by Quah (1993a, 1993b, 1996a, 1996b, 1996c, 1997). He focuses on "distribution dynamics": the evolution of the entire cross-country distribution of income per capita. We will question whether analyses of this kind can speak directly to the convergence hypothesis, but the approach has helped to establish some stylized facts that could be important in assessing the empirical salience of different growth theories.

One strand of this literature takes snapshots of the distribution of income per capita at points in time. For example, Bianchi (1997) tests for multimodality in kernel density estimates of the cross-country distribution of per capita income. He finds evidence of bimodality in densities estimated using Penn World Table data on gross domestic product (GDP) per capita for 1970, 1980 and 1989.[19] He also notes a tendency for the two modes to become more distant from each other over time, supporting the view that the cross-country distribution of per capita income has become increasingly polarized. He finds very little mobility within the distribution; most of the countries nearer to either the upper or lower mode in 1970 are still there in 1989. Henderson, Parmeter and Russell (2007) confirm Bianchi's findings using a longer span of data and more advanced statistical tests.

Others have estimated the density of the cross-country distribution using mixture models, based on weighted sums of component distributions.[20] Multiple components, like multiple modes, can be indicative of multiple steady-states in the dynamic process describing the evolution of per capita income. Paap and van Dijk (1998) fix the number of components at two *a priori*, based on the bimodality of histograms of their data, and then use goodness-of-fit tests to select the shapes of the component distributions. They study mobility between the components by assigning each country to a component in each year according to the maximal conditional probabilities of component "membership," computed using their parameter estimates. They find only limited mobility across components: most of the countries initially assigned to the poorer component remain so throughout the sample period. Pittau, Zelli and Johnson (2008) estimate mixture models of the distribution of GDP per worker at five-year intervals from 1960 to 2000.[21] They use goodness-of-fit and likelihood ratio (LR) tests to conclude that a three-component mixture is the preferred model. As in Paap and van Dijk (1998), they find little mobility between components, and so interpret their results as evidence of the presence of multiple steady-states, contrary to the convergence hypothesis. The results of Davis, Owen and Videras (2007), who fit a mixture model that allows for conditioning on a typical set of proximate growth determinants, suggest that these results are robust to cross-country variation in those variables.

Bloom, Canning and Sevilla (2003) also derive a mixture model for $\log y_{i,t}$. They argue that, if long-run outcomes are determined by fundamental forces alone, the relationship between income levels and exogenous variables ought to be unique. If there are multiple steady-states, so that initial conditions play a role in long-run outcomes, the relationship will not be unique. Instead, under suitable regularity conditions, it will be described by a two component mixture model if there are two steady-states and if large shocks and resultant movements between steady-states are sufficiently infrequent.[22] Using 1985 income data from 152 countries with the absolute value of the latitude of the (approximate) center of each country as the fundamental exogenous variable, they are able to reject the null hypothesis of a single regime model in favor of the alternative of a model with two regimes. The regimes correspond to a high-level ("manufacturing and services") steady-state, in which income does not depend on latitude, and a low-level ("agricultural") steady-state in which income does depend on latitude, perhaps through its influence on climate and agricultural productivity. Further, the probability of being in the high-level steady-state is found to rise with latitude.

Other analyses of the distributions of income and growth have focused on the differences in these distributions across time and across sub-sets of countries. Anderson (2004) uses stochastic dominance methods to compare distributions at different points in time and to construct measures of polarization, arguably the antithesis of convergence. Using nonparametric estimates of the cross-country distribution of per capita income, he finds increased polarization – shifts in probability density mass that increase disparities between relatively rich and relatively poor economies – between 1970 and 1995. Pittau, Zelli and Johnson (2008) reach a

similar conclusion using the Duclos, Esteban and Ray (2004) polarization index. Maasoumi, Racine and Stengos (2007) find that the distributions of growth rates for OECD and non-OECD member countries are persistently different between 1965 and 1995, with the OECD distribution's variance reducing over time whereas the non-OECD distribution appears to be becoming less concentrated.

The methods discussed above permit comparison of distributions at different points in time, but are not explicitly dynamic. In most of these contributions, the process describing the evolution of the cross-country distribution of per capita income over time is not specified or described. Quah (1993a, 1993b, 1996a, 1996b, 1996c, 1997) introduced methods into the growth literature for studying the evolution of distribution dynamics, in order to illuminate issues of mobility, stratification, and polarization that are typically obscured by the standard regression approaches used for testing the convergence hypothesis. In addition, a description of the process governing the evolution of the cross-country income distribution enables analysis of the long-run tendencies implied by that process, through computation of long-run or ergodic distributions.[23]

One way of studying distribution dynamics is to assume that the process describing the evolution of the distribution is described by a time-invariant and first-order Markov chain.[24] Using cross-country per capita income data from the early 1960s through to the mid 1980s, Quah (1993b) takes this approach. He finds that the elements on the main diagonal of the estimated transition matrix are often close to unity, indicating a high degree of persistence, or lack of mobility, within the distribution. Moreover, the implied ergodic mass function is bimodal or "twin-peaked." Together, these findings of persistence and bimodality could be seen as consistent with the presence of multiple basins of attraction or "convergence clubs."

Evidence of bimodality in the long-run cross-country distribution of per capita incomes is also found by Kremer, Onatski and Stock (2001), who update Quah's analysis using more recent data. However, their point estimates imply that most countries will ultimately move to the high-income state, and they are unable to reject the hypothesis that there is a single right-hand peak in the long-run distribution. Quah (2001) observes that the imprecision of their estimates of the ergodic distributions is sufficiently large that it is not possible to reject a wide range of null hypotheses about their shape, including, as it is the point estimate, that of bimodality.

Feyrer (2003) observes that the development traps implied by the "twin peaks" finding could stem from traps in the accumulation of physical or human capital or in total factor productivity (TFP). He uses a combination of Quah's methods and those of development accounting to examine this question, based on data from 1970 to 1989. In particular, he examines whether the possible bimodality in the distribution of GDP per capita can be traced to bimodality in the distributions of aggregate TFP (measured as a residual), the capital-output ratio, or the average level of human capital, measured in the usual way as a Mincerian function of years of schooling. If traps in the accumulation of physical capital are, for example, an important proximate cause of the bimodality in the distribution of output per capita, the distribution of the capital-output ratio should also be bimodal. Feyrer's

estimates imply that the ergodic distributions of both GDP per capita and the productivity residual are bimodal, while those of the accumulable factors of production are not.[25] Consistent with the emphasis of Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999) on productivity differences, he concludes that the proximate cause of the "twin peaks" in the distribution of GDP per capita is bimodality in the distribution of productivity, and accordingly advocates more research on models that emphasize traps in TFP.

Fitting a Markov chain to a continuous variable like GDP per capita requires a discretization of the state space. This is problematic, as it can easily alter the probabilistic properties of the data (Quah, 1996c, 1997, 2001; Bulli, 2001). Reichlin (1999) showed that the inferred dynamic behavior and the long-run implications of that behavior can depend on the discretization scheme that is used. To address this problem, Quah (1996c, 1997) proposed a continuous state space version of the approach that avoids the problems caused by discretization. If the cross-country income distribution at time t has a density function, $f_t(x)$, and if the process describing the evolution of the distribution is time-invariant and first-order Markov, the density at time $t+\tau$, $\tau > 0$, will be given by $f_{t+\tau}(x) = \int_0^\infty g_\tau(x|z) f_t(z) dz$, where $g_\tau(x|z)$ is the τ-period-ahead density of x conditional on z. The function $g_\tau(x|z)$ is the continuous analog of the transition matrix. The implied ergodic (long-run) density function, $f_\infty(x)$, if it exists, solves $f_\infty(x) = \int_0^\infty g_\tau(x|z) f_\infty(z) dz$. Quah (1996c, 1997) uses kernel density methods to estimate various $g_\tau(x|z)$ using cross-country data on output per capita, and finds a general concentration of the mass near points where $x = z$, that is, along the "main diagonal," as well as a tendency for peaks in the plot near the ends of the main diagonal and a trough in the middle. These features imply a lack of mobility within the cross-country distribution of income per capita and a tendency for mass to accumulate in the tails of the long-run distribution.[26] The estimated ergodic densities in Bulli (2001), Johnson (2005) and Fiaschi and Romanelli (2008) are also bimodal and hence support this conclusion. While Quah (2001) observes that there is not yet a theory of inference for these methods, Fiaschi and Romanelli (2008) propose a bootstrap procedure for computation of confidence intervals for the ergodic density, and their results suggest that the bimodality is statistically significant.

These methods have been important in establishing stylized facts concerning the cross-country distribution of per capita output, but there have been relatively few attempts to explore the implications for the empirical relevance of alternative growth theories. Quah (1996c) finds that conditioning on measures of physical and human capital accumulation similar to those used by Mankiw *et al.* (1992), and a dummy variable for the African continent, has little effect on the estimated dynamics of the distribution. This suggests that the heterogeneity revealed by the distributional approaches is, at least in part, due to the existence of convergence clubs rather than heterogeneity in steady-state determinants.[27] Azariadis and Stachurski (2003) derive the form of the $g_\tau(x|z)$ function implied by a stochastic version of the model in Azariadis and Drazen (1990). They estimate the model's parameters and compute forward projections of the sequence of cross-country

income distributions as well as the ergodic distribution implied by the model. Consistent with Quah (1996c, 1997), they find bimodality to be a pervasive feature of the sequence of distributions for about 100 years, even though the ergodic distribution is unimodal. Hence, even if bimodality eventually disappears, it may persist for a long time, as Quah notes in his response to Kremer *et al.* (2001).

There are some possible limitations to the use of distributional dynamics in reaching conclusions about substantive economic questions, especially when convergence is the main focus of interest. It is true that multiple modes in a distribution are consistent with the hypothesis of non-convergence, as they can be a consequence of multiple steady states. But they can also arise in the Solow model if, for example, there are multiple modes in the distribution of investment rates or population growth rates. Hence multimodality is not sufficient for concluding in favor of the existence of convergence clubs. Moreover, the existence of meaningful clubs also requires some degree of immobility within the distribution, so that countries in the vicinity of a mode will tend to remain there for extended periods.

A further point is that, while multimodality and immobility provide evidence of a lack of global convergence in the form of convergence clubs – groups of countries that converge locally but not globally – convergence clubs may be relevant even if the unconditional distribution of GDP per capita is unimodal. We can summarize some of these points by observing that the distribution dynamics literature typically investigates the shape and evolution of the whole cross-country distribution of per capita income at particular points in time, whereas economic debates about convergence are partly about the shape of the long-run or ergodic distribution *for a given country.*

23.5.1 Structural analysis

A final approach to convergence is to take a theoretical model with multiple steady-states and calibrate it to cross-country data. Graham and Temple (2006) carry out this kind of exercise for a two-sector general equilibrium model. The combination of increasing returns to scale in one sector (manufacturing and services) and intersectoral capital and labor mobility gives rise to multiple steady-states. They calibrate the model to data for 127 countries, and find that about a quarter are in a low-output equilibrium. The income differences across the equilibria are sizeable, and imply that multiplicity is capable of explaining up to a quarter of the cross-country dispersion in the logarithm of GDP per worker. Given the importance of structural models in business cycle analysis, it is remarkable how little work of this type has appeared in the convergence literature.

23.6 Time series approaches to convergence

Time series approaches to convergence are based on direct evaluation of the persistence or transience of income per capita differences between economic units, for example between pairs of countries or regions. This approach permits precise statistical definitions of hypotheses about convergence, but has the disadvantage of not being explicitly linked to particular growth theories.

Bernard and Durlauf (1995) define convergence for a set of countries I as occurring if:

$$\lim_{T\to\infty} \text{Proj}(\log y_{i,t+T} - \log y_{j,t+T} \,|\, \Im_t) = 0 \quad \forall i, j \in I, \tag{23.15}$$

where Proj$(a|b)$ denotes the projection of a on b and $\Im_t$ denotes some information set, which will generally include functions of time as well as current and lagged values of $\log y_{i,t}$ and $\log y_{j,t}$. Such a $\Im_t$ would imply a type of unconditional convergence, whereas inclusion of control variables such as investment rates would admit conditional convergence, but this has not been explored in the literature.

Most implementations of this definition have generally focused on the detection of deterministic or stochastic trends in $\log y_{i,t} - \log y_{j,t}$, as the presence of either implies a violation of (23.15). Consequently, time series tests of convergence have typically been implemented by testing for the presence of a unit root in the $\log y_{i,t} - \log y_{j,t}$ process.

Using an approach based on unit root tests, Bernard and Durlauf (1995) find little evidence of convergence in a group of 15 advanced industrialized economies between 1900 and 1989 based on data from Maddison (1982, 1989). Hobijn and Franses (2000) similarly find little evidence of convergence across a group of 112 Penn World Table countries over the period 1960–89.[28] Their work is based on a clustering algorithm to identify groups of converging countries. They find many small clusters, which they view as having distinct steady-states; but their multiplicity, and the absence of controls for structural characteristics, means that these clusters could simply reflect differences in those characteristics, rather than differences in long-run outcomes due to differences in initial conditions. The breadth of the sample used also suggests that the Bernard and Durlauf (1996) argument, about the need for consideration of the substantive economic assumptions that underlie time series methods for studying convergence, is applicable in this case.

One criticism of these tests focuses on the validity of unit root tests in the presence of structural breaks. Perron (1989) has argued that the failure to allow for structural breaks can lead to spurious evidence in support of the presence of a unit root (or, more precisely, can diminish the ability to reject the null of a unit root). Greasley and Oxley (1997) impose breaks exogenously and find convergence for Denmark and Sweden, in contrast to Bernard and Durlauf (1995), who did not allow for breaks. Li and Papell (1999) allow for endogenous trend breaks and find that this reduces, relative to Bernard and Durlauf (1995), the number of country pairs that fail to exhibit convergence.

Carlino and Mills (1993) study US regions and reject convergence except with specifications that allow for a trend break in 1946. However, a trend break violates (23.15), as it implies that some component of $\log y_{i,t} - \log y_{j,t}$ is predictable in the long run. Thus claims that allowing for data breaks produces evidence in favor of convergence invites the question of what is meant by convergence. Note that the sort of violation of (23.15) implied by a trend break is different from the type implied by a unit root, as a break associated with the level of output means that the output difference between two countries is always bounded. The issue of

trend breaks raises the general question of whether convergence dynamics obey a nonlinear process. Chong *et al*. (2008) employ a smooth transition autoregressive process to model output and conclude that OECD member countries are converging to the US level of output per capita. Their analysis is again difficult to interpret in economic terms, as their functional form assumptions do not correspond to any particular economic model and therefore invite questions about the appropriate choice of nonlinear structure.

A second criticism of time series tests has been made by Michelacci and Zaffaroni (2004). They argue that evidence of unit roots in per capita output may be spurious because of a lack of attention to the possibility that dependence in these individual series exhibits long memory, that is, dependence decays at a hyperbolic rather than a geometric rate. If it is the case that the individual series are stationary in levels, the differences between them cannot contain unit roots, so that convergence is occurring. They further argue that long memory can explain the 2% convergence rate found in cross-section regressions. This is an intriguing argument, although Durlauf *et al*. (2005) question the strength of the empirical evidence for long memory as well as its theoretical plausibility. Michelacci and Zaffaroni also do not directly study the behavior of per capita output differences, so it is unclear how to match their analysis with other studies.

The time series and σ approaches to convergence are merged in Evans (1996). He studies the time series properties of σ_t^2, the cross-section variance of $\log y_{i,t}$. He shows that, when there is no cointegration among the series $\log y_{i,t}$, σ_t^2 may be represented as a unit root process with a quadratic time trend, and this suggests a time series test of convergence based on unit root tests applied to a time series for σ_t^2. He uses this test to conclude that there is convergence to a common trend among 13 industrial countries over the period 1870–1989 and among a group of 51 countries over the period 1950–92, although the evidence in the latter case is less conclusive.

Evans (1997) provides a time series approach to estimating rates of convergence. For the contiguous US states over the period 1929–91, he finds that about one-third of the point estimates are negative and about two-thirds of the confidence intervals contain zero. For a sample of 48 countries over the period 1950–90, about half of the point estimates are negative and all but two of the confidence intervals contain zero.

23.6.1 Transitions versus steady-state dynamics

There are important differences between the time series approach to convergence and the β and distribution shape approaches. As argued in Bernard and Durlauf (1996), time series tests assume that the underlying stochastic processes are time invariant, so that countries have transited to an invariant output process. In contrast, cross-section approaches, such as β-convergence and σ-convergence, are motivated by the assumption that countries are in transition to a steady-state, so that the data for a given country at time t are drawn from a different stochastic process than the data at some future time.[29] Bernard and Durlauf further indicate

how convergence under a cross-section test can, in fact, imply a failure of convergence under a time series test, because of these different assumptions. To be clear, it is possible for data from a stationary process to exhibit β-convergence, for example, but it is not clear what the economic interpretation is.

Harvey and Carvalho (2002) and Carvalho and Harvey (2005a, 2005b) propose a number of time series analyses of convergence which distinguish between evaluating whether countries have converged, and whether countries are converging. The idea of this framework is to interpret convergence as involving adjustments of output in a given country to gaps between its past output and that of other countries. It allows one to evaluate the presence of common deterministic trends, as well as tendencies for economies to adjust based on differences with the rest of the world. Relative to the other unit root tests, this approach allows for level differences between economies and so distinguishes between convergence to a common level and convergence to common growth paths. Harvey and Carvalho (2002) use this approach to conclude that there is convergence between the US and Japanese growth paths; Carvalho and Harvey (2005a, 2005b) find both convergence and divergence between Euro zone countries and various pairs of US regions respectively. An additional finding of these papers is that unit root tests of convergence are highly sensitive to the inclusion of time trends and constant terms.

Phillips and Sul (2006) extend the distinction between economies that have converged and economies that are converging, by explicitly addressing the question of invariance of the time series process for output. Their analysis focuses on models of the form:

$$y_{i,t} = b_{i,t}\beta_t + \kappa_{i,t}, \tag{23.16}$$

where β_t is a common trend (deterministic and/or stochastic) and $\kappa_{i,t}$ is a cyclical term. The key advance in this formulation is that $b_{i,t}$, the weights on the common trend, are allowed to vary both with respect to the country and with respect to time. This approach allows one to estimate a transition curve for individual economies which, by tracing out $b_{i,t}\beta_t$, allows for explicit evaluation of how the long-run components of national output co-evolve. Their analysis finds evidence of convergence for the OECD member countries and for US states, but not for a broader sample of countries drawn from the Penn World Table.

23.7 The economics of convergence

The various approaches to convergence that we have discussed are all purely statistical, so it remains to consider convergence as an economic concept, in order to assess what can be learned from econometric studies. In this section we provide some definitions of convergence that, while statistical in nature, can be used to move between the economic notion of convergence and the statistical notions that have been employed to assess it.

Broadly speaking, we take the substantive economic content of the convergence hypothesis to be the claim that initial conditions have no effect on long-run

economic outcomes. As argued in Durlauf *et al.* (2005), the empirical task is then to determine the role of initial conditions in explaining cross-country differences in per capita output. This task is complicated by the role of structural heterogeneity in also explaining those differences; so the empirical literature must disentangle the effects of initial conditions and structural heterogeneity. There are three possibilities: (i) unconditional convergence (to a common long-run level) occurs if differences in per capita incomes are temporary; (ii) conditional convergence occurs if permanent differences reflect only cross-country structural heterogeneity; and (iii) club convergence occurs if initial conditions determine, to some extent at least, long-run outcomes, with countries with similar initial conditions exhibiting similar long-run outcomes.[30]

To formalize these ideas, we associate with economy i initial conditions $\rho_{i,0}$ and say that these initial conditions do not matter in the long run if:

$$\lim_{t\to\infty} \mu(\log y_{i,t}|\rho_{i,0}) \text{ does not depend on } \rho_{i,0}, \tag{23.17}$$

where $\mu(\cdot)$ is a probability measure.[31] Letting $\|\|$ denote a metric for computing the distance between probability measures, we say that countries i and j converge if:

$$\lim_{t\to\infty} \|\mu(\log y_{i,t}|\rho_{i,0}) - \mu(\log y_{j,t}|\rho_{j,0})\| = 0, \tag{23.18}$$

which implies convergence in average income levels in the sense that:

$$\lim_{t\to\infty} E(\log y_{i,t} - \log y_{j,t}|\rho_{i,0}, \rho_{j,0}) = 0. \tag{23.19}$$

This definition can be modified to require that the limiting expected difference between $\log y_{i,t}$ and $\log y_{j,t}$ is bounded if the equality of steady-state growth rates is of interest. This definition is the one that underlies all of the time series approaches to convergence: the differences between Harvey and Carvalho (2002) and Phillips and Sul (2006) and the earlier time series tests of Bernard and Durlauf (1995) and others reflect differences in how this long-run forecast similarity is calculated. This is also consistent with the economic notion of convergence that appears in the neoclassical growth model. Our criticism of some of the cross-section and panel approaches to convergence partially derives from their failure to evaluate this condition fully.

Bernard and Durlauf (1996) suggest a definition of partial convergence that requires contemporaneous income differences be expected to diminish, that is:

$$E(\log y_{i,t} - \log y_{j,t}|\rho_{i,0}, \rho_{j,0}) < \log y_{i,0} - \log y_{j,0}, \tag{23.20}$$

for $\log y_{i,0} > \log y_{j,0}$. Hall, Robertson and Wickens (1997) suggest a definition that requires the variance of output differences to diminish to zero, that is:

$$\lim_{t\to\infty} E((\log y_{i,t} - \log y_{j,t})^2|\rho_{i,0}, \rho_{j,0}) = 0, \tag{23.21}$$

and hence convergence requires output for a pair of countries to behave similarly in the long run. This is a strong requirement since it does not permit $\log y_{i,t} - \log y_{j,t}$ to be stochastic in the long run, as would be the case if the two countries have different short-run shock processes. Hall *et al.* point out the lacuna in the definition (23.19) in the case of long-run deviations whose current direction is not predictable. If, for example, $\log y_{i,t} - \log y_{j,t}$ obeys a random walk with current value zero, then definition (23.19) would hold despite the fact that future output deviations between countries i and j could be large. This problem is avoided by the modified definition:

$$\forall r \geq 0, \qquad \lim_{t \to \infty} E(\log y_{i,t} - \log y_{j,t} | \rho_{i,r}, \rho_{j,r}) = 0. \tag{23.22}$$

This modification has no bearing on the relationship between long-run unforecastability of differences and either theory or the various convergence tests we have described.

The β-convergence concept represents the idea of the long-run irrelevance of initial output per capita by decomposing the growth rate of $y_{i,t}$ into two parts, technological progress and "catching up." In the Solow model a positive rate of conditional convergence, λ, implies that final output is ultimately independent of initial conditions, so that the initial gap between output and its steady-state value ceases to play a role in long-run growth. As we have discussed, a testable implication of a positive λ is a negative β.

In contrast, the literature that focuses on nonlinearities in the growth process seeks to identify sub-groups of countries that obey a common growth model distinct from that obeyed by the countries in other groups. Such behavior is consistent with initial conditions that are relevant even in the long run, as it is consistent with a global growth model with multiple steady-states in which economies with similar initial conditions tend to converge to one another. The convergence approaches that study the behavior of the entire cross-country distribution of output per capita look for evidence of the long-run importance of initial conditions in the form of accumulation points in the distribution, as evinced by intervals with large quantities of probability mass, between which there is little mobility. This behavior is consistent with a dynamic growth process that exhibits multiple basins of attraction and so produces convergence clubs.

The convergence definitions given in (23.19) and (23.20) above make no distinction between the long-run effects of initial conditions and those of structural heterogeneity, and so fail to distinguish conditional convergence from the existence of convergence clubs. This is a serious deficiency for those seeking to use the outcomes of convergence tests to adjudicate between competing theories of economic growth. Long-run effects of cross-country differences in preferences are, for example, consistent with both the neoclassical and endogenous or "new" classes of growth theories. However, neoclassical theories are inconsistent with long-run effects of cross-country differences in initial human and physical capital stocks, whereas this would not be the case for all endogenous growth models. In other words, the finding of a long-run role for initial conditions constitutes evidence in

favor of non-neoclassical models, but care must be taken to control for the long-run effects of differences in preferences and other structural characteristics.

More formally, successful empirical work on the convergence issue requires the distinction between initial conditions $\rho_{i,0}$ and structural characteristics $\theta_{i,0}$. Steady-state effects of the former imply the existence of convergence clubs, but steady-state effects of the latter do not. In order to make this distinction we modify (23.18) and say that countries i and j exhibit convergence if:

$$\lim_{t\to\infty} \|\mu(\log y_{i,t}|\rho_{i,0}, \theta_{i,0}) - \mu(\log y_{j,t}|\rho_{j,0}, \theta_{j,0})\| = 0 \text{ if } \theta_{i,0} = \theta_{j,0}. \tag{23.23}$$

The corresponding modification to the notion of convergence in expected value is:

$$\lim_{t\to\infty} E(\log y_{i,t} - \log y_{j,t}|\rho_{i,0}, \theta_{i,0}, \rho_{j,0}, \theta_{j,0}) = 0 \text{ if } \theta_{i,0} = \theta_{j,0}, \tag{23.24}$$

and the other convergence concepts discussed above can be similarly modified. While conceptually clear, the distinction between initial conditions and structural heterogeneity is potentially difficult in practice. Typically, researchers have treated initial human and physical capital stocks as instances of initial conditions, and other variables as representing structural heterogeneity – for example, those that often appear as controls in cross-country growth regressions, a practice that is problematic if these variables are, in fact, endogenously determined by initial conditions.

Disentangling the respective roles of structural heterogeneity and initial conditions in determining growth performance remains one of the most important challenges for the convergence literature. Economic theory does not always provide a guide to the relevant control variables, let alone the appropriate distinction between variables that capture structural heterogeneity and those that should be classed as initial conditions. It is also important to emphasize that none of the statistical definitions of convergence discussed above is necessarily of any intrinsic interest *per se*; each is useful only to the extent that it can illuminate some economically interesting notions of convergence such as that in (23.24). The failure to distinguish between convergence as an economic concept and convergence as a statistical concept has led to much confusion in the growth literature.

23.8 Conclusions

The empirical convergence literature contains many interesting findings and has helped to identify a number of important generalizations about cross-country growth behavior. At the same time, it has yet to reach any sort of consensus on the deep economic questions for which the statistical analyses were designed. It is not difficult to highlight some of the relevant problems. The fundamentally nonlinear nature of endogenous growth theories renders the conventional cross-section and panel convergence tests inadequate as ways to discriminate between the main classes of theories. Evidence of convergence clubs may simply be evidence of deep nonlinearities in the transitional dynamics towards a unique

steady-state. Time series evidence against divergence does not distinguish between conditional and unconditional convergence. Further, cross-section, panel, and time series approaches to convergence not only yield different results, but are predicated on different views of the nature of transitory versus steady-state behavior of economies, differences that themselves remain hard to test.

None of this is to say that the study of convergence is not a meaningful or productive subject for research. Clearly, considerable progress has been made, and Phillips and Sul (2006) represents a key first step in integrating the transitory and steady-state perspectives. Much remains to be done, and research on convergence should continue to develop in interesting ways. Our belief is that progress is most likely if the economic content of specific versions of convergence is placed at the center of the analysis, so that statistical sophistication is not an end in itself.

Further, we think that more attention should be made to time horizons. Much of the convergence literature has treated the question as a zero–one outcome, whereas it is probably more sensible to ask questions about partial convergence over shorter horizons. While attention to convergence rates addresses this question, it generally has not focused on understanding differences in timing across regions or how timing is affected by various initial conditions. These questions are especially important in assessing anti-poverty policies such as those advocated by the United Nations in the Millennium Development Goals, and they perhaps matter for introducing some much needed modesty into the growth literature.

Finally, we see value in shifting discussions of convergence away from national per capita incomes. Cross-country differences in mortality and morbidity are an obvious context where convergence is of intrinsic interest. In addition, we see some scope for considering convergence for units outside the nation state. This seems clear when one considers the effects of institutions and culture on economic activity. We go so far as to conjecture that important aspects of long-run persistence in national incomes mirror long-term divergence in phenomena that are not nation-specific, which is, of course, an idea that goes back to Max Weber. Thomas Macaulay wrote,[32] comparing the Catholic Church to the British Empire: "she may still exist in undiminished vigor when some traveler from New Zealand shall, in the midst of a vast solitude, take his stand on a broken arch of London Bridge to sketch the ruins of St. Paul's," which well illustrates how all claims of convergence versus divergence or permanence versus transience depend on the choice of context.

Notes

1. Quotations are taken from the Penguin edition.
2. See Durlauf (1996) and the subsequent papers in the July 1996 *Economic Journal*; Durlauf and Quah (1999), Islam (2003), Barro and Sala-i-Martin (2004), and Durlauf, Johnson and Temple (2005) for, *inter alia*, surveys of the convergence literature.
3. Abramovitz (1986) also uses σ-convergence, which we will discuss below.
4. Mankiw (1995, p. 301), for example, argues that for "understanding international experience, the best assumption may be that all countries have access to the same pool of knowledge, but differ by the degree to which they take advantage of this knowledge by investing in physical and human capital." Dowrick and Rogers (2002) argue

that both diminishing returns and technology transfer are important contributors to the convergence process. See also Bernard and Jones (1996) and Barro and Sala-i-Martin (1997).

5. That $\lim_{t\to\infty} y^E_{i,t} = y^E_{i,\infty}$ follows from $\lambda_i > 0$. This long-run independence of $y^E_{i,t}$ from $y^E_{i,0}$ implies that initial conditions do not matter in the long run – an interpretation of convergence that we discuss below.
6. In parallel to equation (23.1), $\lim_{t\to\infty}(y_{i,t} - y^E_{i,\infty}A_{i,0}e^{g_i t}) = 0$, so that again the initial value of output per worker does not affect its long-run value.
7. This assumption is made, for example, by Mankiw *et al.* (1992), who argue that $A_{i,0}$ reflects not just technology, which they assume to be constant across countries, but country-specific influences on growth, such as resource endowments, climate and institutions. They assume these differences vary randomly across countries, independently of the determinants of the steady-state level of output per worker.
8. Barro and Sala-i-Martin (2004, Chs. 11, 12) implement β-convergence tests for a variety of datasets. As pointed out by DeLong (1988), the use of homogeneous groups runs the risk of *ex post* sample selection, especially if the homogeneity relates to final outcomes. In particular, he views the Baumol (1986) finding of unconditional β-convergence over 1870–1979 among a set of countries that were affluent in 1979, as tending to overstate the true degree of convergence. DeLong extends the sample to include countries with similar starting positions in 1870, but which have been less successful since, and this weakens the evidence for convergence.
9. In both Jones and Manuelli (1990) and Kelly (1992), steady-state growth occurs without exogenous technical change, but initially poor economies grow more quickly as β-convergence requires.
10. As Barro and Sala-i-Martin (1992) note, while there is some variation in estimated convergence rates, estimates generally range between 1% and 3%. They attribute this variation to unobserved heterogeneity in steady-state values; but to the extent that it is correlated with variables included in the regressions, this heterogeneity implies the parameter estimates are inconsistent. Panel studies such as Islam (1995), Caselli *et al.* (1996) and Lee, Pesaran and Smith (1998) have found more rapid rates of convergence.
11. For example, Barro and Sala-i-Martin (1991) present results for US states and regions as well as European regions; Barro and Sala-i-Martin (1992) for US states, a group of 98 countries and the OECD; Mankiw *et al.* (1992) for several large groups of countries; Sala-i-Martin (1996a, 1996b) for US states, Japanese prefectures, European regions, and Canadian provinces; Cashin (1995) for Australian states and New Zealand; Cashin and Sahay (1996) for Indian regions; Persson (1997) for Swedish counties; and Shioji (2001) for Japanese prefectures and other geographic units.
12. As Durlauf *et al.* (2005) document, the number of suggested control variables is now almost as large as the number of countries in the world.
13. Den Haan (1995) is an especially sophisticated discussion.
14. See Abramovitz (1986), Baumol (1986), DeLong (1988), Romer (1990) and Temple (1998).
15. An appendix to Durlauf and Johnson (1995) discusses the application of regression tree methods to the issue of locating multiple regimes in growth models. Breiman *et al.* (1984) contains a detailed general treatment of regression tree methods. While these methods suffer from the lack of a well-developed asymptotic theory for testing the number of regimes present in a dataset, they are consistent in the sense that, under relatively weak conditions, the correct model will be revealed as the sample size grows to infinity, if there are a finite number of regimes.
16. Papageorgiou and Masanjala (2004) observe that Durlauf and Johnson's findings could be due to misspecification of the aggregate production function. They estimate a version of the Solow model based on a constant elasticity of substitution (CES) production function rather than the Cobb–Douglas, following findings in Duffy and Papageorgiou (2000), and ask whether or not Durlauf and Johnson's multiple regimes remain under the CES

specification. Using Hansen's (2000) approach to sample splitting and threshold estimation, they find statistically significant evidence of four distinct growth regimes, broadly consistent with those found by Durlauf and Johnson (1995). Johnson and Takeyama (2001) also use the regression tree approach and find evidence of thresholds in US state economic growth behavior defined by variables likely to be proxies for the capital/labor ratio, agglomeration effects and communication effects.

17. For more on GUIDE methods, see Loh (2002).
18. Appendix A of Desdoigts (1999) provides a useful primer on projection pursuit which is developed in Friedman and Tukey (1974) and Friedman (1987).
19. Kernel density estimation is a nonparametric method of estimating density functions that has the attraction of flexibility as it does not impose *a priori* a functional form for the density. See Silverman (1986).
20. The finite mixture model is a semi-parametric alternative to the nonparametric kernel approach to density estimation. See McLachlan and Peel (2000).
21. Tsionas (2000) applies mixture models to US state data while Pittau (2005) and Pittau and Zelli (2006) apply them to EU regional data.
22. This requirement is consistent with the documented lack of mobility within the cross-country distribution of per capita income.
23. Again, multimodality in the ergodic distribution is neither necessary nor sufficient to imply the existence of convergence clubs – not sufficient because it reveals nothing about mobility and not necessary because the existence of multiple stochastic steady-states does not imply multiple modes in the long-run distribution.
24. Applications of Markov chain methods to the study of mobility issues in the social sciences predate the modern empirical growth literature. See, for example, Prais (1955). The time-invariance and first-order Markovian assumptions are testable, with Quah (1993b) presenting an informal test of the latter, while Fingleton (1997) and Bickenbach and Bode (2003) present more formal tests in the cases of EU regions and US states respectively.
25. Johnson (2005) re-examines Feyrer's (2003) results using the continuous state space methods discussed below and finds that, rather than TFP playing an exclusive role in the apparent bimodality in the long-run distribution of GDP per capita, the ergodic distribution of the capital-output ratio is also bimodal.
26. Quah's methods have been widely applied. For example, Andres and Lamo (1995) apply these methods to the OECD; Lamo (2000) to the regions of Spain; Johnson (2000) to US states; Bandyopadhyay (2004) to the Indian states; Andrade *et al.* (2004) to Brazilian municipalities; Ezcurra *et al.* (2005) to EU regions; Fotopoulos (2006) to Greek regions, and Pittau and Zelli (2006) to EU regions. In some cases, the small cross-section dimension of the samples must limit the reliability of the findings. These methods have also been extended to broader notions of distributional dynamics. Fiaschi and Lavezzi (2004) develop an analysis of the joint distribution of income levels and growth rates; their findings are compatible with the existence of multiple equilibria in the sense that countries may become trapped in the lower part of the income distribution.
27. Other efforts to find determinants of intertemporal mobility have produced mixed results. For the OECD countries, Andres and Lamo (1995) condition on the steady-state implied by the Solow model and find little change in the tendency to polarization unless country specific effects are permitted. Lamo (2000) finds only a small increase in mobility for Spanish regions after conditioning on interregional migration flows, while Bandyopadhyay (2004) shows that differences in infrastructure spending and education contribute to polarization between the rich and poor states of India.
28. These findings are echoed in Pesaran (2007), who uses both the Maddison and Penn World Table datasets and employs a convergence definition that explicitly focuses on the probability of large deviations between $\log y_{i,t}$ and $\log y_{j,t}$.

29. Panel studies circumvent this issue by heavy reliance on the adequacy of the log-linear approximation. The assumption is that countries in the sample are sufficiently close to their steady-states that their dynamics can be described by a stationary process. This is potentially problematic for studies that employ a broad cross-section of countries, at least some of which may begin far from their steady-state.
30. This taxonomy is due to Galor (1996), who discusses the relationship between it and the theoretical growth literature. His paper, and those of Azariadis (1996) and Azariadis and Stachurski (2005), give many examples of models in which initial conditions matter for long-run outcomes.
31. The discussion here is in terms of $\log y_{i,t}$, the log level of output per capita in country i at time t; but these definitions could be applied to other variables such as real wages or life expectancy. We use $\log y_{i,t}$ rather than $y_{i,t}$ due to the general interest in the literature in relative rather than absolute inequality.
32. "Von Ranke," *The Edinburgh Review*, 1840. Quotation taken from *Critical and Historical Essays: The Complete Writings of Lord Macaulay Part 4*, reprinted by Kessinger Publishing.

References

Abramowitz, M. (1986) Catching up, forging ahead and falling behind. *Journal of Economic History* **46**, 385–406.

Anderson, G. (2004) Toward an empirical analysis of polarization. *Journal of Econometrics* **122**, 1–26.

Andrade, E., M. Laurini, R. Madalozzo and P. Pereira (2004) Convergence clubs among Brazilian municipalities. *Economics Letters* **83**, 179–84.

Andres, J. and A. Lamo (1995) Dynamics of the income distribution across OECD countries. London School of Economics, Centre for Economic Performance Discussion Paper No. 252.

Atkinson, A. (1970) On the measurement of inequality. *Journal of Economic Theory* **2**, 244–63.

Azariadis, C. (1996) The economics of poverty traps: Part one: Complete markets. *Journal of Economic Growth* **1**, 449–96.

Azariadis, C. and A. Drazen (1990) Threshold externalities in economic development. *Quarterly Journal of Economics* **105**(2), 501–26.

Azariadis, C. and J. Stachurski (2003) A forward projection of the cross-country income distribution. Institute of Economic Research, Kyoto University, Discussion Paper No. 570.

Azariadis, C. and J. Stachurski (2005) Poverty traps. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*. Amsterdam: North-Holland.

Bandyopadhyay, S. (2004) Twin peaks – distribution dynamics of economic growth across Indian states. In A. Shorrocks and R. van der Hoeven (eds.), *Growth, Inequality and Poverty: Prospects for Pro-Poor Growth*. Oxford: Oxford University Press.

Barro, R. (1991) Economic growth in a cross section of countries. *Quarterly Journal of Economics* **106**, 407–43.

Barro, R. and J.-W. Lee (1994) Sources of economic growth (with commentary). *Carnegie-Rochester Conference Series on Public Policy* **40**, 1–57.

Barro, R. and X. Sala-i-Martin (1991) Convergence cross states and regions. *Brookings Papers on Economic Activity* **1**, 107–58.

Barro, R. and X. Sala-i-Martin (1992) Convergence. *Journal of Political Economy* **100**, 223–51.

Barro, R. and X. Sala-i-Martin (1997) Technological diffusion, convergence, and growth. *Journal of Economic Growth* **2**, 1–26.

Barro, R. and X. Sala-i-Martin (2004) *Economic Growth* (second edition). Cambridge, Mass.: MIT Press.

Baumol, W. (1986) Productivity growth, convergence, and welfare: what the long-run data show. *American Economic Review* **76**, 1072–85.

Bernard, A. and S. Durlauf (1995) Convergence in international output. *Journal of Applied Econometrics* **10**, 97–108.
Bernard, A. and S. Durlauf (1996) Interpreting tests of the convergence hypothesis. *Journal of Econometrics* **71**, 161–73.
Bernard, A. and C. Jones (1996) Comparing apples to oranges: productivity convergence and measurement across industries and countries. *American Economic Review* **86**, 1216–38.
Bianchi, M. (1997) Testing for convergence: evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* **12**, 393–409.
Bickenbach, F. and E. Bode (2003) Evaluating the Markov property in studies of economic convergence. *International Regional Science Review* **26**, 363–92.
Bliss, C. (1999a) Galton's fallacy and economic convergence. *Oxford Economic Papers* **51**, 4–14.
Bliss, C. (1999b) Galton's fallacy and economic convergence: a reply to Cannon and Duck. *Oxford Economic Papers* **52**, 420–2.
Bloom, D., D. Canning and J. Sevilla (2003) Geography and poverty traps. *Journal of Economic Growth* **8**, 355–78.
Bond, S., A. Hoeffler and J. Temple (2001) GMM estimation of empirical growth models. Centre for Economic Policy Research Discussion Paper No. 3048.
Breiman, L., J. Friedman, R. Olshen and C. Stone (1984) *Classification and Regression Trees*. Redwood City, Calif.: Wadsworth Publishing.
Bulli, S. (2001) Distribution dynamics and cross-country convergence: new evidence. *Scottish Journal of Political Economy* **48**, 226–43.
Cannon, E. and N. Duck (2000) Galton's fallacy and economic convergence. *Oxford Economic Papers* **53**, 415–19.
Canova, F. (2004) Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review* **45**, 49–77.
Carlino, G. and L. Mills (1993) Are US regional incomes converging? A time series analysis. *Journal of Monetary Economics* **32**, 335–46.
Carree, M. and L. Klomp (1997) Testing the convergence hypothesis: a comment. *Review of Economics and Statistics* **79**, 683–6.
Carvalho, V. and A. Harvey (2005a) Convergence in the trends and cycles of euro zone income. *Journal of Applied Econometrics* **20**, 275–89.
Carvalho, V. and A. Harvey (2005b) Growth, cycles, and convergence in US regional time series. *International Journal of Forecasting* **21**, 667–86.
Caselli, F., G. Esquivel and F. Lefort (1996) Reopening the convergence debate: a new look at cross country growth empirics. *Journal of Economic Growth* **1**, 363–89.
Cashin, P. (1995) Economic growth and convergence across the seven colonies of Australasia: 1861–1991. *Economic Record* **71**, 132–44.
Cashin, P. and R. Sahay (1996) Regional economic growth and convergence in India. *Finance and Development* **33**, 49–52.
Chong, T., M. Hinich, V. Liew and K.-P. Lim (2008) Time series test of nonlinear convergence and transitional dynamics. *Economic Letters*. Forthcoming.
Davis, L.S., A.L. Owen and J. Videras (2007) Do all countries follow the same growth process? Available at SSRN, http://ssrn.com/abstract=1064521.
DeLong, J.B. (1988) Productivity growth, convergence, and welfare: comment. *American Economic Review* **78**, 1138–54.
den Haan, W. (1995) Convergence in stochastic growth models: the importance of understanding why income levels differ. *Journal of Monetary Economics* **35**, 65–82.
Desdoigts, A. (1999) Patterns of economic development and the formation of clubs. *Journal of Economic Growth* **4**, 305–30.
Dowrick, S. and M. Rogers (2002) Classical and technological convergence: beyond the Solow–Swan growth model. *Oxford Economic Papers* **54**, 369–85.
Duclos, J.Y., J. Esteban and D. Ray (2004) Polarization: concepts, measurement, estimation. *Econometrica* **72**, 1737–72.

Duffy, J. and C. Papageorgiou (2000) A cross-country empirical investigation of the aggregate production function specification. *Journal of Economic Growth* **5**, 87–120.
Durlauf, S. (1996) On the convergence and divergence of growth rates. *Economic Journal* **106**, 1016–18.
Durlauf, S. and P. Johnson (1995) Multiple regimes and cross country growth behaviour. *Journal of Applied Econometrics* **10**, 365–84.
Durlauf, S., P. Johnson and J. Temple (2005) Growth econometrics. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*. Amsterdam: North-Holland.
Durlauf, S., A. Kourtellos and A. Minkin (2001) The local Solow growth model. *European Economic Review* **45**, 928–40.
Durlauf, S. and D. Quah (1999) The new empirics of economic growth. In J. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics*. Amsterdam: North-Holland.
Egger, P. and M. Pfaffermayr (2007) On testing conditional sigma-convergence. *Oxford Bulletin of Economics and Statistics*. Forthcoming.
Evans, P. (1996) Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* **20**, 1027–49.
Evans, P. (1997) How fast do economies converge? *Review of Economics and Statistics* **79**, 219–25.
Ezcurra, R., C. Gil, P. Pascual and M. Rapún (2005) Inequality, polarisation and regional mobility in the European Union. *Urban Studies* **42**, 1057–76.
Fernandez, C., E. Ley and M. Steel (2001) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**, 563–76.
Feyrer, J. (2003) Convergence by parts. Mimeo, Dartmouth College.
Fiaschi, D. and A. Lavezzi (2004) Distribution dynamics and nonlinear growth. *Journal of Economic Growth* **8**, 379–401.
Fiaschi, D. and M. Romanelli (2008) Nonlinear dynamics in welfare and the evolution of world inequality. Mimeo, Department of Economics, University of Pisa, http://www-dse.ec.unipi.it/persone/docenti/fiaschi/WorkingPapers.html.
Fingleton, B. (1997) Specification and testing of Markov chain models: an application to convergence in the European Union. *Oxford Bulletin of Economics and Statistics* **59**, 385–403.
Fotopoulos, G. (2006) Nonparametric analysis of regional income dynamics: the case of Greece. *Economics Letters* **91**, 450–57.
Friedman, J. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–66.
Friedman, J. and J. Tukey (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **C23**, 881–90.
Friedman, M. (1992) Do old fallacies ever die? *Journal of Economic Literature* **30**, 2129–32.
Galor, O. (1996) Convergence? Inferences from theoretical models. *Economic Journal* **106**, 1056–69.
Graham, B. and J. Temple (2006) Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* **11**, 5–41.
Greasley, D. and L. Oxley (1997) Time-series tests of the convergence hypothesis: some positive results. *Economics Letters* **56**, 143–7.
Hall, R. and C. Jones (1999) Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* **114**, 83–116.
Hall, S., D. Robertson and M. Wickens (1997) Measuring economic convergence. *International Journal of Finance and Economics* **2**, 131–43.
Hansen, B. (2000) Sample splitting and threshold estimation. *Econometrica* **68**, 575–603.
Harvey, A. and V. Carvalho (2002) Models for converging economies. Mimeo, University of Cambridge.
Henderson, D., C. Parmeter and R. Russell (2007) Modes, weighted modes, and calibrated modes: evidence of clustering using modality tests. *Journal of Applied Econometrics*. Forthcoming.

Hobijn, B. and P.H. Franses, (2000) Asymptotically perfect and relative convergence of productivity. *Journal of Applied Econometrics* **15**, 59–81.
Islam, N. (1995) Growth empirics: a panel data approach. *Quarterly Journal of Economics* **110**, 1127–70.
Islam, N. (1998) Growth empirics: a panel data approach – a reply. *Quarterly Journal of Economics* **113**, 325–9.
Islam, N. (2003) What have we learned from the convergence debate? *Journal of Economic Surveys* **17**, 309–62.
Johnson, P. (2000) A nonparametric analysis of income convergence across the US states. *Economics Letters* **69**, 219–23.
Johnson, P. (2005) A continuous state space approach to "convergence by parts." *Economic Letters* **86**, 317–21.
Johnson, P. and L. Takeyama (2001) Initial conditions and economic growth in the US states. *European Economic Review* **45**, 919–27.
Jones, L. and R. Manuelli (1990) A convex model of equilibrium growth: theory and policy implications. *Journal of Political Economy* **98**, 1008–38.
Kelly, M. (1992) On endogenous growth with productivity shocks. *Journal of Monetary Economics* **30**, 47–56.
Klenow, P. and A. Rodriguez-Clare (1997) The neoclassical revival in growth economics: has it gone too far? In B. Bernanke and J. Rotemberg (eds.), *Macroeconomics Annual 1997*. Cambridge, Mass.: MIT Press.
Klepper, S. (1988) Regression diagnostics for the classical errors-in-variables model. *Journal of Econometrics* **37**, 225–50.
Klepper, S. and E. Leamer (1984) Consistent sets of estimates for regressions with errors in all variables. *Econometrica* **52**, 163–83.
Kocherlakota, N. and K.-M. Yi (1995) Can convergence regressions distinguish between exogenous and endogenous growth models? *Economic Letters* **49**, 211–15.
Kourtellos, A. (2003) A projection pursuit approach to cross-country growth data. Mimeo, University of Cyprus.
Kremer, M., A. Onatski and J. Stock (2001) Searching for prosperity. *Carnegie-Rochester Conference Series on Public Policy* **55**, 275–303.
Lamo, A. (2000) On convergence empirics: some evidence for Spanish regions. *Investigaciones Economicas* **24**, 681–707.
Lee, K., M.H. Pesaran and R. Smith (1997) Growth and convergence in multi country empirical stochastic Solow model. *Journal of Applied Econometrics* **12**, 357–92.
Lee, K., M.H. Pesaran and R. Smith (1998) Growth empirics: a panel data approach a comment. *Quarterly Journal of Economics* **113**, 319–23.
Li, Q. and D. Papell (1999) Convergence of international output: time series evidence for 16 countries. *International Review of Economics and Finance* **8**, 267–80.
Liu, Z. and T. Stengos (1999) Non-linearities in cross country growth regressions: a semiparametric approach. *Journal of Applied Econometrics* **14**, 527–38.
Loh, W.-Y. (2002) Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* **12**, 361–86.
Maasoumi, E., J. Racine and T. Stengos (2007) Growth and convergence: a profile of distribution dynamics and mobility. *Journal of Econometrics* **136**, 483–508.
Maddison, A. (1982) *Phases of Capitalist Development*. New York: Oxford University Press.
Maddison, A. (1989) *The World Economy in the 20th Century*. Paris: OECD.
Mankiw, N.G. (1995) The growth of nations. *Brookings Papers on Economic Activity* **1**, 275–310.
Mankiw, N.G., D. Romer and D. Weil (1992) A contribution to the empirics of economic growth. *Quarterly Journal of Economics* **107**(2), 407–37.
Marris, R. (1982) How much of the slow-down was catch-up? In R.C.O. Matthews (ed.), *Slower Growth in the Western World*. London: Heinemann.

McLachlan, G.J. and D. Peel (2000) *Finite Mixture Models*. New York: Wiley.
Michelacci, C. and P. Zaffaroni (2004) (Fractional) Beta convergence. *Journal of Monetary Economics* **45**, 129–53.
Paap, R. and H. van Dijk (1998) Distribution and mobility of wealth of nations. *European Economic Review* **42**, 1269–93.
Papageorgiou, C. and W. Masanjala (2004) The Solow model with CES technology: non-linearities with parameter heterogeneity. *Journal of Applied Econometrics* **19**, 171–201.
Perron, P. (1989) The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* **57**, 1361–401.
Persson, J. (1997) Convergence across Swedish counties, 1911–1993. *European Economic Review* **41**, 1835–52.
Pesaran, M.H. (2007) A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics* **138**, 312–55.
Phillips, P.C.B. and D. Sul (2006) Economic transition and growth. Mimeo, Yale University.
Pittau, M.G. (2005) Fitting regional income distributions in the European Union. *Oxford Bulletin of Economics and Statistics* **67**, 135–61.
Pittau, M.G. and R. Zelli (2006) Empirical evidence of income dynamics across EU regions. *Journal of Applied Econometrics* **21**, 605–28.
Pittau, M.G., R. Zelli and P. Johnson (2008) Mixture models and convergence clubs. Vassar College Economics Working Paper No. 91, available at http://irving.vassar.edu/VCEWP/VCEWP91.pdf
Prais, S.J. (1955) Measuring social mobility. *Journal of the Royal Statistical Society Series A* **118**, 56–66.
Quah, D. (1993a) Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics* **95**, 427–43.
Quah, D. (1993b) Empirical cross-section dynamics in economic growth. *European Economic Review* **37**, 426–34.
Quah, D. (1996a) Twin peaks: growth and convergence in models of distribution dynamics. *Economic Journal* **106**, 1045–55.
Quah, D. (1996b) Empirics for economic growth and convergence. *European Economic Review* **40**, 1353–75.
Quah, D. (1996c) Convergence empirics across economies with (some) capital mobility. *Journal of Economic Growth* **1**, 95–124.
Quah, D. (1997) Empirics for growth and distribution: stratification, polarization, and convergence clubs. *Journal of Economic Growth* **2**, 27–59.
Quah, D. (2001) Searching for prosperity: a comment. *Carnegie-Rochester Conference Series on Public Policy* **55**, 305–19.
Reichlin, L. (1999) Discussion of "Convergence as distribution dynamics," by Danny Quah. In R. Baldwin, D. Cohen, A. Sapir and A. Venables (eds.), *Market Integration, Regionalism, and the Global Economy*. Cambridge: Cambridge University Press.
Romer, P. (1990) Human capital and growth: theory and evidence. *Carnegie-Rochester Series on Public Policy* **32**, 251–86.
Sala-i-Martin, X. (1996a) The classical approach to convergence analysis. *Economic Journal* **106**, 1019–36.
Sala-i-Martin, X. (1996b) Regional cohesion: evidence and theories of regional growth and convergence. *European Economic Review* **40**, 1325–52.
Sala-i-Martin, X., G. Doppelhofer and R. Miller (2004) Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* **94**, 813–35.
Shioji, E. (2001) Composition effect of migration and regional growth in Japan. *Journal of the Japanese and International Economies* **15**, 29–49.
Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Summers, R. and A. Heston (1988) A new set of international comparisons of real product and price levels estimates for 130 countries, 1950–1985. *Review of Income and Wealth* **34**, 1–25.

Summers, R. and A. Heston (1991) The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* **106**, 327–68.

Tan, C.M. (2008) No one true path: uncovering the interplay between geography, institutions, and fractionalization in economic development. Mimeo, Tufts University.

Temple, J. (1998) Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* **13**, 361–75.

Tsionas, E. (2000) Regional growth and convergence: evidence from the United States. *Regional Studies* **34**, 231–8.

24
The Methods of Growth Econometrics

Steven N. Durlauf, Paul A. Johnson and Jonathan R.W. Temple

Abstract

This chapter provides an overview of current practices in the econometric analysis of economic growth. We describe some of the main methodologies that have been developed to study growth as well as some of the major empirical findings with which they are associated. Further, we explore the relationship between econometric analyses and growth theories. While we argue that there are a number of respects in which growth econometrics is not adequately integrated with growth theory, we believe that substantial methodological progress has been made.

24.1 Introduction

In this chapter we will review some of the methods that have been used in the emerging field of growth econometrics. We define this field as the use of statistical models to explain variation in growth rates and productivity levels across countries or regions. This has long been an active field of research, especially since the mid 1980s, but the position of this field within economics remains somewhat uneasy. There are thoughtful and well-informed observers who regard this form of evidence as essentially inadmissible.

It is important to understand why this is so, and it is true that the field faces many problems, not least the constraints imposed by the available data. We will review these problems in detail, but argue that there are some grounds for optimism. At least in adopting methods appropriate to these datasets and economic questions, genuine progress has been made, and we will highlight areas where further progress seems especially likely. Moreover, the prospects for useful findings seem likely to improve over time, as more and better data become available, and more is learnt about the appropriate methods for analyzing such data.

The abiding interest of the study of aggregate productivity, whether in levels or growth rates, perhaps speaks for itself. Seeking to understand the wealth of nations is one of the oldest and most important research agendas in the entire discipline. At the same time, it is also one of the areas in which genuine progress seems hardest to achieve. The contributions of individual papers can often appear slender. Even when the study of growth is viewed in terms of a collective, incremental endeavor, the various papers cannot easily be distilled into a consensus that would meet the standards of evidence routinely applied in other fields of economics.

Faced with such criticisms, one traditional defence of empirical growth research is phrased in terms of expected payoffs. Each time an empirical growth paper is written, the probability of gaining genuine understanding may be low, but when and where it does emerge, the payoff to that understanding could be vast. Moreover, some contributions can be seen as stepping stones in the development of credible evidence. That gradual process, working towards better methods and more reliable findings, may take years, but could ultimately have a high payoff. From this perspective, even if much of this evidence lacks credibility, the literature is gradually evolving towards methods and findings that should be taken more seriously.

These arguments are plausible, but rely on an important tacit assumption. They all depend on the ability of researchers and policy makers to discriminate between the status of different pieces of evidence – the good, the bad and the ugly – and this process of discrimination carries many difficulties of its own. The accuracy, or otherwise, of such judgments plays a key role in the overall development and intellectual health of any empirical literature. This reinforces the case for building an understanding of the relevant methods, their strengths and limitations, and the ways in which the existing literature is often flawed or inconclusive.

Caution will be needed throughout. Rodriguez and Rodrik (2001) begin their skeptical critique of the evidence on trade policy and growth with an apt quote from Mark Twain: "It isn't what we don't know that kills us. It's what we know

that ain't so." This point applies with some force to almost the entirety of the empirical growth literature. It is well known that some claims have not survived later scrutiny, with Levine and Renelt (1992) an especially famous demonstration of the lack of robustness of some early results. To take a more recent example, few papers in economics have been as directly influential on policy debates as the study of foreign aid and growth by Burnside and Dollar (2000). Yet its claims have been vigorously contested in a series of subsequent papers, and a careful reading of those papers might suggest that the key hypotheses cannot be reliably tested given the limitations of the available data.

Such concerns have not prevented the proposals for growth determinants from increasing with every passing year. Indeed, the number of proposed determinants is now similar to the number of countries for which data are available. It is hard to believe that all these variables are central, yet the range of possibilities, the small number of countries that can be studied, and the arbitrary choices that are often involved in estimating a specific model, all conspire to make learning about economic growth unusually difficult.

These and other difficulties have prompted the field to evolve continuously, and to adopt a wide range of methods. We argue that the statistical tools that have been applied to growth questions are sufficiently rich that they collectively define a distinct field, growth econometrics. This chapter provides an overview of the current state of this field, updating and revising our earlier survey in Durlauf, Johnson and Temple (2005). The chapter will survey the body of econometric and statistical methods that have been brought to bear on growth questions, and provide some assessments of the value of these tools. In keeping with the rest of this book, the focus will predominantly be on the application of econometric methods and techniques and their interpretation, rather than attempting to summarize substantive findings. For an earlier survey with a greater focus on substantive findings, see Temple (1999), and for an earlier evaluation of the econometrics of growth, see Durlauf and Quah (1999).

The techniques that have been used in growth econometrics largely reflect the specialized questions that naturally arise in this context. Consider the identification of empirically salient determinants of growth when the range of potential factors is large relative to the number of observations. The associated model uncertainty is one of the most fundamental problems facing growth researchers. Individual researchers, seeking to communicate the extent of support for particular growth determinants, typically emphasize a single model (or small set of models) and then carry out inference as if that model had generated the data. But there are usually other models that have equally strong claims on our attention, and hence the standard errors will often understate the true degree of uncertainty about the parameters. Moreover, the decision to report one model rather than another is often somewhat arbitrary. The need for a more systematic and objective approach, one that properly accounts for model uncertainty, naturally leads to Bayesian or pseudo-Bayesian approaches to data analysis.

Bayesian approaches seem especially natural for growth econometrics, given the paucity of the available data. This represents a major constraint on the scope for

identifying causal effects. It may seem trivial to say that the main obstacle to understanding growth is the small number of countries in the world, but the problem goes beyond a fundamental lack of variation or information. It also limits the extent to which researchers can address obvious problems, such as measurement error and parameter heterogeneity. Sometimes the problem appears in especially stark form: imagine trying to infer the consequences of democracy for long-run development in poorer countries. The twentieth century provided relatively few examples of stable, multi-party democracies among the poorer nations of the world, and so statistical evidence can make only a limited contribution to this debate, unless one is willing to make exchangeability assumptions about nations that would seem not to be credible.[1] As we discuss later in the chapter, the recent literature has explicitly sought to address this kind of problem by considering the effects of transitions to democracy using within-country variation. This leads to some interesting findings, but the short span of the available data currently precludes the long-term assessment that will often be of most interest.

If there were a larger group of countries to work with, many of the difficulties that face growth researchers could be addressed in ways that are now standard in the microeconometrics literature. For example, Harberger (1987), Solow (1994) and many others have expressed considerable skepticism about any exercise that assumes a common linear model for a heterogeneous set of countries. In principle, these concerns might be addressed by estimating more general models, using interaction terms, nonlinearities or semiparametric methods, so that the marginal effect of a given explanatory variable can differ across countries or over time. The problem is that these solutions will require large samples if the conclusions are to be robust. Similarly, some methods for addressing other problems, such as measurement error, are only useful in samples larger than those available to growth researchers. This helps to explain the need for a flexible approach, and why growth econometrics has evolved in such a pragmatic and eclectic fashion, drawing on a range of statistical methods to a greater extent than is the norm in applied econometrics.

Given the small number of countries in the world, the scope for reliable evidence is likely to rest on the use of time series variation within countries, especially as new data become available. Many empirical growth papers are now based on the estimation of dynamic panel data models with fixed effects, sometimes in conjunction with a time-varying "treatment" variable, such as the advent of democracy or trade reform. The later sections of this chapter will discuss some of the relevant technical issues, and the connection between some of these studies and the microeconometric literature on treatment effects and program evaluation. This connection not only helps to clarify the strengths and limitations of this form of evidence, but also some of the weaknesses of the empirical growth literature generally.

Despite the many difficulties that arise in empirical growth research, we believe genuine progress has been made. Researchers have uncovered stylized facts that growth theories should endeavour to explain, and developed methods to investigate the links between these stylized facts and substantive economic arguments.

They have also helped to establish the clear limits that exist in employing statistical methods to address growth questions. One implication of these limits is that narrative and historical approaches have a lasting role to play in empirical growth analysis, as we will repeatedly emphasize.

The remainder of the chapter is organized as follows. Section 24.2 briefly sketches some of the relevant stylized facts. Section 24.3 describes the relationship between theoretical growth models and econometric frameworks for growth, with a focus on cross-country growth regressions and then an alternative approach, the "levels regression." Section 24.4 describes methods for identifying growth determinants, and a range of questions concerning model specification and evaluation are addressed. Section 24.5 discusses econometric issues that arise according to whether one is using cross-section, time series or panel data. Section 24.6 provides an extended discussion of endogeneity and the associated use of instrumental variable methods. Section 24.7 covers some remaining econometric issues, including the role of outliers and measurement error. Section 24.8 concludes by highlighting some possible directions for future research.

24.2 Stylized facts

The survey by Durlauf *et al.* (2005) and the textbook by Acemoglu (2008) include overviews of stylized facts, concentrating on the period between 1960 and 2000. Some of the relevant facts can be summarized as follows:

1. Over the 40-year period as a whole, most countries have grown richer, but vast income disparities remain. For all but the richest group, growth rates have differed to an unprecedented extent, regardless of the initial level of development.
2. Although past growth is a surprisingly weak predictor of future growth, it is slowly becoming more accurate over time, and so distinct winners and losers are beginning to emerge. The strongest performers are located in East and Southeast Asia, which have sustained growth rates at unprecedented levels. The weakest performers are predominantly located in sub-Saharan Africa, where some countries have barely grown at all, or even become poorer. The record in South and Central America is also distinctly mixed. In these regions, output volatility is high, and dramatic output collapses are not uncommon.
3. For many countries, growth rates were lower in 1980–2000 than in 1960–80, and this growth slowdown is observed throughout most of the income distribution. Moreover, the dispersion of growth rates has increased. A more optimistic reading would also emphasize the growth take-off that has taken place in China and India, home to two-fifths of the world's population and, historically at least, a greater proportion of the world's poor.

Recent observers, such as Collier (2007), have particularly emphasized the emergence of a distinct set of countries, home to perhaps a billion people, where stagnation or slow growth is the norm. These countries appear locked out of

the transitions to modern economic growth that have been seen elsewhere. The idea that countries have divided across two distinct paths is weakly supported by evidence in Durlauf *et al.* (2005) that correlations of growth rates across decades are tending to increase over time. Perhaps more directly and persuasively, it is supported by the finding that the proportion of countries that have stagnated over a 20-year period has gradually increased since the 1960s. It remains to be seen whether the strong world growth and commodity boom of the early 2000s, which has clearly benefited some of the countries in sub-Saharan Africa, will help to overturn this finding.

Even this brief overview of the stylized facts reveals that there is much of interest to be investigated and understood. The field of growth econometrics has emerged through efforts to interpret and understand these facts, partly in the light of predictions made by simple growth models. The complexity of the growth process and the limitations of the available data combine to suggest that scientific standards of proof are unattainable. Perhaps the best this literature can hope for is to constrain what can legitimately be claimed, but that in itself would be an achievement. As a direct consequence of growth econometrics, there are now various claims about the world – for example, that growth is independent of the extent of financial development – that are harder to sustain than would once have been the case.

Researchers such as Levine and Renelt (1991) and Wacziarg (2002) have argued that, seen in this more modest light, growth econometrics can provide a signpost to interesting patterns and partial correlations. Ultimately, this helps to rule out some versions of the world that might otherwise seem plausible, and shift the burden of proof in particular debates. Seen in terms of establishing stylized facts, empirical studies also help to shape the demands made of future theories, and can act as a discipline on quantitative investigations using calibrated models. These are important contributions. In discussing them further, we first describe the models that are usually applied in empirical growth research.

24.3 Cross-country growth regressions: from theory to empirics

The stylized facts of economic growth have led to two major themes in the development of the literature. The first theme is the study of convergence, and we review this work in our companion chapter. The second theme concerns the identification of growth determinants. This has been the more active area in recent research, and will be our central focus in this chapter. Section 24.3.1 provides a general theoretical framework for understanding growth dynamics. The framework is explicitly neoclassical and leads to a model which is the basis for most empirical growth research. Section 24.3.2 examines the relationship between this model and the specification of a growth regression. This also provides relevant background for section 24.3.3, which focuses on the "levels regression" that has become a popular alternative in the recent literature. We will argue that its advantages, relative to growth regressions, have sometimes been exaggerated. Finally, Section 24.3.4 discusses the interpretation of error terms in growth regressions.

24.3.1 Growth dynamics: basic ideas

Our exposition in this section and the next closely follows Durlauf *et al.* (2005). At the heart of the empirical growth literature is a cross-country regression, founded on the one-sector neoclassical growth model. That model implies that, to a first-order approximation, $y^E_{i,t}$, output per efficiency unit of labor, evolves according to:

$$\log y^E_{i,t} = \left(1 - e^{-\lambda_i t}\right) \log y^E_{i,\infty} + e^{-\lambda_i t} \log y^E_{i,0}, \tag{24.1}$$

where $y^E_{i,\infty}$ is the steady-state value of $y^E_{i,t}$ and $\lim_{t\to\infty} y^E_{i,t} = y^E_{i,\infty}$. As is standard, we define output per efficiency unit of labor input as $y^E_{i,t} = Y_{i,t}/(A_{i,t}L_{i,t})$, where $Y_{i,t}$, $L_{i,t}$, and $A_{i,t}$ denote, respectively, the level of output, the labor force, and the level of efficiency in economy i at time t. The labor force is assumed to follow $L_{i,t} = L_{i,0}e^{n_i t}$, where the population growth rate n_i is constant, while $A_{i,t} = A_{i,0}e^{g_i t}$, where g_i is the (constant) rate of labor-augmenting technical progress. The parameter λ_i measures the rate of convergence of $y^E_{i,t}$ to its steady-state value and will typically depend on other parameters in the model. Assuming that $y^E_{i,0} > 0$ and so eliminating the trivial equilibrium $y^E_{i,t} = 0\ \forall t$ when the convergence rate $\lambda_i > 0$, then the value of $y^E_{i,\infty}$ is independent of $y^E_{i,0}$. In this sense, initial conditions do not matter in the long run.

Equation (24.1) expresses growth dynamics in terms of the unobservable $y^E_{i,t}$. In order to describe dynamics in terms of the observable variable, output per labor unit $y_{i,t} = \frac{Y_{i,t}}{L_{i,t}}$, we can use $y_{i,t} = y^E_{i,t}A_{i,t} = y^E_{i,t}A_{i,0}e^{g_i t}$ to write (24.1) as:

$$\log y_{i,t} - g_i t - \log A_{i,0} = \left(1 - e^{-\lambda_i t}\right) \log y^E_{i,\infty} + e^{-\lambda_i t} \left(\log y_{i,0} - \log A_{i,0}\right), \tag{24.2}$$

so that:

$$\log y_{i,t} = g_i t + \left(1 - e^{-\lambda_i t}\right) \log y^E_{i,\infty} + \left(1 - e^{-\lambda_i t}\right) \log A_{i,0} + e^{-\lambda_i t} \log y_{i,0}. \tag{24.3}$$

Defining the growth rate of output per worker between 0 and t as $\gamma_i = t^{-1}\left(\log y_{i,t} - \log y_{i,0}\right)$ and subtracting $\log y_{i,0}$ from both sides of equation (24.3) allows it to be written as:

$$\gamma_i = g_i + \beta_i \left(\log y_{i,0} - \log y^E_{i,\infty} - \log A_{i,0}\right), \tag{24.4}$$

where $\beta_i = -t^{-1}(1 - e^{-\lambda_i t})$. This expression decomposes the growth rate in country i into two distinct components: the first, g_i, measures growth due to technical progress, while the second, $\beta_i \left(\log y_{i,0} - \log y^E_{i,\infty} - \log A_{i,0}\right)$, measures growth due to the gap between initial output per worker and the steady-state value. This second source of growth is one aspect of "catching up" and, as $t \to \infty$, its importance, which reflects the role of initial conditions, diminishes to zero.

Assuming that the rates of technical progress and convergence are constant across countries allows (24.4) to be rewritten as:

$$\gamma_i = g - \beta \log y^E_{i,\infty} - \beta \log A_{i,0} + \beta \log y_{i,0}, \tag{24.5}$$

with the key implication of a negative relationship between initial levels of output and subsequent growth in a cross-section of countries, over any time period. The mechanism is diminishing returns to capital: a country further below its balanced growth path will tend to grow more quickly, other things equal, because a given rate of investment has a larger effect on the growth rates of capital and output.

24.3.2 Cross-country growth regressions

The typical cross-country growth regression, the foundation of the empirical growth literature, is motivated by adding a random error term υ_i to (24.5), giving:

$$\gamma_i = g - \beta \log y^E_{i,\infty} - \beta \log A_{i,0} + \beta \log y_{i,0} + \upsilon_i. \tag{24.6}$$

Operationalization of (24.6) requires empirical analogues for $\log y^E_{i,\infty}$ and $\log A_{i,0}$. Mankiw, Romer and Weil (1992) do this by assuming that aggregate output is described by a three-factor Cobb–Douglas production function $Y_{i,t} = K^{\alpha}_{i,t} H^{\phi}_{i,t} \left(A_{i,t} L_{i,t}\right)^{1-\alpha-\phi}$, where $K_{i,t}$ denotes physical capital and $H_{i,t}$ denotes human capital, assumed to follow the accumulation equations $\dot{K}_{i,t} = s_{K,i} Y_{i,t} - \delta K_{i,t}$ and $\dot{H}_{i,t} = s_{H,i} Y_{i,t} - \delta H_{i,t}$, respectively, where δ denotes the depreciation rate and $s_{K,i}$ and $s_{H,i}$ are the respective (time-invariant) saving rates for physical and human capital and dots above variables denote time derivatives. These assumptions imply that the steady-state value of output per effective worker is:

$$y^E_{i,\infty} = \left(\frac{s^{\alpha}_{K,i} s^{\phi}_{H,i}}{\left(n_i + g + \delta\right)^{\alpha+\phi}} \right)^{\frac{1}{1-\alpha-\phi}}, \tag{24.7}$$

giving a cross-country growth regression of the form:

$$\begin{aligned} \gamma_i = {} & g + \beta \log y_{i,0} + \beta \frac{\alpha+\phi}{1-\alpha-\phi} \log\left(n_i + g + \delta\right) \\ & - \beta \frac{\alpha}{1-\alpha-\phi} \log s_{K,i} - \beta \frac{\phi}{1-\alpha-\phi} \log s_{H,i} - \beta \log A_{i,0} + \upsilon_i. \end{aligned} \tag{24.8}$$

Mankiw *et al.* argue that initial efficiency $A_{i,0}$ should be interpreted as reflecting not just technology, which they assume to be constant across countries, but also country-specific influences such as resource endowments, climate and institutions, assumed to vary randomly in the sense that $\log A_{i,0} = \log A + e_i$, where

e_i is a country-specific shock that is distributed independently of the explanatory variables.[2] Substitution into (24.8) and defining $\varepsilon_i = \upsilon_i - \beta e_i$ gives the regression:

$$\gamma_i = g - \beta \log A + \beta \log y_{i,0} + \beta \frac{\alpha + \phi}{1 - \alpha - \phi} \log (n_i + g + \delta) - \beta \frac{\alpha}{1 - \alpha - \phi} \log s_{K,i} - \beta \frac{\phi}{1 - \alpha - \phi} \log s_{H,i} + \varepsilon_i. \quad (24.9)$$

Note an appealing feature of this regression: although the role of initial income is predicated on diminishing returns to capital, the regression can be estimated without using capital stock data. The measurement of capital stocks, especially for developing countries, is fraught with problems, as discussed in Pritchett (2000b). The specification derived by Mankiw *et al.* (1992) neatly sidesteps some of these problems.

After assuming that $g + \delta = .05$ (based on data from the US and other economies), Mankiw *et al.* use data from 98 countries over the period 1960–85 to obtain estimates of $\hat{\beta} = -.299$ (implying an estimated λ of 0.0142), $\hat{\alpha} = .48$ and $\hat{\phi} = .23$. They are unable to reject the parameter restriction, implicit in (24.9), that the final three slope coefficients sum to zero.

There are many extensions to this "augmented" Solow model that can be characterized as adding control variables, Z_i, to the regression and understood as modeling heterogeneneity in the level of technology at a given instant. In effect, the $g_i - \beta \log A_{i,0}$ terms in (24.4) are replaced with $g - \beta \log A + \pi Z_i - \beta e_i$, giving the regression:

$$\gamma_i = g - \beta \log A + \beta \log y_{i,0} + \beta \frac{\alpha + \phi}{1 - \alpha - \phi} \log (n_i + g + \delta) - \beta \frac{\alpha}{1 - \alpha - \phi} \log s_{K,i} - \beta \frac{\phi}{1 - \alpha - \phi} \log s_{H,i} + \pi Z_i + \varepsilon_i. \quad (24.10)$$

Note that (24.10) does not identify whether the Z_i are correlated with steady-state growth g_i or the initial technology term $A_{i,0}$, so a believer in a common long-run growth rate will not be dissuaded by the finding that particular choices of Z_i help predict growth beyond the Solow regressors. The attribution of the predictive content of Z_i to technology levels versus steady-state growth must largely depend on a researcher's prior beliefs about the long-run process driving the diffusion of technology. It seems plausible, however, that the controls Z_i may sometimes be associated with differences in efficiency growth g_i, rather than simply explaining differences in initial technology levels. Even if all countries have the same rate of technical progress in the long run, that assumption is somewhat implausible over a sample period as short as 20 or 30 years.

The canonical growth regression can be understood as a version of (24.10) in which the embedded parameter restrictions are ignored. A generic representation of the regression is:

$$\gamma_i = \beta \log y_{i,0} + \psi X_i + \pi Z_i + \varepsilon_i, \quad (24.11)$$

where the vector X_i contains a constant, $\log(n_i + g + \delta)$, $\log s_{K,i}$ and $\log s_{H,i}$. The variables spanned by $\log y_{i,0}$ and X_i thus represent the growth determinants suggested by the Solow model, whereas the vector Z_i represents growth determinants that lie outside that model.[3] The distinction between the Solow variables and Z_i is important in understanding the empirical literature. The Solow variables appear in many of the specifications estimated in the literature, reflecting the use of the Solow model as an organizing framework for growth analysis, but choices concerning which Z_i variables to include often vary greatly across studies. This clearly introduces a degree of arbitrariness which will be discussed later in this chapter.

Equation (24.11) represents the baseline for much of growth econometrics and these regressions are sometimes known as "Barro regressions," following the influential early contribution of Barro (1991).[4] This workhorse of empirical growth research has been generalized in a number of dimensions. Some of these extensions reflect the application of (24.11) to time series and panel data settings. Others reflect the use of more general production functions, or allow for nonlinearities and parameter heterogeneity, and we will discuss all these variants below.

24.3.3 Levels regressions

An alternative approach became especially popular after the work of Hall and Jones (1997, 1999). Their work sought to model the cross-section variation in the level of development, rather than the growth rate over a specific time interval. In other words, the dependent variable is the level of gross domestic product (GDP) per capita or GDP per worker, and there is no role for initial GDP on the right-hand side of the regression. In principle, this "levels regression" approach could be attractive for a number of reasons. It seems better suited to theories which emphasize long-run, fundamental sources of differences in development levels, such as geographic characteristics or the historical path taken by institutions. It may also give a direct answer to a key question of interest: "What is the long-run effect of a particular variable on the level of GDP per capita or GDP per worker?"

There are two important limitations of this approach. The first can be seen by contrasting it with the framework for modeling economic growth described above. In that framework, modeling the level of GDP per capita, without allowing for a conditional convergence effect, is akin to assuming that we observe the country in its long-run steady-state. Mankiw *et al.* (1992) initially adopted this approach: having derived the implications of the Solow model for the steady-state level of income, they estimated models with the level of income as the dependent variable. But they also noted that this approach is only valid if countries are distributed randomly around their steady-state positions, and so they moved on to estimate conditional convergence regressions, which did not require that strong assumption. Put differently, the levels regression approach risks omitting a relevant variable, a point that Bhattacharyya (2007) has recently emphasized.

The second limitation is that many candidate explanatory variables are likely to be endogenous to the level of income. A researcher could readily run a regression which "explains" income levels in terms of luxury car ownership, or the number of televisions, but it would be hard to interpret this as a meaningful explanation

of differences in development levels. The same point holds for variables such as institutions or the extent of corruption, where the quality of institutions, and the nature of social norms, may well be partly a function of the level of development. The standard response in the literature has been to use instrumental variables, with Frankel and Romer (1999), Hall and Jones (1999) and Acemoglu, Johnson and Robinson (2001) as three particularly well-known and influential examples. This approach is potentially informative, and has greater claims to identify causal effects than much other work. But it is also open to a range of important criticisms that we will discuss in detail in section 24.6.

Some of these criticisms could be avoided by estimating models in which initial income appears on the right-hand side. In that case, the empirical model may still be used to identify the long-run impact of institutions on the level of development. Recall that in a conditional convergence regression, the explanatory variables are not necessarily explaining long-run growth, but instead determine the long-run steady-state position that countries are converging towards. Put differently, imagine that the hypothesis of interest is the effect of institutional quality on long-run income levels. If we take two countries with the same initial level of income, the country with better institutions should grow more quickly over a given time interval. This is because it must be further below its steady-state growth path; otherwise, its better institutions would not be consistent with initial income levels that are the same.

Hence, growth regressions are often best seen as models of the height of the balanced growth path – that is, as models of long-run level effects. This point is perhaps most easily understood by considering how the analysis in Mankiw *et al.* (1992) would be adapted to include geographical characteristics or measures of institutional quality. In either case, the extension is straightforward, and the growth regression framework can be retained. These points suggest that some of the arguments usually advanced in favour of levels regressions are potentially misguided, and that the conditional convergence regression still has much to recommend it. Based on arguments like these, Bhattacharyya (2004) is an example of a study which revisits the evidence on institutions and development using a conditional convergence specification, rather than a levels regression. One remaining issue concerns whether the convergence specification can identify long-run effects with sufficient precision for the approach to be informative.

24.3.4 Interpreting errors in growth regressions

The development of the relationship between cross-country growth regressions and neoclassical growth theories in section 24.3.2 illustrates the practice, common in the literature, of deriving a deterministic growth relationship and then appending an error term in an *ad hoc* way to capture all aspects of the growth process omitted from the model. One problem with this practice is that some types of errors often have important implications for the asymptotic behavior of the estimator used in the subsequent empirical analysis. Binder and Pesaran (1999) conduct an exhaustive study of this question and conclude, *inter alia*, that if one generalizes the assumption of a constant rate of technical change so that technical change

follows a random walk, the induced non-stationarity in many of the levels series raises attendant unit root questions.

Beyond such issues of asymptotics, the *ad hoc* addition of regression errors described above leaves unanswered the question of the substantive economic assumptions implicitly made by a researcher. Brock and Durlauf (2001a) address this issue using the concept of "exchangeability" and many criticisms of growth regressions can be interpreted as claims that exchangeability has been violated. Loosely, their argument is that researchers working with regressions such as (24.11) typically think of the errors ε_i as interchangeable across observations, so that different patterns of realized errors would be equally likely to be observed if the realizations were permuted across countries. That is, the information available to a researcher about the countries is not informative about the error terms.

This idea can be formalized as F-conditional exchangeability, defined as:

$$\mu\left(\varepsilon_1 = a_1, \ldots, \varepsilon_N = a_N \,|F_1 \ldots F_N\right) = \mu\left(\varepsilon_{\rho(1)} = a_1, \ldots, \varepsilon_{\rho(N)} = a_N \,|F_1 \ldots F_N\right), \tag{24.12}$$

where, for each observation i, F_i is the associated information set available to the researcher and $\rho()$ is an operator that permutes the N indices. In the growth context, F_i may include knowledge of a country's history or culture as well as any more purely "economic" variables that are known. Omitted regressors then, for example, induce exchangeability violations as these regressors are elements of F_i. Parameter heterogeneity similarly leads to non-exchangeability.

Brock and Durlauf argue that exchangeability can be an organizing principle to connect substantive social science knowledge with the error structure. This suggests that it would be good empirical practice if researchers were to question whether or not the errors in their model are genuinely exchangeable and, if not, to determine whether the violation invalidates the purposes for which the regression is being used. As subject-specific knowledge is needed to evaluate the plausibility of exchangeability: this cannot be done in an algorithmic fashion, but instead requires judgments by the analyst.[5]

24.4 Statistical models of the growth process

Although the initial focus of empirical work in this field was the convergence hypothesis, the primary focus of more recent work has been the identification of potential growth determinants. This work may be divided into three main categories, which we discuss in turn in the next three subsections. Section 24.4.1 discusses the analysis of whether specific determinants affect growth, focusing on alternative ways to address the uncertainty about which explanatory variables should be included in a model. Section 24.4.2 explores methods to account for parameter heterogeneity and summarizes some of the relevant evidence. Section 24.4.3 focuses on the analysis of nonlinearities and multiple regimes in the growth process. Models of poverty traps and endogenous growth are often highly nonlinear, or associated with multiple steady-states in the growth process,

with important implications for econometric practice. Observe that in each case, while the associated analyses are often motivated by formal theories, operationally they represent efforts to develop statistical growth models that are consistent with particular specification tests.

24.4.1 Specifying explanatory variables in growth regressions

The first point to emphasize is the sheer number of growth determinants that have been proposed; the number of potential growth determinants now approaches the number of countries available. The table in Appendix B of Durlauf *et al.* (2005) lists many of them, with references to studies that represent either the first use of the variable or an especially well-known use of the variable. That table contains 145 different regressors, the vast majority of which have been found to be significant at conventional levels in at least one study. One reason why so many alternative variables have been identified is due to questions of measurement. For example, even given a specific claim that political freedom might affect growth, there could be many ways to measure such freedoms. But even taking this into account, the multiplicity of possible theories is striking. Durlauf *et al.* organize the empirical literature into 43 distinct growth "theories" (that is, conceptually distinct growth determinants) and it would not be difficult to add further examples.

Moreover, this list does not consider interactions between variables or nonlinear transformations of variables, even though both are often used. The range of possibilities hints at one of the fundamental problems with empirical growth research: a lack of consensus on which growth determinants ought to be included in a statistical model. In this section, we discuss attempts to address this question and to limit the degree of arbitrariness otherwise present.

To fix ideas, let S_i denote the set of regressors always included by the researcher while R_i denotes the set of additional candidate regressors, so that:

$$\gamma_i = \psi S_i + \pi R_i + \varepsilon_i, \tag{24.13}$$

is the putative growth regression. The inclusion of a variable in S does not mean the researcher is certain that it influences growth, only that it is to be included in all models considered. If one takes the regressors that comprise R as fixed, then statements about elements of ψ are straightforward. A frequentist approach to inference will compute an estimate of the parameter ψ with an associated distribution that depends on the data-generating process (DGP). Bayesians would compute a posterior probability density of ψ given the researcher's prior, the data, and the assumption that the linear model is correctly specified. Designating the available data as D and a particular model as m, this posterior may be written as $\mu(\psi | D, m)$.

While extant growth theories can be used to identify candidates for R, the fundamental problem in developing statistical statements about either $\hat{\psi}$ or $\mu(\psi | D, m)$ is that there do not exist good theoretical reasons to favour one particular model and exclude others. As Brock and Durlauf (2001a) point out, growth theories are typically "open-ended" in the sense that different theories are often compatible with one another, rather than mutually exclusive. For example, a theory that institutions matter is not logically inconsistent with a theory that emphasizes the role

of geographic characteristics. A set of K potential growth theories, all of which are logically compatible with combinations of one another, implies that there exist $2^K - 1$ potential specifications of the form (24.13), each one of which corresponds to a particular combination of theories.

One approach to this model uncertainty is to examine robustness with respect to variation in the model specification. This is the idea behind the classic Levine and Renelt (1992) paper which, building on Leamer (1978, 1983) and Leamer and Leonard (1983), used extreme bounds analysis to assess growth determinants. For a model $m \in M$ within the space of possible models, the growth process is specified as:

$$\gamma_i = \psi_m S_i + \pi_m R_{i,m} + \varepsilon_{i,m}, \tag{24.14}$$

where the m subscripts reflect the model specific nature of the parameters and associated errors. Let $\hat{\psi}_m$ denote the point estimate of ψ for every $m \in M$ and let the vector S be composed of a variable of interest, s_I, and other variables which are included in all the specifications considered. Motivated by Leamer (1983), Levine and Renelt (1992) use the following rule: there is strong evidence that s_I affects growth if (and only if) the sign of the associated regression coefficient $\hat{\psi}_{I,m}$ is constant and the coefficient estimate is statistically significant across all $m \in M$. In their analysis, the vector S includes the constant term, initial income, the investment share of GDP, the secondary school enrollment rate, and the population growth rate, as suggested by the augmented Solow model. The possible models are distinguished by alternative combinations of between one and three variables, taken from a set of seven; these correspond to alternative choices of $R_{i,m}$. Applying their rule, they conclude that the only robust growth determinants are initial income and the share of investment in GDP.

These two findings are confirmed in subsequent work by Kalaitzidakis, Mamuneas and Stengos (2000), who allow for potential nonlinearities in (24.14) by considering models of the form:

$$\gamma_i = \psi_m S_i + f_m\left(\pi R_{i,m}\right) + \varepsilon_{i,m}, \tag{24.15}$$

where $f(\cdot)$ is a function allowed to vary across specifications of R. Like Levine and Renelt, they find that initial income and physical capital investment rates are robust determinants of growth, but also find that inflation volatility and a measure of exchange rate distortions are robust, providing an example of how failing to account for nonlinearity in one set of variables can mask the importance of another. A related exercise by Minier (2007) allows for nonlinearities in a parametric way (including squared and interaction terms) and also examines what happens when the Levine–Renelt sample is restricted to the lowest 75% of countries when ranked by initial income, which broadly corresponds to a sample of developing countries. She shows that a wider range of variables is found to be robust, including several fiscal indicators that Levine and Renelt had classed as fragile. An open question here is the extent to which certain kinds of nonlinearity, such as quadratic terms, may be highly sensitive to outliers. Temple (2000b) discussed how an extreme bounds analysis could be combined with robust estimation methods.

More fundamentally, as explained in Brock and Durlauf (2001a) and Brock, Durlauf and West (2003), even a sophisticated extreme bounds approach is somewhat problematic when viewed from a decision-theoretic perspective. Suppose, for example, that interest in ψ_l derives from country i's consideration of a policy change in which a variable $s_{i,l}$ will be increased by one unit. Let $El\left(s_{i,l}, m\right)$ represent the policy maker's expected loss associated with a policy indicator in country i, and suppose that she is only interested in the case where the increase in the indicator will raise growth, so that the policy change is sensible only if $\hat{\psi}_{l,m} > 0$. One can approximate a t-statistic rule, requiring that the coefficient estimate for s_l be statistically significant, as:

$$El\left(s_{i,l}+1, m\right) - El\left(s_{i,l}, m\right) = \left(\hat{\psi}_{l,m} - 2sd\left(\hat{\psi}_{l,m}\right)\right) > 0, \tag{24.16}$$

where $sd\left(\hat{\psi}_{l,m}\right)$ is the estimate of the standard deviation associated with $\hat{\psi}_{l,m}$ and the required significance level is assumed to correspond to a t-statistic of 2. As odd as it seems, this is the form of the loss function implicitly assumed when t-statistics are used to make policy decisions. Extreme bounds analysis requires that (24.16) holds for every $m \in M$ so that $El\left(s_{i,l}\right)$, the expected loss for a policy maker when conditioning only on the policy variable, must have the property that:

$$El\left(s_{i,l}+1\right) - El\left(s_{i,l}\right) > 0 \Rightarrow El\left(s_{i,l}+1, m\right) - El\left(s_{i,l}, m\right) > 0 \ \forall m, \tag{24.17}$$

which means that the policy maker must have minimax preferences with respect to model uncertainty. That is, she will make the policy change only if it yields a positive expected payoff under the least favorable model in M. Such extreme risk aversion seems hard to justify, notwithstanding claims that individuals do indeed assess uncertainty about models in different ways to the uncertainty that arises within models.[6]

Even when one moves away from decision-theoretic considerations, extreme bounds analysis encounters substantial problems.[7] One practical criticism is developed carefully in Hoover and Perez (2004). Using simulations, they show that an extreme bounds analysis can easily lead to the conclusion that many growth determinants are fragile even when they are part of the DGP. Intuitively, adding more and more irrelevant variables to such an analysis always has the potential to overturn any specific finding, including relationships that are part of the true DGP. Hoover and Perez also find that the procedure has poor power properties, in the sense that some regressors that do not matter may spuriously appear to be robust. The first concern, that extreme bounds analysis is excessively conservative, had already led Sala-i-Martin (1997a, 1997b) to propose less demanding criteria. These were justified in essentially heuristic terms, with their statistical properties remaining somewhat uncertain. Again using simulations, Hoover and Perez (2004) found that these newer robustness criteria, although less demanding, could again have poor size properties, in the sense that "true" growth determinants are still likely to fail to be identified.

Dissatisfaction with extreme bounds analysis and its close relatives has led some authors to advocate more systematic methods for model selection. Hendry and Krolzig (2004) and Hoover and Perez (2004) both employ the general-to-specific modeling methodologies associated with the research program of David Hendry, in order to select one version of (24.13) out of the model space. Essentially, these papers use algorithms which select a particular regression model from a space of possible models, through comparisons based on a set of statistical tests. The attractiveness of this approach is closely linked to that of the broader Hendry research program (see Hendry, 1995). We do not provide an extended evaluation here, but note that relying solely on model selection procedures does not possess a clear decision-theoretic justification, which makes it harder to evaluate the output of the procedure in terms of the objectives of a researcher. That said, automated procedures have the important virtue that they can identify sets of models that are well supported by the available data, and may be especially useful as a preliminary step in model-building.

In our judgment, the most promising approach to model uncertainty is the use of model averaging in the Bayesian tradition, and especially the methods developed by Adrian Raftery and co-authors (for example, Raftery, Madigan and Hoeting, 1997). Versions of these methods have been applied to growth data in Brock and Durlauf (2001a), Brock *et al.* (2003), Ciccone and Jarocinski (2007), Durlauf, Kourtellos and Tan (2008), Fernandez, Ley and Steel (2001a), Masanjala and Papageorgiou (2005) and Sala-i-Martin, Doppelhofer and Miller (2004), among others. The basic idea is to treat the identity of the "true" growth model as an unobservable variable about which the researcher is inevitably uncertain.[8] In order to account for this, each element m in the model space M is associated with a posterior model probability $\mu(m|D)$. By Bayes' rule,

$$\mu(m|D) \propto \mu(D|m)\,\mu(m), \tag{24.18}$$

where $\mu(D|m)$ is the likelihood of the data given the model and $\mu(m)$ is the prior model probability. These model probabilities are used to eliminate the dependence of parameter analysis on a specific model. For frequentist estimates, averaging is done across the model-specific estimates $\hat{\psi}_m$ to produce an estimate $\hat{\psi}$ via:

$$\hat{\psi} = \sum_m \hat{\psi}_m \mu(m|D), \tag{24.19}$$

whereas, in the Bayesian context, the dependence of the posterior probability measure of the parameter of interest, $\mu(\psi|D,m)$, on the model choice is eliminated via standard conditional probability arguments, that is:

$$\mu(\psi|D) = \sum_{m\in M} \mu(\psi|D,m)\mu(m|D). \tag{24.20}$$

Brock *et al.* (2003) argue that the strategy of constructing posterior probabilities that are not model-dependent is especially appropriate when the objective of the statistical exercise is to evaluate alternative policy questions, such as whether to

change an element of S_i by one unit. This assumes that the ultimate goal of the exercise is to estimate a parameter, rather than to identify the "best" growth model.

Bayesian model averaging is still relatively new, and many practical questions arise. The implementation of the approach in economics has often closely followed the early work by Raftery (1995) and Raftery *et al.* (1997). One issue concerns the specification of priors on parameters within a model. In line with Raftery's approach, Brock and Durlauf (2001a), Brock *et al.* (2003), and Sala-i-Martin *et al.* (2004) assume a diffuse prior on the model specific coefficients. This has the advantage that, when the errors are normal with known variance, the posterior expected value of ψ, conditional on the data D and model m, is the ordinary least squares (OLS) estimator $\hat{\psi}_m$. One disadvantage of this approach is that, since the diffuse prior on the regression parameters is improper, one has to be careful that the posterior model probabilities associated with the prior are interpretable. But as long as the posterior model probabilities include an appropriate penalty for model complexity, we do not see any conceptual problem in interpreting this approach as strictly Bayesian. Brock and Durlauf (2001a), Brock *et al.* (2003), and Sala-i-Martin *et al.* (2004) all compute posterior model probabilities using Bayesian information criterion (BIC)-adjusted likelihoods. Fernandez *et al.* (2001a) and Masanjala and Papageorgiou (2005) employ proper priors and therefore avoid any such concerns.[9] So far there is only limited evidence that use of improper versus proper priors has important consequences in practice. Masanjala and Papageorgiou compare results using proper priors with the improper priors we have described, and find that the choice is unimportant in their application. Arguably the most important evidence that the conventional approach is problematic can be found in Ciccone and Jarocinski (2007), and we will discuss their study in detail at the end of this section.

Other work has examined the effects of different proper priors. Letting Z_i denote $\left(S_i, R_{i,m}\right)$ and $\eta_m = \begin{pmatrix} \psi_m \\ \pi_m \end{pmatrix}$, Fernandez *et al.* (2001b) propose the use of the Zellner (1986) g-prior:

$$\mu\left(\eta_m \left| m, \sigma_{\varepsilon,m}\right.\right) \propto N\left(0, \sigma^2_{\varepsilon,m}\left(gZ'_mZ_m\right)^{-1}\right), \tag{24.21}$$

with $g = 1/\max\left\{n, k_m^2\right\}$, n denoting the number of observations and k_m denoting the number of regressors in model m. Ley and Steel (2008) extend this analysis by considering the performance of this within-model prior for different model space priors; the model space priors all assume that each variable appears in the true model with equal probability and that variable inclusions exhibit conditional independence as described above. They conclude that, for growth contexts, models of interest typically involve a sufficiently large number of growth determinants such that $k_m^2 > n$, so that $g = k_m^{-2}$. They additionally argue that the frequentist/Bayesian hybrid employed by Sala-i-Martin *et al.* (2004) is well approximated by the $g = n^{-1}$ case and therefore may be criticized on the grounds that this value of g is generally inappropriate. A different approach to within-model parameter priors is proposed by Eicher, Papageorgiou and Raftery (2008). They suggest the use of what Kass and

Wasserman (1995) called a unit information prior; the prior is implicitly defined to ensure that posterior model probabilities are well approximated by the Bayesian information or Schwarz criterion. This paper concludes that the combination of a unit information prior for within-model probabilities and a uniform prior across models produces superior performance against a range of alternatives.

A distinct approach to within-model priors is developed by Magnus, Powell and Prufer (2008), who employ the distinction between variables that are included in all models, S_i, and variables that are model-specific, $R_{i,m}$. For each model they propose estimating a regression based on S_i and $e_{i,m}$, the residual of $R_{i,m}$ when projected against S_i. A Laplace prior is assigned to the $e_{i,m}$ parameters, which corresponds to the idea that a researcher is ignorant as to whether the absolute value of the t-statistic (computed at population values) is greater than 1. They argue that this prior has the advantage that it is not dependent on an arbitrary choice of a parameter, such as occurs because of the need to set g in (24.21).

There is also no consensus on the appropriate specification of the prior model probabilities $\mu(m)$. In the model averaging literature, the usual assumption has been to assign equal prior probabilities to all models in M. This corresponds to assuming that the prior probability that a given variable appears in the "true" model is 0.5 and that the probability that one variable appears in the model is independent of whether others appear. Sala-i-Martin *et al.* (2004) consider modifications of this prior, in which the probability that a given variable appears in the true model is $p < 0.5$ while preserving the assumption that the inclusion probabilities for each variable are independent. The Sala-i-Martin *et al.* probabilities may be written as:

$$p^{\#(m)}\left(1-p\right)^{k-\#(m)}, \tag{24.22}$$

where $\#(m)$ denotes the number of regressors in model m. This prior can be generalized by treating (24.21) as the conditional probability of a model given p and then assigning a prior to p, an idea developed in Brown, Vannucci and Fearn (1998) and applied to growth regressions by Ley and Steel (2008).

The conditional independence assumption may be unappealing given collinearity between regressors. One reason for this is that the different growth regressors are sometimes included as proxies for a common growth theory. Durlauf *et al.* (2008) address this by using dilution priors, due to George (1999), which downweight models that contain potentially "redundant" variables, as when a dataset contains multiple proxies for the same underlying economic concept. This is done by multiplying (24.21) by the determinant of the correlation matrix for the included variables in a given model, which reweights model probabilities so as to downweight those with redundant variables.

The issue of redundant variables is part of a wider set of conceptual problems which arise when a researcher uses a prior which treats all models as equally likely. Brock and Durlauf (2001a), Brock *et al.* (2003) and Durlauf *et al.* (2008) criticize the widespread use of prior model probabilities which assume that the inclusion of one variable should be independent of the inclusion of another. The conceptual problem is analogous to the "red bus/blue bus" problem in discrete choice theory.

Clearly some regressors are similar, such as alternative measures of trade openness, whereas other regressors are quite disparate, such as measures of geography versus institutional quality.

To address this, Brock *et al.* (2003) propose a tree structure to organize model uncertainty for linear growth models. First, they argue that growth models suffer from theory uncertainty. Hence, one can identify alternative classes of models based on what growth theories are included. Second, for each specification of a body of theories to be embedded, they argue there is specification uncertainty. A given set of theories requires determining whether the theories interact, whether they are subject to threshold effects or other types of nonlinearity, and so on. Third, for each theory and model specification, there is measurement uncertainty; to say that climate affects growth does not specify the relevant empirical proxies, such as the number of sunny days or average temperature. Finally, each choice of theory, specification and measurement is argued to suffer from heterogeneity uncertainty, which means that it is unclear which sub-sets of countries obey a common linear model. Brock *et al.* (2003) argue that one should assign priors that account for the interdependence implied by this structure in assigning model probabilities.

We now briefly summarize some of the main findings of model averaging studies. Sala-i-Martin *et al.* (2004) find that four variables have posterior model inclusion probabilities above 0.9, namely initial income, the fraction of GDP from mining, the number of years the economy has been open,[10] and the fraction of the population following Confucianism. Fernandez *et al.* (2001a) also find that four variables have posterior model inclusion probabilities above 0.9, substantially overlapping with Sala-i-Martin *et al.* (2004) despite working with a different model space: initial income, the fraction of the population following Confucianism, life expectancy, and the share of equipment investment in GDP.[11] These findings appear to be somewhat dependent on details of the way in which priors are assigned within and across the model space.[12] Eicher *et al.* (2008) work with the same dataset as Fernandez *et al.* (2001a) and find that the combination of a unit information within-model prior and a uniform model space prior generates 16 different growth variables whose posterior inclusion probabilities are 0.9 or greater. Interestingly, if one reduces the variable inclusion probability so that the expected number of variables in a model is seven (which is the prior used by Sala-i-Martin *et al.*, 2004), only four variables have posterior inclusion probabilities above 0.9; these are the same as those identified by Fernandez *et al.* (2001a).

Durlauf *et al.* (2008) have applied model averaging to an unbalanced panel based on three time periods spanning 1965–94, with a focus on evaluating the robustness of various fundamental growth determinants. They confirm the importance of initial income and investment, and also find a role for population growth and two macroeconomic variables, the share of government consumption (net of defense and education spending) in GDP and the rate of inflation. Their approach differs from those above in that the conventional BIC approximation is applied in the context of two-stage least squares (2SLS), rather than OLS. The development of a rigorous combination of model averaging and instrumental variable methods is an interesting area for further work.

In other applications, Brock and Durlauf (2001a) and Masanjala and Papageorgiou (2005) have employed model averaging to study the reason for the poor growth performance of sub-Saharan Africa. Both papers identify some important differences between Africa and the rest of the world in terms of the relevant growth determinants. The study of Brock *et al.* (2003) takes this idea further by exploring the use of growth regressions in the evaluation of policy recommendations. Specifically, the paper assesses the question of whether a policy maker should favor a reduction of tariffs for sub-Saharan African countries. The analysis assumes that the policy maker has a specific set of mean-variance preferences with respect to the effects of a change in current policies. At first glance, the analysis supports a tariff reduction: it shows that a policy maker with these preferences should be in favor, unless the policy maker has a strong prior that sub-Saharan African countries obey a growth process distinct from the rest of the world. In the latter case, there is sufficient uncertainty about the relationship between tariffs and growth for African countries that a change in trade policy cannot be justified. In statistical terms, this result is a consequence of the strong prior belief in the possibility of heterogeneity. The prior implies that the growth experiences of non-African countries will have little effect on the precision of estimates of marginal effects that are constructed using data for sub-Saharan Africa.

To date, perhaps the most important critique of the Bayesian approach, at least as applied to growth data, is that developed in Ciccone and Jarocinski (2007). Their central point is that agnostic empirical approaches, such as model averaging, appear to be sensitive to modest changes in the data. One of their examples is based on Sala-i-Martin *et al.* (2004) and its application of model averaging to 1960–96 growth determinants, using Penn World Table (PWT) version 6.0 data for output levels and growth rates. When Ciccone and Jarocinski update those results with the revised data provided in PWT version 6.1, they find that the two versions of the PWT lead to disagreement on 13 of 25 determinants of 1960–96 growth that emerge using one version of the data or the other. When they carry out a further exercise, now using the latest PWT 6.2 data for 1960–96, they again find scope for considerable disagreement. They illustrate the potential concerns using a Monte Carlo study, which confirm that the Bayesian approach can be sensitive to data revisions that are modest by the historical standards of revisions to the PWT. Their results imply that a priority for future research should be the development of methods which are less sensitive to small changes in the data, perhaps by reformulating the priors on model coefficients or explicitly allowing for mismeasured data. In this respect, it is worth noting that the growth literature has made relatively little use of the methods for simultaneous model selection and outlier identification that were developed by Hoeting, Raftery and Madigan (1996).

24.4.2 Parameter heterogeneity

The estimation of linear growth models, at best an approximation to the true law of motion of an economy, has generated unease about the statistical foundations of the exercise. It is difficult to sustain the claim that the data for very

different countries are realizations of a common DGP, and many of the modeling assumptions and procedures of the empirical growth literature can appear arbitrary. In the well-known question posed by Harberger (1987): "What do Thailand, the Dominican Republic, Zimbabwe, Greece and Bolivia have in common that merits their being put in the same regression analysis?"

The extent to which this objection is fundamental remains an open question, but there seems to be agreement that, when studying growth, it will be difficult to recover a DGP even if one exists, and the prospects for recovering causal effects are clearly weak. The shortcomings of the relevant economic theory, as well as those of the data and econometric analysis, are considerable. Those who will be satisfied only with the specification and estimation of a structural model, in which parameters are either "deep" and invariant to policy, or correspond to precisely defined causal effects within a coherent theoretical framework, are bound to be disappointed. The more appropriate goal for the growth literature is less ambitious: investigating whether or not particular hypotheses have any support in the data and whether it is possible to rule out some possible claims about the world, or at least shift the burden of proof from one side of a debate to another. In practice, growth researchers are looking for patterns and systematic tendencies that, in combination with historical analysis, case studies, and relevant theoretical models, can increase our understanding of the growth process. A related goal, more difficult than it may appear at first sight, is to communicate the degree of support for any patterns identified by the researcher.

The issue raised by Harberger is essentially that of parameter heterogeneity. Why should we expect disparate countries to lie on a common regression surface? Of course, this criticism could be made of most empirical work in social science: whether the data points reflect the actions and characteristics of individuals and firms, or the aggregations of their choices that are used in macroeconometrics. It is the small sample sizes available to growth researchers that limit the scope for addressing this problem, and mean that Harberger's remark retains some force. In other contexts, an appropriate response would be to use a model that has sufficient flexibility to be a good approximation. But this approach will often be fragile when the sample is rarely greater than 100 observations, as is the case when studying economic growth using cross-country data.

If parameter heterogeneity is present, the consequences are potentially serious, except in the special case where the slope parameters vary randomly across units, and are distributed independently of the variables in the regression and the disturbances. In this case, the coefficient estimate should be an unbiased estimate of the mean of the parameter distribution. However, the assumption of independence will often be unwarranted. For example, when estimating the relationship between growth and investment, the marginal effect of investment will almost certainly be correlated with aspects of the economic environment that should also be included in the regression, such as political stability or the protection of property rights.

Some researchers have allowed greater flexibility in the functional form of their models, often beginning with the canonical Solow regression which, for

comparison purposes, we restate:

$$\gamma_i = k + \beta \log y_{i,0} + \pi_n \log (n_i + g + \delta) + \pi_K \log s_{K,i} + \pi_H \log s_{H,i} + \varepsilon_i. \quad (24.23)$$

For example, Liu and Stengos (1999) estimate a semiparametric partially linear version of this model, namely:

$$\gamma_i = k + f_\beta(\log y_{i,0}) + \pi_n \log (n_i + g + \delta) + \pi_K \log s_{K,i} + f_{\pi_H}(\log s_{H,i}) + \varepsilon_i, \quad (24.24)$$

where $f_\beta(\cdot)$ and $f_{\pi_H}(\cdot)$ are arbitrary functions except for variance smoothness requirements. They find that the value of $f_\beta(\log y_{i,0})$ is only negative when initial per capita income exceeds about \$1,800; below that level, there is less evidence for the conditional convergence effect predicted by the Solow model. Banerjee and Duflo (2003) use a similar approach to study nonlinearity in the relationship between changes in inequality and growth, estimating a version of (24.24) where Solow-type variables and some additional variables enter in a linear way, supplemented by a nonlinear function $f_G(G_{i,t} - G_{i,t-5})$, where $G_{i,t}$ is the Gini coefficient.

Using an alternative method, Durlauf, Kourtellos and Minkin (2001) estimate a version of the augmented Solow model that allows the parameters to vary across countries as functions of initial income:

$$\begin{aligned}\gamma_i = {} & k(y_{i,0}) + \beta(y_{i,0}) \log y_{i,0} + \pi_n(y_{i,0}) \log (n_i + \delta + g) \\ & + \pi_K(y_{i,0}) \log s_{K,i} + \pi_H(y_{i,0}) \log s_{H,i} + \varepsilon_i. \end{aligned} \quad (24.25)$$

Hence each initial income level defines a distinct Solow regression, and thereby shifts the emphasis away from nonlinearity and towards parameter heterogeneity. This study indicates considerable parameter heterogeneity, especially among the poorer countries, and confirms the Liu and Stengos (1999) finding that $\beta(y_{i,0})$ is positive for low $y_{i,0}$ values and negative for higher ones. Durlauf *et al.* (2001) also find that $\pi_K(y_{i,0})$ fluctuates greatly over the range of $y_{i,0}$ values in their sample. The extension by Kourtellos (2003a) uncovers systematic heterogeneity in the parameters on initial literacy and initial life expectancy. A varying coefficient approach is also employed in Mamuneas, Savvides and Stengos (2006), who consider a model in which the coefficient on human capital is allowed to vary with the level of human capital and a measure of trade openness. Constancy of the human capital coefficient is rejected across a range of specifications.

At a minimum, it makes sense for empirical researchers to test for neglected parameter heterogeneity, either using interaction terms or by carrying out diagnostic tests. As an alternative, some authors have used panel data to identify parameter heterogeneity without the imposition of a functional relationship between parameters and various observable variables. An important early contribution along these lines is Canova and Marcet (1995). Defining $s_{i,t}$ as the logarithm of the ratio of a country's per capita income to the time t international aggregate value, and using data either on the regions of Europe or 17 western European countries, they

estimate models of the form:

$$s_{i,t} = a_i + \rho_i s_{i,t-1} + \varepsilon_{i,t}. \tag{24.26}$$

The long-run forecast of $s_{i,t}$ is given by $\frac{a_i}{1-\rho_i}$, with $1 - \rho_i$ being the rate of convergence towards that value. Restricting the parameters a_i and ρ_i to be constant across i gives a standard β-convergence test and yields an estimated annual rate of convergence of approximately 2%, similar to other findings in the literature. But allowing for heterogeneity in these parameters produces a "substantial" and statistically significant dispersion of the implied long-run $s_{i,t}$ forecasts. The positive correlation of those forecasts with the initial values of $s_{i,t}$ implies a dependence of long-run outcomes on initial conditions that contradicts the convergence hypothesis. For the country-level data, differences in initial conditions explain almost half the cross-sectional variation in long-run forecasts. This shows how key findings can be sensitive to the treatment of parameter heterogeneity, and we return to this issue when we discuss panel data models in section 24.5.

24.4.3 Nonlinearity and multiple regimes

We now discuss research that has attempted to disentangle the roles of heterogeneous structural characteristics and initial conditions in determining growth outcomes. These papers employ a variety of statistical methods, but there is considerable agreement in their findings. Many of them indicate the existence of convergence clubs even after accounting for the role of structural characteristics. We have discussed some of this work in our companion chapter on convergence (Chapter 23 in this volume), and here we concentrate on the wider implications of multiple regimes for the statistical methods that should be adopted.

One of the first contributions to this literature was Durlauf and Johnson (1995). They used classification and regression tree (CART) methods to search for nonlinearities.[13] More specifically, the CART procedure identifies sub-groups of countries that obey a common linear growth model based on the Solow variables. These sub-groups are identified by initial income and literacy; a typical sub-group l is defined by countries whose initial income lies within the interval $\vartheta_{l,y} \leq y_{i,0} < \bar{\vartheta}_{l,y}$ and whose literacy rate L_i lies in the interval $\vartheta_{l,L} \leq L_i < \bar{\vartheta}_{l,L}$. The number of sub-groups and the boundaries for the variable intervals that define them are chosen by an algorithm that trades off model complexity (the number of sub-groups) and goodness of fit. Because the procedure uses rules to sequentially split the data into finer and finer sub-groups, it organizes the data into a tree structure, where the branches of the tree ultimately divide the sample into groups of countries that follow distinct regimes.

Durlauf and Johnson (1995) also test the null hypothesis of a common growth regime against the alternative hypothesis of a growth process with multiple regimes. Taking Mankiw *et al.* (1992) as their starting point, and using income per capita and the literacy rate as possible threshold variables, Durlauf and Johnson reject the single regime model. This finding has been confirmed in subsequent research by Papageorgiou and Masanjala (2004). They estimate a version of the

Solow model that is based on a constant elasticity of substitution (CES) production function, building on Duffy and Papageorgiou (2000). By using the Hansen (2000) approach to sample splitting and threshold estimation, they find statistically significant evidence of thresholds in the data. The estimated thresholds divide the sample into four distinct growth regimes that are broadly consistent with those found by Durlauf and Johnson.[14] Relative to the regression tree approach, the Hansen methods have the significant advantage of allowing inference on the level of the estimated threshold.

Another closely related analysis is that of Tan (2004). He employs a procedure known as GUIDE (generalized, unbiased interaction detection and estimation), due to Loh (2002), to identify sub-groups of countries which obey a common growth model. Relative to CART, the GUIDE algorithm has two advantages. First, the algorithm explicitly looks for interactions between explanatory variables when identifying splits. Second, the penalties for model complexity are supplemented with some within-model testing, which reduces the tendency for CART procedures to produce an excessive number of splits in finite samples. Tan (2004) finds strong evidence that measures of institutional quality and ethnic fractionalization define convergence clubs across a wide range of countries. He also finds some evidence that geographic characteristics distinguish the growth process for sub-Saharan Africa from the rest of the world.

Further research has corroborated the evidence of multiple regimes using alternative statistical methods, including projection pursuit.[15] Desdoigts (1999) uses these methods to identify groups of countries with relatively homogeneous growth experiences based on the characteristics and initial conditions of each country. The idea is to find the orthogonal projections of the data into low dimensional spaces that best display some interesting feature of the data; this can be seen as a generalization of principal components analysis. When using principal components analysis, a researcher will typically retain only the components needed to account for "most" of the variation in the data. Similarly, in projection pursuit methods, a researcher will consider as many dimensions as needed to account for "most" of the clustering in the data. Some evidence of their utility can be found in Kourtellos (2003b). Unlike Desdoigts, Kourtellos uses projection pursuit to construct models of the growth process. Formally, he estimates models of the form:

$$\gamma_i = \sum_{l=1}^{L} f_l \left(y_{i,0}\beta_l + X_i\psi_l + Z_i\pi_l \right) + \varepsilon_i. \tag{24.27}$$

Each element in the summation represents a distinct projection. Kourtellos uncovers evidence of two steady-states, including one that corresponds to countries with low initial income and low initial human capital.

Another approach to multiple regimes is employed by Bloom, Canning and Sevilla (2003). This is based on the observation that if long-run outcomes are determined by fundamental forces alone, the relationship between exogenous variables and income levels ought to be unique. If initial conditions play a role there will be multiple relationships, one for each basin of attraction defined by the initial

conditions. If there are two (stochastic) steady-states, and large shocks are sufficiently infrequent,[16] the system will, under suitable regularity conditions, exhibit an invariant probability measure that can be described by a "reduced form" model in which the long-run behavior of $\log y_{i,t}$ depends only on the exogenous variables, m_i, such as:

$$\log y_{i,t} = \log y_1^* (m_i) + u_{1,i,t} \text{ with probability } p(m_i), \tag{24.28}$$

and:

$$\log y_{i,t} = \log y_2^* (m_i) + u_{2,i,t} \text{ with probability } 1 - p(m_i), \tag{24.29}$$

where $u_{1,i,t}$ and $u_{2,i,t}$ are independent, zero mean deviations from the steady-state log means $\log y_1^*(m_i)$ and $\log y_2^*(m_i)$ respectively, and $p(m_i)$ is the probability that country i is in the basin of attraction of the first of the two steady-states. From the perspective of the econometrician, $\log y_{i,t}$ thus obeys a mixture process. The two steady-states associated with (24.28) and (24.29) might be interpreted as a low-income regime or poverty trap, and a high-income or growth regime, respectively. Bloom *et al.* (2003) estimate a linear version of this model using 1985 income data from 152 countries, with the absolute latitude of the country as the fundamental exogenous variable. They are able to reject the null hypothesis of a single regime model in favor of the alternative of a model with two regimes: a high-income steady-state in which income is independent of absolute latitude, and a low-income ("agricultural") steady-state in which income is increasing in absolute latitude. In addition, the probability of being in the high-income steady-state is found to be increasing in absolute latitude.

Adding extra complexity to this model could well be constrained by the small number of countries available. More generally, the empirical investigation of multiple steady-states raises some complex problems for standard methods. One response is to draw more heavily on structural theoretical models as a framework for understanding the data, as in Graham and Temple (2006). Another possibility would be to exploit time series variation in a single country, in order to identify jumps from one equilibrium or steady state to another. But in either case, it is clear that these forms of analysis would have to proceed under strong assumptions, some of which will be difficult to test.

24.5 Time series methods, panel data and event studies

Our discussion now explores alternative ways of modeling growth: time series models, the use of panel data, and studies based on discrete "events" which draw on panel data methods. At the risk of stating the obvious, choices on research design involve significant trade-offs, which depend partly on statistical considerations and partly on the economic context. This means that attempts at universal prescriptions are misguided, and we will try to show the desirability of matching techniques to the economic question at hand. One example, to be discussed below, would be the choice between panel data methods and the estimation of separate

time series regressions for each country. The use of panel data is likely to increase efficiency and allow richer models to be estimated, but at the expense of potentially serious biases if the parameter homogeneity assumptions are incorrect. This trade-off between robustness and efficiency is another running theme of our survey. The scientific solution would be to base the choice of estimation method on a context-specific loss function, but this is clearly a difficult task, and in practice more subjective decisions are involved.

Section 24.5.1 examines the econometric issues that arise in the use of time series data to study growth, emphasizing some of the drawbacks of this approach. Section 24.5.2 discusses the many issues that arise when panel data are employed, an increasingly popular approach to growth questions. We consider the estimation of dynamic models in the presence of fixed effects, and alternatives to standard procedures. Section 24.5.3 describes another increasingly popular approach, namely the use of "event studies" to analyze growth behavior, based on studying responses to major events such as political reform or changes in trade policy.

24.5.1 Time series approaches

At first glance, the most natural way to understand growth would be to examine time series data for each country in isolation. In practice, however, a time series approach runs into substantial difficulties. One key constraint is the available data. For many developing countries, some of the most important data are only available on an annual basis, with limited coverage before the 1960s. Moreover, the listing of annual data in widely used sources and online databases can be misleading. For example, population figures are often based primarily on census data, while measures of average educational attainment are often constructed by interpolating between census observations using school enrollments. The true extent of information in the time series variation may be less than appears at first glance, and conventional standard errors will be misleading.

Some key growth determinants display relatively little time variation. Even where a variable appears to show significant variation, this may not correspond to the concept the researcher originally had in mind. An example would be political stability. Since Barro (1991), researchers have sometimes used the incidence of political revolutions and coups as a measure of political instability. The interpretation of such an index clearly varies depending on the length of the time period used to construct it. If the hypothesis of interest relates to underlying political uncertainty (say, the *ex ante* probability of a transfer of power) then time series observations on political instability would need to be averaged over a long time period. The variation in political instability at shorter horizons casts light on a different hypothesis, namely the direct impact of revolutions and coups, rather than on the effects of *ex ante* political uncertainty.

There are other significant problems with the time series approach. The hypotheses of most interest to growth theorists are mainly about the evolution of potential output, not deviations from potential output, whether business cycles or larger-scale output collapses. Since measured output is a noisy indicator of potential output, it is easy for the econometric modeling of a growth process to be

contaminated by short-run dynamics. These problems are likely to be even more serious in developing countries, where large slumps or crises are not uncommon, and output may deviate for long periods from any previous structural trend. As Pritchett (2000a) emphasizes, output often behaves very differently in developing countries compared to Organization for Economic Cooperation and Development (OECD) member countries, and a major collapse in output is not a rare event. There may be no underlying trend in the sense commonly understood, and conventional time series methods should be applied with caution.

The problem of short-run output instability extends further. It is easy to construct examples where the difference between observed output and potential output is correlated with variables that move up and down at high frequencies, with inflation being one obvious example (Temple, 2000a). At a minimum, this means that any time series or panel data analysis should distinguish carefully between short-run and long-run effects.

Nevertheless, despite these problems, there are some hypotheses for which time series variation may be informative. Jones (1995) and Kocherlakota and Yi (1997) show how time series models might be used to discriminate between different growth theories. More specifically, they develop a statistical test of endogenous growth models based on regressing growth on lagged growth and a lagged policy variable (or the lagged investment rate, as in Jones). Exogenous growth models predict that the coefficients on the lagged policy variable should sum to zero, indicating no long-run growth effect of permanent changes in this variable. In contrast, some endogenous growth models would imply that the sum of coefficients should be non-zero. A simple time series regression then provides a direct test. More formally, as in Jones (1995), for a given country i one can investigate a dynamic relationship for the growth rate $\gamma_{i,t}$, where:

$$\gamma_{i,t} = A(L)\gamma_{i,t-1} + B(L)z_{i,t} + \varepsilon_{i,t}, \tag{24.30}$$

where z is the policy variable or growth determinant of interest, and $A(L)$ and $B(L)$ are lag polynomials assumed to be compatible with stationarity. The hypothesis of interest is whether $B(1) \neq 0$. If the sum of the coefficients in the lag polynomial $B(L)$ is significantly different from zero, this implies that a permanent change in the variable z will affect the growth rate indefinitely. As Jones (1995) explicitly discusses, this test is best seen as indicating whether a policy change affects growth over a long horizon, rather than firmly identifying or rejecting the presence of a long-run growth effect in the theoretical sense of that term. The theoretical conditions under which policy variables affect the long-run growth rate are strict, and many endogenous growth models are best seen as new theories of potentially sizeable level effects.[17]

A related idea is that of Granger causality, where the hypothesis of interest would be the explanatory power of lags of $Z_{i,t}$ for $\gamma_{i,t}$ conditional on lagged values of $\gamma_{i,t}$. Blomstrom, Lipsey and Zejan (1996) carry out Granger causality tests for investment and growth using panel data with five-year sub-periods. They find strong evidence that lagged growth rates have explanatory power for investment rates,

but weaker evidence for causality in the more conventional direction from investment to growth. In a similar vein, Campos and Nugent (2002) find that, once Granger causality tests are applied, the evidence that political instability affects growth may be weaker than usually believed.

A familiar objection to the more ambitious interpretations of Granger causality is that much economic behavior is forward-looking (see, for example, Bils and Klenow, 2000, on the forward-looking nature of educational investments). The movements of stock markets are another obvious instance where temporal sequences can be misleading about causality. Nevertheless, it could be argued that evidence on timing has been under-utilized in the growth literature to date, especially in panel data studies.

An underlying assumption of most studies is that timing patterns and effects will be similar across units such as countries or regions. Potential heterogeneity has only sometimes been acknowledged, as in the observation of Campos and Nugent (2002) that their results are heavily influenced by the African countries in the sample. The potential importance of these factors is also established in Binder and Brock (2004) who, by using panel methods to allow for heterogeneity in country-specific dynamics, find feedbacks from investment to growth beyond those that appear in Blomstrom *et al.* (1996).

Since testing for Granger causality using panel data requires a dynamic model, the use of a standard fixed effects estimator is likely to be inappropriate when individual effects are present. We discuss this further in section 24.5.2. In the context of investment and growth, a comprehensive examination of the associated econometric issues has been carried out by Bond, Leblebicioglu and Schiantarelli (2004). Their work shows that these issues are more than technicalities: unlike Blomstrom *et al.* (1996), they find strong evidence that investment has a causal effect on growth.

24.5.2 Panel data

As we emphasized above, the prospects for reliable generalizations in empirical growth research are often constrained by the limited number of countries available. This constraint makes parameter estimates imprecise, and limits the extent to which researchers can apply more sophisticated methods, such as semiparametric estimators. A natural response to this constraint is to use the within-country variation to multiply the number of observations. Using different episodes within the same country is ultimately the only practical substitute for somehow increasing the number of countries. To the extent that important variables change over time, this appears the most promising way to sidestep many of the problems that face growth researchers. Moreover, as the years pass and more data become available, the prospects for informative work of this kind can only improve.

We first discuss the implementation and advantages of panel data estimators in more detail, and then some of the technical issues that arise in the context of growth. We will use T to denote the number of time series observations in a panel of N countries or regions. At first sight, T should be relatively high in this context, because of the availability of annual data. But the concerns about

time series analysis raised above continue to apply. Important variables are either measured at infrequent intervals, or show little year-to-year variation. Moreover, variation in annual growth rates may give misleading answers about the longer-term growth process. For this reason, growth research using panel data has typically averaged data over five- or ten-year periods. Given the lack of data before 1960, this implies that growth panels not only have relatively few cross-sectional units, but also low values of T, often just five or six once lagged values have been included as explanatory variables or instruments.[18]

Most of the estimated models have been based on the hypothesis of conditional convergence, namely that countries converge to parallel equilibrium growth paths, the levels of which are a function of a few variables. As we saw in section 24.3, a corollary is that an equation for growth (essentially the first difference of log output) should contain some dynamics in lagged output. In this case, the growth equation can be rewritten as a dynamic panel data model in which current output is regressed on controls and lagged output, as in Islam (1995). In statistical terms this rewritten model is identical in all respects, except that the coefficient on initial output (originally β) is now $1 + \beta$:

$$\log y_{i,t} = (1 + \beta) \log y_{i,t-1} + \psi X_{i,t} + \pi Z_{i,t} + \alpha_i + \mu_t + \varepsilon_{i,t}. \tag{24.31}$$

This regression is a panel analogue to the cross-section regression (24.11), but now includes a country-specific effect α_i and a time-specific effect μ_t. The inclusion of time effects is important in the growth context, not least because the means of the log output series will typically increase over time, given productivity growth at the world level. Inclusion of a country-specific effect allows permanent differences in the level of income between countries that are not captured by $X_{i,t}$ or $Z_{i,t}$. In principle, one can also allow the parameters $1+\beta$, ψ, and π to differ across countries or regions.

Standard random effects estimators require that the individual effects α_i are distributed independently of the explanatory variables, and this requirement is clearly violated for a dynamic panel such as (24.31) by construction, given the dependence of $\log y_{i,t}$ on α_i. Hence the vast majority of panel data growth studies use a fixed effects (within-group) estimator. Given their popularity, it is important to understand how these estimators work. In a fixed effects regression there is a full set of country-specific intercepts, one for each country, and inference proceeds conditional on the particular countries observed, a natural choice in this context. Identification of the slope parameters, usually constrained to be the same across countries, relies on variation over time within each country. The "between" variation, namely the variation across countries in the long-run averages of the variables, is not used.

The key strength of this method, familiar from the microeconometric literature, is the ability to address one form of unobserved heterogeneity: any omitted variables that are constant over time will not bias the estimates, even if these omitted variables are correlated with the explanatory variables. Intuitively, the country-specific intercepts can be seen as picking up the combined effects of all

such variables. This is the usual motivation for using fixed effects in the growth context, as discussed in Islam (1995), Caselli, Esquivel and Lefort (1996) and Temple (1999). In more recent work, fixed effects estimators have been used in studying the effects of distinct events or "treatments," such as democratization or trade reform. We will discuss this approach in section 24.5.3.

A particular motivation for the use of fixed effects arises from the Mankiw *et al.* (1992) implementation of the Solow model. As discussed in section 24.3, their version of the model implies that one determinant of the level of the steady-state growth path is the initial level of efficiency ($A_{i,0}$) and cross-section heterogeneity in this variable should usually be regarded as unobservable. Islam (1995) explicitly develops a specification in which this term is treated as a fixed effect, while world growth and common shocks are incorporated using time-specific effects.

The use of panel data methods to address unobserved heterogeneity can bring substantial gains in robustness, but is not without costs. There are times when the question of interest precludes a fixed-effects approach, and sometimes the limitations of the data will make it uninformative. Some variables of interest are measured at only one point in time. Others are highly persistent, and this dependence implies that the amount of useful information in the within-country variation will be limited. At one extreme, some explanatory variables of interest are essentially fixed factors, like geographic characteristics. Here the only available variation is "between-country," and empirical work will have to be based on cross-sections or pooled cross-section time series. Alternatively, a two-stage hybrid of these methods can be used, in which a panel data estimator is used to obtain estimates of the fixed effects, which are then explicitly modeled in a second stage as in Hoeffler (2002).

A common failing of panel data studies based on within-country variation is that researchers do not pay enough attention to the dynamics of adjustment, and the important distinction between short-run and long-run effects. There are many panel data papers on human capital and growth that test only whether a change in school enrollment or years of schooling has an immediate effect on aggregate productivity, which seems an implausible hypothesis. It would be more natural to consider education as having a lagged effect, especially once various possible externalities are considered. Another example, given by Pritchett (2000a), is the use of panels to study inequality and growth. All too often, changes in the distribution of income are implicitly expected to have an immediate impact on growth. Yet many of the relevant theoretical papers highlight long-run effects, associated with the political process for example, and there is a strong presumption that much of the short-run variation in measures of inequality is due to measurement error. In these circumstances, it is hard to see how the available within-country variation can shed much useful light, at least until better data become available.

There is also a more general problem. Since the fixed effects estimator ignores between-country variation, the reduction in bias typically comes at the expense of higher standard errors. Another reason for imprecision is that either of the devices used to eliminate the country-specific intercepts – the within-groups transformation or first-differencing – will tend to exacerbate the effect of measurement error.[19] As a result, it is common for researchers using panel data models with fixed

effects, especially in the context of small T, to obtain imprecise sets of parameter estimates. Given the potentially unattractive trade-off between robustness and efficiency, Barro (1997), Temple (1999), Pritchett (2000a) and Wacziarg (2002) all argue that the use of fixed effects in empirical growth models has to be approached with care. The price of eliminating the misleading component of the between variation – namely, the variation due to unobserved heterogeneity – is that all the between variation is lost. This is costly, because growth episodes within countries inevitably look a great deal more alike than growth episodes across countries, and therefore offer less identifying variation. Restricting the analysis to the within variation eliminates one source of bias, but makes it harder to identify growth effects with any degree of precision. Many of the explanatory variables currently used in growth research are either highly stable over time, or tending to trend in one direction. Without useful identifying variation in the time series data, the within-country approach is in trouble. Moreover, growth is quite volatile at short horizons. It will typically be hard to explain this variation using predictors that show little variation over time, or that are measured with substantial errors. The result has been a number of panel data studies suggesting that a given variable "does not matter," when a more accurate interpretation is that its effect cannot be identified using the data at hand.

Depending on the sources of heterogeneity, even simple recommendations, such as including a complete set of regional dummies, can help to alleviate the biases associated with omitted variables (Temple, 1998). More than a decade of growth research has identified a host of fixed factors that could be used to substitute for country-specific intercepts. A growth model that includes these variables can still exploit the panel structure of the data, and the explicit modeling of the country-specific effects is directly informative about the sources of persistent income and growth differences.

In practice, the literature has focused on another aspect of using panel data estimators to investigate growth. Nickell (1981) showed that within-groups estimates of a dynamic panel data model can be badly biased for small T, even as N goes to infinity. The direction of this bias is such that, in a growth model, output appears less persistent than it should (the estimate of β is too low) and the rate of conditional convergence will be overestimated. The Nickell bias explains why the within-groups estimator is often avoided when estimating dynamic models. The most widely-used alternative is to difference the model to eliminate the fixed effects, and then use 2SLS or generalized method of moments (GMM) to address the correlation between the differenced lagged dependent variable and the induced MA(1) error term. To see the need for instrumental variable procedures, first-difference (24.31) to obtain:

$$\Delta \log y_{i,t} = (1+\beta)\Delta \log y_{i,t-1} + \Delta X_{i,t}\psi + \Delta Z_{i,t}\pi + \Delta \mu_i + \varepsilon_{i,t} - \varepsilon_{i,t-1}, \quad (24.32)$$

and note that (absent an unlikely error structure) the $\log y_{i,t-1}$ component of $\Delta \log y_{i,t-1}$ will be correlated with the $\varepsilon_{i,t-1}$ component of the new composite error term, as is clearly seen by considering equation (24.31) lagged one period.

Hence, at least one of the explanatory variables in the first-differenced equation will be correlated with the disturbances, and instrumental variable procedures are required.

Arellano and Bond (1991), building on work by Holtz-Eakin, Newey and Rosen (1988), developed the GMM approach to dynamic panels in detail, including specification tests and methods suitable for unbalanced panels. Caselli *et al.* (1996) applied their estimator in the growth context and obtained a much faster rate of conditional convergence than found in cross-section studies, consistent with the view that OLS estimates, by ignoring country effects, will yield an upward bias on the lagged dependent variable.

The GMM approach is typically based on using lagged levels of the series as instruments for lagged first differences. If the error terms in the levels equation (ε_{it}) are serially uncorrelated then $\Delta \log y_{i,t-1}$ can be instrumented using $\log y_{i,t-2}$ and as many earlier lagged levels as are available. This corresponds to a set of moment conditions that can be used to estimate the first-differenced equation by GMM. Bond (2002) and Roodman (2006) provide especially accessible introductions to the theory and application of this approach.

In principle, this strategy can alleviate biases due to measurement error and endogenous explanatory variables. In practice, many researchers are sceptical that lags, or "internally generated" instruments, are appropriate choices for instruments. It is easy to see that a variable such as educational attainment may influence output with a considerable delay, so that the exclusion of lags from the growth equation can look arbitrary. More generally, the GMM approach relies on a lack of serial correlation in the error terms of the growth equation (before differencing). This assumption can be tested using the methods developed in Arellano and Bond (1991), and can also be relaxed by an appropriate choice of instruments, but will sometimes be restrictive.

In practice, many applications of these methods have used "too many" moment conditions. The small-sample performance of the GMM panel data estimators is known to deteriorate as the number of moment conditions grows relative to the cross-section dimension of the panel. In that case, the coefficient estimates can be severely biased, and a further consequence is that the power of Sargan-type tests of overidentifying restrictions may collapse, as shown in Bowsher (2002). When tests of the overidentifying restrictions yield p-values near unity, this is an important warning sign that too many moment conditions are in use, and this problem can be seen relatively frequently in the literature. This can be avoided by using only a sub-set of the available lags as instruments, or summing moment conditions over time, while retaining enough overidentifying restrictions to ensure that Sargan-type tests will have some power. Roodman (2007) discusses these issues in more detail.

Another concern is that the explanatory variables may be highly persistent, as is clearly true of output. Lagged levels can then be weak instruments for first differences, and the GMM panel data estimator is likely to be severely biased in short panels. Bond, Hoeffler and Temple (2001) illustrate this point by comparing the Caselli *et al.* (1996) estimates of the coefficient on lagged output with OLS and

within-group estimates. Since the OLS and within-group estimates of β are biased in opposing directions then, leaving aside sampling variability and small-sample considerations, a consistent parameter estimate should lie between these two extremes, as discussed in Nerlove (1999, 2000). Formally, when the explanatory variables other than lagged output are strictly exogenous, we have:

$$p\lim \hat{\beta}_{WG} < p\lim \hat{\beta} < p\lim \hat{\beta}_{OLS}, \tag{24.33}$$

where $\hat{\beta}$ is a consistent parameter estimate, $\hat{\beta}_{WG}$ is the within-groups estimate and $\hat{\beta}_{OLS}$ is the estimate from a pooled OLS regression. For the dataset and model used by Caselli *et al*. (1996), this large-sample prediction is not valid, which raises a question mark over the reliability of the first-differenced GMM estimates. The problem may be one of weak instruments, and unless this can be resolved, it is not difficult to imagine circumstances in which the within-groups estimator, or bias-corrected versions of it, may be preferable to the GMM approach.

One device that can be informative in short panels is to make more restrictive assumptions about the initial conditions. If the observations at the start of the sample are distributed in a way that is representative of steady-state behavior, in a sense that will be made precise below, efficiency gains are possible. Assumptions about the initial conditions can be used to derive a "system" GMM estimator, of the form developed and studied by Arellano and Bover (1995) and Blundell and Bond (1998), and also discussed in Ahn and Schmidt (1995) and Hahn (1999). In this estimator, not only are lagged levels used as instruments for first differences, but lagged first differences are used as instruments for levels, which corresponds to an extra set of moment conditions. Blundell and Bond (1998) provide Monte Carlo evidence that this estimator is more robust than the Arellano–Bond method in the presence of highly persistent series. As also shown by Blundell and Bond, the necessary assumptions can be seen in terms of an extra restriction, namely that the deviations of the initial values of $\log y_{i,t}$ from their long-run (steady-state) values are not systematically related to the individual effects.[20] For simplicity, we focus on the case where there are no explanatory variables other than lagged output. The required assumption on the initial conditions is that, for all $i = 1, \ldots, N$, we have:

$$E\left[\left(\log y_{i,1} - \bar{y}_i\right)\alpha_i\right] = 0, \tag{24.34}$$

where the $\bar{y}_i$ are the long-run values of the $\log y_{i,t}$ series and are therefore functions of the individual effects α_i and the autoregressive parameter β. This assumption on the initial conditions ensures that:

$$E\left[\Delta \log y_{i,2}\alpha_i\right] = 0, \tag{24.35}$$

and this, together with the mild assumption that the changes in the errors are uncorrelated with the individual effects:

$$E\left[\Delta\varepsilon_{i,t}\alpha_i\right] = 0, \tag{24.36}$$

implies $T-2$ extra moment conditions of the form:

$$E\left[\Delta \log y_{i,t-1}(\alpha_i + \varepsilon_{i,t})\right] = 0 \quad \text{for } i = 1, \ldots, N \text{ and } t = 3,\ 4, \ldots, T. \tag{24.37}$$

Intuitively, as is clear from the new moment conditions, the extra assumptions ensure that the lagged first difference of the dependent variable is a valid instrument for untransformed equations in levels, since it is uncorrelated with the composite error term in the levels equation. The additional moment conditions build in some insurance against weak identification, because if the series are persistent and lagged levels are weak instruments for first differences, it may still be the case that lagged first differences will have some explanatory power for levels.[21] Nevertheless, the analysis of Bun and Windmeijer (2007) indicates that weak instrument problems can emerge even for the system GMM estimator, especially when the variance of the country effects is high relative to the variance of the transitory shocks, which may well be the case for growth data.

Moreover, the extra moment conditions are based on assumptions about the initial conditions that are unlikely to command universal assent. In principle, these assumptions can be tested using the incremental Sargan statistic (or C statistic) associated with the additional moment conditions. Yet the validity of the restriction should arguably be evaluated in wider terms, based on some knowledge of the historical forces giving rise to the observed initial conditions. This point, that key statistical assumptions should not always be evaluated only in statistical terms, is one that we will return to later, when discussing the wider application of instrumental variable (IV) methods.

Alternatives to GMM have been proposed. Kiviet (1995, 1999) derives an analytical approximation to the Nickell bias that can be used to construct a bias-adjusted within-country estimator for dynamic panels. The simulation evidence reported in Judson and Owen (1999) and Bun and Kiviet (2001) suggests that this estimator performs well relative to standard alternatives when N and T are small. More recently, Bun and Carree (2005) have developed an alternative bias-adjusted estimator. One serious limitation of the currently available bias-adjusted estimators, relative to GMM, is that they do not address the possible correlation between the explanatory variables and the disturbances due to simultaneity and measurement error. Nevertheless, there is a clear case for implementing these estimators, at least as a complement to other methods.

A further issue that arises when estimating dynamic panel data models is that of parameter heterogeneity. If a slope parameter such as β varies across countries, and the relevant explanatory variable is serially correlated, this will induce serial correlation in the error term. If we focus on a simple case where a researcher wrongly assumes homogeneity in the coefficient on lagged output, or $\beta_i = \beta$ for all $i = 1, \ldots, N$, then the error process for a given country will contain a component that resembles $(\beta_i - \beta) \log y_{i,t-1}$. Hence there is serial correlation in the errors, given the persistence of output. The estimates of a dynamic panel data model will be inconsistent even if GMM methods are applied. This problem was analyzed in more general terms by Robertson and Symons (1992) and Pesaran and Smith (1995)

and has been explored in depth for the growth context by Lee, Pesaran and Smith (1997, 1998). Since an absence of serial correlation in the disturbances is usually a critical assumption for the GMM approach, parameter heterogeneity can be a serious concern. Some of the possible solutions, such as regressions applied to single time series, or the pooled mean group estimator developed by Pesaran, Shin and Smith (1999), have limitations in studying growth for reasons already discussed. An alternative solution is to split the sample into groups that are more likely to share similar parameter values. Groupings by regional location or the initial level of development are a natural starting point.

Perhaps the state of the art in analyzing growth using panel data and allowing for parameter heterogeneity is represented by Phillips and Sul (2003). They allow for heterogeneity in parameters not only across countries, but also over time. Temporal heterogeneity is rarely investigated in panel studies, but may be important.

One drawback of many current panel studies is that some of the necessary decisions, and perhaps especially the construction of the time series observations, can appear arbitrary. There is no inherent reason why five or ten years represent natural spans over which to average observations. Similarly, there is arbitrariness with respect to which time periods are aggregated. A useful endeavor would be the development of tools to ensure that panel findings are robust under alternative ways of assembling the panel from the raw data; it is also possible that the field could draw more heavily on the econometric literature on time aggregation than it does at present.

More fundamentally, the empirical growth literature has not fully addressed the question of the appropriate time horizons over which growth models should be assessed. For example, it remains unclear when business cycle considerations, or instances of output collapses, may be safely ignored. While cross-section studies that examine growth over 30–40-year periods might be exempt from these considerations, it is less clear that panel studies employing five-year averages are genuinely informative about medium-run growth dynamics. As more data become available with the passage of time, concerns over the scope for arbitrary choices can only increase. It will also be important to develop robust methods for inference about long-run effects.

24.5.3 The event study approach

Although we have focused on the limitations of panel data methods, it is clear that the prospects for informative work of this kind should improve over time. The addition of further time periods is valuable in itself, and the history of developing countries in the 1980s and 1990s offers various events that introduce richer time series variation into the data. These events include waves of democratization, macroeconomic stabilization, financial liberalization and trade reform, and panel data methods can be used to investigate their consequences for growth. This can proceed in a similar way to event studies in the empirical finance literature. In event studies, researchers look for systematic changes in asset returns after a discrete event, such as a profits warning. In other fields, before-and-after studies like this have proved an informative way to gauge the effects of inflation stabilization,

as in Easterly (1996), and the consequences of the debt crisis for investment, as in Warner (1992).

The obvious analogue for growth econometrics is to study the time paths of variables such as output growth, investment and total factor productivity (TFP) growth, before and after discrete events such as democratizations. When using fixed effects combined with a binary indicator to capture discrete events, the logic of the approach is similar to the differences-in-differences estimator in the literature on program evaluation. Examples include the work of Giavazzi and Tabellini (2005) on economic and political liberalizations, Henry (2000, 2003) on stock market liberalization, Papaioannou and Siourounis (2007) and Rodrik and Wacziarg (2005) on democratizations, and Wacziarg and Welch (2003) on trade reforms. Depending on the context, one can also study the response of other variables in a way that is informative about the channels of influence. For example, in the case of trade reform, it is natural to study the response of the trade share, as in the work of Wacziarg and Welch.

The rigor of this method should not be exaggerated. As with any other approach to empirical growth, one has to be cautious about inferring a causal effect. This is clear from drawing explicitly on the literature on treatment effects and program evaluation.[22] In the study of growth, "treatments" such as democratization are clearly not exogenously assigned, but are events that have arisen endogenously. This means, for example, that treatment and control groups may differ systematically, either in terms of time-invariant unobservables, or in factors that vary over time. Methods based on fixed effects can address the first of these considerations, but allowing for the second is more complicated. To illustrate the problem, Papaioannou and Siourounis (2007) draw a useful analogy between democratizations and "Ashenfelter's dip" in the program evaluation literature. It is possible that countries experiencing a downturn or weak economic performance are especially likely to democratize, in which case the estimated effect of democratization risks conflating the true effect with the effects of a separate recovery from the pre-treatment "dip."

Moreover, in growth applications, the treatment effects are highly likely to be heterogeneous across countries and over time. They may depend, for example, on whether a policy change is seen as temporary or permanent, as Pritchett (2000a) observes. In these circumstances, the ability to quantify even an average treatment effect is strongly circumscribed. It may still be possible to identify the direction of effects, and here the limited number of observations does have one advantage. With a small number of cases to examine, it is easy for the researcher to present a graphical analysis that allows readers to gauge the extent of heterogeneity in responses, and the overall pattern. Another useful and informative approach, adopted by Papaioannou and Siourounis (2007), is to estimate a model that allows the treatment effect to vary over time, using the methods developed in Laporte and Windmeijer (2005).

There is one remaining problem to note. When growth researchers look at the effects of discrete events, they typically study the effects on serially correlated outcomes such as output or investment. A particular concern in cross-country samples

is that the errors may then be serially correlated, and the standard errors unreliable. Using simulations, Bertrand, Duflo and Mullainathan (2004) showed that differences-in-differences estimators are potentially highly vulnerable to this problem, to the extent that "placebo" interventions are often found to have effects that are statistically significant. This could well be a major concern for the corresponding growth studies carried out to date. In samples with the cross-section dimensions associated with cross-country data, some of the available solutions for calculating panel-robust standard errors may also face problems.

24.6 Endogeneity and instrumental variables

In this section, we consider the use of instrumental variables in cross-section and time-series contexts. One obvious criticism of growth regressions is that they do little to establish directions of causation. As well as reverse causality, there is the standard problem that two variables may be correlated but jointly determined by a third. Variables such as growth and political stability could be seen as jointly determined equilibrium outcomes associated with, say, a particular set of institutions. The usual response has been the use of instrumental variables, for reasons discussed in section 24.6.1, but there are some grounds for caution in their application. These include the difficulty of establishing credible exclusion restrictions (section 24.6.2) and the problems raised by heterogeneous effects in small samples (section 24.6.3).

24.6.1 Concepts of endogeneity

There are many instances in growth research when explanatory variables are clearly endogenously determined in an economic sense. The most familiar example would be a regression that relates growth to the share of investment in GDP. This may tell us that the investment share and growth are associated, but stops short of identifying a causal effect, or explaining why investment varies; presumably it is endogenous to a range of economic variables. When variables are endogenously determined in the economic sense, there is also a strong chance that they will be endogenous in the statistical or technical sense, namely correlated with the disturbances in the structural equation for growth. To give an example, consider what happens if political instability lowers growth, but slower economic growth feeds back into political instability. The OLS estimator will conflate these two effects and yield an inconsistent estimate of the causal effect of instability.[23]

Views on the importance of these considerations differ greatly. One position is that the whole growth research project effectively capsizes before it has even begun. Mankiw (1995) and Wacziarg (2002) have suggested an alternative and more positive view. According to them, one should accept that reliable causal statements are almost impossible to make, but use the partial correlations of the growth literature to rule out some possible hypotheses about the world. Wacziarg uses the example of the negative partial correlation between corruption and growth found by Mauro (1995). Even if shown to be robust, this correlation does not establish that somehow reducing corruption will be followed by higher growth rates. But it does make

it harder to believe some of the earlier suggestions, rarely based on evidence, that corruption could be actively beneficial.

Given the likelihood that variables are inter-related, one response is to model as many as possible of the variables that are endogenously determined. One prominent example is Tavares and Wacziarg (2001), who estimate structural equations for various channels through which democracy could influence development. This approach has some important advantages in both economic and statistical terms. It can be informative about underlying mechanisms in a way that much empirical growth research is not. From a purely statistical perspective, if the structural equations are estimated jointly by methods such as three-stage least squares (3SLS) or full information maximum likelihood (FIML), this is likely to bring efficiency gains. That said, systems estimation is not necessarily the best route: it has the important disadvantage that specification errors in one of the structural equations could contaminate the estimates obtained from the others. Importantly, these specification errors could include invalid exclusion restrictions, a possibility that is often hard to rule out.

24.6.2 Exclusion restrictions

The most common response to endogeneity has been the application of instrumental variable procedures to a single structural equation, with growth as the dependent variable. Appendices C and D in Durlauf *et al.* (2005) describe a wide range of other instrumental variables that have been proposed for the Solow variables and other growth determinants respectively, where the focus has been on the endogeneity of particular variables. Whether these instruments are genuinely plausible is another matter. In our view, the belief that it is easy to identify valid instrumental variables in the growth context is often mistaken. Many applications of instrumental variable procedures in the empirical growth literature are undermined by a failure to address properly the question of whether these instruments are valid, in the sense of being uncorrelated with the error term in a growth regression. When the instrument is invalid, instrumental variables estimates will be inconsistent. Not enough is currently known about the consequences of "small" departures from validity, but there are circumstances in which the 2SLS bias is worse than the OLS bias, especially if the instruments are "weak." It is certainly possible to envisage circumstances under which ordinary least squares would be preferable to instrumental variables on, say, a mean square error criterion.

A common misunderstanding, perhaps based on confusing the economic and statistical versions of "exogeneity," is that predetermined variables, such as geographical characteristics, are inevitably strong candidates for instruments. There is, however, nothing in the predetermined nature of these variables to preclude a direct effect on growth, or the possibility that they are correlated with omitted growth determinants, and hence with the error term. Even if we take the extreme example of geographic characteristics, there are many channels through which these could affect growth, and therefore many ways in which they could be correlated with the disturbances in a growth model. Brock and Durlauf (2001a) use this type of reasoning to criticize the use of instrumental variables in growth economics.

Since growth theories are often mutually compatible, the validity of an instrument requires a positive and explicit argument that it cannot be a direct growth determinant or correlated with an omitted growth determinant. For many of the instrumental variables that have been proposed, such an argument is difficult to construct.

Discussions of the validity of instruments inevitably suffer from some degree of imprecision because of the need to make qualitative and subjective judgments. When one researcher claims that it is implausible that a given instrument is valid, unless this claim is made on the basis of a joint model of the instruments and the variable of original interest, another researcher can always reject the assertion as unpersuasive. To be clear, this element of subjectivity does not mean that arguments about validity are pointless.[24] Rather, one must recognize that not all statistical questions can be adjudicated on the basis of formal tests.

To see how different instruments might be assigned different levels of plausibility, we consider some examples. Hall and Jones (1999) use various indicators of Western European influence as instruments for openness to trade and institutional quality. As subsequent authors, including Acemoglu (2005), have pointed out, these indicators of Western European influence may have affected long-run development through a variety of channels, in which case their usefulness as instruments appears doubtful at best. The case for exclusion is easier to make for the measure of Western European influence constructed by Acemoglu *et al.* (2001), namely the mortality rates of colonial settlers. Glaeser *et al.* (2004) discuss some of the relevant issues in more detail.

As an example where instrument validity may be relatively plausible, consider Cook (2002). He employs measures of the damage caused by World War II as instruments for various growth regressors, such as saving rates. The validity of Cook's instruments again relies on the orthogonality of World War II damage with omitted post-war growth determinants. It may be that levels of wartime damage had consequences for post-war growth performance in other respects, such as institutional change. Nevertheless, that argument would clearly be more involved, and speculative, than would be necessary for some other examples in the growth literature.

To be clear, this discussion is nowhere near sufficient to conclude that one set of instruments is valid and another is not. Our central point is that exclusion restrictions need to rest on careful and explicit arguments. In particular, it is not enough to appeal to a variable being predetermined. The fact that a variable may be exogenous from an economic point of view is not enough to ensure that it is uncorrelated with the disturbances in the structural equation being estimated. This implies that historical information has a vital role to play in evaluating the plausibility of exclusion restrictions.

This discussion of instrumental variables indicates another important, albeit neglected, issue in empirical growth analysis: the relationship between model specification and instrumental variable selection. One cannot discuss the validity of particular instruments independently from the choice of the specific growth determinants under study and decisions about how to specify their relationship with growth. As we noted earlier, an important research question is whether model

uncertainty and instrumental variable selection can be integrated simultaneously into methods for model averaging and model selection. The recent work of Hendry and Krolzig (2005) on automated methods includes an ambitious approach to systematic model selection for simultaneous equation models, in which identifying restrictions are primarily determined by the data.

24.6.3 Instrumental variables and heterogeneous effects

Another issue that arises in applying instrumental variable methods to cross-country data is the possible heterogeneity in the effects of the instrument, and in the marginal effect of the explanatory variable that is instrumented. This is closely related to the idea of "local" average treatment effects. Stock and Watson (2004, sec. 11.7) provide a useful discussion of the issues in the context of 2SLS. Assume that the parameters in the first stage and second stage of 2SLS vary across countries, and are distributed independently of the variables (and instruments) in the model and the error terms. It can then be shown that the probability limit of the 2SLS estimate of the coefficient on the endogenous explanatory variable is not the average causal effect, but a weighted average of the effects for individual countries. The weighted average gives most weight to the countries for which the instrument has the largest effect on the endogenous explanatory variable. A corollary is that, when heterogeneity is present, the estimated effect depends on the choice of instrument. A further consequence is that claims for the exogeneity of the instruments become harder to sustain.

We can illustrate the potential importance of this by considering some of the most influential papers that apply IV methods to cross-country data, using the "levels regressions" approach discussed in section 24.3.3 above. In particular, Acemoglu *et al.* (2001), Frankel and Romer (1999), and Hall and Jones (1999) all study the determinants of income levels using IV methods. The dependent variable is a measure of (log) GDP per capita or per worker, and the explanatory variables include one or more regressors that may determine income levels, but that are themselves likely to be endogenous to the level of development.

What direction of bias should be expected when estimating such models by OLS? The usual expectation would be that policy variables like institutional quality "improve" (that is, move in the direction of promoting development) as GDP per capita increases. Under this view of the world, an OLS estimate of the effect of variables like institutions is likely to overstate their importance. The OLS slope coefficient will be biased away from zero, and correcting for this using IV should lead to a parameter estimate closer to zero, sampling variability aside. But the papers of Acemoglu *et al.* (2001), Frankel and Romer (1999) and Hall and Jones (1999) all have in common the opposite result. Somewhat surprisingly, the IV estimate associated with the variable of interest, such as the quality of institutions, is typically larger in magnitude than the OLS estimate.

There are a number of possible explanations. One is sampling variability, while another is measurement error. But if we take the view that development is likely to encourage improvements in (say) institutions, so that OLS estimates of their effect on output per worker are biased away from zero, the required extent of

measurement error is unlikely to be trivial. It has to be large enough to more than offset the effect of the simultaneity bias acting in the other direction.

A more sceptical view is that the IV coefficient may be larger because either (i) the effects may be heterogeneous; or (ii) the exclusion restrictions used in these papers are not genuinely valid; or some combination of the two. One interesting perspective on the IV results in the cross-country literature is that the instruments might have a more powerful effect on the endogenous explanatory variable (such as institutions) in those countries where the explanatory variable has an especially powerful effect on development outcomes. To give a concrete example, this would be the case if the Acemoglu *et al.* (2001) instrument, namely mortality rates of colonial settlers, had greatest influence on institutional development in sub-Saharan Africa, and if the marginal effect of institutions on development outcomes is most powerful in Africa. Were this true, the estimated effect of institutions will be larger than the average causal effect, because the IV estimate gives more weight to the causal effect for countries where the instrument has greatest influence. This does not invalidate the finding that institutions matter, but does make it harder to generalize about the extent to which they matter.

This is admittedly a speculative hypothesis, but there is some possible supporting evidence in Table 4 of Acemoglu *et al.* (2001). The 2SLS coefficient on institutions is roughly halved when African countries are excluded from the sample, although still precisely estimated. The sensitivity of the 2SLS estimate suggests that the effects of institutions differ substantially across countries. A deeper and more rigorous investigation would be possible when there is at least one additional instrument, so that estimates can be compared across different instrument sets. If the effects of policy are homogeneous across countries, the estimated effect should be the same regardless of the choice of instrument (abstracting from sampling variability). That invariance no longer holds under heterogeneous effects. Then, as we have seen, it is likely that the IV estimate will be influenced by the choice of instrument, and the estimated effect will no longer relate to the whole population.

24.7 Other econometric issues

In this section we consider a range of questions that arise in growth econometrics from the properties of data and errors. Starting with data issues, section 24.7.1 examines how one may handle outliers in growth data. Section 24.7.2 examines the problem of measurement error. This is an important issue since there are good reasons to believe that the quality of the data is sometimes poor for less developed economies. In section 24.7.3 we consider the case where data are missing. Turning to issues of the properties of model errors, section 24.7.4 examines the analysis of heteroskedasticity in growth contexts. Finally, section 24.7.5 addresses the problem of cross-section error dependence.

24.7.1 Outliers

Empirical growth researchers often work with small datasets and estimate relatively simple models. In these circumstances, OLS regressions are almost meaningless

unless they have been accompanied by systematic investigation of the data, including the sensitivity of the results to outlying observations. There are many reasons why some observations may be unrepresentative. It is possible for variables to be measured with error for that particular region or country. Alternatively, the model specified by the researcher may omit a relevant consideration, and so a group of observations will act as outliers. It is inherent in least squares estimators that they are highly sensitive to unrepresentative observations, and the dangers of using OLS were forcibly expressed by Swartz and Welsch (1986, p. 171): "In a world of fat-tailed or asymmetric error distributions, data errors, and imperfectly specified models, it is just those data in which we have the least faith that often exert the most influence on the OLS estimates."

Some researchers respond to this concern by using leverage measures or single-case diagnostics such as Cook's distance statistic. There are well-known problems with these approaches, because where more than one outlier is present, the extent of the influence of one observation can easily be hidden by the presence of another (the "masking" effect). By far the best response is to use a more robust estimator, such as least trimmed squares, at least as a preliminary way of investigating the data.[25] These issues are discussed in more detail in Temple (1998, 2000b).

24.7.2 Measurement error

We now turn to a more general discussion of measurement error. It is clear that measurement errors are likely to be pervasive, especially in data that relate to developing countries, yet relatively few empirical studies of growth consider the impact of measurement error in any detail. This rather casual approach often appeals to the best-known statistical result, which applies to a bivariate model where the independent variable is measured with error.[26] The estimate of the slope coefficient will be biased towards zero, even in large samples, because measurement error induces a covariance between the observable form of the regressor and the error term. This attenuation bias is well known, but sometimes misleads researchers into suggesting that measurement error will only mask effects, a claim that is not true in general. When there are multiple explanatory variables, but only one is measured with error, then typically all the parameter estimates will be biased. Some parameter estimates may be biased away from zero and, although the direction of the bias can be estimated consistently, this is rarely done. When several variables are measured with error, the assumption that measurement error only hides effects is even less defensible.

Where measurement error is present, the coefficients are typically not identified unless other information is used. The most popular solution is to use instrumental variables, if a separate instrument can be found which is likely to be independent of the measurement error. A more complex solution, which does not need an additional variable, is to exploit higher-order sample moments to construct IV estimators, as in Dagenais and Dagenais (1997) and Arcand and Dagenais (2005). The reliability of these procedures in small samples is uncertain, since the use of higher-order moments could make them sensitive to outliers.

Sometimes partial identification is possible, in the sense that bounds on the extent of measurement error can be used to derive consistent estimates of bounds on the slope parameters. Although it can be difficult for researchers to agree on sensible bounds on the measurement error variances, there are easier ways of formulating the necessary restrictions, as discussed by Klepper and Leamer (1984). Their reverse regression approach was implemented by Persson and Tabellini (1994) and Temple (1998), but has rarely been used by other researchers. Another strategy is to investigate sensitivity to varying degrees of measurement error, based on method of moments corrections. Again, this is easy to implement in linear models, and should be applied more routinely than it is at present. Temple (1998) provides a discussion of both approaches in the context of the Mankiw *et al*. (1992) model.

24.7.3 Missing data

Some countries rarely appear in growth datasets, partly by design: it is common to leave out countries with small populations, oil producers, and transition economies. These are countries that seem especially unlikely to lie on a regression surface common to the majority of the OECD member countries or the developing world. Other countries are left out for different reasons. When a nation experiences political chaos, or lacks economic resources, the collection of national accounts statistics will be a low priority. In other cases, countries appear in some studies but not in others, depending on the availability of particular variables of interest.

Missing data can be a serious problem. If a researcher started from a representative dataset and then deleted countries at random, this would typically increase the standard errors but not lead to biased estimates. More serious difficulties arise if countries are missing in a non-random or systematic way, because then parameter estimates are likely to be biased. This problem is given relatively little attention in mainstream econometrics textbooks, despite a large body of research in the statistics literature. A variety of solutions are possible, with the simplest being one form or another of imputation, with an appropriate adjustment to the standard errors. Hall and Jones (1999) and Hoover and Perez (2004) are among the few empirical growth studies to implement this. It may be especially useful when countries are missing from a dataset because a few variables are not observed for their particular cases. It is then easy to justify using other available information to predict the missing data, and thereby exploit the additional information contained in the variables that are observed. Alternative approaches to missing data are also available, based on likelihood or Bayesian methods, which can be extended to handle missing observations.

24.7.4 Heteroskedasticity

It is common in cross-section regressions for the underlying disturbances to have a non-constant variance. As is well known, the coefficient estimates remain unbiased, but OLS is inefficient and the estimates of the standard errors are biased. Most empirical growth research simply uses the heteroskedasticity-consistent standard errors developed by Eicker (1967) and White (1980). These estimates of the standard

errors are consistent but not unbiased, which suggests that alternative solutions to the problem may be desirable. For datasets of the size found in cross-country empirical work, the alternative estimators developed by MacKinnon and White (1985) are likely to have better finite sample properties, as discussed in Davidson and MacKinnon (1993) and supported by simulations in Long and Ervin (2000).

There are at least two other concerns with the routine application of White's heteroskedasticity correction as the only response to heteroskedasticity. The first is that by exploiting any structure in the variance of the disturbances, using weighted least squares, it may be possible to obtain efficiency gains. The second and more fundamental objection is that heteroskedasticity can often arise from serious model misspecification, such as omitted variables or neglected parameter heterogeneity. Evidence of heteroskedasticity should then prompt revisions of the model for the conditional mean, rather than mechanical adjustments to the standard errors. See Zietz (2001) for further discussion and references.

24.7.5 Cross-section error dependence

An unresolved issue in growth econometrics is the treatment of cross-section dependence in model errors. This dependence may have important consequences for inference. As noted by DeLong and Summers (1991) in the growth context, failure to account for cross-sectional error correlation can lead to inaccurate standard errors. Furthermore, there are several reasons to expect cross-sectional error dependence to be present when studying growth. For example, countries that are geographically close together, or trading partners, may well experience common shocks. Output growth may often be related to the growth of large, leading countries within a particular region or world.

The general issue of error dependence has been a focus of recent research, in the context of panel data and panel time series estimators in particular. Whether the effect of dependence is sizeable in practice remains an open question, but one that might be addressed using ideas developed in Baltagi, Song and Koh (2003) and Driscoll and Kraay (1998), among others.

In the context of growth regressions, work on cross-section dependence may be divided into two strands. One concerns tests to identify the presence of cross-section dependence. Pesaran (2004) develops tests that do not rely on any prior ordering; this framework in essence sums the cross-section sample error correlations in a panel and evaluates whether they are consistent with the null hypothesis that the population correlations are zero. Specifically, and recalling that N denotes the cross-section dimension and T the time dimension, he proposes a cross-section dependence statistic CD:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{i,j} \right), \tag{24.38}$$

where $\hat{\rho}_{i,j}$ is the sample correlation between $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$. Pesaran gives conditions under which this statistic converges to a Normal (0,1) random variable (as N and T become infinite) under the null hypothesis of no cross-section correlation. This

test statistic is based on earlier work by Breusch and Pagan (1980) and appears to possess good finite sample properties in comparison to this earlier work. Using a country-level panel, Pesaran (2004) strongly rejects the null of no cross-section dependence for the world as a whole, and within several geographic groupings.

The CD test need not be consistent for some alternatives of interest, however. With this in mind, Pesaran, Ullah and Yamagata (2008) develop a bias-adjusted version of the Breusch and Pagan Lagrange multiplier (LM) test statistic for cross-section error independence, for panels with strictly exogenous regressors and normally distributed errors. This approach retains some power in some circumstances where the CD test does not, but is less robust to departures from normality and the presence of regressors that are only weakly exogenous.

The second strand of research on cross-section error dependence has constructed empirical models that take it explicitly into account. One approach relies on formulating a statistical model of the dependence. Phillips and Sul (2003) model the error term in a growth panel as:

$$\varepsilon_{i,t} = \delta_i \theta_t + u_{i,t}, \tag{24.39}$$

where θ_t and $u_{i,t}$ are independent random variables and $u_{i,t}$ is assumed to be i.i.d. across countries and across time. Pesaran (2006) develops an alternative estimation strategy based on a generalized form of this error structure, one in which θ_t may be a vector. While Phillips and Sul consider how to account for error dependence in a generalized least squares (GLS)-type structure, Pesaran considers ways to filter the individual observations in order to eliminate the dependence. From the perspective of growth dynamics, (24.39) suffers from the problem that it does not account for aspects of the error process associated with growth. In order to account for cross-section dependence in convergence analysis, Phillips and Sul (2007a, 2007b) study the case where δ_i is replaced with $\delta_{i,t}$ in (24.39), arguing that transition dynamics produce time varying coefficients of this type as less advanced economies catch up to more advanced ones.

Another possibility when analyzing error dependence is to treat the problem as one of spatial correlation. This issue has been much studied in the regional science literature, and statisticians in this field have developed spatial analogues of many time series concepts (see Anselin, 2001, 2006, for an overview). Spatial methods may yet have an important role to play in growth econometrics. However, when these methods are adapted from the spatial statistics literature, they raise the problem of identifying the appropriate notion of space. One can imagine many reasons for cross-section correlation. If one is interested in technological spillovers, it may well be the case that in the space of technological proximity, the United Kingdom is closer to the United States than is Mexico. Put differently, unlike the time series and spatial cases, there is no natural cross-section ordering to elements in the standard growth datasets. Following language due to Akerlof (1997), countries are perhaps best thought of as occupying some general socioeconomic-political space defined by a range of factors; spatial methods then require a means to identify their locations.

One approach is pursued by Conley and Ligon (2002). In their analysis, they attempt to construct estimates of the spatial covariation of the residuals ε_i in a cross-section. In order to do this, they construct different measures of socioeconomic distance between countries. They separately consider geographic distance (measured between capital cities) and measures of the costs of transportation between these cities. Once a distance metric is constructed, these are used to construct a residual covariance matrix. Estimation methods for this procedure are developed in Conley (1999). Conley and Ligon (2002) find that allowing for cross-section dependence in this way is relatively unimportant in terms of appropriate calculation of standard errors for growth model parameters. Their methods could be extended to allow for comparisons of different variables as the source for cross-section correlation, as in Conley and Topa (2002) in the context of residential neighborhoods. A valuable generalization of this work would be the modeling of cross-section dependence as a function of multiple variables. Such an analysis would allow further progress on the measurement of distances in socioeconomic space, which may arise through multiple channels.

An alternative approach is to build spillover effects directly into the structure of an empirical model. Easterly and Levine (1997) is an example of a study which incorporates a direct effect of the growth of neighboring countries, but such examples remain rare. Some of the relevant issues have been highlighted in the theoretical literature on social interactions, inspired by empirical problems such as the measurement of peer effects in schools. While a structural approach has advantages, the presence of spillovers has consequences for identification that are not easily resolved, for the reasons explained in Manski (1993) and subsequently discussed in Brock and Durlauf (2001b). These consequences have yet to be fully explored in the context of empirical growth studies. Binder and Pesaran (2001) and Brock and Durlauf (2001b) analyze identification and estimation problems for intertemporal environments that are particularly relevant to the growth context.

24.8 Conclusions: the future of growth econometrics

In this section, we offer some closing thoughts on the most promising directions for empirical growth research. We explicitly draw on previous contributions along these lines, many of which deserve wider currency. It is especially interesting to compare the current state of the field against the verdicts offered in the early survey by Levine and Renelt (1991).

A dominant theme will be that the empirical study of growth requires an eclectic approach, and that the field has been harmed by a tendency for research areas to evolve independently, without enough interaction.[27] This is not simply a question of using a variety of statistical techniques. It also suggests the need for a closer connection between theory and evidence, a willingness to draw on ideas from areas such as trade theory, and more attention to particular features of the countries under study, including the historical context.

Pritchett (2000a) has listed three questions for growth researchers to address:

- What are the conditions that initiate an acceleration of growth or the conditions that set off sustained decline?
- What happens to growth when policies – trade, macroeconomic, investment – or politics change dramatically in episodes of reform?
- Why have some countries absorbed and overcome shocks with little impact on growth, while others seem to have been overwhelmed by adverse shocks?

Although this research agenda is almost ten years old, it retains considerable relevance, not least because it focuses attention on substantive economic issues rather than technicalities. The importance of the first of Pritchett's questions is evident from the many instances where countries have moved from stagnation to growth and vice versa. Hausmann, Pritchett and Rodrik (2005) explicitly model transitions to fast growth ("accelerations") and make clear the scope for informative work of this kind. The second question we have discussed in section 24.5.3, and research in this vein is increasingly prominent. Here, one of the major challenges will be to relax the (sometimes only implicit) assumption that policies are randomly assigned, and to find ways of carrying out inference that are robust in small samples. The third question has been addressed in an important paper by Rodrik (1999).

In all three cases, it is clear that econometric work should be informed by detailed studies of individual countries, such as those collected in Rodrik (2003). Too much empirical growth research proceeds without enough attention to the historical and institutional context. For example, a newcomer to this literature might be surprised at the paucity of work that integrates growth regression findings with, say, the known consequences of the 1980s debt crisis. Another reason for advocating case studies is that much of the empirical growth literature essentially isolates only reduced-form partial correlations. These can be useful, but it is clear that we often need to move beyond this. A partial correlation is more persuasive if it can be supported by theoretical arguments. The two combined are more persuasive if there is evidence of the intermediating effects or mechanisms that are emphasized in the relevant theory. There is plenty of scope for informative work that tries to isolate the mechanisms by which variables such as financial depth, inequality, and political institutions shape the growth process. Wacziarg (2002), in particular, highlights the need for a "structural" growth econometrics, one that aims to recover channels of causation.

A more extreme view is that growth econometrics should be supplanted by the calibration of theoretical models. Klenow and Rodriguez-Clare (1997) emphasize the potential of such an approach. The analysis of Mankiw *et al.* (1992) can be seen partly as a comparison of estimated parameter values with those associated with specific theoretical models, but relatively little of the empirical work that has followed has achieved a similarly close connection between theory and evidence. This has been a recurring criticism of the literature since at least Levine and

Renelt (1991). It would be premature to say that econometric approaches should be entirely replaced by calibration or "quantitative theory," but the two methods could inform each other more often than at present. Calibrated models can help to interpret parameter estimates, not least in comparing the magnitude of the estimates with the implications of plausible models. At the same time, the partial correlations identified in growth econometrics can help to act as a discipline on model-building and can also indicate where model-based quantitative investigations are most likely to be fruitful. This role for growth econometrics is likely to be especially useful in areas where the microeconomic evidence used to calibrate structural models is relatively weak, or standard behavioral assumptions may be flawed.

The need for a tighter connection between theory and evidence is especially apparent in certain areas. Most empirical growth papers are based on the one-sector, closed-economy Solow model, which leaves out aspects of interdependence that are surely important. Howitt (2000) has shown that standard empirical growth models can be reinterpreted in the light of a multi-country theoretical model with a role for technology diffusion. More generally, there is a need for researchers to develop frameworks that are consistent with international flows of goods, capital and knowledge. These issues are partly addressed by the theoretical analyses of Barro, Mankiw and Sala-i-Martin (1995) and Howitt (2000), and empirical work that builds on such ideas deserves greater prominence. Here especially, research that draws on the quantitative implications of specific theoretical models, as in the work of Eaton and Kortum (1999, 2001) on technology diffusion and the role of imported capital goods, could be an important advance.

The neglect of open economy models is just one example of the narrowness with which empirical growth models are often conceived. Much of the empirical literature uses a theoretical framework that was originally developed to explain the long-run growth experiences of the US and other developed nations. This framework is routinely applied to study developing countries, while incorporating few of their distinctive features. A list of these could include the potentially important roles of agricultural employment, dualism, and structural change; a relatively large informal sector, which often accounts for a substantial share of total employment; and periods of extensive state involvement in production, sometimes reflecting the legacy of nationalist movements and many years of socialist economic policies.

The narrowness of focus in existing studies has many limitations. For example, the conventional use of one-sector models depends on unrealistic assumptions about aggregation. It also prevents many relevant and interesting questions from even being asked, including the role of changes in sectoral and occupational structure in productivity growth. Given the scope for more general and more informative models, empirical growth researchers have really only scratched the surface, with a few recent exceptions. Temple (2005, 2006) and Temple and Wößmann (2006) discuss some of the relevant issues, and provide further references.

Some of these issues are closely linked to an especially important research agenda, namely the need to distinguish between different types of growth and

their distributional consequences. For example, the general equilibrium effects of productivity improvements in agriculture may be very different to those in services and industry. Identifying the nature of "shared growth" will require more detailed attention to particular features of developing countries. It will also require an effort to understand how these influence the forms taken by growth, and their distributional consequences. Given that the main source of income for the poor is usually labor income, there is a need to integrate theoretical and empirical growth models with theory and evidence from labor economics, in order to study how growth and labor markets interact. Agénor (2004) and Temple (2005) consider some of the relevant issues. Partly because of the weaknesses of data on income distribution, too little work in growth econometrics has differentiated between types of growth, even though the policy relevance of these questions is clear.

Ideally, research along these various lines will utilize not only statistical methods, but also the power of case studies in generating hypotheses, and in deepening our understanding of the economic, social and political forces at work in determining growth outcomes. Case studies may be especially valuable in at least two areas. The first of these is the study of technology transfer. As emphasized in the survey by Klenow and Rodriguez-Clare (1997), we do not know enough about why some countries are more successful than others in climbing the "ladder" of product quality and technological complexity. What are the relative contributions of human capital, foreign direct investment and trade? In recent years some of these issues have been intensively studied at the microeconomic level, especially the role of foreign direct investment and trade, but there remains work to be done in relating firm and sector-level evidence to aggregate implications.

A second area where case studies may be especially valuable is the study of political economy, in its modern sense. It is a truism that economists, particularly those considering development, have become aware of the need to account for the two-way interaction between economics and politics. A case can be made that the theoretical literature has outpaced the empirical literature in this regard. Analytical case studies of individual countries, drawing on both economic theory and political science, would help to close this gap.

Thus far, we have highlighted a number of limitations of existing work, and directions in which further research seems especially valuable. Some of the issues we have considered were highlighted much earlier by Levine and Renelt (1991). The extent to which limitations have stubbornly persisted over time might lead to pessimism over the long-term prospects of this literature.[28] This also shows that our prescriptions for future research could seem rather pious, since the improvements we recommend are easier said than done. But the literature has also evolved in some interesting and unpredictable ways, and we will end our review by considering some areas in which genuine progress has been made, and where further progress appears likely.

One reason for optimism is the potential of recently developed model averaging methods. These help to address the model selection and robustness issues that have been identified as a major weakness of cross-country growth research since at least Levine and Renelt (1992). By framing the problem explicitly in terms of model

uncertainty, the Bayesian approach can consider many candidate explanatory variables simultaneously, and the extent of their robustness to changes in specification. Researchers can then communicate the degree of support for a particular hypothesis with more faith that the results do not depend on arbitrary modeling choices. Perhaps the main open questions about these methods are those identified by Ciccone and Jarocinski (2007). From their analysis, it is clear that current approaches to Bayesian model averaging may have significant limitations of their own. The development of methods to overcome these should be a research priority.

Another reason for optimism is that the quality of available data is likely to improve over time. The development of new and better data has clearly been one of the main achievements of the empirical growth literature since the early 1990s, and one that was not foreseen by critics of the field. Researchers have developed increasingly sophisticated proxies for drivers of growth that previously appeared resistant to statistical analysis. One approach, pioneered in the growth literature by Knack and Keefer (1995) and Mauro (1995), has been the use of country-specific ratings compiled by international agencies. Such data increasingly form the basis for measures of corruption, government efficiency, and protection of property rights. More recent work, such as that of Kaufmann, Kraay and Zoido-Lobaton (1999a, 1999b) and Kaufmann, Kraay and Mastruzzi (2003), has established unusually comprehensive measures of various aspects of institutional quality. Similarly, research in political science, notably the POLITY project at the University of Maryland, has developed a range of indicators of political institutions that have already played an important role in empirical growth studies.

As more variables become available, the construction of proxies is likely to make increasing use of latent variable methods. These aim to reduce a set of observed variables to a smaller number of indicators that are seen as driving the majority of the variation in the original data, and that could represent some underlying variable of interest. For example, the extent of democracy is not directly observed, but is often obtained by applying factor analysis or extracting principal components from various dimensions of political freedom. There are obvious dangers with this approach, but the results can be effective proxies for concepts that are otherwise hard to measure.[29] Using latent variables makes especially good sense under one view of the proper aims of growth research. It is possible to argue that empirical growth studies will never give good answers to precise hypotheses, but can be informative at a broader level. For example, a growth regression is unlikely to tell us whether the growth effect of inflation is more important than the effect of inflation uncertainty, because these two variables are usually highly correlated. It may even be difficult to distinguish the effects of inflation from the effects of sizeable budget deficits.[30] Instead, a growth regression might be used to address a less precise hypothesis, such as the growth dividend of macroeconomic stability, broadly conceived. In this context, it is natural to use latent variable approaches to measure the broader concept.

Another valuable development is likely to be the creation of rich panel datasets at the level of regions within countries. Regional data offer greater scope for

controlling for some variables that are hard to measure at the country level, such as cultural factors. By comparing experiences across regions, there may also be scope for identifying events that correspond more closely to natural experiments than those found in cross-country data. Work such as that by Besley and Burgess (2000, 2002, 2004), using panel data on Indian states, shows the potential of such an approach. In working with such data more closely, one of the main challenges will be to develop empirical frameworks that incorporate movements of capital and labor between regions: clearly, regions within countries should only rarely be treated as closed economies. Shioji (2001) is an example of how analysis using regional data can take this into account.

Even with better data, at finer levels of disaggregation, the problem of omitted variables can only be alleviated, not resolved. It is possible to argue that the problem applies equally to historical research and case studies, but at least in these instances, the researcher may have some grasp of important forces that are difficult to quantify. Since growth researchers naturally gravitate towards determinants of growth that can be analyzed statistically, there is an ever-present danger that the empirical literature, even taken as a whole, yields a rather partial and unbalanced picture of the forces that truly matter. Even a growth model with high explanatory power, in a statistical sense, has to be seen as a rather provisional set of ideas about the forces that drive growth and development.

This brings us to our final points. We once again emphasize that empirical progress on the major growth questions requires attention to qualitative sources such as historical narratives and studies by country experts. One example we have given concerns the validity of instrumental variables: understanding the historical experiences of various countries seems critical for determining whether exclusion restrictions are plausible. In this regard work such as that of Acemoglu *et al.* (2001, 2002) is exemplary. More generally, nothing in the empirical growth literature suggests that issues of long-term development can be disassociated from the historical and cultural factors that fascinated commentators such as Max Weber, and the examination of these factors must rely at least partly on case studies, or risk missing some of the most interesting and important issues.

These questions have been asked for many decades, and the quest to understand the wealth of nations is as old as the discipline of economics itself. In contrast, growth econometrics is an area of research that is still in its infancy. Researchers in this field have shown flexibility in responding to the specific challenges and questions that arise in this context. They have introduced a number of statistical methods into applied economics, including classification and regression tree algorithms, robust estimation, threshold models and Bayesian model averaging, all of which appear to have wider utility. As with any new literature, especially one tackling questions as complex as these, it is easy to identify significant limitations of the existing evidence, and of the tools that are currently applied. Yet it seems clear to us that significant progress has been and continues to be made, even from the vantage point of our (2005) review. We therefore see good reasons for continued optimism.

Acknowledgments

This chapter updates and extends our earlier survey, Durlauf *et al.* (2005). Durlauf thanks the University of Wisconsin and National Science Foundation for financial support. Johnson thanks the Department of Economics, University of Wisconsin for its hospitality in Fall 2003, during which some of the preparation for this chapter was completed. Temple thanks the Leverhulme Trust for financial support under the Philip Leverhulme Prize Fellowship scheme. Finally, we thank Stephen Bond and William Brock for useful discussions.

Notes

1. See Temple (2000b) and Brock and Durlauf (2001a) for a conceptual discussion of this issue.
2. This independence assumption is sometimes defended either on theoretical grounds or because the parameter estimates are consistent with those predicted by the augmented Solow model.
3. Given the role it plays in the analysis of convergence, the initial income variable $\log y_{i,0}$ is usually distinguished from the other Solow variables.
4. Other early studies include Baumol (1986), DeLong (1988), Grier and Tullock (1989), Kormendi and Meguire (1985) and Marris (1982).
5. Note that while some failures of exchangeability call into question the interpretation of the regression, this is not always the case. For example, a heteroskedastistic error, while violating exchangeability, does not undermine the interpretation of the point estimates of the parameters. See Draper *et al.* (1993) for further discussion of the role of exchangeability in empirical work.
6. See the discussion in Brock *et al.* (2003) of the Ellsberg paradox.
7. For further discussion of extreme bounds analysis, see Temple (2000b) and the references therein.
8. In this discussion, we will assume that one of the models in the model space M is the correct specification of the growth process. When none of the model specifications is the correct one, this naturally affects the interpretation of the model averaging procedure.
9. Fernandez, Ley and Steel (2001b) provide a general analysis of proper model specific priors for model averaging exercises.
10. Sachs and Warner (1995) use this variable as an index of overall openness of an economy.
11. The posterior inclusion probabilities of single variables ignores their interdependence and so may be criticized for reasons similar to those we have raised with respect to model space priors that assume, for inclusion in a given model, conditional independence across variables. Doppelhofer and Weeks (2007) propose ways to measure the jointness of variable inclusion. Letting k and l denote the events "variable r_k appears in the true model" and "variable r_l appears in the true model," and using $\bar{k}$ and $\bar{l}$ to denote the complements of these events, the authors propose the jointness statistic

$$J_{k,l} = \log\left(\frac{\mu\left(k\,|l,D,M\right)}{\mu\left(\bar{k}\,|l,D,M\right)}\cdot\frac{\mu\left(\bar{k}\,\middle|\,\bar{l},D,M\right)}{\mu\left(k\,\middle|\,\bar{l},D,M\right)}\right)$$

to measure the degree of dependence between two candidate variables. The authors find that positive dependence is a common feature of the candidate growth determinants studied by Sala-i-Martin *et al.* (2004).

12. Here we are being imprecise in referring to within-model priors for Sala-i-Martin *et al*. (2004), but note that, following Ley and Steel (2008), their analysis is interpretable in this way.
13. A detailed discussion of regression tree methods appears in Breiman *et al*. (1984). The technical appendix of Durlauf and Johnson (1995) presents a treatment tailored to the specific question of identifying multiple regimes in growth models.
14. Motivated by the debate over trade openness and growth, Papageorgiou (2002) applies Hansen's (2000) methods to the Durlauf and Johnson (1995) data, with the trade share added to the set of variables on which sample splits may occur.
15. Projection pursuit methods are developed in Friedman and Tukey (1974) and Friedman (1987). Appendix A of Desdoigts (1999) provides a useful primer.
16. The assumed rarity of large shocks implies that movements between basins of attraction of each of the steady-states are sufficiently infrequent that they can be ignored in estimation. This assumption is consistent with, for example, the Bianchi (1997) and Paap and van Dijk (1998) findings that there is relatively little mobility within the cross-country income distribution.
17. See Temple (2003) for more discussion of this point and the long-run implications of different growth models.
18. This is true of the many published studies that have used version 5.6 of the Penn World Tables. Now that more recent data are available, there is more scope for estimating panels with a longer time dimension.
19. See Arellano (2003, pp. 47–51) for a more formal treatment of this issue.
20. Note that the long-run values of log output are evolving over time when time-specific effects are included in the model.
21. An alternative approach would be to use small-sample bias adjustments for GMM panel data estimators, such as those described in Hahn, Hausman and Kuersteiner (2001).
22. This connection with the treatment effect literature is sometimes explicitly made, as in Giavazzi and Tabellini (2005), Papaioannou and Siourounis (2007) and Persson and Tabellini (2003). The connection helps to understand the limitations of the evidence, but the scope for resolving the associated identification problems may be limited in cross-country datasets.
23. Although this "reverse causality" interpretation of endogeneity is popular and important, it should be remembered that a correlation between an explanatory variable and the error term can arise for other reasons, including omitted variables and measurement error. As we discuss, it is important to bear a general interpretation of the error term in mind when judging the plausibility of exclusion restrictions in instrumental variable procedures.
24. Put differently, one does not require a precise definition of what makes an instrument valid in order to debate whether a given instrument is valid or not. To take an example due to Taylor (1998), the absence of a precise definition of money does not weaken my belief that the currency in my wallet is a form of money, whereas the computer on which this paper is written is not. To claim such arguments cannot be made is known as the Socratic fallacy.
25. This estimator should not be confused with trimmed least squares and other methods based on deleting observations with large residuals in the OLS estimates. A residual-based approach is inadequate for obvious reasons.
26. This and the following discussion assume classical measurement error. Under more general assumptions, it is usually even harder to identify the consequences of measurement error for parameter estimates.
27. To give a specific example, the macroeconomic literature on international technology differences only rarely acknowledges relevant work by trade economists, including estimates of the Heckscher–Ohlin–Vanek model that suggest an important role for technology differences. See Acemoglu (2008) and Klenow and Rodriguez-Clare (1997) for more discussion.

28. Only now are researchers beginning to engage with some of the issues they raised, such as the varying conditions under which it is appropriate to use international rather than national prices in making productivity comparisons and constructing capital stocks.
29. A relevant question, not often asked, is how high the correlation between the proxy and the true predictor has to be for the estimated regression coefficient on the proxy to be of the "true" sign. Krasker and Pratt (1986, 1987) have developed methods that can be used to establish this under surprisingly general assumptions.
30. As Sala-i-Martin (1991) has argued, various specific indicators of macroeconomic instability should perhaps be seen as symptoms of some deeper, underlying characteristic of a country.

References

Acemoglu, D. (2005) Constitutions, politics and economics: a review essay on Persson and Tabellini's "The economic effects of constitutions." *Journal of Economic Literature* **43**, 1025–48.

Acemoglu, D. (2008) Introduction to modern economic growth. Manuscript, MIT, February.

Acemoglu, D., S. Johnson and J. Robinson (2001) The colonial origins of comparative development: an empirical investigation. *American Economic Review* **91**(5), 1369–401.

Acemoglu, D., S. Johnson and J. Robinson (2002), Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* **117**(4), 1231–94.

Agénor, P.-R. (2004) Macroeconomic adjustment and the poor: analytical issues and cross-country evidence. *Journal of Economic Surveys* **18**(3), 351–408.

Ahn, S. and P. Schmidt (1995) Efficient estimation of models for dynamic panel data. *Journal of Econometrics* **68**, 5–27.

Akerlof, G. (1997) Social distance and economic decisions. *Econometrica* **65**(5), 1005–27.

Anselin, L. (2001) Spatial econometrics. In B. Baltagi (ed.), *A Companion to Theoretical Econometrics*. Oxford: Blackwell.

Anselin, L. (2006) Spatial econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, Volume 1*. Basingstoke: Palgrave Macmillan.

Arcand, J.-L. and M. Dagenais (2005) Errors in variables and the empirics of economic growth. Manuscript, CERDI-CNRS.

Arellano, M. (2003) *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**(2), 277–97.

Arellano, M. and O. Bover (1995) Another look at the instrumental-variable estimation of error-components models. *Journal of Econometrics* **68**, 29–51.

Baltagi, B., S. Song and W. Koh (2003) Testing panel data regression models with spatial error correlation. *Journal of Econometrics* **117**(1), 123–50.

Banerjee, A. and E. Duflo (2003) Inequality and growth: what can the data say? *Journal of Economic Growth* **8**(3), 267–300.

Barro, R. (1991) Economic growth in a cross section of countries. *Quarterly Journal of Economics* **106**(2), 407–43.

Barro, R. (1997) *Determinants of Economic Growth*. Cambridge, Mass: MIT Press.

Barro, R., N.G. Mankiw and X. Sala-i-Martin (1995) Capital mobility in neoclassical models of growth. *American Economic Review* **85**(1), 103–15.

Baumol, W. (1986) Productivity growth, convergence, and welfare: what the long-run data show. *American Economic Review* **76**(5), 1072–85.

Bertrand, M.,E. Duflo and S. Mullainathan (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**(1), 249–75.

Besley, T. and R. Burgess (2000) Land reform, poverty reduction, and growth: evidence from India. *Quarterly Journal of Economics* **115**(2), 389–430.
Besley, T. and R. Burgess (2002) The political economy of government responsiveness: theory and evidence from India. *Quarterly Journal of Economics* **117**(4), 1415–51.
Besley, T. and R. Burgess (2004) Can labor regulation hinder economic performance? Evidence from India. *Quarterly Journal of Economics* **119**(1), 91–134.
Bhattacharyya, S. (2004) Deep determinants of economic growth. *Applied Economics Letters*, **11**(9), 587–90.
Bhattacharyya, S. (2007) Theory and empirics of root causes of economic progress. Ph.D. thesis, Australian National University.
Bianchi, M. (1997) Testing for convergence: evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* **12**(4), 393–409.
Bils, M. and P. Klenow (2000) Does schooling cause growth? *American Economic Review* **90**(5), 1160–83.
Binder, M. and S. Brock (2004) A re-examination of determinants of economic growth using simultaneous equation dynamic panel data models. Mimeo, Johannes Goethe University, Frankfurt.
Binder, M. and M.H. Pesaran (1999) Stochastic growth models and their econometric implications. *Journal of Economic Growth* **4**, 139–83.
Binder, M. and M.H. Pesaran (2001) Life cycle consumption under social interactions. *Journal of Economic Dynamics and Control* **25**(1–2), 35–83.
Blomstrom, M., R. Lipsey and M. Zejan (1996) Is fixed investment the key to growth? *Quarterly Journal of Economics* **111**(1), 269–76.
Bloom, D., D. Canning and J. Sevilla (2003) Geography and poverty traps. *Journal of Economic Growth* **8**, 355–78.
Blundell, R. and S. Bond (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**(1), 115–43.
Bond, S. (2002) Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal* **1**, 141–62.
Bond, S., A. Hoeffler and J. Temple (2001) GMM estimation of empirical growth models. Centre for Economic Policy Research Discussion Paper No. 3048.
Bond, S., A. Leblebicioglu and F. Schiantarelli (2004) Capital accumulation and growth: a new look at the empirical evidence. Nuffield College, Oxford, Working Paper No. 2004-W8.
Bowsher, C. G. (2002) On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* **77**, 211–20.
Breiman, L., J. Friedman, R. Olshen and C. Stone (1984) *Classification and Regression Trees*. Redwood City, Calif.: Wadsworth Publishing.
Breusch, T. and A. Pagan (1980) The Lagrange multiplier test and its application to model specifications in econometrics. *Review of Economic Studies* **47**, 239–53.
Brock, W. and S. Durlauf (2001a) Growth empirics and reality. *World Bank Economic Review* **15**(2), 229–72.
Brock, W. and S. Durlauf (2001b) Interactions-based models. In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Volume 5*. Amsterdam: North-Holland.
Brock, W., S. Durlauf and K. West (2003) Policy evaluation in uncertain economic environments (with discussion). *Brookings Papers on Economic Activity* **1**, 235–322.
Brown, P., M. Vannucci and T. Fearn (1998) Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, **60**, 627–41.
Bun, M.J.G. and Carree, M.A. (2005) Bias-corrected estimation in dynamic panel data models. *Journal of Business and Economic Statistics* **23**(2), 200–10.
Bun, M.J.G. and J. Kiviet (2001) The accuracy of inference in small samples of dynamic panel data models. Tinbergen Institute Discussion Paper No. 2001–006/4.
Bun, M.J.G. and F. Windmeijer (2007) The weak instrument problem of the system GMM estimator in dynamic panel data models. University of Bristol Discussion Paper No. 07/595, March.

Burnside, C. and D. Dollar (2000) Aid, policies, and growth. *American Economic Review* **90**(4), 847–68.
Campos, N. and J. Nugent (2002) Who is afraid of political instability? *Journal of Development Economics* **67**, 157–72.
Canova, F., and A. Marcet (1995) The poor stay poor: non-convergence across countries and regions. Centre for Economic Policy Research Discussion Paper No. 1265.
Caselli, F., G. Esquivel and F. Lefort (1996) Reopening the convergence debate: a new look at cross country growth empirics. *Journal of Economic Growth* **1**(3), 363–89.
Ciccone, A. and M. Jarocinski (2007) Determinants of economic growth: will data tell. Manuscript, Universitat Pompeu Fabra, October.
Collier, P. (2007) *The Bottom Billion*. Oxford: Oxford University Press.
Conley, T. (1999) GMM estimation with cross-section dependence. *Journal of Econometrics* **92**, 1–45.
Conley, T. and E. Ligon (2002) Economic distance and long-run growth. *Journal of Economic Growth* **7**(2), 157–87.
Conley, T. and G. Topa (2002) Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* **17**(4), 303–27.
Cook, D. (2002) World War II and convergence. *Review of Economics and Statistics* **84**(1), 131–8.
Dagenais, M., and D. Dagenais (1997) Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics* **76**(1–2), 193–221.
Davidson, R. and J. MacKinnon (1993) *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
DeLong, J.B. (1988) Productivity growth, convergence, and welfare: comment. *American Economic Review* **78**(5), 1138–54.
DeLong, J.B. and L. Summers (1991) Equipment investment and economic growth. *Quarterly Journal of Economics* **106**(2), 445–502.
Desdoigts, A. (1999) Patterns of economic development and the formation of clubs. *Journal of Economic Growth* **4**(3), 305–30.
Doppelhofer, G. and M. Weeks (2007) Jointness of growth determinants. CESifo Working Paper No. 1978.
Draper, D., J. Hodges, C. Mallows and D. Pregibon (1993) Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A* **156**, 9–37.
Driscoll, J. and A. Kraay (1998) Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics* **80**(4), 549–60.
Duffy, J. and C. Papageorgiou (2000) A cross-country empirical investigation of the aggregate production function specification. *Journal of Economic Growth* **5**, 87–120.
Durlauf, S. and P. Johnson (1995) Multiple regimes and cross country growth behaviour. *Journal of Applied Econometrics* **10**(4), 365–84.
Durlauf, S., P. Johnson and J. Temple (2005) Growth econometrics. In P. Aghion and S.N. Durlauf (eds.), *Handbook of Economic Growth, Volume 1A*. Amsterdam: North-Holland.
Durlauf, S., A. Kourtellos, and A. Minkin (2001) The local Solow growth model. *European Economic Review* **45**(4–6), 928–40.
Durlauf, S., A. Kourtellos and C.M. Tan (2008) Are any growth theories robust? *Economic Journal* **118**(527) 329–46.
Durlauf, S. and D. Quah (1999) The new empirics of economic growth. In J. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics*. Amsterdam: North-Holland.
Easterly, W. (1996) When is stabilization expansionary? *Economic Policy* **22**, 67–98.
Easterly, W. and R. Levine (1997) Africa's growth tragedy: policies and ethnic divisions. *Quarterly Journal of Economics* **112**(4), 1203–50.
Eaton, J. and S. Kortum (1999) International technology diffusion: theory and measurement. *International Economic Review* **40**(3), 537–70.
Eaton, J. and S. Kortum (2001) Trade in capital goods. *European Economic Review* **4**(7), 1195–235.

Eicher, T., C. Papageorgiou and A. Raftery (2008) Determining growth determinants: default priors and predictive performance in Bayesian model averaging. Mimeo, University of Washington.
Eicker, F. (1967) Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*. Berkeley: University of California.
Fernandez, C., E. Ley and M. Steel (2001a) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**(5), 563–76.
Fernandez, C., E. Ley and M. Steel (2001b) Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**(2), 381–427.
Frankel, J. and D. Romer (1999) Does trade cause growth? *American Economic Review* **89**(3), 379–99.
Friedman, J. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–66.
Friedman, J. and J. Tukey (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **C23**, 881–90.
George, E.I. (1999) Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. *Bayesian Statistics* **6**, 175–77.
Giavazzi, F. and G. Tabellini (2005) Economic and political liberalizations. *Journal of Monetary Economics* **52**(7), 1297–330.
Glaeser, E., R. La Porta, F. Lopes-de-Silanes and A. Shleifer (2004) Do institutions cause growth? *Journal of Economic Growth* **9**(3), 271–303.
Graham, B. and J. Temple (2006) Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* **11**(1), 5–41.
Grier, K. and G. Tullock (1989) An empirical analysis of cross national economic growth, 1951–80. *Journal of Monetary Economics* **24**(2), 259–76.
Hahn, J. (1999) How informative is the initial condition in the dynamic panel model with fixed effects? *Journal of Econometrics* **93**(2), 309–26.
Hahn, J., J. Hausman and G. Kuersteiner (2001) Bias corrected instrumental variables estimation for dynamic panel models with fixed effects. Mimeo, MIT.
Hall, R. and C. Jones (1997) Levels of economic activity across countries. *American Economic Review Papers and Proceedings* **87**(2), 173–7.
Hall, R. and C. Jones (1999) Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* **114**(1), 83–116.
Hansen, B. (2000) Sample splitting and threshold estimation. *Econometrica* **68**(3), 575–603.
Harberger, A. (1987) Comment. In S. Fischer (ed.), *Macroeconomics Annual 1987*. Cambridge, Mass.: MIT Press.
Hausmann, R., L. Pritchett and D. Rodrik (2005) Growth accelerations. *Journal of Economic Growth* **10**(4), 303–29.
Hendry, D. (1995) *Dynamic Econometrics*. New York: Oxford University Press.
Hendry, D. and H.-M. Krolzig (2004) We ran one regression. *Oxford Bulletin of Economics and Statistics* **66**(5), 799–810.
Hendry, D. and H.-M. Krolzig (2005) The properties of automatic GETS modelling. *Economic Journal* **115**(502), C32–61.
Henry, P. (2000) Do stock market liberalizations cause investment booms? *Journal of Financial Economics* **58**(1–2), 301–34.
Henry, P. (2003) Capital account liberalization, the cost of capital, and economic growth. *American Economic Review* **93**(2), 91–6.
Hoeffler, A. (2002) The augmented Solow model and the African growth debate. *Oxford Bulletin of Economics and Statistics* **64**(2), 135–58.
Hoeting, J., A.E. Raftery and D. Madigan (1996) A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis* **22**, 251–70.

Holtz-Eakin, D., W. Newey and H. Rosen (1988) Estimating vector autoregressions with panel data. *Econometrica* **56**(6), 1371–95.
Hoover, K. and S. Perez (2004) Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics* **66**(5), 765–98.
Howitt, P. (2000) Endogenous growth and cross-country income differences. *American Economic Review* **90**(4), 829–46.
Islam, N. (1995) Growth empirics: a panel data approach. *Quarterly Journal of Economics* **110**(4), 1127–70.
Jones, C. (1995) Time series tests of endogenous growth models. *Quarterly Journal of Economics* **110**(2), 495–525.
Judson, R. and A. Owen (1999) Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters* **65**, 9–15.
Kalaitzidakis, P., T. Mamuneas and T. Stengos (2000) A non-linear sensitivity analysis of cross country growth regressions. *Canadian Journal of Economics* **33**(3), 604–17.
Kass, R. and L. Wasserman (1995) A reference Bayesian test for nested hypotheses and its relation to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–34.
Kaufmann, D., A. Kraay and M. Mastruzzi (2003) Governance matters III: governance indicators for 1996–2002. Mimeo, World Bank.
Kaufmann, D., A. Kraay and P. Zoido-Lobaton (1999a) Aggregating governance indicators. World Bank Policy Research Department Working Paper No. 2195.
Kaufmann, D., A. Kraay and P. Zoido-Lobaton (1999b) Governance matters. World Bank Policy Research Department Working Paper No. 2196.
Kiviet, J. (1995) On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* **68**(1), 53–78.
Kiviet, J. (1999) Expectations of expansions for estimators in a dynamic panel data model: some results for weakly exogenous regressors. In C. Hsiao (ed.), *Analysis of Panels and Limited Dependent Variable Models: In Honour of G.S. Maddala*. Cambridge: Cambridge University Press.
Klenow, P. and A. Rodriguez-Clare (1997) Economic growth: a review essay. *Journal of Monetary Economics* **40**, 597–617.
Klepper, S. and E. Leamer (1984) Consistent sets of estimates for regressions with errors in all variables. *Econometrica* **52**(1), 163–83.
Knack, S. and P. Keefer (1995) Institutions and economic performance: cross country tests using alternative institutional measures. *Economics and Politics* **7**(3), 207–27.
Kocherlakota, N. and K.-M. Yi (1997) Is there endogenous long-run growth? Evidence from the United States and the United Kingdom. *Journal of Money, Credit and Banking* **29**(2), 235–62.
Kormendi, R. and P. Meguire (1985) Macroeconomic determinants of growth: cross country evidence. *Journal of Monetary Economics* **16**(2), 141–63.
Kourtellos, A. (2003a) Modeling parameter heterogeneity in cross-country growth regression models. Mimeo, University of Cyprus.
Kourtellos, A. (2003b) A projection pursuit approach to cross-country growth data. Mimeo, University of Cyprus.
Krasker, W. and J. Pratt (1986) Bounding the effects of proxy variables on regression coefficients. *Econometrica* **54**(3), 641–55.
Krasker, W. and J. Pratt (1987) Bounding the effects of proxy variables on instrumental variables coefficients. *Journal of Econometrics* **35**(2/3), 233–52.
Laporte, A. and F. Windmeijer (2005) Estimation of panel data models with binary indicators when treatment effects are not constant over time. *Economics Letters* **88**(3), 389–96.
Leamer, E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley.
Leamer, E. (1983) Let's take the con out of econometrics. *American Economic Review* **73**(1), 31–43.

Leamer, E. and H. Leonard (1983) Reporting the fragility of regression estimates. *Review of Economics and Statistics* **65**(2), 306–17.
Lee, K., M. Pesaran and R. Smith (1997) Growth and convergence in multi country empirical stochastic Solow model. *Journal of Applied Econometrics* **12**(4), 357–92.
Lee, K., M.H. Pesaran and R. Smith (1998) Growth empirics: a panel data approach: a comment. *Quarterly Journal of Economics* **113**(1), 319–23.
Levine, R. and D. Renelt (1991) Cross-country studies of growth and policy: methodological, conceptual, and statistical problems. World Bank PRE Working Paper No. 608.
Levine, R. and D. Renelt (1992) A sensitivity analysis of cross-country growth regressions. *American Economic Review* **82**(4), 942–63.
Ley, E. and M. Steel (2008) On the effects of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*. Forthcoming.
Liu, Z. and T. Stengos (1999) Non-linearities in cross country growth regressions: a semiparametric approach. *Journal of Applied Econometrics* **14**(5), 527–38.
Loh, W.-Y. (2002) Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* **12**, 361–86.
Long, J. and L. Ervin (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician* **54**(3), 217–24.
MacKinnon, J. and H. White (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–25.
Magnus, J., O. Powell and P. Prufer (2008) A comparison of two model averaging techniques with an application to growth empirics. CentER Discussion Paper No. 2008–39, Tilburg University.
Mamuneas, T., A. Savvides and T. Stengos (2006) Economic development and the return to human capital: a smooth coefficient semiparametric approach. *Journal of Applied Econometrics* **21**(1), 111–32.
Mankiw, N.G. (1995) The growth of nations. *Brookings Papers on Economic Activity* **1**, 275–310.
Mankiw, N.G., D. Romer and D. Weil (1992) A contribution to the empirics of economic growth. *Quarterly Journal of Economics* **107**(2), 407–37.
Manski, C. (1993) Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* **60**(3), 531–42.
Marris, R. (1982) How much of the slow-down was catch-up? In R.C.O. Matthews (ed.), *Slower Growth in the Western World*. London: Heinemann.
Masanjala, W. and C. Papageorgiou (2005) Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian model averaging. Mimeo, Louisiana State University.
Mauro, P. (1995) Corruption and growth. *Quarterly Journal of Economics* **110**(3), 681–713.
Minier, J. (2007) Nonlinearities and robustness in growth regressions. *American Economic Review* **97**(2), 388–92.
Nerlove, M. (1999) Properties of alternative estimators of dynamic panel models: an empirical analysis of cross-country data for the study of economic growth. In C. Hsiao (ed.), *Analysis of Panels and Limited Dependent Variable Models: In Honour of G.S. Maddala*. Cambridge: Cambridge University Press.
Nerlove, M. (2000) Growth rate convergence, fact or artifact? An essay on panel data econometrics. In E. Ronchetti (ed.), *Panel Data Econometrics: Future Directions: Papers in Honour of Professor Pietro Balestra*. Amsterdam: North-Holland.
Nickell, S. (1981) Biases in dynamic models with fixed effects. *Econometrica* **49**(6), 1417–26.
Paap, R. and H. van Dijk (1998) Distribution and mobility of wealth of nations. *European Economic Review* **42**(7), 1269–93.
Papageorgiou, C. (2002) Trade as a threshold variable for multiple regimes. *Economics Letters* **71**(1), 85–91.
Papageorgiou, C. and W. Masanjala (2004) The Solow model with CES technology: nonlinearities with parameter heterogeneity. *Journal of Applied Econometrics* **19**(2), 171–201.

Papaioannou, E. and G. Siourounis (2007) Democratization and growth. *Economic Journal*. Forthcoming.
Persson, T. and G. Tabellini (1994) Is inequality harmful for growth? *American Economic Review* **84**(3), 600–21.
Persson, T. and G. Tabellini (2003) *The Economic Effects of Constitutions*. Cambridge, Mass.: MIT Press.
Pesaran, M.H. (2004) General diagnostic tests for cross-section dependence in panels. Mimeo, University of Cambridge.
Pesaran, M.H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74**(4), 967–1012.
Pesaran, M.H., Y. Shin and R. Smith (1999) Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association* **94**(446), 621–34.
Pesaran, M.H. and R. Smith (1995) Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* **68**(1), 79–113.
Pesaran, M.H., A. Ullah and T. Yamagata (2008) A bias-adjusted LM test of error cross-section independence. *Econometrics Journal* **11**, 105–27.
Phillips, P. and D. Sul (2003) Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* **6**, 217–59.
Phillips, P. and D. Sul (2007a) Transition modeling and econometric convergence tests. *Econometrica* **75**(6), 1771–855.
Phillips, P. and D. Sul (2007b) Economic transition and growth. Mimeo, University of Auckland.
Pritchett, L. (2000a) Understanding patterns of economic growth: searching for hills among plateaus, mountains, and plains. *World Bank Economic Review* **14**(2), 221–50.
Pritchett, L. (2000b) The tyranny of concepts: CUDIE (cumulated, depreciated, investment effort) is not capital. *Journal of Economic Growth* **5**(4), 361–84.
Raftery, A. (1995) Bayesian model selection in social research. *Sociological Methodology* **25**, 111–63.
Raftery, A., D. Madigan and J. Hoeting (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**(437), 179–91.
Robertson, D. and J. Symons (1992) Some strange properties of panel data estimators. *Journal of Applied Econometrics* **7**(2), 175–89.
Rodriguez, F. and D. Rodrik (2001) Trade policy and economic growth: a user's guide. In B. Bernanke and K. Rogoff (eds.), *Macroeconomics Annual 2000*. Cambridge, Mass.: MIT Press.
Rodrik, D. (1999) Where did all the growth go? External shocks, social conflict, and growth collapses. *Journal of Economic Growth* **4**(4), 385–412.
Rodrik, D. (ed.) (2003) *In Search of Prosperity: Analytic Narratives on Economic Growth*. Princeton: Princeton University Press.
Rodrik, D. and R. Wacziarg (2005) Do democratic transitions produce bad economic outcomes? *American Economic Review* **95**(2), 50–5.
Roodman, D. (2006) How to do xtabond2: an introduction to "Difference" and "System" GMM in Stata. CGD Working Paper No. 103, December.
Roodman, D. (2007) A short note on the theme of too many instruments. CGD Working Paper No. 125, August.
Sachs, J. and A. Warner (1995) Economic reform and the process of global integration (with discussion). *Brookings Papers on Economic Activity* **1**, 1–118.
Sala-i-Martin, X. (1991) Growth, macroeconomics, and development: comments. In O. Blanchard and S. Fischer (eds.), *Macroeconomics Annual 1991*. Cambridge, Mass.: MIT Press.
Sala-i-Martin, X. (1997a) I just ran 4 million regressions. National Bureau of Economic Research Working Paper No. 6252.
Sala-i-Martin, X. (1997b) I just ran 2 million Regressions. *American Economic Review* **87**(2), 178–83.

Sala-i-Martin, X., G. Doppelhofer and R. Miller (2004) Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* **94**(4), 813–35.
Shioji, E. (2001) Public capital and economic growth: a convergence approach. *Journal of Economic Growth* **6**(3), 205–27.
Solow, R. (1994) Perspectives on growth theory. *Journal of Economic Perspectives* **8**, 45–54.
Stock, J. and M. Watson (2004) *Introduction to Econometrics*. Boston: Addison-Wesley.
Swartz, S. and R. Welsch (1986) Applications of bounded-influence and diagnostic methods in energy modeling. In D. Belsley and E. Kuh (eds.), *Model Reliability*. Cambridge, Mass.: MIT Press.
Tan, C.M. (2004) No one true path to development: uncovering the interplay between geography, institutions, and ethnic fractionalization in economic development. Mimeo, Tufts University.
Tavares, J. and R. Wacziarg (2001) How democracy affects growth. *European Economic Review* **45**(8), 1341–78.
Taylor, C. (1998) *Socrates*. New York: Oxford University Press.
Temple, J. (1998) Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* **13**(4), 361–75.
Temple, J. (1999) The new growth evidence. *Journal of Economic Literature* **37**(1), 112–56.
Temple, J. (2000a) Inflation and growth: stories short and tall. *Journal of Economic Surveys* **14**(4), 395–426.
Temple, J. (2000b) Growth regressions and what the textbooks don't tell you. *Bulletin of Economic Research* **52**(3), 181–205.
Temple, J. (2003) The long-run implications of growth theories. *Journal of Economic Surveys* **17**(3), 497–510.
Temple, J. (2005) Dual economy models: a primer for growth economists. *Manchester School* **73**(4), 435–78.
Temple, J. (2006) Aggregate production functions and growth economics. *International Review of Applied Economics* **20**(3), 301–17.
Temple, J. and L. Wößmann (2006) Dualism and cross-country growth regressions. *Journal of Economic Growth* **11**(3), 187–228.
Wacziarg, R. (2002) Review of Easterly's the elusive quest for growth. *Journal of Economic Literature* **40**(3), 907–18.
Wacziarg, R. and K. Welch (2003) Trade liberalization and growth: new evidence. National Bureau of Economic Research Working Paper No. 10152.
Warner, A. (1992) Did the debt crisis cause the investment crisis? *Quarterly Journal of Economics* **107**(4), 1161–86.
White, H. (1980) A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–38.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno deFinetti*. Amsterdam: North-Holland.
Zietz, J. (2001) Heteroskedasticity and neglected parameter heterogeneity. *Oxford Bulletin of Economics and Statistics* **63**(2), 263–73.

25
The Econometrics of Finance and Growth

Thorsten Beck

Abstract

This chapter reviews different econometric methodologies to assess the relationship between financial development and growth. It illustrates the identification problem, which is at the center of the finance and growth literature, using the example of a simple ordinary least squares estimation. It discusses cross-sectional and panel instrumental variable approaches to overcome the identification problem. It presents the time series approach, which focuses on the forecast capacity of financial development for future growth rates, and differences-in-differences techniques that try to overcome the identification problem by assessing the differential effect of financial sector development across states with different policies or across industries with different needs for external finance. Finally, it discusses firm- and household-level approaches that allow analysts to dig deeper into the channels and mechanisms through which financial development enhances growth and welfare, but pose their own methodological challenges.

25.1 Introduction

Economists have discussed over the past 100 years whether or not financial development has a causal impact on economic development. Theory suggests that effective financial institutions and markets that help overcome market frictions introduced by information asymmetries and transaction costs can foster economic growth through several channels. Specifically, they help (i) ease the exchange of goods and services by providing payment services, (ii) mobilize and pool savings

from a large number of investors, (iii) acquire and process information about enterprises and possible investment projects, thus allocating society's savings to its most productive use, (iv) monitor investments and exert corporate governance, and (v) diversify and reduce liquidity and intertemporal risk. However, other models show that higher returns from better resource allocation may depress saving rates, resulting in overall growth rates actually slowing with more effective financial markets and institutions.[1]

While the finding of a positive correlation between indicators of financial development and economic growth cannot settle this debate, advances in computational capacity and availability of large cross-country datasets with relatively large time dimensions have enabled researchers to rigorously explore the relationship between financial development and economic growth. Further, as more disaggregated datasets have become available, the finance and growth literature has proceeded from using country-level data, to using industry- and firm-level data, to more recently using household data. While the cross-country literature has developed more sophisticated models to address biases introduced by measurement error, reverse causation and omitted variables, the progress to firm- and household-level data allows not only additional ways to address these biases, but also tests of the specific channels through which finance might enhance economic growth.

The econometrics of finance and growth can be summarized in the following simple regression model:

$$g(i,t) = y(i,t) - y(i,t-1) = \alpha + \beta_i f(i,t) + C(i,t)\gamma_i + \mu(i) + \varepsilon(i,t), \qquad (25.1)$$

where y is the log of real GDP per capita or of another measure of welfare, g is the growth rate of y, f is an indicator of financial development, C is a set of conditioning information, μ and ε are error terms, i is the observational unit – be it a country, an industry, a firm or a household – and t is the time period. While ε is a white-noise error with a mean of zero, μ is a country-specific element of the error term that does not necessarily have a mean of zero. The explanatory variables are measured either as an average over the sample period or as an initial value. The sign and significance of the coefficient β_i is at the center of the debate. As discussed in the remainder of this chapter, the estimate of β_i can be biased for a variety of reasons, among them measurement error, reverse causation and omitted variable bias. While the cross-country literature assumes $\beta_i = \beta$, with some research supporting this assumption (Loayza and Ranciere, 2006), the time series literature does not impose this restriction. Further, several industry- and firm-level studies test whether β varies across industries or firms with different characteristics, utilizing interaction terms.

This chapter is concerned with an unbiased, consistent and efficient estimator of β_i.[2] In this context, we abstract from a number of other problems in the finance and growth literature. First, this chapter does not cover problems arising from the lack of appropriate data, although we are concerned about measurement error in the financial indicators and the bias this introduces in the estimation. Second, while

we are concerned about the bias introduced by the potential reverse causation from growth to finance, we are not concerned about this reverse causation *per se*, that is, we do not discuss in depth the literature focusing on the impact of economic growth on financial development and bidirectional causality. Finally, this chapter does not intend to be a fully-fledged survey of the empirical finance and growth literature, as is Levine (2005), but rather focuses on studies with methodological contributions.

While this chapter is concerned about estimating the relationship between finance and growth, some remarks about measuring financial development might be useful. While the theoretical literature links specific functions of the financial system to economic growth, data limitations have forced researchers to focus on variables capturing the size, activity or efficiency of specific financial institutions or markets. The first generation of papers in the finance and growth literature have built on aggregate data on financial institutions, mainly banks, available for 30–40-year periods for a large number of developed and developing countries. Such indicators include monetization variables, such as the ratio of M2 or M3 to gross domestic product (GDP), or financial depth indicators, such as the ratio of private credit (outstanding claims of financial institutions on the private sector) to GDP. Later papers have added indicators of the size and liquidity of stock markets, albeit available for fewer countries and shorter time periods. Indicators for the efficiency and competitiveness of financial systems, non-bank financial institutions such as institutional investors and, most importantly, the outreach of financial systems, are available for only a few countries and often do not have a time dimension.[3] Within-country studies allow researchers to utilize more micro-based data or focus on specific policy interventions or reforms.

The remainder of the chapter is structured as follows. Section 25.2 illustrates the identification problem, which is at the center of the finance and growth literature, using the example of a simple ordinary least squares (OLS) estimation of regression (25.1). Section 25.3 discusses instrumental variable (IV) approaches using cross-sectional and panel data. Section 25.4 discusses time series approaches, and section 25.5 differences-in-differences techniques. Section 25.6 discusses the use of firm- and household-level data and the methodological challenges this implies. Section 25.7 concludes and looks forward to new research directions.

25.2 Correlation versus causality: the identification problem

Goldsmith (1969) was the first to empirically show the positive correlation between financial development and GDP per capita, using data on the assets of financial intermediaries relative to GNP and data on the sum of net issues of bonds and securities plus changes in loans relative to GNP for 35 countries over the period 1860–1963. Such a correlation, however, does not control for other factors that are associated with economic growth and might thus be driven by other country characteristics correlated with both finance and growth. Second, such a correlation does not provide any information on the direction of causality between finance and growth. The early finance and growth literature has therefore used standard

cross-country OLS regressions that build on an augmented Barro growth regression as in (25.1), with data for each country averaged over the sample period, assuming $\beta_i = \beta$ and $\gamma_i = \gamma$ for all countries, and including the lagged dependent variable as a control variable:

$$g(i) = y(i,t) - y(i,t-1) = \alpha + \beta f(i) + C(i)\gamma + \delta y(i,t-1) + \varepsilon(i). \tag{25.2}$$

Unlike regression (25.1), regression (25.2) has thus only a cross-country, but not a time series, dimension. The log of initial income per capita is included to control for convergence predicted by the Solow–Swan growth models. Including other country characteristics, such as initial levels of human or physical capital, and policy variables, such as government consumption or trade openness, in a set of conditioning information allows testing for an independent partial correlation of finance with growth. The coefficient β is of interest for finance and growth researchers, who interpret a positive and significant coefficient as evidence for a positive partial correlation between finance and growth.

Running this cross-country regression for a sample of 77 countries over the period 1960–89, King and Levine (1993) found a positive and significant relationship between several financial development indicators and GDP per capita growth. Their study focuses mostly on monetization indicators and indicators measuring the size and relative importance of banking institutions. Using initial values of financial development confirms their finding. Levine and Zervos (1998) expanded the analysis to include measures of stock market development and found a positive partial correlation of both stock market and bank development with GDP per capita growth over the period 1976–94.[4] Interestingly, they found a positive and significant link between liquidity of stock markets – as measured by a turnover indicator or value traded to GDP – and economic growth, but no robust relationship between the size of stock markets and economic growth. The empirical relationship between finance and growth, however, is not only statistically, but also economically, significant. Levine and Zervos found that a one standard deviation change in stock market liquidity and banking sector development explains an annual GDP per capita growth difference of 0.8 and 0.7 percentage points, respectively, adding up to a total difference in GDP per capita of 31% over the 18-year sample period.

OLS estimates, however, are only consistent if the following orthogonality conditions hold:

$$E[C(i)'\varepsilon(i)] = 0; \quad E[y(i,t-1)'\varepsilon(i)] = 0; \quad E[f(i)'\varepsilon(i)] = 0. \tag{25.3}$$

A violation of this condition can arise for several reasons. First, the presence of an unobserved country-specific effect $\mu(i)$ – as in regression (25.1) – results in a positive correlation of the lagged dependent variable with the error term as, unlike the error term $\varepsilon(i)$, $\mu(i)$ does not have a mean of zero, so that:

$$E[y(i,t-1)'(\mu(i) + \varepsilon(i))] \neq 0. \tag{25.4}$$

Omitted variable bias can also arise if other explanatory variables are correlated with the unobserved country-specific effect or if explanatory variables that should

be included in regression (25.2) are (i) not included and (ii) correlated with included explanatory variables, so that:

$$E[C(i)'(\mu(i) + \varepsilon(i))] \neq 0. \tag{25.5}$$

Second, reverse causation from GDP per capita growth to financial development or another explanatory variable could violate the orthogonality condition and thus bias the estimator of β if $\varepsilon(i)$ and $\nu(i)$ are correlated with each other, as would occur if:

$$f(i) = \lambda y(i, t-1) + \nu(i). \tag{25.6}$$

Third, one of the explanatory variables could be mismeasured, so that:

$$f^*(i) = f(i) + u(i), \tag{25.7}$$

where f^* is the true level and f is the measured level of financial development. This could result in attenuation bias if the measurement error is correlated with f.

Several simple approaches to overcome these biases have been suggested. First, controlling for other country traits and policies can help minimize the omitted variable bias and allow testing for the robustness of the finance and growth link (Levine and Renelt, 1992). However, the number of observations, and thus degrees of freedom, severely limits this approach in a typical cross-country regression. Second, several studies have used initial values of financial development, rather than values averaged over the same period as GDP per capita growth. If the true time span over which an improvement in financial development results in higher growth is shorter than the sample period used in the regression, then using initial values might reduce biases stemming from reverse causation. On the other hand, using initial values does not correct for biases introduced by omitted variables, measurement error or the inclusion of the lagged dependent variable, and implies a loss of information to be used in the estimation. Third, using panel regressions with fixed country effects would eliminate any time-invariant omitted variable bias and time-invariant measurement bias. However, the correlation between the transformed lagged dependent variable and the transformed error term will make the fixed-effect estimator biased, and this bias is only eliminated as the number of time periods goes towards infinity, which is certainly not the case for the typical growth regression with fewer than 40 annual data points. Finally, fixed-effect regressions also have the conceptual shortcoming that they effectively limit the analysis to within-country variation in growth and financial development by differencing-out cross-country variation.

25.3 The IV approach

The classical approach in cross-country growth regressions to overcome the biases related to OLS is to identify an instrument that helps isolate that part of the variation in the endogenous variable that is not associated with reverse causation,

omitted variables and measurement error. Following the seminal work by La Porta *et al.* (1997, 1998), who identified variation in countries' legal origin as an historical exogenous factor explaining current variation in countries' level of financial development, an extensive literature has utilized this variable to extract the exogenous component of financial development.

To overcome biases related to the inclusion of the lagged dependent variable and omitted variable bias, while at the same time controlling for reverse causation and measurement error, researchers have utilized dynamic panel regressions using lagged values of the explanatory endogenous variables as instruments. Finally, to control for country heterogeneity in the finance–growth relationship, researchers have utilized pooled mean group estimators. We will discuss each methodology in turn.

25.3.1 Cross-sectional regressions

Underlying IV estimation is the following specification:

$$g(i) = y(i,t) - y(i,t-1) = \alpha_1 + \beta_1 f(i) + C(i)\gamma_1 + \delta_1 y(i,t-1) + \varepsilon(i) \tag{25.8}$$

$$f(i) = \alpha_2 + Z(i)\beta_2 + C(i)\gamma_2 + \delta_2 y(i,t-1) + \nu(i) \tag{25.9}$$

$$f^*(i) = f(i) + u(i), \tag{25.10}$$

where C are the included exogenous and Z the excluded exogenous control variables; the latter are also referred to as IVs which allow us to extract the exogenous component of $f(i)$ that is not correlated with $\varepsilon(i)$, that is, $E[Z(i)'\varepsilon(i)] = 0$, and $E[Z(i)'u(i)] = 0$.[5] Estimating regression (25.8) with instruments can help alleviate biases arising from reverse causation, omitted variables and measurement error.

Regression (25.8) is typically estimated with a two-stage least squares (2SLS) estimator. Unlike the OLS estimator, the 2SLS estimator only uses the variation in the explanatory variables that is correlated with the instrument and therefore uses less information than the OLS estimator. If OLS is consistent, it is therefore more efficient than IV, whereas if OLS is inconsistent, the IV estimator is both consistent and efficient.[6]

The 2SLS estimator can also be derived as a generalized method of moments (GMM) estimator that minimizes a set of orthogonality conditions (Hansen, 1982). In the case where there are more excluded exogenous than endogenous variables, a weighting matrix has to be used. While the 2SLS estimator uses a weighting matrix constructed under the assumption of homoskedasticity, the weighting matrix of the GMM estimator is constructed as the inverse of the variance-covariance matrix, thus assigning different weights to the orthogonality condition, according to their variances. While the 2SLS estimator is thus consistent, it is inefficient as it does not use all the available information. On the other hand, the GMM estimator relies on asymptotic characteristics and therefore suffers from a finite-sample bias as the optimal weighting matrix is a function of fourth moments (Hayashi, 2000).[7]

Using legal origin as an instrument for financial development, Levine (1998, 1999) finds a positive relationship between finance and economic growth. Researchers have also used other historical and exogenous country characteristics

as instruments for financial development, such as settler mortality and latitude, to proxy for geographic conditions, ethnic fractionalization, religious composition of the population, and years since independence (McCraig and Stengos, 2005). Guiso, Sapienza and Zingales (2004) use sub-national variation in historical bank restriction indicators across 20 Italian regions and its 103 provinces as IVs to assess the impact of financial development and competition on economic growth and other real sector outcomes.

IV regressions depend on the quality of the IVs, independent of whether 2SLS or GMM is applied. As discussed above, these instruments are typically exogenous country characteristics, such as geographic traits, or based on historical experience, such as legal origin. The challenge is to identify the economic mechanisms through which the IVs influence the endogenous variable – financial development – while at the same time assuring that the instruments are not correlated with growth directly. An extensive literature has discussed the historic determinants of financial sector development and the channels through which, for example, legal origin has helped shape current financial sector development,[8] but there are also several formal econometric conditions to be fulfilled in order for an instrument to be valid. First, the exogenous variables cannot be correlated with error terms, that is, $E[Z(i)'\varepsilon(i)] = 0$ (orthogonality or exogeneity condition). Second, the excluded exogenous instruments have to explain the variation in the endogenous variables after controlling for the included exogenous variables, that is, the F-test for $Z(i)$ in (25.9) is rejected at conventional levels (relevance condition).

The orthogonality condition is typically tested with the Sargan (1958) test of overidentifying restrictions (OIR) if there are more instruments than explanatory variables, that is: $\hat{\varepsilon}'Z(Z'Z)^{-1}Z'\hat{\varepsilon}/\widehat{\sigma^2}$, where $\widehat{\sigma^2} = (\hat{\varepsilon}'\hat{\varepsilon})/n$ and $\hat{\varepsilon}$ is the vector of residuals from estimating regression (25.8). This test can easily be calculated from a regression of the IV regression residuals on included and excluded exogenous variables. It is distributed as χ^2 with $(J - K)$ degrees of freedom under the null hypothesis that the residuals are not correlated with the exogenous variables, where J is the number of instruments and K is the number endogenous variables.[9] Hansen's (1982) J-test is a generalization of the Sargan OIR test to the GMM context and is the value of the GMM objective function evaluated at the efficient GMM estimator: $\hat{\varepsilon}'Z(Z'\hat{\Omega}Z)^{-1}Z'\hat{\varepsilon}$, where $\hat{\Omega}$ is the estimated variance-covariance matrix of the residuals from regression (25.8). As with the Sargan test, Hansen's test is distributed as χ^2 with $(J - K)$ degrees of freedom.

The test of OIR, however, is relatively weak. First, the test only assesses the validity of any additional instruments, that is, it cannot be performed if the number of excluded exogenous variables is the same as the number of endogenous variables. Further, the test tends to reject the null hypothesis of valid instruments too often in small samples (Murray, 2006). Most importantly, the test over-rejects if the instruments are weak, that is, if they do not explain the endogenous variables in the first stage.

The second condition of instrument relevance can be tested in different ways. First, one can use an F-test of the joint significance of the instruments in (25.9);

the critical values of this F-test for IV estimation, however, are larger than for OLS estimation; for the case of a single endogenous variable, Staiger and Stock (1997) show, using Monte Carlo simulations, that for most specifications and independent of the degrees of freedom, a critical value of 10 is sufficient to reject the null hypothesis, and Stock and Yogo (2005) derive critical values for this F-test for the case of several endogenous variables, with the critical values increasing with the number of instruments.[10] Second, one can use a partial R^2 of the first-stage regression (25.9) that takes into account the intercorrelation among the instruments (Shea, 1997). Specifically, Godfrey (1999) shows that this statistic for endogenous regressor *i* is:

$$\frac{\hat{\sigma}_i^{OLS}}{\hat{\sigma}_i^{IV}}\left[\frac{\left(1-R_{IV}^2\right)}{\left(1-R_{OLS}^2\right)}\right],$$

where $\hat{\sigma}_i$ is the estimated asymptotic variance of the coefficient *i*. This measure thus tests for the relevance of the individual instruments, unlike the F-test, which tests for the overall relevance.

Weak instruments can bias the IV results towards OLS and turn them inconsistent. Further, weak instruments can result in an over-rejection of the overidentification test discussed above. If instruments are both invalid and irrelevant, the bias thus increases in a multiplicative way.[11]

Most of the cross-country finance and growth papers utilizing IVs find that the IV estimator of β_1 is higher than the OLS estimator.[12] Manipulating regressions (25.8), (25.9) and (25.10), one can show that this implies:

$$\hat{\delta}_2 + \hat{\rho}\frac{\widehat{\sigma(\upsilon)}}{\widehat{\sigma(\varepsilon)}} < \hat{\beta}_1\left(1-\hat{\beta}_1\hat{\delta}_2\right)\frac{\widehat{\sigma(u)}}{\widehat{\sigma(\varepsilon)}}, \tag{25.11}$$

where ρ is the correlation between ε and υ, and the other parameters are taken from regressions (25.8), (25.9) and (25.10). There are several possible explanations for this finding and thus for inequality (25.11) to hold (Kraay and Kaufman, 2002). First, there could be negative reverse causation ($\delta_2 < 0$), which would bias the OLS estimator of the β_1 coefficient downwards. Given empirical studies showing the positive relationship between economic and financial development, this explanation seems rather unlikely (Harrison, Sussman and Zeira, 1999). A second explanation that makes inequality (25.11) hold is that omitted variables are correlated with growth and finance with opposite signs ($\rho < 0$), an explanation for which, again, little evidence exists. A third – and most commonly adopted – explanation relies on attenuation bias, where measurement error in financial development ($\widehat{\sigma(u)}$) biases the OLS estimate downwards and makes inequality (25.11) hold. Critically, however, if the IVs are positively correlated with omitted variables and the exclusion condition is thus violated, the IV estimator of β_1 is biased upwards. This is of concern, as a few IVs, such as historical country traits, have been

used for many different institutional variables in the context of growth regressions (Pande and Udry, 2006). Specifically, legal origin has been shown to be associated with an array of institutional arrangements, ranging from financial markets over general regulatory approaches, to labor market institutions. A significant correlation between institutional variables left out of the regressions and the IVs can therefore also result in an upwardly biased IV estimator of β_1.

25.3.2 Dynamic panel analysis

While the cross-sectional IV regressions address biases related to omitted variables, reverse causation and measurement error, they do face several limitations. First, cross-country studies using cross-sectional IV regressions typically control only for the endogeneity and measurement error of financial development, but not of other explanatory variables entering the growth regressions. Second, in the presence of country-specific omitted variables, the lagged dependent variable is correlated with the error term if it is not instrumented.

As an alternative to cross-sectional IV regressions, researchers have therefore used dynamic panel regressions of the following format:

$$g(i,t) = \alpha + \beta f(i,t) + C^{(1)}(i,t)\gamma_1 + C^{(2)}(i,t)\gamma_2 + \delta y(i,t-1) + \mu(i) + \lambda(t) + \varepsilon(i,t), \tag{25.12}$$

where $C^{(1)}$ represents a set of exogenous explanatory variables, $C^{(2)}$ a set of endogenous explanatory variables, and λ a vector of time dummies. Note that β is still assumed to be constant across countries, a restriction that we will relax further below.

Unlike the cross-sectional regressions, which use external instruments, that is, variables that are completely external to the second-stage regression, the dynamic panel regressions use internal instruments, that is, lagged realizations of the explanatory variables. While this method does not control for full endogeneity, it does control for weak exogeneity, which means that current realizations of f or variables in $C^{(2)}$ can be affected by current and past realizations of the growth rate, but must be uncorrelated with future realizations of the error term. Thus, under the weak exogeneity assumption, future innovations of the growth rate do not affect current financial development.

In order to address the different biases in regression (25.12), Arellano and Bond (1991) suggest first-differencing the regression equation to eliminate the country-specific effect, as follows:[13]

$$\Delta g(i,t) = \beta \Delta f(i,t) + \Delta C^{(1)}(i,t)\gamma_1 + \Delta C^{(2)}(i,t)\gamma_2 + \delta \Delta y(i,t-1) + \Delta\lambda(t) + \Delta\varepsilon(i,t), \tag{25.13}$$

where $\Delta x(t) = x(t) - x(t-1)$. This procedure solves the omitted variable bias, as described above, but introduces a correlation between the new error term, $\Delta\varepsilon(i,t)$, and the lagged dependent variable, $\Delta y(i,t-1)$. To address this correlation and the endogeneity and measurement problems, Arellano and Bond suggest using lagged

values of the explanatory variables in levels as instruments for current differences of the endogenous variables. Under the assumptions that there is no serial correlation in the error term ε and that the explanatory variables f and $C^{(2)}$ are weakly exogenous, one can use the following moment conditions to estimate regression (25.13):

$$\begin{aligned} E[f(i,t-s)'\Delta\varepsilon(i,t)] &= 0, \quad \text{for each } t = 3,\ldots,T, s \geq 2 \\ E\left[C^{(2)}(i,t-s)'\Delta\varepsilon(i,t)\right] &= 0, \quad \text{for each } t = 3,\ldots,T, s \geq 2 \\ E[y(i,t-s)'\Delta\varepsilon(i,t)] &= 0, \quad \text{for each } t = 3,\ldots,T, s \geq 2. \end{aligned} \tag{25.14}$$

Using these moment conditions, Arellano and Bond propose a two-step GMM difference estimator. In the first step, the error terms are assumed to be both independent and homoskedastic across countries and over time, while in the second step, the residuals obtained in the first step are used to construct a consistent estimate of the variance-covariance matrix, thus relaxing the assumptions of independence and homoskedasticity. Simulations, however, have shown very modest efficiency gains from using the two-step as opposed to the one-step estimator, while the two-step estimator tends to underestimate the standard errors of the coefficient given that the two-step weight matrix depends on estimated parameters from the one-step estimator (Bond and Windmeijer, 2002).

There are several conceptual and econometric shortcomings with the difference estimator. First, by first-differencing we lose the pure cross-country dimension of the data. Second, differencing may decrease the signal-to-noise ratio, thereby exacerbating measurement error biases (see Griliches and Hausman, 1986). Finally, Alonso-Borrego and Arellano (1999) and Blundell and Bond (1998) show that, if the lagged dependent and the explanatory variables are persistent over time, that is, have very high autocorrelation, then the lagged levels of these variables are weak instruments for the regressions in differences.[14] Simulation studies show that the difference estimator has a large finite-sample bias and poor precision.

To address these conceptual and econometric problems, Arellano and Bover (1995) suggest an alternative estimator that combines the regression in differences with the regression in levels. Using Monte Carlo experiments, Blundell and Bond (1998) show that the inclusion of the level regression in the estimation reduces the potential biases in finite samples and the asymptotic imprecision associated with the difference estimator. Using the regression in levels, however, does not directly eliminate the country-specific effect μ. Lagged differences of the explanatory variables can be used as instruments for the levels of the endogenous explanatory variables under the assumption that the correlation between μ and the levels of the explanatory variables is constant over time, such that:

$$\begin{aligned} E[f(i,t+p)'\mu(i)] &= E[f(i,t+q)'\mu(i)], \quad \text{for all } p \text{ and } q \\ E\left[C^{(2)}(i,t+p)'\mu(i)\right] &= E\left[C^{(2)}(i,t+q)'\mu(i)\right], \quad \text{for all } p \text{ and } q. \end{aligned} \tag{25.15}$$

Under this assumption, lagged differences are valid instruments for the regression in levels, and the moment conditions for the regression in levels are as follows:

$$
\begin{aligned}
E[\Delta f(i,t-s)'(\varepsilon(i,t)+\mu(i))] &= 0, \quad \text{for each } t=3,\ldots,T, s=2 \\
E\left[\Delta C^{(2)}(i,t-s)'(\varepsilon(i,t)+\mu(i))\right] &= 0, \quad \text{for each } t=3,\ldots,T, s=2 \\
E[\Delta y(i,t-s)'(\varepsilon(i,t)+\mu(i))] &= 0, \quad \text{for each } t=3,\ldots,T, s=2. \qquad (25.16)
\end{aligned}
$$

The system thus consists of the stacked regressions in differences and levels, with the moment conditions in (25.14) applied to the first part of the system, the regressions in differences, and the moment conditions in (25.16) applied to the second part, the regressions in levels.[15] As with the difference estimator, the model is estimated in a two-step GMM procedure.

The consistency of the GMM estimator depends both on the validity of the instruments (exclusion condition) and the assumption that the error term, ε, does not exhibit serial correlation. Arellano and Bond (1991) propose two tests to examine these assumptions. The first is a Sargan test of OIR, which is constructed in a similar manner to the cross-sectional test discussed above. In the context of the system estimator, one can also compute a "difference-in-Sargan" test, the C-statistic (Eichenbaum, Hansen and Singleton, 1988), to test the orthogonality condition of a sub-set of instruments, such as the instruments applied to the level regressions. The C-statistic is computed as the difference of two Sargan/Hansen statistics, the one for the regression using the full set of instruments and the one using a smaller set of instruments. The C-statistic is distributed as χ^2 with the degrees of freedom equal to the number of instruments dropped from the second regression.

The second test examines the assumption of no serial correlation in the error terms, specifically whether the differenced error term is second-order serially correlated as, by construction, the error term $\Delta\varepsilon(i,t)$ from the difference regression is first-order serially correlated and we cannot use the error terms from the regression in levels since they include the country-specific effect μ. This test is based on the standardized average residual autocovariances and, under the null hypothesis of no second-order serial correlation, has a standard normal distribution.

Rousseau and Wachtel (2000) use the difference estimator with annual data over the period 1980–95 across 47 countries and find a positive link between indicators of bank and stock market development and economic growth.[16] Using five-year averages over the period 1960–95 across 74 countries, Beck, Levine and Loayza (2000) and Levine, Loayza and Beck (2000) use both the difference and the system estimator and find a positive and significant relationship between indicators of financial intermediary development and GDP per capita growth, with the specification tests referred to above confirming the validity of both instruments and econometric model.[17] Beck *et al.* (2000) also find that the effect of finance on growth is through productivity growth, while there is no robust relationship between financial development and capital accumulation when controlling for biases due to simultaneity, omitted variables and measurement error.

The dynamic panel estimators have typically been applied to panels with few time periods and many countries. Further, the instrumental variable matrix Z is typically constructed with separate columns for instruments in different time periods, resulting in a quadratic increase in the number of columns of Z as the number of time periods increases (Roodman, 2007). This results in an overfit of the endogenous variables, biasing the coefficient estimates towards OLS estimates and biasing the Sargan/Hansen test for joint validity of the instruments towards over-accepting the null hypothesis (Bowsher, 2002). In order to avoid overfitting, one can limit the number of lags used in the difference regression or combine instruments into smaller sets, effectively imposing the constraint that instruments of each lag distance have the same coefficient when projecting regressors onto instruments (Beck and Levine, 2004; Roodman, 2007). In this case, the orthogonality conditions for the difference regressions are:

$$\begin{aligned} E[f(i,t-s)'\Delta\varepsilon(i,t)] &= 0, \quad \text{for each } s \geq 2 \\ E\left[C^{(2)}(i,t-s)'\Delta\varepsilon(i,t)\right] &= 0, \quad \text{for each } s \geq 2 \\ E[y(i,t-s)'\Delta\varepsilon(i,t)] &= 0, \quad \text{for each } s \geq 2, \end{aligned} \tag{25.17}$$

and the orthogonality conditions for the levels regressions are:

$$\begin{aligned} E[\Delta f(i,t-s)'(\varepsilon(i,t)+\mu(i))] &= 0, \quad \text{for each } s \geq 2 \\ E\left[\Delta C^{2}(i,t-s)'(\varepsilon(i,t)+\mu(i))\right] &= 0, \quad \text{for each } s \geq 2 \\ E[\Delta y(i,t-s)'(\varepsilon(i,t)+\mu(i))] &= 0, \quad \text{for each } s \geq 2. \end{aligned} \tag{25.18}$$

Given that data on financial sector indicators for a broad cross-section of countries are only available for a 25–40-year period, most studies split the sample period into non-overlapping five-year periods, thus controlling for business cycle effects, while at the same time having a reasonable number of time periods. An alternative to splitting the sample period into a number of five-year periods is to utilize overlapping five-year periods, as proposed by Bekaert, Harvey and Lundblad (2005), thus allowing researchers to increase the number of time periods in the panel. In order to control for the MA(4) character of the data, the weighting matrix of the GMM estimator has to be adjusted accordingly.

Both the cross-sectional and the dynamic panel regressions discussed up to now assume a homogeneous relationship between finance and growth across countries, that is, $\beta_i = \beta$. At the other extreme, the time series approach, discussed in the next section, assumes complete country heterogeneity, but relies on a sufficiently large time series of data. When both cross-country and time series dimension are sufficiently large, Pesaran, Smith and Im (1995) show that a consistent mean coefficient across countries is the unweighted average of the coefficients from independent country regressions (mean group (MG) estimator). The pooled mean group (PMG) estimator, introduced by Pesaran, Shin and Smith (1999), is in between

these two extremes of cross-country and time series approaches, as it imposes the same coefficient across countries on the long-run coefficients, but allows the short-run coefficients and intercepts to be country-specific. Loayza and Ranciere (2006) use the PMG estimator on a sample of 75 countries and annual data over the period 1960–2000 and find a positive long-run relationship between financial development and growth, while the mean short-run coefficient on current financial development enters negatively.[18] Using the Hausman test that compares the MG with the PMG model, they cannot reject the hypothesis that the long-run coefficients on finance are the same in a cross-country panel growth regression. This is also evidence that the assumption that $\beta_i = \beta$ in the cross-country estimations discussed so far is a valid one, as long as the focus is on the long-term relationship between financial development and economic growth.

25.4 The time series approach

The use of higher-frequency data, often limited to one or a few countries, and the concept of causality are the main differences between the time series approach and the cross-country approach discussed in the previous section. First, the time series approach relies on higher-frequency data, mostly yearly, to gain econometric power, while the cross-country approach typically utilizes multi-year averages.[19] Further, the time series approach relaxes the somewhat restrictive assumption of the finance–growth relationship being the same across countries, that is, $\beta_i = \beta$, and allows country heterogeneity of the finance–growth relationship; most studies therefore focus their analysis on a few countries with long time series data. The time series approach also directly addresses biases introduced by the persistence and potential unit root behavior of financial development, as we will see in the following.

Second, and more importantly, different causality concepts underlie the two approaches. The time series approach relies on the concept of Granger causality, as first developed by Granger (1969). A time series X is said to Granger-cause Y if, controlling for lagged Y values, lagged X values provide statistically significant information about the current value of Y. Granger causality tests are tests of forecast capacity; that is, to what extent does one series contain information about the other series? Unlike the cross-country panel regressions discussed earlier, this concept therefore does not control for omitted variable bias by directly including other variables or by controlling with instrumental variables. Rather, by including a rich lag structure, which is lacking in the cross-sectional approach, the time series approach hopes to capture omitted variables. The cross-country approach, on the other hand, estimates the empirical relationship between finance and growth controlling for the different biases discussed in section 25.2, including the omitted variable bias, by extracting an exogenous component of finance that is related to growth only through finance.

In the context of the finance and growth literature, finance is said to Granger-cause GDP per capita if the inclusion of past values of finance in a regression of GDP per capita on its lags and the conditioning information set reduces the mean

squared error *mse*. Formally:

$$mse[y(t+s)/y(t), y(t-1), \ldots] > mse[y(t+s)/y(t), y(t-1), \ldots, f(t), f(t-1), \ldots], \tag{25.19}$$

where the null hypothesis of no Granger causality is typically tested using F-tests on current and lagged values of f. Most studies test for bidirectional Granger causality using the following vector autoregression (VAR) system:

$$Y(t) = \alpha_1 Y(t-1) + \alpha_2 Y(t-2) + \cdots + \alpha_j Y(t-j) + \mu(t), \tag{25.20}$$

where Y is a vector comprising both GDP per capita and finance, as well as possibly other macroeconomic variables, and μ is a vector of error terms. Jung (1986) finds evidence for Granger causality from finance to GDP per capita for a sample of 56 countries, with some evidence of reverse Granger causality in the case of developed countries.

Testing for Granger causality between finance and GDP per capita using a levels VAR has the shortcoming that both finance and GDP per capita are non-stationary variables in most countries, as shown by standard tests for unit roots, such as the augmented Dickey–Fuller (ADF) and Phillips and Perron (PP) tests, but stationary in first differences. However, only if two (or more) non-stationary series are co-integrated, that is, if some linear combination of the series is stationary, can one use a levels VAR to test for Granger causality (Toda and Phillips, 1993, 1994). Co-integration thus implies a long-run equilibrium relationship between finance and GDP per capita. As in the case of Granger causality, cointegration does not directly control for omitted variable or measurement biases, but rather exploits the long time series of data to assess whether there is a stable relationship between these two variables.

If the vector Y is cointegrated, regression (25.20) can be rewritten in the vector error correction (VEC) form (Engle and Granger, 1987):

$$\Delta Y(t) = \alpha_1 \Delta Y(t-1) + \alpha_2 \Delta Y(t-2) + \cdots + \gamma \delta' Y(t-1) + \mu(t), \tag{25.21}$$

where the vector γ of error correction coefficients (loading factors) indicates the direction and speed of adjustment of the respective dependent variable to temporary deviations from the long-run relationship, while δ is the cointegrating vector. If there exists a non-zero cointegrating vector such that $\delta' Y(t)$ is stationary, the variables in Y are considered cointegrated. Testing for cointegration of the vector $Y(t)$ therefore is equivalent to a test that $\delta' Y(t)$ is stationary. If we can reject the null hypothesis that $\delta' Y(t)$ is stationary, we can also reject the null hypothesis that $Y(t)$ is cointegrated. In the case of two variables, this implies testing the residuals from a regression of $y(1,t)$ on $y(2,t)$ or $y(2,t)$ on $y(1,t)$ for stationarity. While the standard ADF test can be applied, the critical values are not the same as the test is performed on estimated residuals (Engle and Yoo, 1987). If there is no unit root, the two variables are cointegrated. In the case of more than two variables, inferences on the number and coefficients of the cointegrating vectors can be based

on Johansen's (1991) full information maximum likelihood approach. Johansen (1988) and Johansen and Juselius (1990) show that the maximum likelihood estimator of γ and δ can be derived as a solution of a generalized eigenvalue problem, and likelihood ratio tests, based on these eigenvalues, can be used to test hypotheses on the number of cointegrating vectors.[20] The number of linear independent cointegrating vectors is equal to the rank of the matrix δ. Alternatively, one can test the hypothesis of a specific known cointegrating vector (Horvath and Watson, 1995), as done by Neusser and Kugler (1998).

Demetriades and Hussein (1996) and Luintel and Khan (1999) use the VEC specification and test for weak exogeneity of finance to GDP per capita by testing the null hypothesis that the corresponding loading factor in the GDP per capita regression in (25.21) is zero, while they follow Toda and Phillips' (1993) suggestion and use the product of the loading factor and the cointegrating parameter to test for long-run causality. While Demetriades and Hussein (1996) find evidence for bidirectional causality and reverse causation from income to finance across a sample of 16 developing countries with at least 27 annual observations, with results varying substantially from country to country, Luintel and Khan (1999) find consistent evidence for bidirectional causality across a sample of ten developing countries with at least 36 years of data.

In the case of a cointegrating relationship between finance and GDP per capita, however, a levels VAR as in (25.20) can be used to test for short-term Granger causality, with conventional F-test statistics applying (Toda and Phillips, 1993, 1994; Sims, Stock and Watson, 1990),[21] and the VEC representation in (25.21) to estimate the adjustment speed γ. Rousseau and Wachtel (1998) use both the VAR specification of (25.20) and the VEC specification of (25.21) to determine the direction of causality between economic and financial development for five industrialized countries for the period 1870–1929. Specifically, using the VEC specification of (25.21), they find a cointegrating relationship for all five countries, while Granger causality tests suggest that finance leads GDP per capita in all five countries.[22] In addition, Neusser and Kugler (1998) apply the Granger and Lin (1995) test to measure the strength of causality from finance to GDP per capita at frequency zero, that is, in the long term, which is a function of the correlation of the errors in a bivariate VEC model and the adjustment coefficient vector γ.

In order to gain degrees of freedom, as unit root and cointegration tests have low power in the case of short time series, several studies have expanded the time series approach to panel data (Neusser and Kugler, 1998; Christopoulos and Tsionas, 2004). Averaging individual Dickey–Fuller unit root tests yields the Im, Pesaran and Shin (2003) test, while combining p-values from individual ADF tests yields the Maddala and Wu (1999) test, both of which allow testing for a unit root in panels. To establish cointegration relationships in a panel, Pedroni (1997) suggests estimating the cointegrating regression by OLS separately for each country before a unit root test similar to the PP test is applied to the stacked residuals. Further, the VEC specification (25.21) can be extended to a panel with country-specific fixed effects to test for both long- and short-run relationships between finance and GDP per capita. Christopoulos and Tsionas (2004) find evidence for cointegration and

long-run Granger causality from finance to GDP per capita for a sample of ten developing countries for the period 1970–2000, both for individual countries and for the panel. Unlike other studies in the time series tradition, they also confirm their findings by applying dynamic panel regression techniques using lagged values as instruments in the panel version of (25.21).

Using Geweke's (1982) measure of linear dependence, Calderon and Liu (2003) compute the relative strength of Granger causality from finance to GDP per capita, from GDP per capita to finance and the instantaneous feedback between finance and GDP per capita. Specifically, using variance-covariance matrices calculated under different restrictions on the system (25.20) allows calculating a measure of the overall strength of the relationship between the two variables and the three different sources. They find a stronger effect from finance to GDP per capita than for the reverse effect for developing countries, which increases when they average data over longer time periods. While they consider the linear decomposition in the context of panel regressions, with data averaged over five-year periods, they do not assess the finance–GDP per capita relationship at different frequencies.

25.5 Differences-in-differences estimations

While the cross-country IV approach focuses on identifying instruments to overcome the different biases found in an OLS regression, and the time series approach focuses on the forecast capacity of finance in a VAR including GDP per capita, the differences-in-differences technique can be understood as a "smoking gun" or controlled treatment approach. Specifically, traditional differences-in-differences estimation consists of comparing the difference between the treatment and the control groups before and after a treatment, such as a policy change, thus controlling for other confounding influences on growth.[23]

The seminal paper in this literature is Jayartne and Strahan (1996), who exploit the fact that states across the US deregulated intrastate branch restrictions at different times over the period 1970–1995 and relate this policy change to subsequent state-level growth. In this case the treatment and control groups are in flux; at any point in time, the treatment group consists of states that have deregulated, while the control group consists of those states that have not deregulated yet. By controlling for state- and year-specific effects, this approach effectively measures the impact of deregulation on state-level growth relative to the average state-level growth rate over the sample period and relative to the average growth rate in the US in this specific year. The specification is:

$$g(i,t) = \alpha(i) + \lambda(t) + \beta d(i,t) + C(i,t)\gamma + \delta y(i,t-1) + \varepsilon(i,t), \quad i = 1,\ldots,49; \\ t = 1976,\ldots,1994, \tag{25.22}$$

where $\alpha(i)$ is a vector of state dummies, $\lambda(t)$ a vector of year dummies, $C(i,t)$ a vector of time-varying state characteristics and d the treatment variable, which is branch deregulation in the case of Jayaratne and Strahan (1996), who found a positive and significant coefficient β, thus suggesting that branch deregulation led

to higher growth.[24] They also find evidence for a large economic effect of branch deregulation, explaining an annual growth difference of at least 0.5 percentage points, compared to an average annual growth rate across states of 1.6%. Consistent with cross-country results, they also find evidence that the finance–growth nexus worked through improved lending efficiency rather than more lending and investment.

The differences-in-differences estimator reduces, but does not eliminate, the biases of reverse causation and omitted variables. Specifically, any omitted variable has to be time-variant in order to bias the results, because otherwise it would be picked up by the state dummies. Further, by considering sub-national variation, differences-in-differences estimation is less subject to biases introduced by unobserved heterogeneity across countries and measurement error is reduced as the focus is on one specific policy measure, implemented in the same way but at different times across sub-national units.[25] On the other hand, the events in different states, such as branch deregulation, were not independent from each other, but rather came in waves, which might bias the estimate of β (Huang, 2008). Further, the concern of reverse causation can only be addressed by utilizing instrumental variables or by showing that the decision to implement the policy change across states is not correlated with future growth rates, as was done by Jayaratne and Strahan (1996).

Apart from the problem of endogeneity, serial correlation of the error terms in differences-in-differences estimations can lead to underestimation of standard errors, as shown by Bertrand, Duflo and Mullainathan (2004).[26] This problem increases with the number of time periods and the persistence of the dependent variable and is exacerbated by the fact that the treatment variable, for example, branch deregulation, shows little change across states, at most one change from zero to one. Using Monte Carlo simulation, Bertrand, Duflo and Mullainathan show that collapsing data to before and after-treatment[27] or allowing for correlation within states (clustering) are solutions that resolve the problem of underestimated standard errors.

Going even more local, Huang (2008) uses county-level data from contiguous counties only separated by a state border in cases where one state deregulated at least three years earlier than the other. This helps reduce concerns of omitted variables, as one can assume a very similar structure of two contiguous counties and also helps reduce concerns of reverse causation, as expected higher future growth of a specific county is unlikely to affect state-level political decisions.[28, 29]

A somewhat related differences-in-differences approach is suggested by Rajan and Zingales (1998), who conjecture that the effect of financial development should vary by sector or industry according to the financing need of each sector or industry. They thus assess the finance and growth link by focusing on a specific channel through which financial development should foster economic development, that is, the channeling of society's savings to industries with the highest demand for external finance. Specifically, they use variation across industries in their dependence on external finance and variation across countries in their level of financial development to assess the impact of finance on industry growth, and apply the

following specification:[30]

$$g(i,k) = \alpha(i) + \lambda(k) + \beta(External(k)^{*}f(i)) + \gamma Share(i,k) + (Industry(k)^{*}Country(i))\delta + \varepsilon(i,k), \quad (25.23)$$

where g is growth of value added in industry k in country i; α and λ are vectors of country and industry dummies; *Share* is the initial share of industry k's value added in total manufacturing value added of country i; *External* is the external dependence of industry k; f is a measure of financial development in country i; *Industry* is a vector of other industry characteristics that do not vary across countries; and *Country* is a vector of other country characteristics that do not vary across industries. By including industry and country specific effects, the coefficient β measures the differential growth impact of financial development on high-dependence industries relative to low-dependence industries. When redefining this exercise in terms of a controlled experiment, we could see industries (rather than states) as the treated objects, some of which (high external dependence) are subjected to the treatment of financial development. In a sample of 41 countries and 36 manufacturing industries, Rajan and Zingales (1998) find robust evidence for a significant and positive β, which is even stronger when focusing on young firms in the computation of external dependence. To gauge the economic significance, Rajan and Zingales assess the growth difference between the industries at the 75th and 25th percentiles of external dependence in the countries at the 75th and 25th percentiles of their financial development indicator. Their results suggest that the annual growth difference between Machinery (75th percentile of external dependence) and Beverages (25th percentile of external dependence) is 1.3 percentage points higher in Italy (75th percentile financial development) than in Philippines (25th percentile financial development). This compares to an average industry growth rate of 3.4%, and thus is a relatively large effect.

As in the case of Jayaratne and Strahan (1996), regression (25.23) does not control for biases due to omitted variables or reverse causation. Rajan and Zingales (1998) address concerns about the endogeneity of the treatment, that is, of financial development, by focusing on the smallest 50% of industries in terms of initial value added in each country, as it is less likely that the financial sector develops in response to the smallest industries. They address the omitted variable bias by including other interaction terms between industry and country characteristics that can explain cross-country, cross-industry growth variation and utilizing instrumental variables for financial development.[31] Critically, the differences-in-differences estimator depends on the assumption that there are industry-inherent characteristics that do not vary across countries and that they are properly measured by the data in the US (von Furstenberg and von Kalckreuth, 2006, 2007).

25.6 Firm- and household-level approaches

While the three approaches discussed so far, cross-country instrumental variable regressions, VAR models and differences-in-differences estimation, have tried to

address the different biases resulting from the standard OLS cross-country growth regression, a fourth approach has used disaggregated firm and, more recently, household-level data to assess the impact of access to financial services on firm growth and household welfare. The advantage of using micro-level data is that it allows more clearly the disentangling and testing of the mechanisms and channels through which financial development enhances economic growth. A disadvantage is that it focuses on the direct effect of finance on firm growth and household welfare but commonly does not consider spillover effects on other firms and households and therefore does not allow for individual effects to be added up to an aggregate growth effect.[32]

Further, as in the case of cross-country regressions, biases due to omitted variables, measurement error and reverse causation have to be addressed. This section discusses several studies using micro-data that assess whether easier access to finance is associated with faster firm growth and higher household welfare. Unlike the previous section, this section does not introduce new methodologies, but rather discusses methodological challenges stemming from the use of micro-, as opposed to country-level data.

25.6.1 Firm-level approaches

The different approaches discussed in this section consist of relating firm-level growth or investment to country-level financial development measures. As in the case of cross-country regressions, however, this implies controlling for biases stemming from reverse causation and omitted variables. A first approach, suggested by Demirgüç-Kunt and Maksimovic (1998), compares firm growth to an exogenously given benchmark. Specifically, they calculate for each firm in an economy the rate at which it can grow, using (i) only its internal funds or (ii) its internal funds and short-term borrowing, based on the standard "percentage of sales" financial planning model (Higgins, 1977). Given a set of simplifying assumptions, the external financing needs *EFN* at time t of a firm growing at rate $g(t)$ is given by:[33]

$$EFN(t) = g(t)^{*}Assets(t) - [1 - g(t)]^{*}Earnings(t)^{*}b(t), \tag{25.24}$$

where $b(t)$ is the fraction of the firm's earnings that are retained for reinvestment at time t. Assuming that the firm retains all its earnings, that is, $b(t) = 1$, the internally financed growth rate $IG(t)$ is the maximum growth rate that can be financed with internal resources only, that is:

$$IG(t) = ROA(t)/[1 - ROA(t)]. \tag{25.25}$$

Demirgüç-Kunt and Maksimovic (1998) then regress the percentage of firms in a country that grow at rates exceeding $IG(t)$ on financial development, other country characteristics and averaged firm characteristics in a simple OLS set-up and show, for a sample of 8,500 firms across 30 countries, that the proportion of firms growing beyond the rate allowed by internal resources is higher in countries with better developed banking systems and more liquid stock markets.[34]

An alternative approach to assess the impact of access to finance on firm growth is the use of firm-level survey data, as done by Beck, Demirgüç-Kunt and Maksimovic (2005), who use firm-level survey data for over 4,000 firms in 54 countries to run the following regression:

$$g(i,k) = \alpha + \beta_1 o(i,k) + \beta_2 f(i) + \beta_3 o(i,k)^* f(i) + C^{(1)}(i,k)\gamma_1 + C^{(2)}(i)\gamma_2 + \varepsilon(i,k), \tag{25.26}$$

where g is sales growth of firm k in country i over the period 1996–99, $C^{(1)}$ is a set of firm-level control variables, $C^{(2)}$ is a set of country-level control variables, o is the financing obstacle as reported by the firm and f is a country-level financial development indicator. The financing obstacle is the response by the firm to the question of whether financing is an obstacle to its operation and growth, and responses are coded as no obstacle (1) minor obstacle (2), moderate obstacle (3), and major obstacle (4). While β_1 indicates the relationship between the reported financing obstacle and firm growth, β_3 indicates whether this relationship varies across countries with different levels of financial development. Beck, Demirgüç-Kunt and Maksimovic find a negative and significant coefficient on β_1 and a positive and significant coefficient on β_3, suggesting that firms reporting higher financing obstacles experience slower sales growth, but that this relationship is less strong in countries with better developed financial systems. Further, using triple interaction terms, they show that the mitigating effect of financial development on the relationship between financing obstacles and firm growth is stronger for small firms than for large firms.

Another methodology consists of assessing the relationship between country-level financial development and firms' financing constraints derived from a structural investment model, such as the Euler equation (Love, 2003; Laeven, 2003). Specifically, the Euler equation derives the optimal investment decision as the point where the marginal cost of today's investment is equal to the discounted marginal cost of postponing investment until the next period, which includes the marginal product of capital, the adjustment cost and the price of investment tomorrow. In the absence of credit market constraints, firms' investment decisions should thus be independent of firms' cashflow holdings, while the investment decisions of credit constrained firms should be a positive function of available cash. Financial sector development, on the other hand, should reduce the dependence of firms' investment on cash holdings. To test for the presence of credit market constraints and the impact of financial development on the relationship between credit market constraints and investment, the following regression is used:

$$\begin{aligned} I(k,t) = {} & \alpha(k) + \lambda(t) + \beta_1 Cash(k,t-1) + \beta_2 Cash(k,t-1)^* f(i,t) \\ & + C^{(1)}(k,t)\gamma_1 + C^{(2)}(k,t-1)\gamma_2 + \varepsilon(i,k,t), \end{aligned} \tag{25.27}$$

where I is investment, $\alpha(k)$ is a vector of firm dummies, $\lambda(t)$ a vector of time dummies, *Cash* is liquid assets relative to total assets, $C^{(1)}$ and $C^{(2)}$ are sets of current and lagged firm-level control variables, such as investment-to-capital ratios

and sales-to-capital ratios, and the i refers to countries. The existence of credit constraints implies $\beta_1 > 0$, while the alleviating role of financial sector development implies $\beta_2 < 0$. As regression (25.27) poses similar problems in terms of the different biases identified in section 25.2 for cross-country growth regressions, most studies use the dynamic panel techniques suggested by Arrellano and Bond (1991) and Arrellano and Bover (1995) to control for these biases. Using data for 5,000 firms across 36 countries, Love (2003) shows that financial development reduces firms' dependence on cash holdings for investment, while Laeven (2003) shows, for a sample of 400 firms across 13 countries, that financial liberalization helped reduce small firms' financing dependence on internal cash, while it adversely affected large firms' financing possibilities. The effect of financial development and liberalization is also economically significant. Love (2003) shows that firms' financing constraints – as measured by the cost of capital – in countries with low levels of financial development are twice as high as in countries with average levels of financial development, while Laeven (2003) shows that financial liberalization had a significant economic effect on firms' financing constraints, reducing small firms' constraints by 80%.

25.6.2 Household-level approaches

While the availability of financial information for listed companies and survey data for non-listed companies has resulted in a rapid expansion of firm-level studies, the lack of comparable data for households has impeded similar research for the effect of access to finance on household welfare until recently. As in the case of aggregate and firm-level studies, the identification problem prevents inference from cross-sectional household surveys with data on welfare and access to finance variables. A final and very recent technique therefore uses controlled experiments with households and/or micro-entrepreneurs, whose financing constraints are randomly alleviated and who are then compared to a control group whose constraints were not alleviated. The challenges of these studies are less in estimation techniques than in the proper identification of treatment and control groups and of the experimental treatment itself. In the following, we will discuss three examples.

First, Pitt and Khandker (1998) use household survey data to assess the impact of micro-credit on household welfare across several programs in Bangladesh. However, as in the case of cross-country regressions, omitted variable bias and reverse causation would bias the result of simple OLS estimation, as illustrated by the following system:

$$y(i,j) = C(i,j)\alpha_1 + \beta f(i,j) + \eta(i) + \varepsilon(i,j) \tag{25.28}$$

$$f(i,j) = C(i,j)\alpha_2 + Z(i,j)\delta + \mu(i) + \nu(i,j), \tag{25.29}$$

where y is a measure of household welfare of household i in village j, f is the amount of credit obtained by a household, C is a vector of household characteristics, and Z is a set of household or village characteristics that serve as instruments for the endogenous credit variable. μ and η are unobservable village characteristics, that are correlated with household welfare and credit, respectively. Correlations

between μ and η and between ε and ν can result in a biased OLS estimate of β in (25.28). These correlations can arise because micro-credit program placement is non-random, often related with specific village characteristics, such as poverty levels. Further, unmeasured household and village characteristics can influence both the demand for micro-credit and household outcomes y. Pitt and Khandker (1998) therefore use the exogenously imposed restriction that only farmers with less than a half-acre of land are eligible to borrow from micro-finance institutions in Bangladesh as an exclusion condition to compare eligible and non-eligible farmers in program and non-program villages. Using survey data for 1,800 households and treating landownership as exogenous to welfare outcomes, they exploit the discontinuity in access to credit for households above and below the threshold and find a positive and significant effect of credit on household consumption expenditures. Morduch (1998), however, shows that mistargeting, that is, allowing farmers with landholdings above the threshold to access micro-credit, violates the exclusion condition, and that different econometric techniques exploiting the landholding restriction lead to different findings.

Coleman (1999) exploits the fact that future micro-credit borrowers are identified before the roll-out of the program in Northern Thailand and can thus exploit the differences between current and future borrowers and non-borrowers in both treated and to-be-treated villages.[35] His model is:

$$y(i,j) = C^{(1)}(i,j)\alpha + \beta p(i,j) + C^{(2)}(j)\gamma + \delta M(i,j) + \varepsilon(i,j) \tag{25.30}$$

where y is an array of measures of household welfare, $C^{(1)}$ is a set of observable household and $C^{(2)}$ a set of observable village characteristics, M is dummy that takes the value one for current and future borrowers and p is a dummy that takes the value one for villages that already have access to credit programs. M can be thought of as a proxy for unobservable household characteristics that determine whether a household decides to access credit or not, whereas β measures the impact of the credit program by comparing current and prospective borrowers. Coleman (1999) does not find any robustly significant estimate of β and therefore rejects the hypothesis that micro-credit helps households in this sample and this institutional setting.

A final example is Karlan and Zinman (2006), who use a sample of marginally rejected applicants of a South African consumer credit institution. They convinced the credit institution to provide loans to a randomly chosen sub-set of these borrowers. Surveying both treatment and control groups six and twelve months after providing credit to the treatment group, they find that borrowers were more likely to retain wage employment and less likely to experience hunger in their household and be impoverished:

$$y(i) = C(i)\alpha + \beta p(i) + \varepsilon(i), \tag{25.31}$$

where y is an indicator of household welfare, C is a vector of household characteristics and p is the treatment dummy that takes the value 1 if the individual surveyed has received a loan.

While controlled experiments can assess the effect of access to credit (or other financial services) on the growth of micro-enterprises or household welfare, there are shortcomings to this methodology. First, they are very costly to conduct. Second, they are environment-specific and it is not clear whether the results will hold in a different environment with a different sample population. Third, the controlled experiments, as they have been undertaken up to now, do not consider any spillover effects of access to credit by the treated individuals or enterprises to other individuals or enterprises in the economy.

25.7 Concluding remarks

The finance and growth literature has come a long way from simple correlation and OLS regressions to dynamic panel regressions and the use of firm- and household-level data. While each of the different methodologies and aggregation levels has its shortcomings, the body of evidence accumulated over the past 15 years provides a strong case for a relationship between financial development and economic growth that is not driven by omitted variables, measurement error or reverse causation.

While the profession has made great progress in measuring financial development, especially by moving towards micro-data, this chapter has focused on methodological advances to overcome the biases illustrated by a simple cross-country OLS regression. Most importantly, overcoming endogeneity and simultaneity biases with a proper identification strategy has been the main challenge for researchers. While the cross-country literature has focused on finding external and internal instruments, the time-series literature has exploited high-frequency data, a rich lag structure, and the forecast capacity of finance for GDP per capita. Differences-in-differences approaches address the identification challenge by assessing natural experiments, exploiting either exogenous policy reforms or inherent industry characteristics that result in a differential impact of financial development.

Using firm- and household-level data allows a deeper look into the mechanisms through which finance enhances firm growth and household welfare and thus provides additional evidence, but poses its own set of identification challenges. While many of the methodologies used at the cross-country-level, such as instrumental variables or differences-in-differences, can also be applied at the firm and household level, randomized controlled experiments with households and micro-entrepreneurs open new and exciting research opportunities, as they allow researchers to subject households and micro-enterprises to a specific treatment under the control of the researcher.

Different methodologies imply different aggregation levels. While assessing the finance and growth relationship on a more disaggregated level might allow better controlling for different biases – such as measurement error when considering a specific policy change on the sub-national level or simultaneity bias when using household data in a controlled randomized experiment – this has to be balanced with the limited extent to which we can draw policy conclusions from such a specification. Further, using firm- or household-level data does not properly control

for spillover effects, are often very costly exercises, and do not lend themselves easily to compute the aggregate growth effect of financial development. While randomized experiments have the advantage that they are the cleanest exercise possible, as they are controlled by researchers, they might not properly mimic the real world, and might not allow inferences outside the geographic and institutional experiment area.

While a wide array of cross-country techniques has been applied to the finance and growth field, some techniques have not been used yet, such as identification through heterogeneity in structural shocks (Rigobon, 2003). Further, it is easy to predict that there will be further advances in GMM techniques that control better for country heterogeneity and in techniques to assess the finance and growth relationship at different frequencies. As before, the finance and growth literature will benefit in the years to come from methodological advances in neighboring fields, especially in growth econometrics. Merging VAR and cross-country techniques – two literatures which have moved mostly parallel to each other up to now – also promises further methodological insights.

More important than these advances at the aggregate level, however, will be advances at the micro-level, and specifically on two fronts. First, randomized experiments involving both households and micro- and small enterprises will shed light on the effect of access to finance on household welfare and firm growth. One of the challenges to overcome will be to include spillover effects and thus move beyond partial equilibrium results to aggregate results. Second, further studies evaluating the effect of specific policy interventions can give insights into which policy reforms are most effective in enhancing financial development and positive real sector outcome.[36] Advances in both areas, however, will depend on the collection of micro-based data on access to and use of financial services.

Acknowledgments

I am grateful to George Clarke, Aart Kraay, Luc Laeven, Ross Levine, Kerry Patterson and Peter Rousseau for comments and useful discussions.

Notes

1. See Levine (1997, 2005) for surveys of the theoretical literature.
2. For a broader survey on the econometrics of growth regressions, see Durlauf, Johnson and Temple (2005).
3. See Beck, Demirgüç-Kunt and Levine (2000) for an overview of different cross-country indicators of financial development and Beck *et al.* (2008) for a discussion of the different dimensions of financial development, such as depth, efficiency and reach. See World Bank (2007) for a discussion of financial outreach indicators.
4. Other early finance and growth studies using cross-sectional OLS regressions include Atje and Jovanovic (1993) and De Gregorio and Guidotti (1995).
5. Most of the papers using this approach assume that only financial development is an endogenous variable and thus treat all control variables as exogenous.
6. The literature has developed several tests to resolve the issue of OLS versus IV, including the Hausman test.

7. The presence of heteroskedasticty can be examined with a test proposed by Pagan and Hall (1983).
8. See Beck and Levine (2005) for an overview.
9. An alternative test was developed by Basmann (1960) and does not impose the overidentifying restrictions.
10. In the case of several endogenous variables, the Stock and Yogo test also requires each instrument to predict primarily just one of the endogenous variables.
11. For further discussion on weak instruments and how to deal with them, see Murray (2006) and Baum, Schaffer and Stillman (2003).
12. Most papers in the literature, however, do not formally test whether the difference between the OLS and the IV estimate is significant, which could be done with a Hausman test.
13. Alternatively, one can use the forward orthogonal deviation transformation.
14. Formal unit root tests as discussed in section 25.4 are not feasible in this context, as there are too few observations.
15. Given that lagged levels are used as instruments in the difference regressions, only the most recent difference is used as an instrument in the level regressions, as using additional differences would result in redundant moment conditions (Arellano and Bover, 1995).
16. Rousseau and Wachtel (2000) was also the first paper to combine dynamic panel techniques with vector autoregression techniques discussed in the next section.
17. Other papers using dynamic panel techniques include Rioja and Valev (2004a, 2004b) and Benhabib and Spiegel (2000). The latter, however, assume exogeneity of financial development and weak exogeneity only for capital accumulation, but not the other control variables.
18. This negative short-run coefficient is consistent with the finding of the banking crisis literature. See, for example, Demirgüç-Kunt and Detragiache (1999).
19. It is important to note, however, that the power of such high-frequency tests depends on the span of the time series rather than the number of observations.
20. Specifically, the "trace" test can be used to test the hypothesis of r against zero cointegrating vectors, while the "λ-max" or maximum eigenvalue test can be used to test the hypothesis of $r + 1$ cointegrating vectors against r cointegrating vectors.
21. Specifically, Toda and Phillips (1993, 1994) and Sims, Stock and Watson (1990) show that in the case of cointegrated series the conventional Wald statistic converges to a χ^2 distribution.
22. Following this approach, Rousseau and Sylla (2005) use data for the US over the period 1850–1997, Bell and Rousseau (2001) use data for India, and Xu (2000) uses data for 43 countries over the period 1960–93; all find robust evidence for a leading role of finance.
23. While we treat such exogenous policy changes in the context of differences-in-differences estimations, one could also use them as instruments for financial development in the context of regular cross-sectional regressions (Guiso, Sapienza and Zingales, 2004).
24. Following the model of Jayartne and Strahan (1996), Dehejia and Lleras-Muney (2007) show that, over the period 1900–40 across states of the US, regulatory changes that allowed branching accelerated the mechanization of agriculture and spurred growth in manufacturing, while the introduction of deposit insurance had negative consequences.
25. On the other hand, focusing on one country reduces the policy relevance of its findings, as the relationship might vary across countries with different economic and institutional settings. Further, sub-national variation might not be independent from each other given the higher mobility of capital and labor within rather than across countries.
26. Bertrand, Duflo and Mullainathan (2004) find over-rejection of the null hypothesis using randomly assigned placebo treatments in Monte Carlo simulation.
27. Specifically, this would imply regressing growth on state and year fixed effects and other time-varying control variables, taking the residuals and averaging them for the period

before and after the treatment for each state. The estimate of the treatment can then be obtained from a regression of this two-period state panel on the treatment dummy.

28. This argument, however, is only valid if there is sufficient variation in growth across different counties within the state.
29. Given the lack of randomness of the sample relative to the population, Huang (2008) constructs critical values from a distribution of the effects of fictitious placebo treatments on county pairs on non-event borders, taking into account spatial correlation across counties along the same borders. Only if 95% of all placebo treatments result in a growth difference below a certain value can this value be considered a significant growth difference for a real world treatment at the 5% significance level.
30. Rajan and Zingales (1998) compute the industry-level dependence on external finance from data of listed firms in the US, that is, firms that should have the least problems in raising external finance and thus face a perfectly elastic supply curve, to get measures of industry-level demand for external finance. They conjecture that demand for external finance measured in this way proxies for the industry-inherent demand for external finance, rather than country- or firm-specific characteristics, in the US.
31. The differences-in-differences approach of Rajan and Zingales (1998) has subsequently been used by many other researchers interested in the linkage between financial development and growth and specific mechanisms and channels, including Beck and Levine (2002), Beck (2003), Beck *et al*. (2008), Braun and Larrain (2005), Claessens and Laeven (2003), Fisman and Love (2003), and Raddatz (2006).
32. Indirect effects of financial development can be very important, as shown by Beck, Levine and Levkov (2007), who find that the main channel through which branch deregulation across US states led to lower income inequality was through labor market effects rather than through providing increased access to finance.
33. The three simplifying assumptions are as follows. First, the ratio of assets used in production to sales is constant. Second, the firm's profits per unit of sales are constant. Finally, the economic deprecation rate equals the accounting depreciation rate.
34. Subsequently, this technique has been applied by Demirgüç-Kunt and Maksimovic (2002) and Guiso, Sapienza and Zingales (2004), among others.
35. This technique is also referred to as "pipeline matching" (Goldberg and Karlan, 2005).
36. One example assessing the effect of different legal reforms is Haselmann, Pistor and Vig (2005).

References

Alonso-Borrego, C. and M. Arellano (1999) Symmetrically normalised instrumental-variable estimation using panel data. *Journal of Business & Economic Statistics* **17**(1), 36–49.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**(2), 277–97.

Arellano, M. and O. Bover (1995) Another look at the instrumental-variable estimation of error-components models. *Journal of Econometrics* **68**(1), 29–52.

Atje, R. and B. Jovanovic (1993) Stock markets and development. *European Economic Review* **37**(2-3), 632–40.

Basmann, R.L. (1960) On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* **55**(292), 650–9.

Baum, C., M. Schaffer and S. Stillman (2003) Instrumental variables and GMM: estimation and testing. *Stata Journal* **3**(1), 1–31.

Beck, T. (2003) Financial dependence and international trade. *Review of International Economics* **11**, 296–311.

Beck, T., A. Demirgüç-Kunt, L. Laeven and R. Levine (2008) Finance, firm size, and growth. *Journal of Money, Credit, and Banking* **40**, 1379–405.
Beck, T., A. Demirgüç-Kunt and R. Levine (2000) A new database on financial development and structure. *World Bank Economic Review* **14**, 597–605.
Beck, T., A. Demirgüç-Kunt and V. Maksimovic (2005) Financial and legal constraints to firm growth: does firm size matter? *Journal of Finance* **60**, 137–77.
Beck, T., E. Feijen, A. Ize and F. Moizeszowicz (2008) Measuring financial development. Mimeo, World Bank.
Beck, T. and R. Levine (2002) Industry growth and capital allocation: does having a market- or bank-based system matter? *Journal of Financial Economics* **57**, 107–31.
Beck, T. and R. Levine (2004) Stock markets, banks and growth: panel evidence. *Journal of Banking and Finance* **28**, 423–42.
Beck, T. and R. Levine (2005) Legal institutions and financial development. In C. Menard and M. Shirley (eds.), *Handbook of New Institutional Economics*. Dordrecht: Kluwer.
Beck, T., R. Levine and A. Levkov (2007) Big bad banks? The impact of U.S. branch deregulation on income distribution. World Bank Policy Research Working Paper No. 3340.
Beck, T., R. Levine and N. Loayza (2000) Finance and the sources of growth. *Journal of Financial Economics* **58**, 261–300.
Bekaert, G., C. Harvey and C. Lundblad (2005) Does financial liberalization spur economic growth? *Journal of Financial Economics* **77**, 3–55.
Bell, C. and P.L. Rousseau (2001) Post-independence India: a case of finance-led industrialization? *Journal of Development Economics* **65**, 153–75.
Benhabib, J. and M. Spiegel (2000) The role of financial development in growth and investment. *Journal of Economic Growth* **5**, 341–60.
Bertrand, M., E. Duflo and S. Mullainathan (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 249–75.
Blundell, R. and S. Bond (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115–43.
Bond, S. and E. Windmeijer (2002) Finite sample inference for GMM estimators in linear panel data models. Cemmap Working Paper 04/02, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London.
Bowsher, C. (2002) On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* **77**, 211–20.
Braun, M. and B. Larrain (2005) Finance and the business cycle: international, inter-industry evidence. *Journal of Finance* **60**, 1097–128.
Calderon, C. and L. Liu (2003) The direction of causality between financial development and economic growth. *Journal of Development Economics* **72**, 321–34.
Christopoulos, D. and E. Tsionas (2004) Financial development and economic growth: evidence from panel unit root and cointegration tests. *Journal of Development Economics* **73**, 55–74.
Claessens, S. and L. Laeven (2003) Financial development, property rights and growth. *Journal of Finance* **58**, 2401–36.
Coleman, B. (1999) The impact of group lending in northeast Thailand. *Journal of Development Economics* **60**, 105–42.
De Gregorio, J. and P. Guidotti (1995) Financial development and economic growth. *World Development* **23**, 433–48.
Dehejia, R. and A. Lleras-Muney (2007) Financial development and pathways of growth: state branching and deposit insurance laws in the United States, 1900–1940. *Journal of Law and Economics* **50**, 239–72.
Demetriades, P.O. and K.A. Hussein (1996) Does financial development cause economic growth? Time-series evidence from 16 countries. *Journal of Development Economics* **51**, 387–441.

Demirgüç-Kunt, A. and E. Detragiache (1999) Financial liberalization and financial fragility. In B. Pleskovic and J. Stiglitz (eds.), *Proceedings of the World Bank Annual Conference on Development Economics*. Washington, DC: World Bank.
Demirgüç-Kunt, A. and V. Maksimovic (1998) Law, finance and firm growth. *Journal of Finance* **53**, 2107–37.
Demirgüç-Kunt, A. and V. Maksimovic (2002) Funding growth in bank-based and market-based financial systems: evidence from firm-level data. *Journal of Financial Economics* **65**, 337–63.
Durlauf, S.N., P.A. Johnson and J.R.W. Temple (2005) Growth econometrics. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*. Amsterdam: North-Holland.
Eichenbaum, M.S., L.P. Hansen and K.J. Singleton (1988) A time-series analysis of representative agent models of consumption and leisure. *Quarterly Journal of Economics* **103**, 51–78.
Engle, R.W. and C.W.J. Granger (1987) Cointegration and error correction: representation, estimation, and testing. *Econometrica* **55**, 251–76.
Engle, R.W. and S.B. Yoo (1987) Forecasting and testing in cointegrated systems. *Journal of Econometrics* **35**, 143–59.
Fisman, R.J. and I. Love (2003) Trade credit, financial intermediary development, and industry growth. *Journal of Finance* **58**, 353–74.
Geweke, J. (1982) Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association* **77**, 304–24.
Godfrey, L.G. (1999) Instrument relevance in multivariate linear models. *Review of Economics and Statistics* **81**, 550–2.
Goldberg, N. and D. Karlan (2005) The impact of microfinance: a review of methodological issues. Mimeo, Yale University.
Goldsmith, R.W. (1969) *Financial Structure and Development*. New Haven, Conn.: Yale University Press.
Granger, C. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–38.
Granger, C. and J. Lin (1995) Causality in the long run. *Econometric Theory* **11**, 530–6.
Griliches, Z. and J. Hausman (1986) Errors in variables in panel data. *Journal of Econometrics* **31**, 93–118.
Guiso, L., P. Sapienza and L. Zingales (2004) Does local financial development matter? *Quarterly Journal of Economics* **119**(3), 929–69.
Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.
Harrison, P., O. Sussman and J. Zeira (1999) Finance and growth: new evidence. Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System.
Haselmann, R., K. Pistor and V. Vig (2005) How law affects lending. Columbia Law and Economics Working Paper No. 285.
Hayashi, F. (2000) *Econometrics* (first edition). Princeton: Princeton University Press.
Higgins, R.C. (1977) How much growth can a firm afford? *Financial Management* **6**, 316.
Horvath, M. and M. Watson (1995) Testing for cointegration when some of the cointegrating vectors are prespecified. *Econometric Theory* **11**, 894–1014.
Huang, R. (2008) Did branching deregulation accelerate growth? *Journal of Financial Economics*. Forthcoming.
Im, S.K., H.M. Pesaran and Y. Shin (2003) Testing for unit roots in heterogeneous panels. *Journal of Econometrics* **11**, 53–74.
Jayaratne, J. and P.E. Strahan (1996) The finance–growth nexus: evidence from bank branch deregulation. *Quarterly Journal of Economics* **111**, 639–70.
Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231–54.
Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551–80.

Johansen, S. and K. Juselius (1990) Maximum likelihood estimation and inference on cointegration – with applications to the demand for money. *Oxford Bulletin of Economics and Statistics* **52**, 169–210.
Jung, W. (1986) Financial development and economic growth: international evidence. *Economic Development and Cultural Change* **34**, 333–46.
Karlan, D. and J. Zinman (2009) Expanding credit access: using randomized supply decisions to estimate the impacts. *Review of Financial Studies*. Forthcoming.
King, R.G. and R. Levine (1993) Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics* **108**, 717–38.
Kraay, A. and D. Kaufmann (2002) Growth without governance. *Economia* **3**(1) (Fall), 169–78.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R.W. Vishny (1997) Legal determinants of external finance. *Journal of Finance* **52**, 1131–50.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R.W. Vishny (1998) Law and finance. *Journal of Political Economy* **106**, 1113–55.
Laeven, L. (2003) Does financial liberalization reduce financing constraints? *Financial Management* **32**, 5–34.
Levine, R. (1997) Financial development and economic growth: views and agenda. *Journal of Economic Literature* **35**, 688–726.
Levine, R. (1998) The legal environment, banks, and long-run economic growth. *Journal of Money, Credit, and Banking* **30**, 596–620.
Levine, R. (1999) Law, finance, and economic growth. *Journal of Financial Intermediation* **8**, 8–35.
Levine, R. (2005) Finance and growth: theory and evidence. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*. Amsterdam: North-Holland.
Levine, R., N. Loayza and T. Beck (2000) Financial intermediation and economic growth: causes and causality. *Journal of Monetary Economics* **46**, 31–77.
Levine, R. and D. Renelt (1992) Sensitivity analysis of cross-country growth regressions. *American Economic Review* **82**, 942–63.
Levine, R. and S. Zervos (1998) Stock markets, banks, and economic growth. *American Economic Review* **88**, 537–58.
Loayza, N., and R. Ranciere (2006) Financial development, financial fragility, and growth. *Journal of Money, Credit, and Banking* **38**, 1051–76.
Love, I. (2003) Financial development and financing constraints: international evidence from the structural investment model. *Review of Financial Studies* **16**, 765–91.
Luintel, K.B. and M. Khan (1999) A quantitative reassessment of the finance–growth nexus: evidence from a multivariate VAR. *Journal of Development Economics* **60**, 381–405.
Maddala, G.S. and S. Wu (1999) A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* **61**, 631–52.
McCraig, B. and T. Stengos (2005) Financial intermediation and growth: some robustness tests. *Economics Letters* **88**, 306–12.
Morduch, J. (1998) Does microfinance really help the poor? New evidence from flagship programs in Bangladesh. Mimeo, Princeton University.
Murray, M. (2006) Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* **20**, 111–32.
Neusser, K. and M. Kugler (1998) Manufacturing growth and financial development: evidence from OECD countries. *Review of Economics and Statistics* **80**, 638–46.
Pagan, A.R. and D. Hall (1983) Diagnostic tests as residual analysis. *Econometric Reviews* **2**, 159–218.
Pande, R. and C. Udry (2006) Institutions and development: a view from below. Mimeo, Yale University.
Pedroni, P. (1997) Panel cointegration: asymptotic and finite sample properties of the pooled time series tests with an application to the PPP hypothesis: new results. Indiana University.
Pesaran, H., Y. Shin and R. Smith (1999) Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association* **94**, 621–34.

Pesaran, H., R. Smith and K. Im (1995) Dynamic linear models for heterogeneous panels. In L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data*. Dordrecht: Kluwer Academic Publishers.

Pitt, M.M. and S.R. Khandker (1998) The impact of group-based credit programs on poor households in Bangladesh: does the gender of participants matter? *Journal of Political Economy* **106**, 958–96.

Raddatz, C. (2006) Liquidity needs and vulnerability to financial underdevelopment. *Journal of Financial Economics* **80**(3), 677–722.

Rajan, R. and L. Zingales (1998) Financial dependence and growth. *American Economic Review* **88**, 559–87.

Rigobon, R. (2003) Identification through heteroskedasticity. *Review of Economics and Statistics* **85**(4), 777–92.

Rioja, F. and N. Valev (2004a) Does one size fit all? A reexamination of the finance and growth relationship. *Journal of Development Economics* **74**(2), 429–47.

Rioja, F. and N. Valev (2004b) Finance and the sources of growth at various stages of economic development. *Economic Inquiry* **42**(1), 127–40.

Roodman, D. (2007) A short note on the theme of too many instruments. CDGEV Working Paper 125, Center for Global Development, Washington, DC.

Rousseau, P.L. and R. Sylla (2005) Emerging financial markets and early U.S. growth. *Explorations in Economic History* **42**, 1–16.

Rousseau, P.L. and P. Wachtel (1998) Financial intermediation and economic performance: historical evidence from five industrial countries. *Journal of Money, Credit, and Banking* **30**, 657–78.

Rousseau, P.L. and P. Wachtel (2000) Equity markets and growth: cross-country evidence on timing and outcomes, 1980–95. *Journal of Banking and Finance* **24**, 1933–57.

Sargan, J.D. (1958) The estimation of economic relationships with instrumental variables. *Econometrica* **26**, 393–415.

Shea, J. (1997) Instrument relevance in multivariate linear models: a simple measure. *Review of Economics and Statistics* **79**, 348–52.

Sims, C.A., J. Stock and M.W. Watson (1990) Inference in linear time series models with some unit roots. *Econometrica* **58**, 113–44.

Staiger, D. and J.H. Stock (1997) Instrumental variables regressions with weak instruments. *Econometrica* **65**, 557–86.

Stock, J.H. and M. Yogo (2005) Testing for weak instruments in IV regressions. In D.W.K. Andrews and J.H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. New York: Cambridge University Press.

Toda, H. and P. Phillips (1993) Vector autoregression and causality. *Econometrica* **61**, 1367–93.

Toda, H. and P. Phillips (1994) Vector autoregression and causality: a theoretical overview and simulation study. *Econometric Review* **13**, 259–85.

von Furstenberg, G.M. and U. von Kalckreuth (2006) Dependence on external finance: an inherent industry characteristic? *Open Economies Review* **17**, 541–59.

von Furstenberg, G.M. and U. von Kalckreuth (2007) Dependence on external finance: examining the measure and its properties. *Économie Internationale* **111**, 55–80.

World Bank (2007) *Finance for All? Policies and Pitfalls in Expanding Access*. Washington, DC: World Bank.

Xu, Z. (2000) Financial development, investment and economic growth. *Economic Inquiry* **38**, 331–44.

This page intentionally left blank

Part IX
Spatial Econometrics

This page intentionally left blank

26
Spatial Hedonic Models

Luc Anselin and Nancy Lozano-Gracia

Abstract

In this chapter, we focus on some econometric aspects related to a sub-set of hedonic house price models, which we refer to as spatial hedonic models. In these, the locational aspects of the observations are treated explicitly, and the estimation of the models is an application of spatial econometrics. As defined in Anselin (2006), spatial econometrics "consists of a sub-set of econometric methods that is concerned with spatial aspects present in cross-sectional and space-time observations." These methods focus in particular on two forms of so-called spatial effects in econometric models, referred to as spatial dependence and spatial heterogeneity. In this chapter we provide a review of the principles underlying the hedonic house price model, and continue to extensively discuss spatial econometric aspects due to spatial models and spatial data specific to house price applications. We review and discuss the treatment of spatial dependence (including space-time dynamics) and spatial heterogeneity with selective illustrations from the empirical literature.

26.1 Introduction

Hedonic pricing models have become a common tool in applied microeconomics, going back to the classic contributions of Lancaster (1966) and Rosen (1974) (for a recent review, see, e.g., Malpezzi, 2002). The ability of hedonic models to relate the price of a product to the relative contributions of different characteristics has led to a wide range of applied econometric work using these specifications. An important class of applications pertains to house price models, in which characteristics of the property, the neighborhood and other amenities are included in an econometric specification for the sales price or assessed value of a housing unit. Such models are now routinely used in mass appraisal exercises as well as in the valuation of non-market amenities that contribute to the price of the house. The rationale for the latter is that, in an efficient market, superior amenities (such as clean air, access to parks or beaches, and views) should be capitalized into the value of the house. In other words, *ceteris paribus*, houses with superior amenities should be more expensive and the price differential constitutes a measure for the value of the amenity as expressed through market transactions.

In this chapter, we focus on some econometric aspects related to a sub-set of hedonic house price models, which we refer to as *spatial hedonic models*. In these, the locational aspects of the observations are treated explicitly, as an application of *spatial econometrics*. As defined in Anselin (2006), spatial econometrics "consists of a sub-set of econometric methods that is concerned with spatial aspects present in cross-sectional and space-time observations." These methods focus, in particular, on two forms of so-called *spatial effects* in econometric models, referred to as *spatial dependence* and *spatial heterogeneity* (Anselin, 1988).

As outlined in Anselin, spatial dependence or spatial autocorrelation is a special case of cross-sectional dependence in which the structure of the covariation between observations at different locations is subject to a spatial ordering. This ordering is related to the relative positioning, distance or spatial arrangement of the observations in geographic space, or, more generally, in (social) network space. This type of dependence differs from time series dependence in that it is both two-dimensional as well as multidirectional. This implies a simultaneous feedback between observations ("I am my neighbor's neighbor"), which requires the application of specialized techniques that are not simply an extension of time series methods to two dimensions.

Spatial heterogeneity is a special instance of structural instability, which can be observed or unobserved. The spatial aspect of this issue is that spatial structure provides the basis for the specification of the heterogeneity. This may inform models for spatial structural change (referred to as spatial regimes), heteroskedasticity, or spatially varying and random coefficients.

Spatial patterns in the housing market are expected to arise from a combination of spatial heterogeneity and spatial dependence (Anselin, 1998). For example, spatial heterogeneity may originate from spatially differentiated characteristics of demand, supply, institutional barriers, or racial discrimination. This systematic variation in the behavior of economic agents across space warrants special attention, since any model that imposes homogeneity will be misspecified.

Spatial autocorrelation may appear when either the prices or characteristics of houses that are closer together are more similar to each other than those from houses that are farther apart. Alternatively, it may also stem from measurement problems in explanatory variables, omitted variables, and other forms of model misspecification (Baumont, 2004). A major class of such misspecifications pertains to so-called neighborhood effects, which are typically unobserved and modeled as part of the error term. Importantly, spatial heterogeneity and spatial autocorrelation may be observationally equivalent (Anselin, 2001a), which may lead to difficulties in isolating the two effects in practice. Spatial autocorrelation may also result from spatial heterogeneity not being modeled correctly (Anselin and Griffith, 1988; Baumont, 2004).

The consequences of ignoring spatial autocorrelation and spatial heterogeneity when they are, in fact, present in the data-generating process have been widely discussed in the literature and has led to the separate field of spatial econometrics (Anselin, 1988; Anselin and Bera, 1998). A recent comprehensive review of the field can be found in Anselin (2006). Also, after some initial work by Dubin (1988) and Can (1990, 1992), among others, the explicit consideration of spatial effects through the application of spatial econometrics has become more commonplace in empirical studies of housing and real estate markets. Reviews of the basic specifications and estimation methods applied to these spatial hedonic models are provided in Anselin (1998), Basu and Thibodeau (1998), Pace *et al*. (1998), Dubin *et al*. (1999), Gillen *et al*. (2001), and Pace and LeSage (2004), among others.

The literature on hedonic models is vast, both theoretical as well as empirical. We do not attempt to review this in the current chapter, but instead focus on the methodological aspects related to the implementation of spatial hedonic house price models in empirical studies. We illustrate how different spatial econometric approaches have been applied and their implications for model specification, estimation and interpretation. We do not attempt to provide a comprehensive review of the empirical literature, but consider a wide range of articles, illustrative of the different perspectives taken in applied work.

We begin the remainder of the chapter by setting the stage with a brief review of the principles underlying the hedonic house price model, followed by an extensive discussion of spatial econometric aspects due to spatial models and spatial data, specific to house price applications (for a more comprehensive technical review, see Anselin, 2006). We then review in turn the treatment of spatial dependence (including space-time dynamics) and spatial heterogeneity, with selective illustrations from the empirical literature. We close with a discussion of policy implications and some conclusions.

26.2 Hedonic house price models

In this section, we first outline the main properties of the hedonic price model in the context of specifications for house prices. Next, we discuss some important features of the estimation and identification of such models. We postpone the discussion of specific spatial aspects to the next section.

26.2.1 General framework

In the second half of the 1960s, a new branch of utility theory evolved from the pioneering work of Lancaster (1966), in which utility was defined as a function of the characteristics of a good. Initially, the focus was primarily on consumer models, until Rosen (1974) generalized this to a model of market equilibrium that took into account both consumers and producers. In this, the individual's utility becomes a function of the characteristics of a commodity, and producer costs depend on the characteristics of the good.

The hedonic price equation defines a market equilibrium after all interactions between supply and demand have taken place. A considerable literature has built upon this basic model (for a review, see, e.g., Malpezzi, 2002), and active research pertaining to both theoretical and econometric aspects continues apace (e.g., Ekeland *et al.*, 2004).

In the specific context of house price models, the basic hedonic specification assumes that the utility of a household or an individual is a function of a composite good x, a vector of location specific environmental characteristics q, a vector of structural characteristics S, a vector of social and neighborhood characteristics N, and finally a vector of locational characteristics L (Freeman, 1999). One of the main assumptions of the hedonic model is that preferences are weakly separable in housing and its characteristics. This implies that the demand for housing characteristics can be written as a function of "expenditure and prices within the group alone" (Deaton and Muellbauer, 1980, p. 124). Specifically, this implies that housing demand can be written as a function of the house price, with the prices of all other goods represented by a composite good x as the numeraire. Also, perfect information is assumed, in the sense that a consumer perceives all relevant house characteristics and takes them into consideration when purchasing a house.

The decision problem involved in a house purchase then consists of maximizing utility subject to the usual income constraint. Formally, for house i:

$$\begin{aligned} Max \quad & U\left(x, q_i, S_i, N_i, L_i\right) \\ \text{s.t.} \quad & M = P_i + x, \end{aligned} \tag{26.1}$$

where M is the household income, P_i is the price of house i, and the composite good x is the numeraire.

For each characteristic of interest, the first-order condition defines the marginal willingness to pay (MWTP) for changes in the levels of such a characteristic. For

example, for environmental characteristic q_i, this would yield:

$$\frac{\partial U/\partial q_i}{\partial U/\partial x} = \partial P_i/\partial q_i. \tag{26.2}$$

The hedonic price function is an equilibrium price equation where the price of house i is defined as a function of the house characteristics:

$$P_i = P(q_i, S_i, N_i, L_i). \tag{26.3}$$

With estimates for the coefficients in this function to hand, it then becomes possible to estimate the individual's marginal willingness to pay for any characteristic that enters the utility function. As shown in equation (26.2), differentiating equation (26.3) with respect to the characteristic of interest yields the desired result.

The marginal willingness to pay can be interpreted as a "price" for the characteristic and exploited to construct an inverse demand function. To accomplish this, the MWTP is "observed" at different levels of the characteristic, say q_i, and combined with additonal demand shifter variables C in a demand function:

$$MWTP_i = f(q_i, C_i). \tag{26.4}$$

This allows for the analysis of non-marginal changes in the characteristic.

26.2.2 Estimation

The operational implementation of a hedonic analysis consists of two stages: the estimation of a hedonic price function and the construction of the inverse demand function for house characteristics. In most applied work, the second stage is not carried out.

In the first stage, a hedonic price function is specified in terms of the relevant characteristics of the house, typically a combination of individual house features (size, number of rooms, amenities such as air conditioning, etc.), environmental characteristics, neighborhood characteristics and location. Different functional forms can be used, either linear or nonlinear, the most commonly being linear, semi-log and log-log. Alternatively, a flexible Box–Cox approach can be taken. In a much cited study, Cropper *et al*. (1988) carried out a large number of simulations to assess the sensitivity of the results to functional specification. They found that when there are omitted variables or when proxies are used in the absence of a measure of the real variable, simpler functional forms such as linear or semi-log perform better than more complex forms. In most applied work, this is the path taken.

The estimates of the parameters in the price function yield the marginal price for each characteristic as the partial derivative of the function with respect to the characteristic. Depending on the functional form, this marginal price may change with the level of the characteristic. It is interpreted as the marginal willingness to pay for the characteristic.

The second stage of a hedonic analysis is to relate the marginal willingness to pay to different levels of the characteristic in order to yield an inverse demand function

for the characteristic. The general form of this expression is as in equation (26.4). This second stage is especially important when interest centers on the effect of non-marginal changes in the characteristic.

One important problem associated with this second stage is identification (Palmquist, 1991; Freeman, 1999). This is primarily a result of the fact that the prices used in the second-stage inverse demand function are not actually observed, but derived from the coefficient estimates in the first stage (the hedonic price function). Also, the MWTP is computed as a function of at least one of the explanatory variables used in the price specification. In order to be able to separately identify the inverse demand function, it is necessary that additional variables (demand shifters) be included in the second stage.

Furthermore, the attribute variable (amount of the characteristic) is an explanatory variable in both the first and second stage and is thus endogenous. This follows because the selection of a point on the hedonic price function simultaneously determines the level of the characteristic and the marginal price associated with it. This endogeneity must be taken into account in the second stage estimation, typically by using instrumental variables. The instruments should be truly exogenous to the model, which in practice may be difficult to establish.

One solution to the identification problem, suggested by Palmquist (1991), is to use information from spatially or temporally distinct sub-markets. The rationale behind this is that the heterogeneity between the sub-markets provides sufficient variability to identify a demand function for the characteristic. This is estimated by taking marginal prices from different sub-markets as well as demand shifters that vary sufficiently across the sub-markets (e.g., income).

Apart from the identification issue for the demand function (the second stage), sub-markets should also be explicitly considered when spatial heterogeneity invalidates the assumption of a single equilibrium. After defining meaningful spatially delineated sub-markets, a separate hedonic price function should be estimated for each.

In practice, interest has focused on the first stage of the hedonic model and on estimation of MWTP for various characteristics. An important body of applications pertains to the valuation of amenities, such as environmental quality, access to open space and views. Very few spatial econometric applications have carried out the second stage of estimating the inverse demand function. Recent exceptions are Beron *et al*. (2004) and Brasington and Hite (2005).

26.3 Spatial models

Increasingly, applied econometric work dealing with hedonic house price models has taken an explicit spatial econometric perspective. Some recent examples include Basu and Thibodeau (1998), Dubin *et al*. (1999), Bell and Bockstael (2000), Bowen *et al*. (2001), Bourassa *et al*. (2003), Kim *et al*. (2003), Beron *et al*. (2004), Pace and LeSage (2004), Brasington and Hite (2005), Anselin and Le Gallo (2006), Neill *et al*. (2007) and Anselin and Lozano-Gracia (2008). This work focuses, in particular, on the treatment of market interactions or unobserved neighborhood effects

through the incorporation of spatial dependence and on the use of model specifications that allow for spatial heterogeneity in the form of submarkets. Several different model formulations and estimation methods have been applied, reflecting the richness of the spatial methodology (Anselin, 2006).

The motivation for incorporating spatial effects into the specification of a hedonic house price model is based on two main concerns. One, which we refer to as *substantive*, is that the model form is intended to capture either interaction effects, market heterogeneity, or both. The other is more pragmatic and we refer to it as a *nuisance*, in that spatial autocorrelation in omitted variables, or unobserved externalities and heterogeneities, are relegated to the error term. In dealing with spatial dependence, these two perspectives are reflected in the *lag* and *error* models (Anselin, 1988). In addressing spatial heterogeneity, varying coefficient models and spatial regimes reflect substantive models, whereas various specifications for heteroskedasticity deal with nuisance effects.

The econometric treatment of these two types of effects differs considerably. Substantive models require a new class of estimation methods and specification tests, whereas nuisance models are simply special cases of a non-spherical error variance-covariance matrix. The consequences of ignoring these effects differ as well. Omitting substantive effects when they should be included results in model misspecification. Consequently, the estimates of the remaining parameters will be biased and inconsistent, and inference may be spurious. In a hedonic context, this implies that conclusions about the marginal price of specific characteristics (e.g., environmental improvements) may be wrong. On the other hand, nuisance effects are primarily a problem of efficiency. Ignoring those effects when present will yield biased estimates of standard errors in a traditional ordinary least squares (OLS) regression if the proper adjustments are not carried out. This will yield biased t-tests and misleading indications of precision. Since the coefficient estimates in hedonic models are used in further calculations (e.g., of marginal willingness to pay), it remains important to have correct measures of standard errors in order to properly address uncertainty in a policy decision making context.

One additional complexity with spatial models is that spatial dependence and spatial heterogeneity are often difficult to distinguish in a cross-sectional setting. The properties of specification tests and estimators developed for one type of effect are affected by the presence of the other type. In practice, one typically addresses one type of spatial effect first, carries out specification tests for remaining problems and subsequently addresses those if warranted.

In this section, we review the main model specifications and estimation methods that have been applied in hedonic studies. Here, we only focus on the basic properties and do not intend to duplicate the extensive methodological reviews provided in Anselin and Bera (1998) and Anselin (2006), among others. We also limit our discussion to the most commonly used specifications. For spatial (and space-time) dependence, these are the lag and error models, as well as some recently suggested semiparametric approaches. For spatial hetereogeneity, we cover the treatment of discrete (regimes) and continuous (varying coefficients) spatial heterogeneity.

26.3.1 Spatial dependence

26.3.1.1 Spatial lag model

The spatial lag specification is characterized by the inclusion of a new variable on the right-hand side of the equation. This variable, referred to as a spatially lagged dependent variable (Anselin, 1988) captures the spatial interaction effect as a weighted average of neighboring observations. This is most commonly applied in a linear form, as:

$$y = \rho Wy + X\beta + u, \tag{26.5}$$

where y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times k$ matrix of observations on explanatory variables, W is an $n \times n$ spatial weights matrix, u is an $n \times 1$ vector of independent and identically distributed (i.i.d.) error terms, ρ is the spatial autoregressive coefficient, and β is a $k \times 1$ vector of regression coefficients.

The $n \times n$ spatial weights matrix defines the neighbor set for each individual location. Its positive elements w_{ij} are non-zero when observations i and j are *neighbors*, and zero otherwise. By convention, self-neighbors are excluded, such that the diagonal elements of W are zero. In addition, in practice the weights matrix is typically row-standardized, such that $\sum_j w_{ij} = 1$. Many different definitions of the neighbor relation are possible, and there is little formal guidance on the choice of the "correct" spatial weights.[1] The term Wy in equation (26.5) is referred to as a spatially lagged dependent variable, or spatial lag. For a row-standardized weights matrix, it consists of a weighted average of the values of y in neighboring locations, with weights w_{ij}.

As stated in Anselin and Bera (1998), there are two main interpretations for a significant spatial autoregressive coefficient ρ. First, this may suggest a contagion process or the presence of spatial spillovers. However, this interpretation is valid only if the process takes place at the spatial unit used in the analysis, and is supported by a theoretical model. In the context of spatial hedonic models, this is often difficult to maintain, since it is unlikely that economic agents simultaneously take into account the prices of neighboring units. An alternative explanation for a significant spatial autocorrelation coefficient is the existence of a mismatch between the observed spatial unit and the true spatial scale of the process being studied.

The theoretical motivation for a spatial lag specification is based on the literature on interacting agents and social interaction. For example, a spatial lag follows as the equilibrium solution of a *spatial reaction function* (Brueckner, 2003) that includes the decision variable of other agents in the determination of the decision variable of an agent (see also Manski, 1993, 2000). In hedonic models, however, where a purely cross-sectional setting is more common, it is often difficult to maintain such a theoretical motivation, since it would imply that buyers and sellers simultaneously take into account prices obtained in other transactions.

An alternative interpretation is provided by focusing on the reduced form of the spatial lag model:

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} u, \tag{26.6}$$

where, under standard regularity conditions, the inverse $(I - \rho W)^{-1}$ can be expressed as a power expansion:

$$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \ldots. \tag{26.7}$$

The reduced form thus expresses the house price as a function of its own characteristics (X), but also of the characteristics of neighboring properties, (WX, W^2X), albeit subject to a distance decay operator (the combined effect of powering the spatial autoregressive parameter and the spatial weights matrix). In addition, omitted variables, both property-specific as well as related to neighboring properties, are encompassed in the error term. In essence, this reflects a scale mismatch between the property location and the spatial scale of the attributes that enter into the determination of the equilibrium price.

From a purely pragmatic perspective, one can also argue that the spatial lag specification allows for a *filtering* of a strong spatial trend (similar to detrending in the time domain), i.e., ensuring the proper inference for the β coefficients when there is insufficient variability across space. Formally, the spatial filter interpretation stresses the estimation of β in:

$$y - \rho Wy = X\beta + u. \tag{26.8}$$

In most spatial hedonic applications, the use of a spatial lag specification follows as the result of a specification search based on specialized Lagrange multiplier tests that indicate the preference of this alternative over an error specification (Anselin and Bera, 1998; Anselin, 2001a; Florax *et al*., 2003). In such instances, the selection of this model is mostly pragmatic, without necessarily implying a theoretical model of social interaction.

The most commonly applied estimation method for the parameters of the spatial lag model is maximum likelihood, following the principles outlined by Ord (1975) (for additional technical details, see, e.g., Anselin, 1988; Anselin and Bera, 1998; Anselin, 2006). More recently, an instrumental variables or spatial two stage least squares approach has gained greater popularity, because it lends itself more readily to application in the large datasets characteristic of hedonic studies. Early results were given in Anselin (1988) and Kelejian and Robinson (1993), but more recently interest has focused on formal proofs of asymptotic properties and the choice of optimal instruments, e.g., in Lee (2003, 2007), Das *et al*. (2003) and Kelejian *et al*. (2004). Bayesian estimation of spatial autoregressive models is covered in LeSage (1997).

We leave a more detailed discussion of specific applications of spatial hedonic models to section 26.5.

26.3.1.2 Spatial error model

From a theoretical viewpoint, a spatial error specification is the more natural way to include spatial effects in a hedonic model. Unobserved neighborhood effects will be shared by housing units in the same area and naturally lead to spatially correlated error terms. This results in a non-diagonal error variance-covariance matrix:

$$\text{Var}[uu'] = \text{E}[uu'] = \Sigma, \tag{26.9}$$

where $\Sigma \neq \mathbf{I}$, with $\mathbf{I}$ as the identity matrix. Typically, Σ contains "nuisance" parameters that need to be estimated consistently. This, in turn, yields consistent estimates for the regression coefficients by means of a feasible generalized least squares (FGLS) estimation. The interpretation of the nuisance parameters is very different from the spatial autoregressive coefficient in the spatial lag model, in that there is no particular relation to a substantive model of spatial interaction. These parameters are only included in order to obtain better estimates for the regression slope coefficients.

The particular structure of Σ follows from a spatial ordering of the observations (e.g., as argued in Kelejian and Robinson, 1992). In practice, the most commonly used specification assumes a spatial autoregressive process for the error terms:

$$y = X\beta + \varepsilon, \tag{26.10}$$

with:

$$\varepsilon = \lambda W\varepsilon + u, \tag{26.11}$$

with $u \sim i.i.d.$, and λ as the spatial autoregressive coefficient.

The resulting error variance-covariance matrix is as follows:

$$E[\varepsilon\varepsilon'] = \sigma^2[(I - \lambda W)(I - \lambda W')]^{-1}. \tag{26.12}$$

A commonly used alternative in hedonic analyses is to base the structure of the error variance-covariance matrix on principles from geostatistics. Early work by Dubin (1988) (see also Dubin, 1992; Basu and Thibodeau, 1998; Dubin *et al.*, 1999; Miltino *et al.*, 2004) suggested a so-called *direct representation* for the elements of the variance-covariance matrix.

In this approach, the covariance between each pair of error terms is specified as an inverse function of the distance between them. Formally:

$$\mathrm{E}[\varepsilon_i\varepsilon_j] = \sigma^2 f(d_{ij}, \phi), \tag{26.13}$$

with $\varepsilon_i, \varepsilon_j$ as the regression error terms, σ^2 the error variance, and d_{ij} the distance separating i and j. The function f should be a distance *decay* function that ensures a positive definite covariance matrix. This requires $\partial f/\partial d < 0$ and $|f(d_{ij}, \phi)| \leq 1$, with $\phi \in \Phi$ as a $p \times 1$ vector of parameters on an open sub-set Φ of $\mathbb{R}^p$. This approach is closely related to the variogram models used in geostatistics, and requires assumptions of stationarity and isotropy (see Cressie, 1993, for an extensive review). The complete variance-covariance is then:

$$\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2\Omega(d_{ij}, \phi). \tag{26.14}$$

A commonly used specification is based on a negative exponential distance decay:

$$\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2[I + \gamma\Psi], \tag{26.15}$$

with the off-diagonal elements of Ψ being $\Psi_{ij} = e^{-\phi d_{ij}}$, and γ as a non-negative scaling parameter. In order to facilitate interpretation and specification testing, the diagonal elements of Ψ are often set to zero (the variance is captured by the term $\sigma^2 I$). The distance metric and parameter space must be such that the elements of $e^{-\phi d_{ij}}$ yield a valid spatial correlation matrix (see Anselin, 2001a, for technical details).

Estimation in parametric spatial error models is most commonly based on the ML principle (see Anselin, 1988; Dubin, 1988). Due to computational limitations in very large datasets, recent attention has shifted to alternatives, such as the generalized moments (GM) and generalized method of moments (GMM) estimators suggested by Kelejian and Prucha (1998, 1999). An early application of this approach to a hedonic specification can be found in Bell and Bockstael (2000) (see also section 26.5 for further examples). Generalization of this approach to an error structure that contains both spatial autocorrelation and heteroskedasticity can be found in recent papers by Lin and Lee (2005) and Kelejian and Prucha (2006).

A different approach is to avoid the parametric specification of spatial covariance as a function of a distance metric and to use a nonparametric perspective. This is an extension to the spatial domain of the principle behind the heteroskedasticity and autocorrelation consistent covariance matrix estimation of Newey and West (1987) and Andrews (1991), among others.

As in the direct representation approach, the spatial covariance is a function of the distance separating two observations, but the functional form is left unspecified. For example, for the regression error terms:

$$\mathrm{E}[\varepsilon_i \varepsilon_j] = f(d_{ij}), \tag{26.16}$$

where d_{ij} is a "proper" positive and symmetric distance metric (for regularity conditions on the distance metric, see Conley, 1999; Kelejian and Prucha, 2007).

This estimator follows essentially the same principle as in the time series domain by adding up sample spatial autocovariances. In order to ensure positive definiteness of the estimator, a kernel is applied to the cross-products. For example, in the recent paper by Kelejian and Prucha (2007), a general covariance matrix estimator takes the form:

$$\hat{V} = n^{-1} \sum_i \sum_j x_i x_j' \hat{\varepsilon}_i \hat{\varepsilon}_j K(d_{ij}/d), \tag{26.17}$$

where $K()$ is a kernel function and d a suitable cutoff distance. This yields a so-called heteroskedastic and spatial autocorrelation consistent, or HAC, estimator. A recent application of this approach to spatial hedonic models can be found in Anselin and Lozano-Gracia (2008).

Arguably, the treatment of spatially structured omitted variables may be addressed without resorting to a spatial error. The most commonly used technique in the empirical literature is to address this by means of spatial fixed-effects, e.g., by including a dummy variable for a larger spatial area that individual housing

units belong to, such as a census tract or block group. This rests on the assumption that the spatial range of the unobserved heterogeneity/dependence is specific to each spatially delineated unit. In practice, there may indeed be spatial units (such as school districts) where such a spatial fixed effects approach is sufficient to correct the problem. However, the nature of omitted neighborhood variables tends to be complex, as is the definition of the correct "neighborhood," and in many instances the fixed-effects approach will be insufficient to remove all residual spatial autocorrelation.

26.3.1.3 Other models of spatial dependence

In addition to the familiar spatial lag and spatial error models just outlined, a number of other techniques have been adopted to deal with spatial effects in hedonic house price functions. We briefly review here semiparametric approaches.

An alternative way to account for space in a hedonic regression is to incorporate it directly in the hedonic price function in the form of a trend surface, while maintaining the assumption of constant marginal prices across space. In a parametric approach, this would consist of including a polynomial in the X, Y coordinates of the observations as explanatory variables in the hedonic equilibrium equation. This could also be combined with a fixed-effects approach in the form of dummy variables for administrative units, such as zip code or census tracts.

A semiparametric alternative, first discussed by Clapp *et al*. (2002), includes, in addition to the usual hedonic variables, a nonparametric function $f(X_c, Y_c)$ of the location of the observations. This is to model the omitted spatial variables in the mean function, rather than relegating them to the error term. Formally, this yields:

$$P = X\beta + f(X_c, Y_c) + \varepsilon. \tag{26.18}$$

This function may be estimated using standard nonparametric techniques such as local polynomial regression. The Nadaraya–Watson estimator is the method most frequently used in the literature.

Clapp *et al.* suggest estimating this model in an iterative fashion consisting of two main steps. First, the parameters of all house characteristics are estimated using OLS. In a second step, the residuals from this regression are fitted using Bayesian or local polynomial regression techniques. The first step thus yields OLS residuals $\hat{\eta}^0$ as:

$$\hat{\eta}^0 = P - X\hat{\beta}^0. \tag{26.19}$$

In a first iteration, these residuals are then smoothed using a Nadaraya–Watson estimator, as:

$$\bar{\eta}^0 = \sum_{i=1}^{q} \frac{K_h(X_{c_i} - X_0)K_h(Y_{c_i} - Y_0)(\hat{\eta}^0)}{\sum_{i=1}^{q} K_h(X_{c_i} - X_0)K_h(Y_{c_i} - Y_0)}. \tag{26.20}$$

Then, the estimated residuals from a first iteration are obtained as:

$$\hat{\epsilon^0} = \hat{\eta^0} - \bar{\eta^0}. \tag{26.21}$$

In the next iteration, the linear part of the model is estimated again using OLS, but now using $\hat{Y}^1$ as the dependent variable, where:

$$\hat{Y}^1 \equiv X\hat{\beta}^0 + \hat{\varepsilon}^0. \tag{26.22}$$

Iterations between the parametric and nonparametric portions of the model continue until the $\hat{\beta}$s change by no more than 5%.

A different estimation approach for this model is described in Gibbons (2003) and Day *et al*. (2007), who use *spatially filtered* variables to consistently estimate the β coefficients, following the steps first outlined in Robinson (1988). To implement this, the model in equation (26.18) is rewritten as:

$$P - E[P|X_c, Y_c] = (X - E[X|X_c, Y_c])\beta + \varepsilon. \tag{26.23}$$

Estimates for the marginal prices are obtained by using the spatially weighted means of the explanatory variables. First, the conditional means $E[P|X_c, Y_c]$ and $E[X|X_c, Y_c]$ are estimated using nonparametric regression. Estimates of these functions are then substituted into equation (26.23) and, following Robinson (1988), consistent estimates for the β coefficients are obtained from OLS on equation (26.23).

Day *et al*. (2007) consider an additional complication and allow for the presence of spatial error autocorrelation in the form of a spatial autoregressive process (as in equation 26.11). Rather than applying OLS, estimates for the β in equation (26.23) are obtained using the Kelejian and Prucha (1999) GM approach.

Alternatively, Clapp *et al*. (2002) utilize Bayesian methods to remove any remaining spatial autocorrelation from the model. In standard Bayesian fashion, the error terms are specified as consisting of two components; one being spatial, the other a white-noise process. Formally:

$$\hat{\varepsilon}(c_i) = \delta(c_i) + \psi(c_i), \tag{26.24}$$

in which $\delta(c_i)$ is assumed to come from a stationary Gaussian spatial process with mean 0 and spatial covariance function $cov(\delta(c), \delta(c')) = \sigma^2 exp(-\psi\|c - c'\|)$ and ψ is assumed $i.i.d \sim N(0, \tau)$. Bayesian fitting is then applied to the first stage residuals using Gibbs sampling combined with a Metropolis–Hasting procedure to account for remaining spatial autocorrelation in the residuals.

26.3.1.4 Models for space-time dependence

The temporal dimension has not received much attention in spatial hedonic models. There are both theoretical as well as practical reasons for this omission. First, using data for several time periods would require the assumption that the marginal prices stay constant through time. While this assumption may seem appropriate for a short period of time, it is unlikely to hold when several years are considered. As a result, hedonic analyses have tended to favor pure cross-sectional approaches.

Furthermore, explicitly including both the temporal and spatial dimension requires complex estimation methods. Most applications have used Bayesian methods to tackle this complexity. Some examples of spatio-temporal hedonic analyses

include Pace *et al*. (1998), Pace *et al*. (2000), Gelfand *et al*. (2004), Sun *et al*. (2005) and Huang *et al*. (2006).

As an example, consider Pace *et al*. (1998) and Pace *et al*. (2000). They propose a spatio-temporal specification in which a filtering approach is carried out across both dimensions. The model is:

$$(I - W)Y = (I - W)X\beta + u, \tag{26.25}$$

where W is a row-standardized $N \times N$ lower triangular spatio-temporal weights matrix. The weight matrix W is used to filter out the spatio-temporal correlation and the model is then estimated for the uncorrelated variables $(I - W)X$ and $(I - W)Y$.

Data are ordered by time period, so that the first row in the data refers to the earliest observation. Therefore, only previous sales are allowed to influence current house prices. This explicit relation between time periods contrasts with the approach taken in cross-sectional studies, where all sales are assumed to have taken place during the same period. As a result, in a cross-sectional set-up, there is a simultaneous feedback between all house prices in the sample, even though technically the sales may have taken place at different times during the period (so, conceivably, a later sale could affect an earlier sale). The explicit inclusion of space-time correlation allows for a more realistic model of the actual timing of real estate transactions.

In this approach, W is not assumed *a priori* but rather defined through a flexible form, where $(I - W) = (I - \phi_S S - \phi_T T - \phi_{TS} TS - \phi_{ST} ST)$, with T and S respectively defined as temporal and spatial weight matrices. The expanded definition of $(I - W)$ is then plugged back into equation (26.25) and the model is estimated using Bayesian methods for the unfiltered variable Y.

Gelfand *et al*. (2004) introduce time and space in the model by allowing the coefficients to change over time. They define three possible forms for the error term in a separable form, which avoids explicit specification of space-time interactions:

$$U(s,t) = \alpha(t) + W(s) + \varepsilon(s,t), \tag{26.26}$$

$$U(s,t) = \alpha_s(t) + \varepsilon(s,t), \tag{26.27}$$

$$U(s,t) = W_t(s) + \varepsilon(s,t), \tag{26.28}$$

with $\varepsilon(s,t)$ as *i.i.d.*$N(0, \sigma_\varepsilon^2)$ error terms, $\alpha_s(t)$ as the temporal effect and $W(s)$ as the spatial effect.

Equation (26.26) provides a structure that is additive in spatial and temporal effects. In contrast, equation (26.27) suggests that the temporal effects are local, changing from one site to the other. Finally, the form for the error term suggested in equation (26.28) pertains to the case where the spatial effects are specific to the time period. Gelfand *et al*. (2004) then outline a fully Bayesian approach and specify the associated likelihoods together with the prior distributions for all parameters in the model.

A simpler spatio-temporal approach is introduced in Huang *et al.* (2006). Here, an expanded version of a spatial lag model is estimated using ML, under the assumption that spatial correlation remains constant through time. Furthermore, after the inclusion of the spatially lagged dependent variable, the error terms are assumed to be correlated only in time. This spatio-temporal version of the spatial lag model takes the following form:

$$y^* = \rho(W \otimes I_T)y^* + X^*\beta + u^*, u^* \sim N(0, \sigma_u^2), \tag{26.29}$$

with $y^* = (I_N \otimes Q)y$, $X^* = (I_N \otimes Q)X$, and $u^* = (I_N \otimes Q)u$. Q is a transformation matrix that "removes the effect" of the $AR(1)$ process in the residuals following a method suggested in Judge *et al.* (1988) and Hsieh *et al.* (2001). If the $AR(1)$ correlated residuals are defined as:

$$u_{it} = \lambda u_{i,t-1} + \upsilon_{it}, \tag{26.30}$$

with $\upsilon_{it} \sim N(0, \sigma_\upsilon^2)$, $E[u_{it}u'_{is}] = \sigma_\upsilon^2\Omega(\lambda)$, and $t \neq s$, then:

$$\Omega(\lambda) = I_n \otimes \frac{1}{1-\lambda^2}\begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{T-1} \\ \lambda & 1 & \lambda & \cdots & \lambda^{T-2} \\ \lambda^2 & \lambda & 1 & \cdots & \lambda^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda^{T-1} & \lambda^{T-2} & \lambda^{T-3} & \cdots & 1 \end{bmatrix}, \tag{26.31}$$

and Q is the transformation matrix such that $\Omega^{-1} = I \otimes (Q(\lambda)'Q(\lambda))$. In particular, Q takes the following form:

$$Q(\lambda) = \begin{bmatrix} \sqrt{1-\lambda^2} & 0 & 0 & \cdots & 0 & 0 \\ -\lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda & 1 \end{bmatrix}. \tag{26.32}$$

This Q is used to define the transformed model in equation (26.29), which can then be estimated using ML. Huang *et al.* (2006) conclude that introducing these spatial and temporal correlations improves the goodness-of-fit of the model.

26.3.2 Spatial heterogeneity

26.3.2.1 Discrete spatial heterogeneity

Discrete spatial heterogeneity is a special case of structural instability in the model specification (functional form and/or parameters), where the form of the instability follows a spatial structure, referred to as *spatial regimes* (Anselin, 1988, 1990). In practice, this often occurs when structural breaks can be observed between different sub-regions, such as core and peripheral regions or urban and rural areas. In

hedonic models, discrete spatial heterogeneity is taken into account in the form of separate models for *sub-markets*. For example, inelasticities in supply and demand may lead to market segmentation that results in spatial heterogeneity in the form of varying marginal prices (Goodman and Thibodeau, 1998). Sub-markets can be defined spatially or non-spatially but, in practice, preference goes to delineations that follow clearly observed spatial boundaries (for examples, see section 26.6.1).

More formally, discrete spatial heterogeneity can be expressed as:

$$y_i = f_i(X_i, \beta_i, \epsilon_i), \tag{26.33}$$

with i as an index corresponding to a given discrete sub-set of the data. This general formulation includes as special cases functional instability, $f_i \neq f_j$ (e.g., linear model for one region, log-linear in another), parameter variation, $\beta_i \neq \beta_j$ (e.g., different parameter values for house characteristics in different sub-markets), as well as heteroskedasticity, $Var[\epsilon_i] \neq Var[\epsilon_j]$.

These examples represent fairly standard methodological issues that can readily be addressed and do not require an explicit spatial econometric treatment. However, in practice, in many instances heterogeneity and spatial dependence occur together, or, are difficult to identify separately (Anselin and Griffith, 1988). For example, spatial spillover may not be constrained to each specific spatial sub-set of the data, but may reach across the boundaries. In those cases, the treatment of the heterogeneity becomes complicated by the presence of spatial dependence, and extensions of the standard spatial lag and error models and associated specification tests are in order. One example of such tests is the so-called spatial Chow test of coefficient stability, which is an extension of the standard case that incorporates spatial dependence (Anselin, 1990).

In spatial hedonic models, attention has focused primarily on the delineation of sub-markets, and the acknowledgment of spatial dependence between sub-markets has only received limited attention. We provide specific examples in section 26.6.1.

26.3.2.2 Continuous spatial heterogeneity

As an alternative to considering discrete spatial sub-sets of the data, heterogeneity can be viewed as a smooth continuous process of varying parameters. One of the earliest applications of this perspective to spatial analysis was in the so-called *spatial expansion method* proposed by Casetti in the early 1970s (see, e.g., Casetti, 1972, 1997).

Spatial expansion is a special case of a varying coefficients model and also shows great similarity to the approach taken in multi-level modeling (e.g., Goldstein, 1995). Using Casetti's terminology, the first step is a so-called initial equation, which is a simple linear regression specification for each observation i:

$$y_i = \sum_k x_{ki}\beta_{ki} + \epsilon_i, \tag{26.34}$$

for k explanatory variables, including a constant term. Next, an expansion equation expresses the variability of the regression coefficient over i as a function

of additional explanatory variables z_{hi} (again, including a constant term):

$$\beta_{ki} = \sum_h z_{hi}\gamma_h, \tag{26.35}$$

where the γ_h are an additional set of parameters. Note that, unlike the standard multilevel model, the observational unit for the x and z variables is the same (i). The combination of the initial model with the expansion equation yields the so-called final equation, which contains the original explanatory variables, as well as interaction variables, the product of each x_{ki} with all the z_{hi}:

$$y_i = \sum_k x_{ki}\left(\sum_h z_{hi}\gamma_h\right) + \epsilon_i. \tag{26.36}$$

A slight generalization is obtained when the expansion equation includes an error term, which yields a heteroskedastic model (Anselin, 1992). A common problem in the implementation of this approach is a high degree of multicollinearity.

In spatial hedonic specifications, the expansion method is used to model heterogeneity in the form of so-called neighborhood drift (Can, 1992). This may also account for some omitted variables at the neighborhood level and therefore reduce the intensity of the spatial autocorrelation problem. However, it is important to note that, even if a parametric drift is introduced, spatial autocorrelation and heterogeneity may remain as a problem.

An alternative to the parametric specification of the expansion equation is a nonparametric approach, in which the variability of the model parameters is determined by the data. The best known among these approaches is arguably the geographically weighted regression (GWR), popularized in the work of Fotheringham and collaborators (for an overview, see Fotheringham *et al.*, 2002).

GWR is essentially a special case of a local regression model (LRM) (e.g., as proposed in Cleveland and Devlin, 1988), in which the weighting scheme that determines the variability in the parameters is based on the spatial closeness of observations (for examples of spatial hedonic applications, see Pavlov, 2000; Gelfand *et al.*, 2003; Cho *et al.*, 2006; Kestens *et al.*, 2006, among others). In this approach, a model parameter is defined as a function of the location of individual observations. In addition, a weighting scheme is designed such that greater weight is given to locations that are closer in space. To illustrate this approach, consider a hedonic model specified as:

$$P = \beta_0(c) + \sum_k X_k\beta_k(c) + \varepsilon, \tag{26.37}$$

where c is a vector of X_c, Y_c coordinates that define the location of the data points. The parameters are estimated by minimizing a weighted residual sum of squares:

$$\min_{p,q} \sum_i \left\{ W_i(c)\left[P - \beta_0(c) - \sum_k X_k\beta_k(c)\right]^2 \right\}, \tag{26.38}$$

where $W_i(c)$ are the weights that depend on the location (c). The solution to the minimization problem in equation (26.38) yields the standard weighted least squares expression, in matrix notation:

$$\hat{\beta} = (X'WX)^{-1}(X'Wy). \tag{26.39}$$

In this expression, W is not a spatial weights matrix, but a matrix that extracts the observations used in the estimation of the parameter for each location i. Several approaches have been suggested, such as a straightforward k-nearest neighbors weighting scheme (Pavlov, 2000), or a kernel smoother. The latter is the common approach taken in GWR, where a Nadaraya–Watson-type kernel smoother ensures that those observations near the point where the parameters are being estimated have more influence than those observations further away. Using such a kernel function in GWR then yields the function to be minimized as:

$$\min_{p,q} \sum_i \left\{ K_h(d_{0i}) \left[P - \beta_0(c) - \sum_k X_k \beta_k(c) \right]^2 \right\}, \tag{26.40}$$

where $K_h(\cdot) = K(\cdot/h)$, K is a given kernel function and h a bandwidth parameter.

Common choices for the kernel are the bi-square function:

$$K(t) = \begin{cases} (1-t^2)^2, & \text{if } |t| \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{26.41}$$

and the Gaussian kernel:

$$K(t) = exp\left(-\frac{1}{2}t^2\right). \tag{26.42}$$

Since the term "GWR" was first introduced in Brunsdon *et al*. (1996), an extensive set of papers has been published treating various theoretical issues related to model estimation, specification testing and cross-validation (see, among others, Fotheringham *et al*., 1998, 2002; Paez *et al*., 2002a, 2002b). Specific empirical applications to spatial hedonic specifications are reviewed in section 26.6.2.

26.4 Methodological challenges

Spatial hedonic analysis not only considers the specification of spatial relationships in the model, but also the estimation of relevant parameters on the basis of *spatial data*. In this section, we briefly point out some important methodological issues that need to be accounted for, especially the problem of spatial scale and the treatment of endogeneity.

26.4.1 Spatial scale

Spatial scale is important in the empirical implementation of hedonic models in a number of ways. The standard assumption is that the spatial units of observation match the process under consideration. However, with spatial data, this is not necessarily the case, and errors due to aggregation or interpolation need to be

accounted for. In addition, it is not always obvious how the sample of observations relates to a population (or super-population), which has repercussions for the type of asymptotics that can be applied.

26.4.1.1 Spatial observations

The theoretical framework outlined in section 26.2 implies that the proper estimation of the model parameters should be based on observations for individual transactions. In practice, this is not always possible and many studies instead rely on spatially aggregated data for units such as block groups, census tracts, and even counties (e.g., Brasington and Hite, 2005; Capozza *et al.*, 2005; Chay and Greenstone, 2005; Huang *et al.*, 2006). This leads to the problem of *ecological inference*, also known in geography as the modifiable areal unit problem, or in the statistical literature as the change of support problem (Gotway and Young, 2002). As shown in Anselin (2002), the parameters of spatial models estimated at an aggregate level (in particular the spatial autoregressive coefficient) do not correspond to those at the individual level. Consequently, estimates of hedonic specifications based on such aggregate units have only a tenuous basis in micro-theory and rely on a notion of representative agents (representative housing units) that may be highly unrealistic. The crucial aspect determining the extent of the problem is the intra-unit heterogeneity. If housing units, their characteristics, or the profiles of the household units that occupy them, vary considerably within a spatial unit, then an aggregate analysis based on a mean or median characteristic will not be very meaningful.

A second issue related to the change of support problem occurs when observations on some housing characteristics are not available for each individual unit. For example, in many instances, data on socioeconomic variables related to the households, such as income and education, cannot be obtained at the micro-level, but instead are proxied by spatial aggregates, such as the median income or percentage high school graduates at the census tract. All individual observations in the same census tract thus share the same value for these explanatory variables. At the very least, this leads to heteroskedastic error terms, but it may also result in more serious specification problems, as pointed out in Moulton (1990).

In other instances, the change in support problem manifests itself in a mismatch between the location and scale at which observations are collected for specific explanatory variables and the location of the housing units. A common example is the use of interpolated values for environmental variables related to air quality (e.g., ozone), which are typically collected at a small number of monitoring stations. In Anselin and Le Gallo (2006), the effect of applying different interpolation methods on the resulting estimates of MWTP for air quality is assessed. In a comparison of Thiessen polygons, inverse distance weighting, kriging and splines, the geostatistical kriging method yielded the best results in terms of model fit. More importantly, the differences in both coefficient estimates as well as in the calculations of MWTP were significant between the various interpolation methods, suggesting that greater attention to this aspect of the data is warranted.

26.4.1.2 Spatial sampling

The statistical foundations for the analysis of spatial hedonic models derive from two very different paradigms, related to the way in which the sampling of

observations is conceptualized (for details, see Anselin, 2002). The most widely used framework is referred to as lattice analysis, due to the fact that considerable early work in this area pertained to regularly spaced or gridded observations (e.g., Besag, 1974). In lattice analysis, the observations are discrete spatial units that exhaust the space, such as contiguous census tracts or counties. The main distinction is that the notion of interpolation is not supported, since observations are available on all spatial units in the "population." Asymptotics are based on a notion of expanding domain, i.e., growing the sample by adding additional units at the edge. In contrast, in so-called geostatistical analysis (Cressie, 1993), observations are a sample from a continuous surface. The main objective is to extract the characteristics of the continuous surface, so that interpolation (spatial prediction) can be carried out. The proper asymptotics are referred to as infill asymptotics and can be conceived of as increasing the density of sampling. Importantly, the two forms of asymptotics are different and properties that hold under one do not necessarily hold under the other (see, e.g., the discussion in Lahiri, 1996).

Hedonic house price studies are typically based on a sample of individual sales transactions or appraisals, and seldom include the full population. This is only the case in analyses for aggregate spatial units, such as census tracts or counties. Because of the nature of the sales sample, a geostatistical perspective should be the preferred approach. However, in practice, most empirical work is couched in a lattice perspective, using the standard spatial lag and spatial error specifications with spatial weights derived from contiguity or nearest neighbor criteria. This aspect is seldom highlighted, but it does raise potential problems in terms of the asymptotics necessary to obtain the consistency of estimators. For example, when a simple contiguity weights matrix is used between the locations of sales transactions (e.g., by using Thiessen polygons to define neighboring sales), a lattice approach assumes that the observed houses are the only houses in the population. In other words, the effect of the *neighbors* should be interpreted as the effect of neighboring *sales*, but not of neighboring properties. When the sales only constitute a sample of all transactions, the underlying assumption becomes that the effect of sampled neighbors is the same as that of the unobserved true neighbors. The validity of this assumption rests on the degree of spatial homogeneity of the housing market, in terms of both house and household characteristics. Without further information, this is very difficult to verify in practice.

26.4.2 Endogeneity

The issue of endogeneity in the estimation of demand equations that arise from a nonlinear hedonic price schedule is a familiar problem (see, e.g., Palmquist, 2005). Much less common is the focus on endogeneity in the estimation of the hedonic price equation itself. In the context of spatial hedonic models, this has received some attention, specifically in the study of the valuation of the contribution of air quality (and, to a lesser extent, of school quality).

There are two different perspectives on the endogeneity problem. In one, attention focuses on a specific house characteristic and the degree to which this is truly exogenous. For example, this would not be the case if air quality is correlated with

other unobserved characteristics of the house. Alternatively, if the house purchase decision is taken jointly with the assessment of environmental quality, endogeneity would also result. Similarly, sorting by house purchasers when there is heterogeneity in their preference functions associated with different pollution levels would result in endogeneity of the air quality variable. This aspect was treated extensively in a recent paper by Chay and Greenstone (2005), in the context of an application where air quality is measured by total suspended particles. They suggest the use of instrumental variables to obtain consistent estimates. While considerable care is taken in addressing these specification problems, the model itself is estimated at a fairly aggregate spatial scale of US counties.

Bayer *et al*. (2006) follow Chay and Greenstone (2005) by considering the possibility that local air pollution is correlated with unobserved local characteristics. They address this form of endogeneity by using the contribution of distant sources to local air pollution as an instrument. However, this study is also carried out at the spatially aggregate county level, and could therefore suffer from ecological fallacy.

The potential correlation of specific house or household characteristics with unobserved errors is considered by Gibbons (2003). Using a semiparametric model, the potential endogeneity of educational composition is accounted for by using the postcode-sector proportion of households in social housing as an instrument for educational composition.

In the second perspective, endogeneity follows as a consequence of an "errors in variables" problem. This is a special case of the change in support problem due to the limited number of sample points for air pollution. As a result, the "observations" of air quality at individual house locations are actually the result of a statistical spatial interpolation process with its own prediction error. In Anselin (2001b), it was pointed out that the spatial structure of the prediction error is likely to lead to correlation with the overall model disturbance term and thus to the familiar simultaneity bias. Anselin and Lozano-Gracia (2008) elaborate on this idea and estimate a spatial lag hedonic price equation using spatial two-stage least squares (2SLS), including additional instrumental variables to address the endogeneity of the air quality variable. Specifically, they use the components of a polynomial in the coordinates of the house locations as instruments.

Irrespective of the actual source of the endogeneity, the use of instrumental variables for some of the characteristic variables will yield consistent estimates. However, in the absence of optimal instruments, the precision of these estimates (and of the resulting computations of MWTP and other related measures) may be improved upon. This remains a subject of future investigation.

26.5 Empirical evidence: spatial dependence

The first attempts to incorporate spatial considerations into empirical hedonic house price studies consisted of including distance from the central business district as an explanatory variable in the model specification. While appropriate for monocentric cities, this is less suitable for polycentric areas, such as the Los Angeles metropolitan area. This resulted in several empirical studies reporting either

insignificant or positive signs for distance decay, a finding not compatible with theory (Dubin, 1992).

Dubin (1988) introduced the concept of spatial autocorrelation into the treatment of hedonic house price models. Her approach was based on geostatistical principles, in which the structure of spatial autocorrelation follows from an estimated theoretical semi-variogram.[2] As argued in section 26.4.1.2, the geostatistical approach is conceptually most suited to the analysis of a sample of house sales transactions. In spite of this, it has seen relatively few applications in applied spatial hedonic work. Other than the work of Dubin and co-authors (e.g., Dubin, 1992, 1998; Case *et al*., 2004), some notable examples include articles by Thibodeau (Basu and Thibodeau, 1998; Gillen *et al*., 2001), Miltino *et al*. (2004), and Bourassa *et al*. (2007), as listed in in Table 26.1. An overview of the major methodological issues is given in Dubin *et al*. (1999).

The bulk of applied work in spatial hedonic house price analysis takes a lattice data perspective and employs the standard spatial lag and spatial error models. An overview of several illustrative studies is given in Tables 26.2 and 26.3, for cases where, respectively, the spatial lag and spatial error specifications were the primary focus of attention. Topics covered range from the simple definition of the hedonic equilibrium and explanation of price differentials to the valuation of environmental benefits, accessibility to transportation systems, wildfire risk, and the impact of preservation policies.

Around the same time as Dubin's article appeared, Can (1990) was one of the first to consider the implications of spatially autocorrelated errors in the estimation of spatial regression models using the lattice perspective. Specifically, she allowed coefficients of the structural characteristics to vary across observations in a spatial lag specification. Using a sample of 577 house sales for 1980 in Columbus, Ohio, she concluded that a linear neighborhood quality drift expansion model is the most appropriate hedonic price specification. Later on, Can (1992) used data from 563 single-family house sales in 1980 for Franklin County to obtain heteroskedastic consistent estimators for a spatial autoregressive model based on bootstrapping techniques.

In recent years, the application of spatial econometric techniques in empirical hedonic studies has become more widespread. Most analyses still rely on ML

Table 26.1 Spatial dependence: geostatistics

Article	*Source*
Dubin (1988)	*Review of Economics and Statistics*
Dubin (1992)	*Regional Science and Urban Economics*
Basu and Thibodeau (1998)	*Journal of Real Estate Finance and Economics*
Dubin (1998)	*Journal of Real Estate Finance and Economics*
Dubin *et al*. (1999)	*Journal of Real Estate Literature*
Case *et al*. (2004)	*Journal of Real Estate Finance and Economics*
Miltino *et al*. (2004)	*Journal of Real Estate Finance and Economics*
Bourassa *et al*. (2007)	*Journal of Real Estate Finance and Economics*

Table 26.2 Spatial dependence: lag

Article	*Source*
Can (1990)	*Economic Geography*
Can (1992)	*Regional Science and Urban Economics*
Bowen *et al*. (2001)	*Growth and Change*
Gawande and Jenkins-Smith (2001)	*Journal of Env. Economics and Management*
Kim *et al*. (2003)	*Journal of Env. Economics and Management*
Miltino *et al*. (2004)	*Journal of Real Estate Finance and Economics*
Capozza *et al*. (2005)	*Real Estate Economics*
Hunt *et al*. (2005)	*Ecological Economics*
Anselin and Le Gallo (2006)	*Spatial Economic Analysis*
Armstrong and Rodríguez (2006)	*Transportation*
Huang *et al*. (2006)	*American Journal of Agricultural Economics*
Donovan *et al*. (2007)	*Land Economics*
Neill *et al*. (2007)	*Southern Economic Journal*
Richards *et al*. (2007)	*Journal of Agricultural and Res. Economics*
Anselin *et al*. (2008)	*World Bank Working Paper*
Anselin and Lozano-Gracia (2008)	*Empirical Economics*

Table 26.3 Spatial dependence: error

Article	*Source*
Pace and Gilley (1997)	*Journal of Real Estate Finance and Economics*
Bell and Bockstael (2000)	*Review of Economics and Statistics*
Legget and Bockstael (2000)	*Journal of Env. Economics and Management*
Beron *et al*. (2004)	*Advances in Spatial Econometrics*
Brasington (2004)	*Journal of Real Estate Finance and Economics*
Case *et al*. (2004)	*Journal of Real Estate Finance and Economics*
Rodríguez and Targa (2004)	*Transport Reviews*
Boxall *et al*. (2005)	*Resource and Energy Economics*
Brasington and Hite (2005)	*Regional Science and Urban Economics*
Rogers (2006)	*Land Economics*
Day *et al*. (2007)	*Environmental and Resource Economics*
Donovan *et al*. (2007)	*Land Economics*
Hui *et al*. (2007)	*Building and Environment*
Munroe (2007)	*Environment and Planning B: Planning and Design*
Neill *et al*. (2007)	*Southern Economic Journal*
Noonan (2007)	*Economic Development Quarterly*
Osland *et al*. (2007)	*Journal of Real Estate Research*
Richards *et al*. (2007)	*Journal of Agricultural and Resource Economics*

estimation (e.g., Kim *et al*., 2003; Brasington, 2004; Capozza *et al*., 2005; Hunt *et al*., 2005; Armstrong and Rodríguez, 2006; Hui *et al*., 2007), although alternative methods are being increasingly considered as well. For example, general method of moments estimators were applied in Bell and Bockstael (2000), Legget and Bockstael (2000), Anselin and Le Gallo (2006), Munroe (2007), Anselin *et al*.

(2008) and Anselin and Lozano-Gracia (2008), among others. A comparative perspective is offered in Bell and Bockstael (2000), using about 1,000 observations on parcel data for Anne Arundel County, Maryland. Overall, the average difference between estimates for the characteristics based on ML and GM methods is less than 5%. However, the estimates for the (error) spatial autoregressive coefficient differ between 10% and 29%.

In general, studies using a spatial econometric approach show significant differences in the estimates of marginal prices, in particular when employing a spatial lag specification. For example, Pace and Gilley (1997) find that the simultaneous spatial autoregressive (SAR) model ML estimator obtains a much better fit than the OLS estimator using data from 506 sales in Boston SMSA (Standard Metropolitan Statistical Area). Pace and Gilley (1998) also compare a SAR with OLS and a grid adjustment model and conclude that, by going from the OLS to the SAR specification, the estimated residuals fall by 44%.

For a spatial error model, the issue is not consistency of the estimates, but precision. Here again, a spatial approach seems to pay off. For example, Legget and Bockstael (2000), using data on coastal properties from Ann Arundel County in Maryland, suggest that the significance of the coefficients improves considerably. This allows them to confirm the relationship between residential prices and water quality with more confidence relative to the estimates obtained from OLS.

An important sub-set of empirical studies focuses on the way in which environmental quality, and air quality in particular, becomes capitalized into the house price. Considerable differences between the results of spatial and non-spatial estimates are observed in these studies as well. For example, Kim *et al*. (2003) found that the marginal price for air quality estimated using spatial 2SLS for a spatial lag model was half the size of the estimate obtained through OLS. Using a survey of 609 owner-occupied households in Seoul, Korea, they estimated a hedonic price equation in which air pollution is introduced as NO_x and SO_2, obtained from readings from 20 monitoring stations. The air pollution measures were interpolated to allocate a value to each of 78 residential sub-districts. An important contribution made in Kim *et al*. (2003) is to spell out the estimation of the marginal benefit in a spatial lag model. They note that it does not only include the direct effect seen in traditional OLS applications, but also a so-called *spatial multiplier* effect that captures the "induced effects of a neighborhood's housing characteristics change."

Beron *et al*. (2004) go one step further and consider the welfare effects of non-marginal changes in air pollution by estimating the second stage of the hedonic model. They compare the welfare estimates from a 10% reduction in air pollution between a standard regression using OLS and SAR-based models. Using single-family home sales records for four counties in the South Coast Air Basin (Los Angeles, Orange, San Bernardino and Riverside Counties), they estimate a semi-log form of the hedonic price equation for six different years (1980, 1983, 1986, 1989, 1992 and 1995). Air quality is measured as the annual average of PM10 (airborne particulate matter resulting from the burning of fossil fuels, such as petrol in cars) at each of 40 monitoring stations, and interpolated using the geostatistical kriging technique. Interestingly, an additional random resampling is carried

out to eliminate the effects of spatial autocorrelation, which reduces the ultimate sample size to 51,110 observations. The WTP estimates obtained from the different models ranged between $15,719 and $34,154 for the SAR model and between $15,639 and $30,489 for the OLS-based models. Beron *et al.* (2004) suggest that, even though a spatial model for the first stage of the hedonic model provides a statistically superior specification, it does not reduce the variability seen in estimated benefits derived from different model specifications. While it is clear that both spatial heterogeneity and spatial dependence violate the assumption of spherical error terms, the implications for the empirical results of the hedonic model of taking these misspecifications into account may vary from one case to the other.

The wide range seen in the estimated WTP for reductions in air quality calls for further research in this area. Additional insight remains to be gained into the sensitivity of the WTP to different specifications. In addition, many studies do not obtain standard errors for the calculated welfare effects, which makes meaningful comparisons difficult.

Further attention to non-marginal changes in environmental quality is given in Brasington and Hite (2005), who found that the demand for environmental quality is considerably more inelastic in spatial models than non-spatial models suggest. They use house sales data in Ohio for 1991, aggregated to the census block group (CBG) level, to estimate a hedonic specification that considers the presence of spatial autocorrelation in both first and second stages of the model.[3] The original number of house transactions considered is 44,255. However, since the study is carried out at the CBG level, the effective sample size consists of the number of CBG in the study, 5,051. Brasington and Hite estimate a model that includes both a spatial lag and a spatially correlated error term, with the environmental variable measured as the distance to the nearest hazard.[3]

They show that the explanatory power of the spatial model is higher and, most importantly, this model suggests a more inelastic demand function than the other specifications. For example, using the non-spatial model, the estimated consumer surplus loss from a decrease of half a mile in the median distance to a hazard is $2,276 per household. In contrast, the spatial model gives an estimate of $3,278. These results suggest that ignoring spatial characteristics may lead to underestimating the consumer surplus loss of a reduction in environmental quality. However, since no standard errors are reported, it is difficult to assess the significance of the difference in estimates.

Other recent examples of spatial econometric hedonic applications include Anselin and Le Gallo (2006), Donovan *et al.* (2007), Hui *et al.*, (2007), Munroe (2007), Anselin and Lozano-Gracia (2008) and Anselin *et al.* (2008), among others.

Anselin and Le Gallo (2006) point out that the spatial interpolation method used to create some of the explanatory variables included in hedonic models determines to a great extent the estimates of marginal prices. They recommend the kriging interpolator as the preferred method. Donovan *et al.* (2007) use a spatial lag model to estimate the effect of wildfire risk on house prices. This study follows an interesting approach to define the weights matrix, in which a correlogram is used to determine the extent of spatial correlation. They conclude that ignoring spatial

autocorrelation leads to biases for the estimated risk measures that range between 37% and 167%.

Munroe (2007) combines geovisualization of single-family residential site prices, exploration of univariate and bivariate measures of spatial autocorrelation, and spatial econometric estimation of a hedonic model, to identify factors that are likely to affect the land value in Mecklenburg County, North Carolina.

Finally, Anselin and Lozano-Gracia (2008) consider the endogeneity from an errors in variable problem of the interpolated air pollution variable. Furthermore, they provide the first empirical application of the HAC estimator suggested in Kelejian and Prucha (2007). By reporting standard errors and 95% confidence intervals, they compare the estimates from spatial and non-spatial models. Although statistical differences are mainly seen between estimates that ignore (or not) endogeneity, they point out that the spatial models allow for a distinction between direct and multiplier effects in the estimation of benefits associated with marginal changes in house characteristics, which is not possible in the standard non-spatial specification.

An interesting set of studies has focused on comparing the performance of geostatistical and lattice spatial econometric models, with particular attention given to the estimates in the hedonic price equation and out of sample prediction. For example, Miltino *et al*. (2004) compare a conditional autoregressive model (CAR) and a SAR model with a geostatistical model and a linear mixed effects model. The parameters of these four specifications are estimated using a relatively small sample, consisting of 293 dwelling sales in Pamplona, Spain. Coefficient estimates are very similar across models. Using AIC and BIC information criteria, the CAR model seems to be a better alternative than the SAR model, but differences do not appear to be significant. Miltino *et al*. (2004) suggest that using a linear mixed effect model may be a better alternative, because this type of model avoids the problems associated with the selection of an appropriate weights matrix. A similar comparison for a larger dataset remains to be carried out.

Case *et al*. (2004) find superior out of sample prediction performance for three spatial models relative to OLS using 50,000 house sales observations for Fairfax County, Virginia. Bourassa *et al*. (2007) extend this work by also including an OLS model that includes dummy variables for different sub-markets. This is assessed for 4,880 house sales from Auckland, New Zealand. Specifically, they compare the performance of a geostatistical model, lattice spatial models, and a standard model with dummy variables (spatial fixed effects) for mass appraisal purposes. The addition of sub-market dummy variables seems to outperfom both geostatistical as well as lattice models in terms of out of sample predictions. However, these conclusions may need to be put into perspective, since the sub-market models considered did not include spatial effects (see also section 26.6). To some extent then, the spatial and dummy variable specifications are not directly comparable. Also, for lattice models, the notion of out-of-sample prediction is complicated when some observations are removed from the original dataset. Since this alters the specification of the weights matrix, additional uncertainty is introduced, which needs to be taken into account (see also section 26.4.1.2).

Overall, there is ample empirical evidence that properly accounting for spatial dependence by means of spatial econometric methods pays off in terms of superior estimates of marginal price coefficients as well as welfare effects. However, several methodological issues still require further investigation.

26.6 Empirical evidence: spatial heterogeneity

Following the same organization as the methodological discussion, spatial heterogeneity can be classified as either discrete or continuous. In this review of some of the empirical literature, we follow the same distinction. An overview of the empirical applications considered is given in Table 26.4.

26.6.1 Discrete heterogeneity: sub-markets

Heterogeneity in the determination of house prices across spatial sub-units of the market is commonly approached by considering this as a special case of market segmentation (see Stevenson, 2004, for a review). This segmentation may result from the existence of market barriers or other types of market imperfections across space.

Table 26.4 Spatial Heterogeneity

Article	*Source*
Goodman (1981)	*Journal of Regional Science*
Can (1990)	*Economic Geography*
Can (1992)	*Regional Science and Urban Economics*
Allen *et al*. (1995)	*Journal of Real Estate Finance and Economics*
Goetzmann and Spiegel (1997)	*Journal of Real Estate Finance and Economics*
Goodman and Thibodeau (1998)	*Journal of Housing Economics*
Bourassa *et al*. (1999)	*Journal of Housing Economics*
Pavlov (2000)	*Real Estate Economics*
Fotheringham *et al*. (2002)	*Geographically Weighted Regression*
Bourassa *et al*. (2003)	*Journal of Housing Economics*
Fik *et al*. (2003)	*Real Estate Economics*
Gelfand *et al*. (2003)	*Journal of the American Statistical Association*
Goodman and Thibodeau (2003)	*Journal of Housing Economics*
Theriault *et al*. (2003)	*Property Management*
Day *et al*. (2004)	*CSERGE Working Paper*
Ugarte *et al*. (2004)	*Spatial and Spatiotemporal Econometrics*
Brasington and Hite (2005)	*Regional Science and Urban Economics*
Cho *et al*. (2006)	*Journal of Agricultural and Resource Economics*
Farber and Yeates (2006)	*Canadian Journal of Regional Science*
Kestens *et al*. (2006)	*Journal of Geographical Systems*
Bitter *et al*. (2007)	*Journal of Geographical Systems*
Bourassa *et al*. (2007)	*Journal of Real Estate Finance and Economics*
Long *et al*. (2007)	*Working Paper: Center for Spatial Analysis McMaster University, Hamilton, ON*
Bourassa *et al*. (2008)	*Swiss Finance Institute Research Paper*
Wang *et al*. (2008)	*Environment and Planning A*

The main methodological problem is then how to delineate the submarkets. In the literature, this has been approached in a number of different ways. Examples include the use of political boundaries, such as census tract, zip code zone or county boundaries (see, e.g. Goodman, 1981; Goetzmann and Spiegel, 1997; Brasington and Hite, 2005), or subjectively determined areas defined by real estate agents or appraisers (e.g., Bourassa *et al*., 2003, 2007). Alternatives rely on the application of statistical techniques, such as principal component and cluster analysis (Bourassa *et al*., 1999, 2003, 2008), including model-based clustering (Day *et al*., 2004), hierarchical models (Goodman and Thibodeau, 1998, 2003), and mixtures of linear models (Ugarte *et al*., 2004). In addition to geographical boundaries, physical characteristics of a property, such as the number of rooms, the lot and floor area, and the type of property, have also been used to define sub-markets.

In practice, most attempts to account for this type of spatial heterogeneity include dummy variables for each sub-market in the hedonic specification, rather than estimating a separate hedonic equilibrium for each sub-market (e.g., Bourassa *et al*., 2007). The inclusion of sub-market dummies may improve the predictive power of the model, which is usually the stated objective. However, it does not fully account for parameter heterogeneity, which may be important for identification purposes.

Representative examples of the prior selection of sub-markets are the work of Bourassa and co-authors. For example, in Bourassa *et al*. (2003), sub-markets defined by real estate agents are compared to those derived from the application of principal components and cluster analysis for 8,421 house sales transactions in Auckland, New Zealand, in 1996. Specifically, factor scores obtained using principal components are used in *k*-means cluster analysis, which results in sub-markets that do not impose contiguity. In terms of out of sample prediction, the geographically defined sub-markets used by appraisers outperform those based on statistical criteria. Also, for mass appraisal purposes, urban boundaries seem to be better as definitions of spatial sub-markets.

Further comparative evidence is provided in Bourassa *et al*. (2008), where alternative specifications of sub-markets are evaluated using a sample of 13,000 housing transactions for Louisville, Kentucky. Districts defined by local property tax assessment, as well as a classification of census tracts generated by principal components and cluster analysis, are used to derive sub-markets. For the purposes of mass appraisal, both the use of sub-market dummy variables as well as geostatistical methods increase the predictive accuracy of hedonic models.

In contrast to this spatial fixed effects approach, Allen *et al*. (1995) suggest that the differences between the sub-markets identified may be treated as random effects. The specific application is a study of an aggregate residential rental market. Through this approach, the authors allow for the possibility of individuals considering more than one property type in their choice set, while still considering an aggregated rental market, instead of modeling each sub-market independently.

Goodman and Thibodeau (1998), on the other hand, suggest that submarkets should not be imposed but specified explicitly using a hierarchical approach. For example, they use 28,939 single-family property transactions in Dallas, Texas,

between 1995 and 1997 and derive aggregates of school zones as sub-markets. The procedure starts by estimating a hierarchical model for two adjacent school zones. Then, if the coefficient associated with the sub-market is significant, those school zones are considered to pertain to different sub-markets. If, on the other hand, the coefficients are not significant, the two zones are merged. One by one, each school zone is added until all zones have been included. To avoid sub-market definitions that are path dependent, sensitivity checks are included of how the final sub-market definition depends on the starting point.

Comparative evidence is provided in the evaluation of the hierarchical approach to a sub-market definition using zip codes and census tracts in Goodman and Thibodeau (2003). In terms of prediction accuracy, the Goodman and Thibodeau (1998) hierarchical approach outperforms the other two methods.

In more recent work, Ugarte *et al*. (2004) propose the use of a mixture of linear models. This first provides a classification of the observations into groups (sub-markets), and then estimates the parameters for the hedonic price equilibrium in each group. The data are allowed to determine the group structure and coefficients are estimated jointly. A linear mixed model with random effects is estimated by means of nonparametric ML. However, due to the small number of properties considered (293 in Pamplona, Spain), the degree of generality of this approach remains a topic for further research.

In the empirical literature, the discussion of sub-market definition and evaluation of the performance of different methods has tended to focus primarily on the out-of-sample predictive precision. The effect of different sub-market definitions on the substantive interpretation of the value and MWTP of specific house (and environmental) characteristics largely remains to be investigated.

26.6.2 Continuous spatial heterogeneity: spatially varying coefficients

Early efforts to allow for continuous variation of the parameters in a hedonic specification were based on applications of Casetti's expansion method. This was first implemented by Can (1990) as a way to address spatial heterogeneity in house prices at the neighborhood level. In her specification, the marginal price of housing attributes are assumed to vary as a function of neighborhood characteristics. Specifically, Can defined a composite neighborhood quality (NQ) index and introduced linear and quadratic functions of NQ as the expansion equation. The hedonic price equilibrium is then specified as:

$$P = \alpha + \sum_k \beta_k S_k + \varepsilon, \tag{26.43}$$

where the S_k are the structural characteristics of the house and β_k is a function of a neighborhood variable NQ. For example, for the linear case:

$$\beta_k = \beta_{k0} + \beta_{k1} NQ + u. \tag{26.44}$$

Other examples of applications of the expansion method to hedonic specifications can be found in Can (1992), Theriault *et al*. (2003), Fik *et al*. (2003) and

Kestens *et al*. (2006). However, their feasibility in practice is typically limited due to the severe problems of multicollinearity that follow from the interaction terms.

Continuous spatial heterogeneity as implemented in GWR has seen several applications to hedonic specifications. Fotheringham's text contains multiple examples of hedonic price models (Fotheringham *et al*., 2002), but others have applied this methodology as well, such as Cho *et al*. (2006), Kestens *et al*. (2006), Farber and Yeates (2006), Long *et al*. (2007), Bitter *et al*. (2007) and Wang *et al*. (2008). Related applications to repeat house sales models include McMillen (2003, 2004).

Pavlov (2000) suggests that the spatial varying coefficient model forms an alternative approach to dealing with sub-markets that outperforms several competing specifications in terms of cross-validation residuals. The models evaluated include a standard linear regression and a linear regression that includes dummy variables for zip codes, as well as a parametric model including a quadratic polynomial of the X, Y coordinates of the points in the data.

Other applications focus more on the parameter instability related to specific characteristics in the hedonic specification, such as environmental quality. For example, Cho *et al*. (2006) investigate whether public open space is capitalized into nearby residential property values. They use an original dataset that includes over 22,000 single-family housing sales transactions between 1998 and 2002 in Knox County, Tennessee. Of this total sample, 15,500 transactions were randomly selected for analysis. GWR estimation of a hedonic specification that includes proximity to water bodies and parks suggests considerable variation in the marginal prices of both amenities along different regions of the county. The resulting local marginal prices would be obscured in a global model that assumes a constant marginal price across the whole region.

A number of studies have compared the performance of the parametric expansion method to the nonparametric GWR. For example, using data on 761 single-family properties sold between 1993 and 2001 in Quebec City, Canada, as well as information on the profiles of buyers, Kestens *et al*. (2006) compare the results from a spatial expansion method and GWR. They also suggest that introducing detailed household-profile data into the hedonic specification helps in explaining spatial heterogeneity while at the same time reducing spatial dependence. Household income, previous tenure status and age of the buyer show a significant effect on house prices.

A similar comparison is carried out by Farber and Yeates (2006), who consider global specifications, such as a standard linear and a SAR model (both estimated using OLS) relative to local models such as GWR. Using an adjusted version of R^2 as the criterion, they conclude that GWR obtains the best fit. Similarly, Bitter *et al*. (2007), using data for over 10,000 single-family house sales in Tucson, Arizona, find that GWR outperforms the expansion method in terms of both explanatory power and predictive accuracy.[4] On the other hand, Kestens *et al*. (2006) suggest that, in their application, GWR and the expansion method have similar explanatory power.

Long *et al*. (2007) assess the difference in predictive accuracy between moving windows regression (MWR), GWR, kriging, and moving windows kriging (MWK)

using data on 33,494 transactions of single-family detached houses sold between January 2001 and December 2003 in the City of Toronto; 10% of this sample is taken out for out-of-sample validation. The MWR is an alternative to the locally weighted regression (LWR) (and therefore GWR) in which a window (neighborhood) is defined at every location. All points inside the window, and only those points, are used to estimate the parameters for each observation. Therefore, in contrast to LWR/GWR, a constant weight is given to all observations within the window, ignoring observations outside the window. Similarly, MWK uses a local spatial covariance structure that varies at every point in the sample (Haas, 1990). In this example, MWR and GWR provide the most accurate results in terms of prediction, while MWK produces very poor results in terms of out-of-sample prediction.

Overall, the explicit incorporation of spatial heterogeneity in hedonic specifications illustrates the need to better understand the nature of market segmentation and the complex interactions between location and the value of individual house characteristics.

26.7 Concluding remarks

The contribution of spatial econometrics to hedonic analysis is not limited to improving the quality and precision of the estimates obtained, as reviewed in the previous sections. Spatial econometric methods also provide additional insight for policy analysis. In this concluding section, we focus on two aspects in particular, the notion of the spatial multiplier and its implications for the interpretation of welfare effects, and the use of spatially explicit simulations to assess the impact of non-marginal changes in characteristics.

As outlined in detail in section 26.2.1, the marginal implicit price derived from the hedonic price equilibrium may be interpreted as a measure of a household's marginal utility. Therefore, the derivative of the hedonic price equilibrium equation with respect to the characteristic of interest forms the basis for the estimation of MWTP.

In a non-spatial log-linear model, this MWTP equals the estimated coefficient for the characteristic of interest z_k times the price (P), or:

$$\widehat{MWTP}_{z_k} = \frac{\partial P}{\partial z_k} = \hat{\beta}_{z_k} P. \tag{26.45}$$

As shown in Kim *et al*. (2003), in a spatial lag model this is no longer the case. Instead, a spatial multiplier effect needs to be accounted for to accurately compute the MWTP. Specifically, in the case of a uniform change in the amenity across all observations, the MWTP can be shown to be:

$$\widehat{MWTP} = \hat{\beta}_{z_k} P \left(\frac{1}{1 - \widehat{\rho}} \right), \tag{26.46}$$

with $\widehat{\rho}$ as the estimate of the spatial autoregressive coefficient.

The distinction between equations (26.45) and (26.46) is important in light of the points recently raised by Small and Steimetz (2006). They considered a different interpretation of welfare effects between the *direct* valuation in equation (26.45) and the multiplier effect included in (26.46). In their view, the multiplier effect should only be considered as part of the welfare calculation in the case of a technological externality associated with a change in amenities. In the case of a purely pecuniary externality, the direct effect is the only correct measure of welfare change. Only in a spatial lag specification is it possible to distinguish between these two effects.

The use of an analytical approach to deriving MWTP in impact analysis is limited in a number of ways. It is constrained by the functional specification of the hedonic model. Specifically, if nonlinearities are introduced in the hedonic model, such nonlinearities will be transfered to possible dependencies and/or nonlinearities in the MWTP itself. Also, the analytical derivation of MWTP is limited to marginal changes in the characteristics. For non-marginal changes, the inverse demand function needs to be estimated, which is often difficult in practice. Moreover, nonlinear specifications require a value for the price and/or characteristics in the calculation, which is typically simplified by using a mean or median value.

An alternative to the analytical derivation is to use a simulation approach, as outlined in Anselin and Le Gallo (2006). The essence of the approach is that valuation is based on the computation of predicted values for individual observations given their actual household characteristics. In essence this boils down to a discrete approximation to the notion of marginal willingness to pay. A major advantage of the simulation approach is that it allows greater flexibility, both in the specification of the type of policy experiment as well as in the valuation. Since the valuation is computed for individual house observations, the results can be obtained for any desired level of spatial aggregation, such as by county or zip code (for an extensive application, see Anselin *et al.*, 2008).

In a non-spatial linear model, the change in predicted values can be expressed as follows:

$$\hat{p}_0 - \hat{p}_{-k} = (z_0 - z_{-k})\beta_k, \tag{26.47}$$

where $\hat{p}_0 - \hat{p}_{-k}$ is the change in valuation and $(z_0 - z_{-k})$ is the change in characteristic k. If a spatially lagged dependent variable is included in the hedonic price equilibrium, the change in valuation resulting from a change in characteristic k would turn out to be $(I - \hat{\rho}W)^{-1}(z_0 - z_{-k})\beta_k$. Any nonlinearities of the hedonic price function would also be reflected in this approximation to the change in valuation.

By ignoring this multiplier effect and looking at individual or private benefits only, underestimation of the overall social welfare from changes in house characteristics may result, such as reductions in air pollution, greater access to public facilities, and public services. In decision making under strict efficiency rules, this may lead to an underinvestment in such characteristics.

In sum, a spatial econometric approach yields more efficient estimates of policy effects of interest, allows for a distinction between direct and indirect welfare effects

and forms the basis for flexible policy simulation experiments. While the field has made significant progress to date, several important methodological issues remain to be addressed satisfactorily. Foremost among these are the treatment of spatial scale, endogeneity and sub-market heterogeneity. It is hoped that the current chapter has raised awareness of these issues and will stimulate further progress.

Notes

1. For a more extensive discussion, see Anselin (2002, pp. 256–60) and Anselin (2006, pp. 909–10).
2. Commonly used theoretical specifications include the negative exponential, spherical and Gaussian semi-variograms. A detailed discussion of specific functional forms is given in Dubin *et al*. (1999).
3. Note that, rather than estimating the more traditional inverse demand function, Brasington and Hite (2005) estimate a direct demand function in the second stage of the hedonic model.
4. Bitter *et al*. (2007) also introduce a spatially lagged variable into the GWR specification, but it is doubtful that the OLS estimation procedure used would yield consistent estimates of this lag term.

References

Allen, M.T., T.M. Springer and N.G. Waller (1995) Implicit pricing across residential rental submarkets. *Journal of Real Estate Finance and Economics* **11**, 137–51.

Andrews, D.W. (1991) Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–58.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990) Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* **30**, 185–207.

Anselin, L. (1992) Spatial dependence and spatial heterogeneity: model specification issues in the spatial expansion paradigm. In J.P. Jones and E. Casetti (eds.), *Applications of the Expansion Method*, pp. 334–54. London: Routledge.

Anselin, L. (1998) GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research* **9**(1), 113–33.

Anselin, L. (2001a) Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference* **97**, 113–39.

Anselin, L. (2001b) Spatial effects in econometric practice in environmental and resource economics. *American Journal of Agricultural Economics* **83**(3):705–10.

Anselin, L. (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* **27**(3), 247–67.

Anselin, L. (2006) Spatial econometrics. In T. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*, pp. 901–69. Basingstoke: Palgrave Macmillan.

Anselin, L. and A. Bera (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah and D.E. Giles (eds.), *Handbook of Applied Economic Statistics*, pp. 237–89. New York: Marcel Dekker.

Anselin, L. and D.A. Griffith (1988) Do spatial effects really matter in regression analysis. *Papers, Regional Science Association* **65**, 11–34.

Anselin, L. and J. Le Gallo (2006) Interpolation of air quality measures in hedonic house price models: spatial aspects. *Spatial Economic Analysis* **1**, 31–52.

Anselin, L. and N. Lozano-Gracia (2008) Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics* **34**, 5–34.

Anselin, L., N. Lozano-Gracia, U. Deichmann and S.V. Lall (2008) Valuing access to water – a spatial hedonic approach applied to Indian cities. Working Paper No. 4533, World Bank, Washington, DC.

Armstrong, R. and D. Rodríguez (2006) An evaluation of the accessibility benefits of commuter rail in eastern Massachusetts using spatial hedonic price functions. *Transportation* **33**, 21–43.

Basu, S. and T.G. Thibodeau (1998) Analysis of spatial autocorrelation in house prices. *Journal of Real Estate Finance and Economics* **170**(1), 61–85.

Baumont, C. (2004) Spatial effects in housing price models: do housing prices capitalize urban development policies in the agglomeration of Dijon (1999)? Working Paper, Université de Bourgogne.

Bayer, P., N. Keohane and C. Timmins (2006) Migration and hedonic valuation: the case of air quality. Working Paper No. 12106, National Bureau of Economic Research.

Bell, K. and N. Bockstael (2000) Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *Review of Economics and Statistics* **82**(1): 72–82.

Beron, K.J., Y. Hanson, J.C. Murdoch and M.A. Thayer (2004) Hedonic price functions and spatial dependence: implications for the demand for urban air quality. In L. Anselin, R.J. Florax and S.J. Rey (eds.), *Advances in Spatial Econometrics: Methodology, Tools and Applications*, pp. 267–81. Berlin: Springer-Verlag.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36**(2), 192–236.

Bitter, C., G. Mulligan and S. Dallérba (2007) Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems* **9**, 7–27.

Bourassa, S., E. Cantoli and M. Hoesli (2007) Spatial dependence, housing submarkets, and house price prediction. *Journal of Housing Economics* **12**, 12–28.

Bourassa, S.C., E. Cantoni and M. Hoesli (2008) Predicting house prices with spatial dependence: impacts of alternative submarket definitions. Research Paper 08-01, Swiss Finance Institute.

Bourassa, S., F. Hamelink, M. Hoesli and B. MacGregor (1999) Defining residential submarkets. *Journal of Housing Economics* **8**, 160–83.

Bourassa, S., M. Hoesli and V. Peng (2003) Do housing submarkets really matter. *Journal of Real Estate Finanace and Economics* **35**, 143–60.

Bowen, W., B.A. Mikelbank and D.M. Prestegaard (2001) Theoretical and empirical considerations regarding space in hedonic housing price model applications. *Growth and Change* **32**(4), 466–90.

Boxall, P., W. Chan and M. McMillan (2005) The impact of oil and natural gas facilities on rural residential property values: a spatial hedonic analysis. *Resource and Energy Economics* **27**, 248–69.

Brasington, D.M. (2004) House prices and the structure of local government: an application of spatial statistics. *Journal of Real Estate Finance and Economics* **29**(2), 211–31.

Brasington, D.M. and D. Hite (2005) Demand for environmental quality: a spatial hedonic analysis. *Regional Science and Urban Economics* **35**, 57–82.

Brueckner, J.K. (2003) Strategic interaction among governments: an overview of empirical studies. *International Regional Science Review* **26**(2), 175–88.

Brunsdon, C., A. Fotheringham and M. Charlton (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* **28**, 281–98.

Can, A. (1990) The measurement of neighborhood dynamics in urban house prices. *Economic Geography* **66**, 254–72.

Can, A. (1992) Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics* **22**, 453–74.

Capozza, D., R. Israelsen and T. Thomson (2005) Appraisal, agency and atypicality: evidence from manufactured homes. *Real Estate Economics* **33**(3), 509–37.

Case, B., J. Clapp, R. Dubin and M. Rodriguez (2004) Modeling spatial and temporal house price patterns: a comparison of four models. *Journal of Real Estate Finance and Economics* **29**(2), 167–91.

Casetti, E. (1972) Generating models by the expansion method: applications to geographical research. *Geographical Analysis* **4**, 81–91.

Casetti, E. (1997) The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review* **20**, 9–33.

Chay, K.Y. and M. Greenstone (2005) Does air quality matter? Evidence from the housing market. *Journal of Political Economy* **113**(2), 376–424.

Cho, S., J. Bowker and W. Park (2006) Measuring the contribution of water and green space amenities to housing values: an application and comparison of spatially weighted hedonic models. *Journal of Agricultural and Resource Economics* **31**(3), 485–507.

Clapp, J., H.-J. Kim and A. Gelfand (2002) Predicting spatial patterns of house prices using LPR and Bayesian smoothing. *Real Estate Economics* **30**, 79–105.

Cleveland, W.S. and S.J. Devlin (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

Conley, T.G. (1999) GMM estimation with cross-sectional dependence. *Journal of Econometrics* **92**, 1–45.

Cressie, N. (1993) *Statistics for Spatial Data*. Wiley InterScience.

Cropper, M.L., L.B. Decl and K.E. McConnell (1988) On the choice of functional form for hedonic price functions. *Review of Economic and Statistics* **70**, 668–75.

Das, D., H.H. Kelejian and I.R. Prucha (2003) Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Papers in Regional Science* **82**, 1–27.

Day, B., I. Bateman and I. Lake (2004) Nonlinearity in hedonic price equations: an estimation strategy using model-based clustering. Working Paper, Centre for Social and Economic Research of the Global Environment, University of East Anglia.

Day, B., I. Bateman and I. Lake (2007) Beyond implicit prices: recovering theoretically consistent and transferable values for noise avoidance from a hedonic property price model. *Environmental and Resource Economics* **37**, 211–32.

Deaton, A. and J. Muellbauer (1980) *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Donovan, G.H., P.A. Champ and D.T. Butry (2007) Wildfire risk and housing prices: a case study from Colorado Springs. *Land Economics* **83**(3), 217–33.

Dubin, R.A. (1988) Estimation of regression coefficients in the presence of spatially autocorrelated errors. *Review of Economics and Statistics* **70**, 466–74.

Dubin, R.A. (1992) Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics* **22**, 433–52.

Dubin, R.A. (1998) Predicting house prices using multiple listings data. *Journal of Real Estate Finance and Economics* **17**(1), 35–59.

Dubin, R., R. Kelley Pace and T.G. Thibodeau (1999) Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature* **7**, 79–95.

Ekeland, I., J.J. Heckman and L. Nesheim (2004) Identification and estimation of hedonic models. *Journal of Political Economy* **112**(1), S60–109.

Farber, S. and M. Yeates (2006) A comparison of localized regression models in a hedonic house price context. *Canadian Journal of Regional Science* **29**(3), 405–19.

Fik, T., D. Ling and G. Mulligan (2003) Modeling spatial variation in housing prices: a variable interaction approach. *Real Estate Economics* **31**(4), 623–46.

Florax, R.J., H. Folmer and S.J. Rey (2003) Specification searches in spatial econometrics: the relevance of Hendry's methodology. *Regional Science and Urban Economics* **33**(5), 557–79.

Fotheringham, A., C. Brunsdon and M. Charlton (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A* **30**, 1905–27.

Fotheringham, A., C. Brunsdon and M. Charlton (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationship*. Chichester: John Wiley.
Freeman, A.M.I. (1999) *The Measurement of Environmental and Resource Values*. Washington, DC: Resources For the Future.
Gawande, K. and H. Jenkins-Smith (2001) Nuclear waste transport and residential property values: estimating the effects of perceived risks. *Journal of Environmental Economics and Management* **42**, 207–33.
Gelfand, A.E., M.D. Ecker, J.R. Knight and C.F. Sirmans (2004) The dynamics of location in home price. *Journal of Real Estate Finance and Economics* **29**(2), 149–66.
Gelfand, A.E., K. Hyon-Jung, C. Sirmans and S. Banerjee (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association* **98**(462), 387–97.
Gibbons, S. (2003) Paying for good neighbours: estimating the value of an implied educated community. *Urban Studies* **40**(4), 809–33.
Gillen, K., T.G. Thibodeau and S. Wachter (2001) Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics* **23**(1), 5–30.
Goetzmann, W.N. and M. Spiegel (1997) A spatial model of housing returns and neighborhood substitutability. *Journal of Real Estate Finance and Economics* **14**, 11–31.
Goldstein, H. (1995) *Multilevel Statistical Models* (second edition). London: Edward Arnold.
Goodman, A.C. (1981) Housing submarket within urban areas: definitions and evidence. *Journal of Regional Science* **21**(2), 175–85.
Goodman, A.C. and T.G. Thibodeau (1998) Housing market segmentation. *Journal of Housing Economics* **7**, 121–43.
Goodman, A.C. and T.G. Thibodeau (2003) Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics* **12**, 181–201.
Gotway, C.A. and L.J. Young (2002) Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632–48.
Haas, T.C. (1990) Kriging and automated variogram modeling within aÊ moving window. *Atmospheric Environment Part A – General Topics* **24**(7), 1759–69.
Hsieh, W., E. Irwin and D. Forster (2001) Evidence of county-level urbanization spillovers from a space-time model of land use change. Working Paper, Department of Agricultural Economics, Ohio State University, Columbus.
Huang, H., G. Miller, B. Sherrick and M. Gomez (2006) Factors influencing Illinois farmland values. *American Journal of Agricultural Economics* **88**(2), 458–70.
Hui, E., C. Chau, L. Pun and M. Law (2007) Measuring the neighboring and environmental effects on residential property value: using spatial weighting matrix. *Building and Environment* **42**, 2333–43.
Hunt, L., P. Boxall, J. Englin and W. Haider (2005) Remote tourism and forest management: a spatial hedonic analysis. *Ecological Economics* **53**, 101–13.
Judge, G., R. Hill, W. Griffiths, H. Lutkephol and T. Lee (1988) *Introduction to the Theory and Practice of Econometrics* (second edition). New York: John Wiley.
Kelejian, H.H. and I.R. Prucha (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* **17**(1), 99–121.
Kelejian, H.H. and I.R. Prucha (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* **40**(2), 509–33.
Kelejian, H.H. and I.R. Prucha (2006) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Working Paper, Department of Economics, University of Maryland, College Park.
Kelejian, H.H. and I.R. Prucha (2007) HAC estimation in a spatial framework. *Journal of Econometrics* **140**(1), 131–54.
Kelejian, H.H., I.R. Prucha and Y. Yuzefovich (2004) Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: large and small sample results.

In J.P. LeSage and R. Kelley Pace (eds.), *Advances in Econometrics. Volume 18: Spatial and Spatiotemporal Econometrics*, pp. 163–98. Oxford: Elsevier Science.
Kelejian, H.H. and D.P. Robinson (1992) Spatial autocorrelation: a new computationally simple test with an application to per capita country police expenditures. *Regional Science and Urban Economics* **22**, 317–33.
Kelejian, H.H. and D.P. Robinson (1993) A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* **72**, 297–312.
Kestens, Y., M. Thériault and F.D. Rosiers (2006) Heterogeneity in hedonic modelling of house prices: looking at buyers household profiles. *Journal of Geographical Systems* **8**, 61–96.
Kim, C.W., T. Phipps and L. Anselin (2003) Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of Environmental Economics and Management* **45**, 24–39.
Lahiri, S. (1996) On the inconsistency of estimators under infill asymptotics for spatial data. *Sankhya A* **58**, 403–17.
Lancaster, K.J. (1966) A new approach to consumer theory. *Journal of Political Economy* **74**, 132–56.
Lee, L.-F. (2003) Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews* **22**, 307–35.
Lee, L.-F. (2007) GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* **137**, 489–14.
Legget, C.G. and N.E. Bockstael (2000) Evidence of the effects of water quality on residential land prices. *Journal of Environmental Economics and Management* **39**, 124–44.
LeSage, J.P. (1997) Bayesian estimation of spatial autoregressive models. *International Regional Science Review* **20**, 113–29.
Lin, X. and L.-F. Lee (2005) GMM estimation of spatial autoregressive models with unknown heteroskedasticity. Working Paper, Ohio State University, Columbus.
Long, F., A. Paez and S. Farber (2007) Spatial effects in hedonic price estimation: a case study in the city of Toronto. Working Paper No. WP020, Center for Spatial Analysis, McMaster University, Hamilton, Ontario.
Malpezzi, S. (2002) Hedonic pricing models: a selective and applied review. In K. Gibb and A. O'Sullivan (eds.), *Housing Economics: Essays in Honour of Duncan Maclennan*, pp. 67–89. Oxford: Blackwell Science Ltd.
Manski, C.F. (1993) Identification of endogenous social effects: the reflexion problem. *Review of Economic Studies* **60**, 531–42.
Manski, C.F. (2000) Economic analysis of social interactions. *Journal of Economic Perspectives* **14**(3), 115–36.
McMillen, D.P. (2003) Neighborhood house price indexes in chicago: a Fourier repeat sales approach. *Journal of Economic Geography* **3**, 57–73.
McMillen, D.P. (2004) Employment subcenters and home price appreciation rates in metropolitan Chicago. In J.P. LeSage and R. Kelley Pace (eds.), *Advances in Econometrics. Volume 18: Spatial and Spatiotemporal Econometrics*, pp. 237–57. Oxford: Elsevier Science.
Miltino, A.F., M.D. Ugarte and L. Garcia-Reinaldos (2004) Alternative models for describing spatial dependence among dwelling selling prices. *Journal of Real Estate Finance and Economics* **29**(2), 193–209.
Moulton, B.R. (1990) An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* **72**, 334–8.
Munroe, D.K. (2007) Exploring the determinants of spatial pattern in residential land markets: amenities and disamenities in Charlotte, NC, USA. *Environment and Planning B: Planning and Design* **34**, 336–54.
Neill, H.R., D.M. Hassenzahl and D.D. Assane (2007) Estimating the effect of air quality spatial versus traditional hedonic price models. *Southern Economic Journal* **73**(4), 1088–111.
Newey, W.K. and K.D. West (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–8.

Noonan, D.S. (2007) Finding an impact of preservation policies: price effects of historic landmarks on attached homes in Chicago, 1990–9. *Economic Development Quarterly* **21**(1), 17–33.

Ord, J.K. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* **70**, 120–6.

Osland, L., I. Thorsen and J.P. Gitlesen (2007) Housing price gradients in a region with one dominating center. *Journal of Real Estate Research* **29**(3), 321–46.

Pace, R. Kelley, R. Barry, J.M. Clapp and M. Rodriguez (1998) Spatial autocorrelation and neighborhood quality. *Journal of Real Estate Finance and Economics* **17**(1), 15–33.

Pace, R. Kelley, R. Barry, O.W. Gilley and C.F. Sirmans (2000) Simple spatial-temporal forecasting with an application to real estate prices. *International Journal of Forecasting* **16**, 229–46.

Pace, R. Kelley and O.W. Gilley (1997) Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics* **14**(3), 333–40.

Pace, R. Kelley and O.W. Gilley (1998) Generalizing the OLS and the grid estimator. *Real Estate Economics* **26**, 331–47.

Pace, R. Kelley and J.P. LeSage (2004) Spatial statistics and real estate. *Journal of Real Estate Finance and Economics* **29**, 147–8.

Paez, A., T. Uchida and K. Miyamoto (2002a) A general framework for estimation and inference of geographically weighted regression models, 1: location-specific kernel bandwidths and a test for local heterogeneity. *Environment and Planning A* **34**, 733–54.

Paez, A., T. Uchida and K. Miyamoto (2002b) A general framework for estimation and inference of geographically weighted regression models, 2: spatial association and model specification tests. *Environment and Planning A* **34**, 883–904.

Palmquist, R.B. (1991) Hedonic methods. In J.B. Braden and C.D. Kolstad (eds.), *Measuring the Demand for Evironmental Quality*, pp. 77–120. Amsterdam: North-Holland.

Palmquist, R.B. (2005) Property value models. In K. Mäler and J. Vincent (eds.), *Handbook of Environmental Economics, Volume 2*, pp. 763–819. Amsterdam: North-Holland.

Pavlov, A. (2000) Space-varying regression coefficients: a semi-parametric approach applied to real estate markets. *Real Estate Economics* **28**(2), 249–83.

Richards, T.J., P.M. Patterson and S.F. Hamilton (2007) Fast food, adiction and market power. *Journal of Agricultural and Resource Economics* **32**(3), 425–47.

Robinson, P. (1988) Root-n-consistent semi-parametric regression. *Econometrica* **56**, 931–54.

Rodríguez, D. and F. Targa (2004) Value of accessibility to bogotas bus rapid transit system. *Transport Reviews* **24**(5), 587–610.

Rogers, W.H. (2006) A market for institutions: assessing the impact of restrictive covenants on housing. *Land Economics* **82**(4), 500–12.

Rosen, S.M. (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* **82**, 534–57.

Small, K.A. and S. Steimetz (2006) Spatial hedonics and the willingness to pay for residential amenities. Working Paper No. 05–06–31, University of California, Irvine.

Stevenson, S. (2004) New empirical evidence on heteroskedasticity in hedonic house models. *Journal of Housing Economics* **13**, 136–53.

Sun, H., Y. Tu and S.-M. Yu (2005) A spatio-temporal autoregressive model for multi-unit residential market analysis. *Journal of Real Estate Finance and Economics* **31**(2), 155–87.

Theriault, M., F.D. Rosiers, P. Villeneuve and Y. Kestens (2003) Modelling interactions of location with specific value of housing attributes. *Property Management* **21**(1), 25–48.

Ugarte, M.D., T. Goicoa and A.F. Miltino (2004) Searching for housing submarkets using mixtures of linear models. In J.P. LeSage and R. Kelley Pace (eds.), *Advances in Econometrics. Volume 18: Spatial and Spatiotemporal Econometrics*. pp. 259–79. Oxford: Elsevier Science.

Wang, N., C.L. Mei and X.-D. Yan (2008) Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. *Environmental and Planning A* **40**, 986–1005.

27

Spatial Analysis of Economic Convergence

Sergio J. Rey and Julie Le Gallo

Abstract

This chapter reviews some of the major econometric models, approaches and issues related to the spatial dimensions of economic convergence and inequality. Key themes concern the implications of spatial dependence (autocorrelation) and heterogeneity for the specification, estimation, and interpretation of convergence models on the one hand and, on the other, the treatment of these spatial effects in the analysis of distributional dynamics and the application of related exploratory data analysis methods. We draw linkages between recent contributions in the mainstream econometric literature and developments in spatial econometrics and regional science We identify a number of areas where cross-fertilization between these fields would be beneficial.

27.1 Introduction

The last two decades have witnessed a renaissance in the field of growth econometrics. Defined as the set of statistical tools for the study of growth (Durlauf *et al.*, 2005), growth econometrics can be organized along two dimensions. In the first, which we refer to as the regression approach, predictions from formal neoclassical (and other) growth theories have been tested using cross-sectional, time series and panel data econometric specifications. The second group of methods departs from the representative economy assumption underlying most of the regression approaches and instead examines the entire distribution of incomes. Here the focus shifts to how different parts of the distribution may behave over time, and to questions of whether there are changes in modality and shape of the distribution and concerns with the intradistributional dynamics and mixing. These methods are distinct not only in their focus, but also in the specific statistical techniques that are employed.

A prominent trend in the growth literature has been a decidedly regional turn, where the focus has shifted from cross-country analyses to examine the nature of convergence as it may operate at the sub-national scale (cf. Barro and Sala-i-Martin, 1992; Carlino and Mills, 1993; Neven and Gouyette, 1995; Sala-i-Martin, 1996; Rey and Montouri, 1999; Fingleton, 2004). Early in this regional turn, there was some appreciation for the challenges that regional data posed for the application of standard growth models. Given that regions typically display a greater deal of openness than is the case for national economies, various forms of regional interdependencies, such as labor and capital flows along with trade, take on increased importance at the finer spatial scale. Yet there was also a perception that there was a lack of econometric methods for modeling regional interdependencies. When confronted with evidence of spatial dependence and heterogeneity in regional growth sets, the strategy adopted by many researchers has been to remain in the closed-economy model but to adjust the closed model for these spatial effects.

Paradoxically, during this period of renewal and resurgence of interest in growth econometrics, there was a similar burst of activity in the fields of spatial econometrics and spatial statistics (for example, Anselin, 1988; Cressie, 1993; Anselin *et al.*, 2004). Indeed, as Arbia (2006, p. 3) has noted: "until a few years ago, spatial econometric methods were well developed in the literature but the drama was that no one used them in the mainstream applied economic analysis!"

Towards the close of the century this began to change, with a number of studies adopting methods and tools of spatial analysis to the question of regional, or sub-national, economic convergence beginning to appear in the literature. The intersection of these two literatures has generated a large, and growing, body of empirical studies, as well as the identification of a number of challenging methodological issues and some advances in the modeling and analysis of spatial growth and convergence. The growing recognition of the unique characteristics of spatial data is reflected by Durlauf *et al.* (2005), who note: "The problem of spatial correlation has been much studied in the regional science literature, and statisticians in this field have developed spatial analogues of many time series concepts ..."

While this recognition is promising, it is also indicative of the need for further interaction between the growth econometric and spatial analysis literatures. As we develop more fully in what follows, spatial correlation is not a simple spatial analogue to temporal correlation, nor are the methods that have been developed in the spatial literature simple extensions of time series methods. Moreover, spatial correlation/dependence is but one of several types of spatial effects encountered in the analysis of geographically referenced data.

This chapter explores the intersection of the growth empirics and spatial analysis literature. Our objectives for doing so are threefold. First, while there have been some efforts at providing overviews of regional convergence research, these have tended to appear in the regional science and economic geography literature (e.g., Magrini, 2004; Abreu *et al*., 2005; Rey and Janikas, 2005) and not the wider growth literature. Because of this the amount of cross-fertilization between the traditional growth econometrics literature and the spatial analysis is not as advanced as it could be. Thus we hope that by updating previous reviews we can bring this work to a wider audience.

At the same time, existing reviews have divided the literature into those studies adopting a confirmatory approach to formal growth modeling on the one hand and, on the other, the "atheoretical" exploratory literature. In our view this is an artificial distinction as we see both literatures as complementary, and thus our second objective is to explore the potential synergies between these two fields. Finally, there are a number of outstanding methodological issues that require further attention in order for the field of regional convergence analysis to move forward. We identify these and suggest an agenda for future research.

The chapter is organized as follows. Section 27.2 provides a selective survey of the treatment of space in empirical econometric work on convergence. Section 27.3 examines methods for distributional dynamics and related exploratory data analysis and their application to spatially referenced data. Section 27.4 concludes with a summary evaluation of progress to date and the identification of possible directions for future research.

27.2 Space and econometric modeling of convergence

27.2.1 Models of economic growth

The primary basis for the analysis of spatial effects in empirical convergence studies has been cross-country growth regressions, based on the seminal studies by Barro and Sala-i-Martin (1992) and Mankiw *et al*. (1992). The main prediction of the neoclassical growth models is that the growth rate of an economy is positively related to the distance that separates it from its own steady-state. Let us take as a starting point the canonical form for such regressions:

$$\frac{1}{T}\log(y_{i,t_0+T}/y_{i,t_0}) = \alpha + \beta \log(y_{i,t_0}) + X_{1,i}\delta_1 + X_{2,i}\delta_2 + \varepsilon_i, \qquad (27.1)$$

where $y_{i,t}$ is the per capita income of country or region i at time t, and $X_{1,i}$ is a set of additional structural regressors suggested by the Solow growth model (population growth, technological change and physical and human capital savings rates).

They are transformed in ways implied by the model (see Durlauf and Quah, 1999, for additional insights into this specification). $X_{2,i}$ is a set of additional control variables capturing differences in aggregate production functions and ε_i is an error term with the following properties: $\varepsilon_i \sim i.i.d.(0, \sigma_\varepsilon^2)$.

In this specification, the average growth rate in per capita income over the period t_0 to $t_0 + T$ is related to the initial level for income y_{i,t_0} and a set of steady-state determinants ($X_{1,i}$ and $X_{2,i}$). There is conditional convergence if the estimate of β is significantly negative, with a convergence speed equal to $b = -\log(1 + T\beta)/T$ and a half-life equal to $\tau = -\log(2)/\log(1+\beta)$. The concept of unconditional convergence is defined when all the economies are assumed to be structurally similar, that is, that they are characterized by the same steady-state, and differ only by their initial conditions. This assumption is tested with the cross-sectional model, including only the initial per capita income as an explanatory variable.

As frequently pointed out in the growth econometrics literature, the tests of conditional convergence face several problems, such as robustness with respect to choice of control variables, multicollinearity, heterogeneity, endogeneity, and measurement problems (Durlauf and Quah, 1999; Temple, 1999; Durlauf *et al.*, 2005). Durlauf *et al.* question the assumption usually made on the error terms. Indeed, by assuming that they are independent and identically distributed (i.i.d.), the researcher thinks of them as interchangeable across observations. This is the concept of exchangeability: "different patterns of realized errors are equally likely to occur if the realizations are permuted across countries. In other words, the information available to a researcher about the countries is not informative about the error terms" (2005, p. 36). They show that many econometric problems highlighted in growth regressions can be interpreted as violations of exchangeability. Parameter heterogeneity (discussed below) or omitted regressors induce non-exchangeability.

The assumption of constant returns to scale, on which neoclassical theory is based, has been challenged by new economic growth and new economic geography theories. Consequently, Fingleton and McCombie (1998) suggest alternative theoretical frameworks that allow increasing returns to scale, especially when one deals with regions. At the heart of this approach is Verdoorn's law, based on Kaldor's second law, which has been traditionally estimated as a linear relationship between the exponential growth rate of labor productivity (p) and output (q):

$$p = b_0 + b_1 q + \varepsilon, \tag{27.2}$$

where $\varepsilon \sim i.i.d.(0, \sigma_\varepsilon^2)$. In equation (27.2), the coefficient b_0 is the autonomous rate of productivity and the coefficient b_1 is called the Verdoorn coefficient. Its estimated value is consistently about one-half when the model is fitted to various data on manufacturing productivity growth and output growth. This implies that a one-percentage-point increase in output growth induces an increase in the growth of employment of one-half of a percentage point and an equivalent increase in the growth of productivity. The increasing returns implied by Verdoorn's law have been illustrated by Fingleton (2000) using a static Cobb–Douglas production

function model. While exchangeability problems have not been discussed in this specific context, it is also relevant given the type of assumption made on the errors terms.

27.2.2 The cross-sectional approach to growth and convergence

We focus in this section on another possible violation of the exchangeability assumption: spatial dependence in the error terms (Ertur and Koch, 2007). Indeed, in the cross-sectional context, the observational units are spatially organized and the i.i.d. assumption may therefore be overly restrictive. Various alternative specifications are appropriate to deal with different forms of spatial dependence (Rey and Montouri, 1999). They are best described if we rewrite equations (27.1) and (27.2) in matrix form:

$$y = X\gamma + \varepsilon, \tag{27.3}$$

where y is the $(N \times 1)$ vector containing the observations on the dependent variable, X is the matrix containing the observations on all explanatory variables including the constant term, ε is the $(N \times 1)$ vector of error terms, the properties of which we describe below, and α and γ are the unknown parameters to be estimated. In the convergence case, y contains the vector of average growth rates of per capita income between date t_0 and $t_0 + T$ and X contains the initial log per capita income and all the other control variables. In the Verdoorn case, y contains the labor productivity growth rate. Several spatial econometric specifications have been used to control for spatial dependence in growth econometrics models: the spatial lag model, the spatial error model and the spatial Durbin model.[1] In the spatial lag model, or mixed regressive spatial autoregressive (AR) model, a spatially lagged dependent variable Wy is added to the right-hand side of the regression specification:

$$y = \rho Wy + X\gamma + \varepsilon, \tag{27.4}$$

where W is an $(N \times N)$ spatial weights matrix, usually row-standardized, and ρ is the spatial autoregressive parameter. In a convergence context, for instance, this specification allows measuring how the growth rate in a region may relate to the ones in its surrounding regions (as defined in W) after conditioning on the starting levels of per capita income and the other variables. Unlike the time series case, the spatial lag term is endogenous, since it is always correlated with ε (Anselin, 1988). Therefore, this specification must be estimated using instrumental variables (IVs) or, assuming that ε follows a multivariate normal distribution with zero mean and a constant scalar diagonal variance-covariance matrix $\sigma^2 I_N$, by maximum likelihood (ML). In the former case, Kelejian and Prucha (1999) show that the low-order spatial lags of the exogenous variables can be used as instruments for Wy. Of course, if additional endogenous variables are present in the specification, this approach can easily be extended by adding additional instruments.[2]

The spatial error model is a special case of a non-spherical error covariance matrix in which the spatial error process is based on a parametric relation between a location and its neighbors. Two specifications have commonly been used in spatial

econometrics: the autoregressive and the moving average specifications. It is interesting to note that only the former has been extensively used for the modeling of spatial error dependence in convergence or Verdoorn's law models. A spatial autoregressive specification for the $(N \times 1)$ error vector can be expressed as:

$$\varepsilon = \lambda W \varepsilon + u, \tag{27.5}$$

where λ is the spatial autoregressive parameter and $u \sim i.i.d.(0, \sigma_u^2 I_N)$. Note that this model can be rewritten in another form, called the spatial Durbin model: if (27.3) is premultiplied by $(I - \lambda W)$, we get $(I - \lambda W)y = (I - \lambda W)X\gamma + u$. Hence:

$$y = \lambda W y + X\gamma + \delta WX + u, \tag{27.6}$$

with $\delta = -\lambda\gamma$. These restrictions can be tested by the common factor test (Burridge, 1981). If it cannot be rejected, then model (27.6) reduces to model (27.5). In this model, the average growth rate of a region i is influenced by the average growth rate of neighboring regions, by the initial per capita income of neighboring regions and the spatial lags of the other explanatory variables.

Conversely, the spatial moving average specification can be expressed as:

$$\varepsilon = \widetilde{\lambda} W u + u, \tag{27.7}$$

where $\widetilde{\lambda}$ is the spatial autoregressive parameter and $u \sim i.i.d.(0, \sigma_u^2 I_N)$. Both models can be estimated by ML, under the normality assumption, or generalized methods of moments (GMM) (see Kelejian and Prucha, 1999, for the AR case and Fingleton, 2008, for the moving average (MA) case). These two specifications differ in the terms of the range of spatial dependence in the variance-covariance matrix and of the diffusion process they imply. More precisely, in the first case, the variance-covariance matrix for ε is $\Omega = \sigma_u^2 (A'A)^{-1}$ with $A = I_N - \lambda W$. While W may be sparse, the inverse term $(A'A)^{-1}$ is not. As a consequence, a random shock at one location i is transmitted to all other locations of the sample: the spillovers are global. Rey and Montouri (1999) and Le Gallo *et al*. (2005) illustrate this property in the context of a β-convergence model and show how a random shock in one US state or in one European region impacts upon the per capita income growth rates of all the regions in the sample. Conversely, in the MA case, the variance-covariance matrix does not involve a matrix inverse: $\widetilde{\Omega} = \sigma_u^2 \widetilde{A}'\widetilde{A}$. Therefore, the spillovers remain local: a shock at location i will only affect the directly interacting locations as given by the non-zero elements in W. Finally, higher-order spatial models have been investigated by Kosfeld *et al*. (2006).

This basic framework has been extensively used to analyze the convergence patterns among several sets of countries and regions: convergence among US states (Rey and Montouri, 1999; Lee, 2004; Garrett *et al*., 2007), European regions (Fingleton, 1999; Maurseth, 2001; Carrington, 2003; Le Gallo *et al*., 2003; Arbia, 2006; Le Gallo and Dall'erba, 2006), Brazilian states (Lall and Shalizi, 2003; Magalhães *et al*., 2005; Azzoni and Silveira-Neto, 2006), Spanish regions (Villaverde, 2005, 2006) and Turkish regions (Gezici and Hewings, 2004; Yildirim and Öcal, 2006). Similarly,

these spatial econometric models have been fitted to Verdoorn's law specifications by Bernat (1996), Fingleton and McCombie (1998), Pons-Novell and Viladecans-Marsal (1999) and Dall'erba *et al*. (2008).[3] Several methodological issues should be kept in mind when dealing with these models.

First, in the absence of sound theoretical foundations for the specific form taken by spatial autocorrelation in these models, most of these papers apply a classical "specific to general" specification search approach outlined in Anselin and Rey (1991) and Anselin and Florax (1995) to discriminate between the two forms of spatial dependence – spatial error autocorrelation or spatial lag. Several Lagrange multiplier (LM) tests are used for that purpose. Florax *et al*. (2003) show by means of Monte Carlo simulation that this classical approach outperforms Hendry's "general to specific" approach. Note that this classical approach has several drawbacks, including the problem of multiple comparisons highlighted by Savin (1984): the significance levels of the sequence of tests conducted in this section are unknown.

Second, particular attention should be given to the interpretation of the coefficients in the spatial lag model, as compared to those of the spatial error model or the simple model estimated by ordinary least squares (OLS). Indeed, in these latter models, the marginal effect of one explanatory variable x_k corresponds to the associated parameter β_k. Conversely, the spatial lag model (27.4) can be rewritten as $y = (I - \rho W)^{-1}(\rho W y + X\gamma + \varepsilon)$. Since (in most cases) the elements of the row-standardized weights matrix W are less than one, a Leontief expansion of the matrix inverse follows as: $(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + ...$ Consequently, the growth rate of per capita income (in the convergence context) or of labor productivity (in the Verdoorn context) of one region i is not only affected by a marginal change of the explanatory variables of region i but is also affected by marginal changes of the explanatory variables in the other regions, more importantly so for closer regions. As a consequence, the estimated coefficients in a spatial lag model include only the direct marginal effect of an increase in the explanatory variables, excluding all indirect induced effects, while in the standard model estimated by OLS, they represent the total marginal effect. It is therefore not relevant to compare OLS and ML or two-stage least squares (2SLS) estimates for a spatial lag. This aspect has so far been overlooked in the spatial econometrics literature in general (Pace and LeSage, 2007) and in the spatial growth empirics literature in particular (Abreu *et al*., 2005) and should be kept in mind when drawing inference on determinants of the economic growth process using, for instance, the computationally feasible means of summarizing the output into direct and indirect impacts of each variable of interest suggested by Pace and LeSage (2007).

Third, endogeneity in cross-country regression models is a problem that is commonly encountered as output growth, investment rates, and so on, in a particular period are likely to be jointly determined. Caselli *et al*. (1996) note that "At a more abstract level, we wonder whether the very notion of exogenous variables is at all useful in a growth framework (the only exception is perhaps the morphological structure of a country's geography)." In a non-spatial context, they deal with this issue using the Arellano and Bond (1991) GMM procedure. Similarly, while Verdoorn's law is usually treated as a single equation and estimated via OLS,

there is a debate on the assumed endogeneity and exogeneity of this specification (Kaldor, 1975; Rowthorn, 1975a, 1975b).

In a spatial context, as mentioned previously, the case of a spatial lag model with additional endogenous variables is straightforward since it can be estimated by 2SLS. However, the estimation of a model with a spatial error process and endogenous variables is not possible with the usual ML approach. In this case, Fingleton and Le Gallo (2008) have extended the Kelejian and Prucha (1998) feasible generalized spatial two-stage least squares (FGS2SLS) estimator to account for endogenous variables due to system feedback, given an AR or a MA error process. Angeriz *et al*. (2008) use this strategy in the Verdoorn context. Alternatively, rather than modeling the error process, Kelejian and Prucha (2007) have suggested a nonparametric heteroskedastic and autocorrelation consistent (HAC) estimator of the variance-covariance matrix in a spatial context (SHAC), which can be computed for general regression models allowing for endogenous regressors, their spatial lags and exogenous regressors. This methodology may also prove useful in spatial econometric growth studies. Coupled with this is an extensive taxonomy of simultaneous equation frameworks for spatial process models, recently suggested by Rey and Boarnet (2004), that appears well-suited to convergence research.

Finally, the problem of model uncertainty has been raised by Brock and Durlauf (2001), Fernandez *et al*. (2001), Doppelhofer *et al*. (2004) and Sala-i-Martin (1997). This problem can arise from several sources. First, the selection of appropriate conditioning variables is a difficult issue in convergence models and involves a trade-off between the arbitrary selection of a small number of variables, which may imply some omitted variables bias, and the introduction of a larger set of variables with a number of econometric problems such as endogeneity or multicollinearity. Second, as is typical in all regression models, we also face parameter uncertainty. Fernandez *et al*. (2001) and Doppelhofer *et al*. (2004) employ Bayesian model averaging techniques that can accommodate both model and parameter uncertainty. In a convergence framework, they find that the posterior distribution of β computed across alternative specifications assigns all probability mass to the negative half interval. This results in strong support to the convergence hypothesis. However, LeSage and Fischer (2007) point out that in a spatial setting an additional source of model uncertainty arises: one also has to specify the appropriate weights matrix W that defines connectivity between regions. In other words, the estimates and associated inferences in spatial growth regressions are not only conditional on the set of explanatory variables employed but also on the chosen spatial weights matrix. Extending the approach of LeSage and Parent (2007), LeSage and Fischer (2007) derive a Bayesian model comparison approach that simultaneously specifies the spatial weights structure and the explanatory variables in spatial Durbin models, with an application to the convergence process among European regions.

27.2.3 Dealing with heterogeneity

In this section, we deal with heterogeneity problems and how the literature has considered them in conjunction with spatial autocorrelation. We consider the

convergence literature here as these problems have been mainly encountered in this context. First, unobserved heterogeneity is one of the most common problems related with conditional convergence models, in particular due to technological differences between economies. Given the difficulty in accounting for technological differences in a cross-sectional framework (Islam, 2003), an alternative tactic is to resort to the panel data approach to eliminate unobservable economy-level heterogeneity. For that purpose, a dynamic panel data model with individual fixed effects has been suggested by Islam (1995). Panel data techniques have several advantages. They allow controlling for unobserved heterogeneity in the initial level of technology and omitted variables that are persistent over time. Moreover, endogeneity bias, common in convergence equations as stated above, may also be rectified by estimating the panel data convergence model using Arrelano and Bond's (1991) GMM procedure.

Inclusion of spatial autocorrelation in this context has investigated been by few authors to date. One possibility is to get rid of spatial autocorrelation in order to apply the usual estimators. This strategy has been adopted by Badinger *et al*. (2004), who propose a two-step estimation procedure. First, they apply the filtering technique suggested by Getis and Griffith (2002) that separates the spatially correlated component from the data. Second, standard GMM estimators are used to provide inference on convergence. However, the properties of the estimators obtained in this two-step procedure are unknown. Moreover, this approach assumes that spatial autocorrelation is only a nuisance, whereas the following section shows how spatial autocorrelation can be considered as a component of the growth process in its own right. In their suggestion to deal with spatial autocorrelation directly, Arbia and Piras (2005) analyze the European growth process by including a spatial lag variable or a spatial error term in a convergence model with region fixed effects. The spatial parameters are assumed to be fixed over time and the model is estimated using ML. One drawback of this method is that consistent estimation of the individual fixed effects is not possible as $N \to \infty$, due to the incidental parameter problem (Anselin *et al*., 2008).

Heterogeneity may also concern the regression parameters. While absolute β-convergence is frequently rejected for large samples of countries and regions, it is usually accepted for more restricted samples of economies belonging to the same geographical area. This observation can be linked to the presence of convergence clubs: there is not only one steady-state to which all economies converge. Rather, there may be multiple, locally stable, steady-state equilibria. Therefore, a convergence club is a group of economies whose initial conditions are near enough to converge toward the same long-term equilibrium. It is noteworthy that the hypothesis $\beta < 0$ can be consistent with non-converging alternatives, such as a threshold growth model with multiple steady-states (Azariadis and Drazen, 1990). The determination of those clubs, when they exist, is then a critical issue. In this respect, some select *a priori* criteria, such as initial per capita gross domestic product (GDP) cut-offs (Durlauf and Johnson, 1995). On the contrary, endogenous methods of club detection are quite diverse and include regression trees (*ibid*.), projection pursuit methods (Desdoigts, 1999), Bayesian methods that identify a

mixture distribution for the predictive density of per capita output (Canova, 2004), among others.

It is noteworthy that Durlauf and Johnson (1995), by endogenizing the splitting using the regression tree method, point out the geographic homogeneity within each group. This is even more bound to happen in a regional context, as regional economies are often characterized by strong geographic patterns. In Europe, the core–periphery pattern has frequently been underlined as representative of a form of *spatial heterogeneity*. In a regression context, spatial heterogeneity can be reflected by spatially varying coefficients, that is, structural instability, and/or by spatially varying variances across observations. With this observation in mind, several attempts aimed at analyzing spatial convergence clubs have been suggested. A first strand of papers just use *a priori* spatial regimes. For example, Neven and Gouyette (1995) define two regimes: Northern and Southern European regions. Ramajo *et al*. (2008) split their sample of European regions between the EU cohesion-fund countries (Ireland, Greece, Portugal and Spain) and all the others. Using a model with groupwise heteroskedasticity, two spatial regimes and spatial dependence, they show that, over 1981–96, there was a faster conditional convergence speed in regions belonging to cohesion countries than in the rest of the regions. A similar strategy has been adopted by Roberts (2004) on a sample of British counties, by distinguishing between northern and southern counties.

Others explicitly take into account the spatial dimension of the data and use exploratory spatial data analysis to detect spatial regimes. These techniques are described in more detail in the following section. We note here that Ertur *et al*. (2006) use Moran scatterplots (Anselin, 2006), based on the initial per capita GDP of a sample of European regions, to determine spatial clubs. Two clubs are constructed this way: HH (High–High) regions and LL (Low–Low) regions, corresponding respectively to regions with High (Low) initial per capita GDP surrounded by regions with High (Low) initial per capita GDP. Atypical regions (that is, regions classified as High–Low or Low–High) are eliminated from the sample as they are not numerous enough to constitute a club. Alternatively, Le Gallo and Dall'erba (2006), Dall'erba and Le Gallo (2008) and Fischer and Stirböck (2006) prefer the use of Getis–Ord statistics (Ord and Getis, 1995), applied to initial per capita GDP, that lead to a two-way partitioning of the sample: spatial clusters of high values of per capita GDP (corresponding to positive values of the statistic) and spatial clusters of low values of per capita GDP (corresponding to negative values of the statistics). We can think of these methods as being "semi-endogenous," as the number of clubs is fixed (four in the case of Moran scatterplots and two in the case of Getis–Ord statistics) but the economies are endogenously allocated to the clubs.

The endogenous detection of convergence clubs in data characterized by spatial autocorrelation remains a serious problem, as the properties of the methods already suggested (regression trees, and so on) remain unknown in the presence of spatial autocorrelation. A first step in this direction is described in the paper by Basile and Gress (2005), who suggest a semiparametric spatial autocovariance specification that simultaneously takes into account the problems of nonlinearities and spatial dependence. To that end, Liu and Stengos' (1999) nonparametric specification is

extended by allowing a spatial lag term or a spatial error process. Another methodological problem is the fact that spatial heterogeneity (representative of spatial convergence clubs) and spatial dependence may be observationally equivalent in a cross-section (Abreu *et al.*, 2005). Indeed, a cluster of high-growth regions may be the result of spillovers from one region to another or it could be due to similarities in the variables affecting the regions' growth. Moreover, standard tests of structural instability and heteroskedasticity are not reliable in the presence of spatial autocorrelation. Therefore, as Rey and Janikas (2005) note, the existing specification search procedures should be extended to be able to distinguish between spatial dependence and spatial heterogeneity, while formal specification search strategies for spatial heterogeneity have yet to be suggested.

Rather than partitioning the cross-sectional sample into regimes based on structural characteristics, parameter heterogeneity might also be country- or region-specific. For example, Durlauf *et al.* (2001) allow the Solow growth model to vary according to a country's initial income by using the varying coefficient model suggested by Hastie and Tibshirani (1993). In a spatial context, Ertur *et al.* (2007) argue that similarities in legal and social institutions, as well as culture and language, might create spatially local uniformity in economic structures, which lead to situations where rates of convergence are similar for observations located nearby in space. One possibility is to use geographically weighted regression (GWR) (Fotheringham *et al.*, 2004), which is a locally linear, nonparametric estimation method aimed at capturing, for each observation, the spatial variations of the regression coefficients. For that purpose, a different set of parameters is estimated for each observation by using the values of the characteristics taken by the neighboring observations. For the conditional β-conditional convergence model (equation 27.1), this procedure allows estimation of the set of unknown parameters (β and coefficients associated with the other structural characteristics) for each economy of the sample:

$$\frac{1}{T}\log(y_{i,t_0+T}/y_{i,t_0}) = \alpha_i + \beta_i \log(y_{i,t_0}) + X_{1,i}\delta_{1i} + X_{2,i}\delta_{2i} + \varepsilon_i. \tag{27.8}$$

This model is estimated using weighted least squares with the weights being specific to each observation: for an observation i, the weights are a continuous and monotone decreasing function of the distance between observation i and all other observations. This method is useful for identifying the nature and patterns of spatial heterogeneity over the observations and the results of a GWR (local estimated coefficients, local t-statistics and measures of quality of fit) can be mapped. In a convergence context, this method has been used by Bivand and Brunstad (2003) for a sample of European regions and by Eckey *et al.* (2007) for 180 labor market regions in Germany. Eckey *et al.* show that the German labor market regions are moving at different speeds towards their steady-states, with the value of the half-life increasing from North to South.

While useful for capturing heterogeneity in growth experiences in a sample of economies, inference in this context is problematic. Indeed, Wheeler and Tiefelsdorf (2005) show that the local regression estimates are potentially collinear even if

the underlying exogenous variables in the data-generating process are uncorrelated. This collinearity can degrade coefficient precision in GWR and lead to counter-intuitive signs for some regression coefficients. Another methodological problem has been pointed out by Paez *et al*. (2002) and Pace and LeSage (2004). Indeed, one of the motivations of this approach is that, if spatial autocorrelation only arises due to inadequately modeled spatial heterogeneity, GWR can potentially eliminate this problem. However, this is not necessarily the case as substantive spatial interactions may coexist with parameter heterogeneity, as we will show in the next section. Therefore, Pace and LeSage (2004) have generalized GWR to allow simultaneously for spatial parameter heterogeneity and spatial autocorrelation: the spatial autoregressive local estimation (SALE). Formally, estimates are produced using n-models, where n represents the number of cross-sectional sample observations, using the locally linear spatial autoregressive model:

$$U(i)y = \rho_i U(i)Wy + U(i)X\gamma_i + U(i)\varepsilon, \tag{27.9}$$

where $U(i)$ represents an $(N \times N)$ diagonal matrix containing distance-based weights for observation i that assign the weight of one to the m nearest neighbors to observation i and weights of zero to all other observations. The product $U(i)y$ then represents an $(m \times 1)$ sub-sample of observed per capita income rates associated with the m observations nearest in location to observation i. The other products are interpreted in a similar fashion. As $m \to N$, $U(i) \to I_n$, and the local estimates approach the global estimates from (27.4) as the sub-sample size increases. This model is estimated by recursive ML for $\varepsilon_i \to N(0, \sigma_i^2 U(i)I_N)$. This approach has been implemented by Ertur *et al*. (2007) for a sample of 138 European regions for the period 1980–95. Moreover, they also control for non-constant variances with a Bayesian spatial autoregressive local estimation. They show that, while the mean of the estimates for ρ is near zero, there are still a number of regressions for which spatial dependence estimates take on large and significant values. Country-level differences are also obvious for the different estimates of the convergence parameter β, with negative and significant values across EU regions in Spain, Portugal and some French regions.

27.2.4 Theoretical foundations of spatial effects

As pointed out by Islam (2003), the specifications of growth regressions used in initial studies of β-convergence were not derived from theoretical growth models. Only at a subsequent stage were the regression specifications formally derived from the neoclassical growth models by Barro and Sala-i-Martin (1992) and Mankiw *et al*. (1992). The literature focusing on the relationships between space and growth has evolved quite similarly. All the studies surveyed above included spatial effects in an *ad hoc* way, allowing for spatial autocorrelation and/or spatial heterogeneity in conditional β-convergence models or Verdoorn models in order to obtain a better fit and consistent estimates. Spatial autocorrelation in this context may reflect spatial spillovers arising between economies but can also be the result of some model misspecification or omitted variables. Recently, some authors have tried to

provide sound theoretical foundations for the inclusion of spatial dependence in β-convergence or Verdoorn models. We review some of this recent work here.

Ertur and Koch (2007) show how a spatial Durbin version of the β-convergence model (equation 27.6) can be obtained from a theoretical growth model with Arrow–Romer externalities and spatial externalities that imply inter-economy technology interdependence. They start with an aggregate Cobb–Douglas production function for economy i at time t that exhibits constant returns to scale in labor and reproducible physical capital:

$$Y_i(t) = A_i(t)K_i^{\alpha}(t)L_i^{1-\alpha}(t) \tag{27.10}$$

$$\text{where } A_i(t) = \Omega(t)k_i^{\Phi}(t)\prod_{j\neq i}^{N}A_j^{\gamma w_{ij}}(t),$$

where $Y_i(t)$ is output, $K_i(t)$ is the level of reproducible physical capital, $L_i(t)$ is the level of labor, N is the number of economies and $A_i(t)$ is the aggregate level of technology of economy i at time t. This level depends on three terms. First, as in the standard Solow growth model, Ertur and Koch (2007) assume that some proportion of technological progress is exogenous and identical in all countries: $\Omega(t) = \Omega(0)e^{\mu t}$, where μ is the constant growth rate. Second, each economy's aggregate level of technology increases with the aggregate level of physical capital per worker $k_i(t) = K_i(t)/L_i(t)$, and the parameter Φ (with $0 \leq \Phi < 1$) describing the strength of home externalities generated by physical capital accumulation. Finally, as these externalities may spill over to neighboring economies, it is also assumed that there is technological interdependence generated by the level of spatial externalities γ (with $0 \leq \gamma < 1$). The w_{ij} are the usual terms of the spatial weights matrix.

This specification yields a spatially augmented β-convergence model, similar to a non-constrained spatial Durbin model, where the growth rate of real income per worker not only depends on its own saving rate and population growth rate, but also depends on the same variables in the neighboring economies and on the growth rate of its neighboring economies weighted by their speed of convergence. Interestingly, complete parameter heterogeneity can be allowed for when the speed of convergence is not assumed to be identical across economies. Ertur and Koch (2007) estimate this heterogeneous model on a set of countries using the SALE model as in equation (27.9). Other such attempts to motivate theoretically the presence of spatial dependence have been suggested. For instance, López-Bazo *et al*. (2004) assume that the spatial externalities originate from physical and human capital accumulation rather than knowledge, yielding a β-convergence model with a spatial lag term as in (27.4). Egger and Pfaffermayr (2006) decompose the speed of convergence term into three components: one measuring the speed of convergence net of spillovers, and the other two accounting for the importance of spatial spillovers.

A similar path has recently been followed to provide theoretical foundations for the spatial versions of Verdoorn's law. In particular, Fingleton (2000) shows how Verdoorn's law including a spatial lag term can be motivated by inter-regional

technology diffusion. Indeed, take as a point of departure a static Cobb–Douglas production function given by:

$$Q = A_0 \exp(\lambda t) K^{\alpha} E^{\beta}, \tag{27.11}$$

in which A_0 is the level of technology at time 0, λ is the growth of total factor productivity or exogenous technical change, Q, K and E are the levels of output, capital, and employment at time t; α and β are elasticities. Fingleton (2000) assumes that the technical progress made by firms as a result of the growth of capital per worker (p) is not fully internalized but spills over to benefit other firms and individuals. Two forms of technology spillovers are envisaged: one occurs as a result of intraregional technical change and one occurs as a result of extraregional technical change in neighboring regions:

$$\lambda = \lambda^{*} + \phi p + \kappa W p. \tag{27.12}$$

λ in one region is then proportional to p in the same region and, through the matrix product Wp, is also a function of capital accumulation occurring within the neighbors for each region, as specified by W. Taking equation (27.11) in logs, differentiating with respect to time and assuming that capital stock growth and output growth are approximately the same,[4] he shows that the reduced equation can be written as:

$$p = \rho W p + b_0 + b_1 q + Xb + \varepsilon, \tag{27.13}$$

where X contains other determinants of labor productivity growth, such as the initial level of technology gap between each region and the leading technology region, and measures of peripherality and urbanization. In subsequent papers, Fingleton (2001a, 2001b) shows that the spatial lag version of Verdoorn's law is, in fact, also consistent with assumptions that underpin new economic geography with, as a starting point, a Cobb–Douglas production function for the level of output produced by manufacturers that depends on the input of manufacturing labor efficiency units, on composite intermediate services and on the input of land. Increasing returns are modeled via the product variety theory emphasized by Dixit and Stiglitz (1977) and the rate of technical progress is assumed to be as in equation (27.11).

27.3 Exploratory spatial data analysis of convergence

While a great deal of work has been done on confirmatory econometric analysis of growth and convergence, a number of criticisms have been pointed at the relatively restrictive nature of the underlying theoretical frameworks, their inability to account for empirical regularities in growth datasets and the general problem of using cross-section regressions to explain time-averaged growth rates with which to make inferences about growth dynamics (Quah, 1993b). In response to some of the criticisms, a number of researchers have adopted novel methods of exploratory

data analysis (EDA) to examine growth datasets. EDA has its roots in the work of John Tukey (Tukey, 1977), who defined the field as a set of statistical methods designed to detect and describe patterns, trends, and relationships in data. EDA methods often rely on interactive statistical graphics to support different types of interrogation of the data.

In this section we provide an overview of the main branches of EDA methods that have appeared in the convergence literature. These follow from some of the particular challenges posed by the study of growth and convergence in a spatial context. On the one hand, traditional EDA techniques often rest on the same restrictive assumptions regarding random sampling that we encountered in early econometric work on regional convergence. There has been much recent work developing spatially explicit methods of EDA designed to take the spatial characteristics of the data into account. These methods fall under the heading of *exploratory spatial data analysis* (ESDA) (Anselin, 1996).

At the same time, the dynamic characteristics of growth datasets pose interesting challenges for the use of ESDA methods, since the latter have primarily been designed for cross-sectional datasets. The second branch of the exploratory literature we review consists of efforts designed to extend ESDA methods to the dynamic context. We refer to this branch of the literature as *exploratory space-time data analysis* (ESTDA).

The use of ESDA and ESTDA methods for convergence and growth analysis relies on a number of computational tools as well as interactive statistical graphics and maps for data exploration. In what follows, we draw on examples using the statistical package STARS: Space-Time Analysis of Regional Systems (Rey and Janikas, 2006), which has implemented a number of these methods for spatial convergence analysis.[5]

27.3.1 Exploratory spatial data analysis

27.3.1.1 Spatial σ-convergence

The point of departure in the EDA branch of the convergence literature has been the entire distribution of regional incomes itself, with a focus on a number of characteristics of this distribution. The earliest studies examined the level of dispersion in the distribution and its evolution over time. Labeled as σ-convergence, the typical approach is to consider the cross-sectional variance (or standard deviation) of the incomes:

$$\hat{\sigma}_t^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_{i,t} - \bar{y}_t)^2, \tag{27.14}$$

where $y_{i,t}$ is the per capita income or product of economy i in time period t and $\bar{y}_t = \frac{1}{n} \sum_{i=1}^{n} y_{i,t}$. Implicit in the application of this measure is its theoretical relationship to β-convergence:

$$\sigma_t^2 = (1-\beta)^2 \sigma_{t-1}^2 + \sigma_\epsilon^2, \tag{27.15}$$

where α, β, and σ_ϵ^2 are as defined in equation (27.1). If $0 < \beta < 1$ the difference equation is stable and β-convergence gives rise to the steady-state variance:

$$\sigma_*^2 = \frac{\sigma_\epsilon^2}{1-(1-\beta)^2}. \tag{27.16}$$

The cross-sectional variance in incomes will fall with increases in β (in the stable range) but rise with the initial σ_ϵ^2.

There are a number of limitations of σ-convergence for studying the dynamics of the distribution. First, it focuses on only the second moment of the distribution and is thus silent on other moments, such as skewness and kurtosis, which can be important from a substantive perspective. Second, the sample variance would also mask any multimodality or twin-peakedness in the distribution, which tends to be a common finding at the international scale. In addition to being silent on the morphology of the distribution, measures of σ-convergence provide no insight on the degree of intradistributional mixing and mobility.

While these criticisms are generally well known, there are also several lesser known problems with the application of σ-convergence to spatially referenced data. The first is a spatial identification problem. This arises from the sample variance being what is known as a "whole map" statistic. More specifically, given a map (spatial distribution) of n incomes $y_{i,t} : i = 1, 2, \ldots, n$, with sample variance $\hat{\sigma}_t^2$, there are $n!$ spatial permutations of the map that would have the same sample variance.

The second difficulty with σ-convergence in a spatial context relates to the i.i.d. assumption on ϵ. As is the case for the confirmatory econometric modeling of convergence, the presence of spatial dependence in the error term of the model complicates the analysis. The impact of spatial dependence on the interpretation of σ-convergence has been examined by Egger and Pfaffermayr (2006) and Rey and Dev (2006), who show that, in addition to the β coefficient and the initial variance level, the value of the sample variance will also reflect the level and structure of the spatial dependence. More specifically, if the underlying data generating process is the spatial lag specification, then the sample variance for the income levels will be sensitive to the value of the spatial lag parameter and the specific structure of the spatial weights matrix. These additional sources of dynamics complicate the interpretation of the sample variance as a measure of σ-convergence.

27.3.1.2 Markov chain models

Quah (1993a, 1996a) has adopted a discrete Markov chain approach to study the evolution of income distributions. Using international data for 1962–84, Quah discretizes the income distribution in each period into k classes, and the probability of an economy transitioning between each pair of classes is estimated from the income series for the country economies. An application of this approach to the lower 48 state incomes for the US over the period 1929–2000 is reported in Table 27.1. The income values are standardized to the national value each year and the class cut-offs are taken from the quintiles of the relative distribution for the first year in the sample. Thus, for a state that had an income level below 66% of

Table 27.1 Markov transitions: US state incomes 1929–2000

	$t+1$				
t	0.661	0.884	1.031	1.309	1.934
0.661	0.893	0.107	0.000	0.000	0.000
0.884	0.013	0.913	0.073	0.001	0.000
1.031	0.000	0.054	0.872	0.074	0.000
1.309	0.000	0.001	0.065	0.913	0.021
1.934	0.000	0.000	0.000	0.112	0.888

Note: The row (column) headings are the upper bounds for the quintiles of relative incomes normalized by the national average. The values in the body of the table are the empirical transition probabilities of moving from class i in period t to class j in period $t+1$.

Table 27.2 Ergodic US state income distribution

x	0.661	0.884	1.031	1.309	1.934
$P(x)$	0.029	0.233	0.314	0.358	0.067

Note: Column headings are the upper bounds for the quintiles of relative incomes normalized by the national average.

the national average, the probability of moving up into the next higher income class during one year was 0.107. At the other end of the distribution, states with incomes greater than 193% of the national level moved down one income class with a probability of 0.112.

Based on an estimate of the transition probability matrix, one can in turn generate an estimate of the ergodic distribution for the regional incomes. For the US case, the long-run distribution implied by these transitional dynamics is reported in Table 27.2. The tendency towards convergence is clearly evident in this distribution, as the extreme classes lose mass over what they had in the quintile distribution at the beginning of the period (1929).

While the Markov chain is an innovative approach to the exploration of distributional dynamics, the estimates of the transition probabilities rest on several assumptions, such as order, time-homogeneity and independent transitions. Independent transitions mean that the spatial context facing a given economy is not taken into account when estimating the probability of movement out of a particular income class.

27.3.1.3 Spatial Markov

Rey (2001) extends the discrete Markov approach to consider this context by estimating transition matrices subject to the spatial lag of the income values for an economy. This is done for the US example in Table 27.3, where the same five income

Table 27.3 US state incomes: spatial Markov transition matrix

		$t+1$				
	t	0.661	0.884	1.031	1.309	1.934
Lag class 0.661	0.661	0.985	0.015	0.000	0.000	0.000
	0.884	0.032	0.871	0.097	0.000	0.000
	1.031	0.000	0.222	0.778	0.000	0.000
	1.309	0.000	0.000	0.000	0.000	0.000
	1.934	0.000	0.000	0.000	0.000	0.000
Lag class 0.884	0.661	0.842	0.158	0.000	0.000	0.000
	0.884	0.022	0.940	0.038	0.000	0.000
	1.031	0.000	0.046	0.881	0.073	0.000
	1.309	0.000	0.000	0.321	0.679	0.000
	1.934	0.000	0.000	0.000	0.000	1.000
Lag class 1.031	0.661	0.857	0.143	0.000	0.000	0.000
	0.884	0.007	0.916	0.077	0.000	0.000
	1.031	0.000	0.056	0.873	0.071	0.000
	1.309	0.000	0.000	0.060	0.936	0.004
	1.934	0.000	0.000	0.000	0.231	0.769
Lag class 1.309	0.661	0.000	0.000	0.000	0.000	0.000
	0.884	0.000	0.795	0.192	0.014	0.000
	1.031	0.000	0.050	0.870	0.080	0.000
	1.309	0.000	0.002	0.059	0.904	0.035
	1.934	0.000	0.000	0.000	0.137	0.863
Lag class 1.934	0.661	0.000	0.000	0.000	0.000	0.000
	0.884	0.000	0.000	0.000	0.000	0.000
	1.031	0.000	0.000	0.889	0.111	0.000
	1.309	0.000	0.000	0.024	0.929	0.048
	1.934	0.000	0.000	0.000	0.048	0.952

classes are used to estimate transition matrices for subsets of the states at different points in time. The sub-setting is based on the spatial lag of the incomes. For example, focusing on the first matrix, we see that poor states that were surrounded by poor states (i.e., those with incomes less than 66% of the national average) moved out of the bottom class with a probability of 0.015. However, if this is contrasted with other states in the lower class, but those who had neighbors that were slightly better off (i.e., in the second class), the probability of moving out of the bottom class now jumps to 0.158. The impact of the spatial context can also be seen for the higher-income states, as the wealthiest states, when surrounded by similarly wealthy states, have a probability of remaining in the upper class of 0.952. However, when a wealthy state is surrounded by states with average incomes in the next lower class, the probability of remaining in the upper class of the distribution drops to 0.863.

The spatial Markov approach has been applied by Le Gallo (2004) to a sample of European regions and by Hammond (2004) to US labor markets. This approach has

also been adopted to explore the question of regional heterogeneity. Bickenbach and Bode (2003) do this by estimating transition matrices for the groups of states in each of the Bureau of Economic Analysis (BEA) regions of the US. They also suggest formal tests of regional homogeneity and find strong evidence that the transitional dynamics are not homogeneous across the US space economy.

While the spatial Markov approach offers an interesting extension of the classic Markov chain to the geographical context, there are a number of methodological issues associated with the approach that require further investigation. The first issue surrounds the data-intensive requirements needed to estimate a spatial Markov matrix. Rather than having to estimate k^2 transition probabilities as in the classic approach, when one is conditioning on k classes for the spatial lag, the number of probabilities to estimate grows to k^3. This can lead to many zero, or small count, observations for specific types of transitions, which in turn means a loss of precision in the estimation of those probabilities.

One solution to this degrees of freedom problem is to combine the thin count cells with other cells and estimate the transition probabilities for the new aggregated cells. This could be done in several ways. One approach would be to keep the number of classes for the spatial lag fixed, and then collapse the same group of cells within each of the k conditional transition matrices. This would result in k conditional matrices of order $r \times c$, with $r \leq k$ and $c \leq k$. An alternative approach would be to keep the number of income classes fixed at k but reduce the number of classes for the spatial lag, essentially aggregating together cells across the conditional transition matrices. In this case one would have $l \leq k$ conditional transition matrices of order $k \times k$. Of course, a third option would be to aggregate both the number of classes for the income variable as well as the spatial lag. The choice between these alternatives is not inconsequential, as the first approach would trade a loss in resolution of the income class transitions for maintaining detail in the effect of spatial context (that is, the spatial lag). In the second approach, more detail on the class transitions is gained at the expense of a coarser view of spatial context effects. To date, however, the relative merits of these different approaches remain unexplored.

A final issue with the spatial Markov approach is that it focuses only on the transitions of individual economies within the income distribution subject to the level of income in the surrounding economies in the initial period. It does not also treat the transition of the spatial lag, but uses that as the conditioning variable. Joint consideration of the transitions of the spatial lag and the income of an economy would require estimating k^4 transition probabilities (assuming k classes for incomes and the spatial lag). Clearly this would exacerbate the degrees-of-freedom problem.

27.3.1.4 Spatial rank mobility

Rank mobility measures attempt to address the problems with discretization to provide a more comprehensive picture of regional income mobility (Boyle and McCarthy, 1997; Webber *et al.*, 2005). One classic measure is Kendall's rank

correlation statistic:

$$\tau = \frac{n_c - n_d}{(n^2 - n)/2}, \tag{27.17}$$

where n_c is the number of concordant pairs of economies over a given interval. Concordant pairs are those that maintain the same relative pair-wise ranking in two periods. n_d is the number of discordant pairs, where a discordant pair are two economies whose relative ranks are reversed over the period. If all pairs are concordant in a given period then $n_c = (n^2 - n)/2,\ n_d = 0$ and $\tau = 1$. Conversely, if all pairs are discordant, then $\tau = -1$.

To introduce a geographical dimension to this rank mobility measure, Rey (2004b) suggested a decomposition of the concordance measure as:

$$n_c = n_{c,r} + n_{c,o}, \tag{27.18}$$

where $n_{c,r}$ is the number of concordant pairs in which the economies involved were geographical neighbors, while $n_{c,o}$ includes those discordant pairs that involve non-neighboring economies. Together with a similar spatial decomposition of the discordant pairs, spatial versions of the rank correlation statistic can be defined as:

$$\tau_r = \frac{n_{c,r} - n_{d,r}}{(n_r^2 - n_r)/2}, \tag{27.19}$$

for the neighboring pairs, and:

$$\tau_o = \frac{n_{c,o} - n_{d,o}}{(n_o^2 - n_o)/2}, \tag{27.20}$$

for the non-neighboring pairs of economies. The traditional measure of rank mobility can be related to these spatial versions:

$$\tau = \psi\tau_r + (1 - \psi)\tau_o, \tag{27.21}$$

where:

$$\psi = \frac{\omega_r}{\omega}, \tag{27.22}$$

with ω_r being the number of neighboring pairs of economies, and $\omega = (n^2 - n)/2$ being the number of all pairs.

Figure 27.1 summarizes the results of a spatial rank mobility analysis of state incomes. The right-hand figure shows the definition of the neighborhood sets, here based on BEA regions. The left-hand figure shows the z-value for the τ test statistic standardized against its expected value based on spatial permutations of the income values. The statistic is significantly negative in the earlier periods, indicating that the amount of intraregional concordance was below what would be expected if there was no spatial structure to income growth. In other words, the differences in

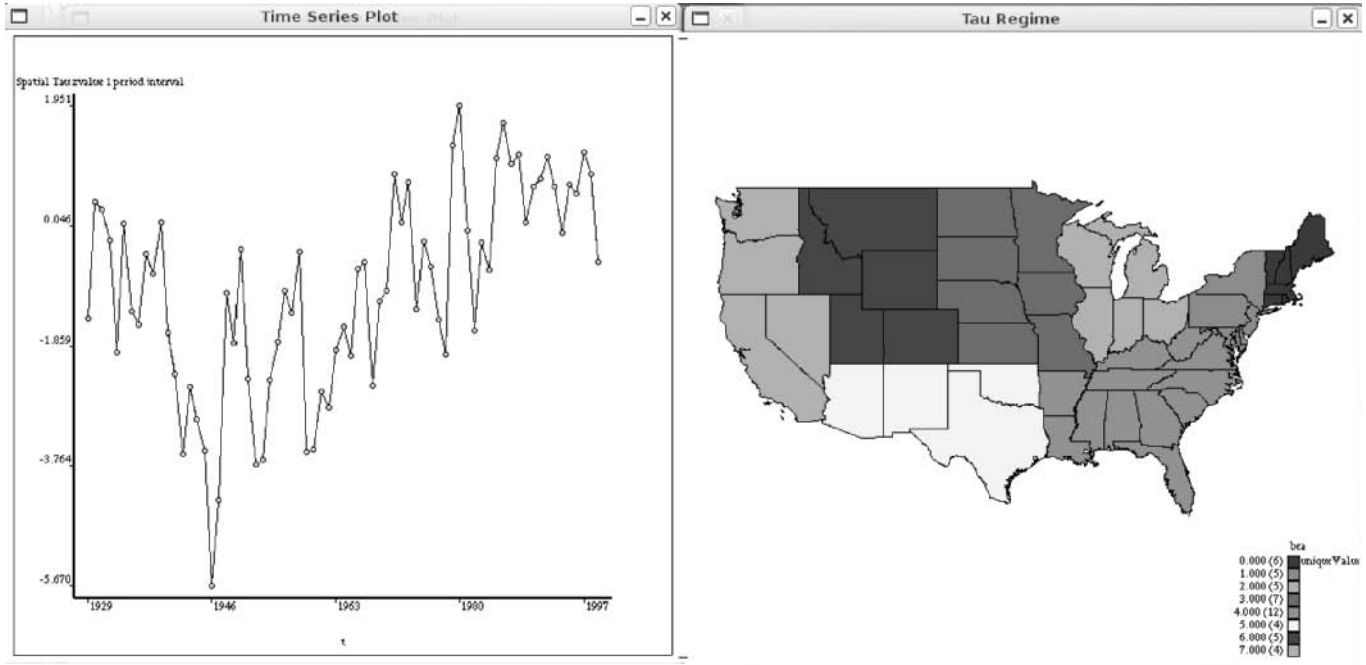


Figure 27.1 Spatial rank mobility

income levels between states from the same region are smaller (and the discordance in rank changes over time greater) than is the case for states from different regions. This form of spatial dependence weakens over time as the spatial rank mobility statistic lessens in absolute value.

27.3.2 Exploratory space-time data analysis

The previous sets of ESDA methods can be seen as attempts to extend a-spatial methods of exploratory convergence analysis to include a spatial component. A second set of methods have been developed that extend exploratory spatial data methods to the dynamic context.

27.3.2.1 Spatial transitions

Since full estimation of a square spatial Markov transition matrix is likely to be infeasible in most data contexts, alternative approaches towards analyzing the movements of the income level of an economy and that of its neighbors have been suggested. These are based on the notion of a local indicator of spatial association (LISA) statistic originally developed by Anselin (1995).

The LISA statistics can be visualized in a Moran scatterplot (Anselin, 1996), depicted in the right-hand panel of Figure 27.2. In the first quadrant are located relatively high-income economies that are neighbored by high-income economies reflected in the spatial lag (HH). Quadrant two contains lower-income (L) economies with wealthier neighbors (LH), while quadrant three might be considered spatial poverty zones since it contains poor economies with poor neighbors. Finally, quadrant four has the "diamonds in the rough" economies – those with high incomes and poor neighbors. These statistics have been extensively used to analyze spatial distributions in several samples of European regions (López-Bazo *et al.*, 1999; Le Gallo and Ertur, 2003; Ertur and Koch, 2007) or other sub-sets of regions (Ying, 2000; Mossi *et al.*, 2004; Patacchini and Rice, 2007).

The spatial concentration of the low values is seen in the map on the left where the user has selected a sub-set of the Southern states. In response, the positions of those states in the scatterplot are indicated by solid black circles. The degree of spatial association for this local cluster (dashed line) is contrasted against the global pattern (solid line). The user can move the lasso around on the map to explore the stability of the spatial clustering over regions of the map. Alternatively, the scatterplot could serve as the origin view and the lasso moved over regions of the plot to reveal the location of selected observations in geographical space, as shown in Figure 27.3.

A number of summary measures of distributional mobility have appeared in the literature. One such measure, due to Shorrocks (1978), is based on the estimates of the classic Markov transition matrix:

$$SI = \frac{k - \sum_i p_{ii}}{k - 1}, \tag{27.23}$$

which is bounded on the interval $[0, k/(k-1)]$, with the lower bound indicating a complete lack of mobility and the upper bound maximum mobility.[6] This captures

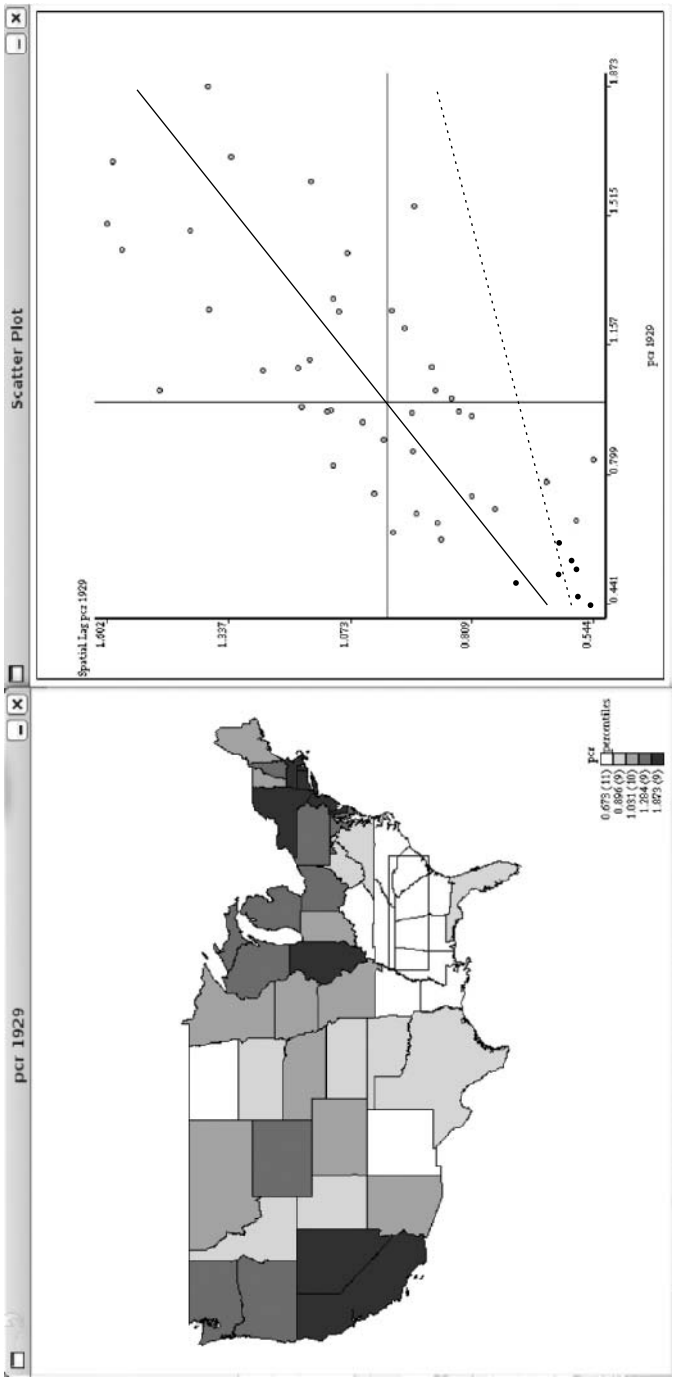


Figure 27.2 Moran scatterplot

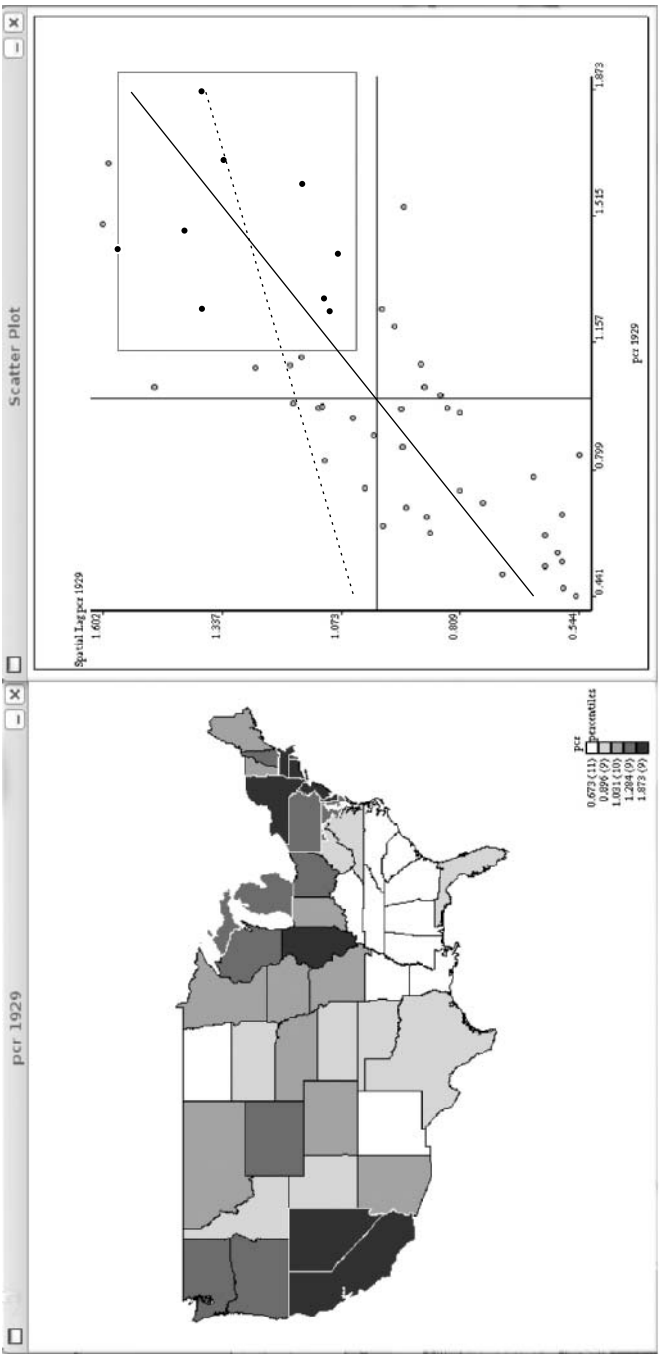


Figure 27.3 Moran scatterplot as origin

class mobility within the distribution and, as such, it is subject to the criticisms associated with the discretization of the income distribution to define the classes. Chief among these is that moves that cross a class boundary are the only kind recorded, while other moves of the same magnitude but remaining within the class are not recorded. Since it is based on the classic Markov approach, the measure also ignores any spatial dependence. However, one could construct a test statistic based on values for this index taken from the specific conditional matrices of the spatial Markov approach to explore whether mobility was affected by spatial context.

To incorporate a spatial dimension to the mobility analysis, Rey (2001) has suggested analyzing the movement of a LISA statistic over a given time interval and noting whether a particular LISA statistic remains in the same quadrant or transitions to a different quadrant. In any period there are four possible states for a LISA – HH, LH, LL, HL, so between any two periods there are 16 different spatial transitions that are possible. These can be summarized in a LISA Markov transition matrix, as is done for the US case in Figure 27.4.

Applying the summary mobility index (27.23) to the traditional Markov transition matrix generates a value of 0.130, which is also found for the case of the LISA Markov transitions, suggesting similar rates of mobility for the spatial versus a-spatial transitions. However, a closer inspection of the two tables reveals some subtle but important differences. In the classic Markov model, the probability of remaining in an extreme (poorest or richest) class is lower than the probability of remaining in the intermediate (2nd or 4th quintile) classes, reflecting a tendency away from polarization in the extreme tails of the distribution. By contrast, for the

	$t+1$		
t	HH	LH	LL
HH	0.929	0.037	0.004
LH	0.079	0.859	0.060
LL	0.005	0.020	0.947
HL	0.053	0.002	0.070

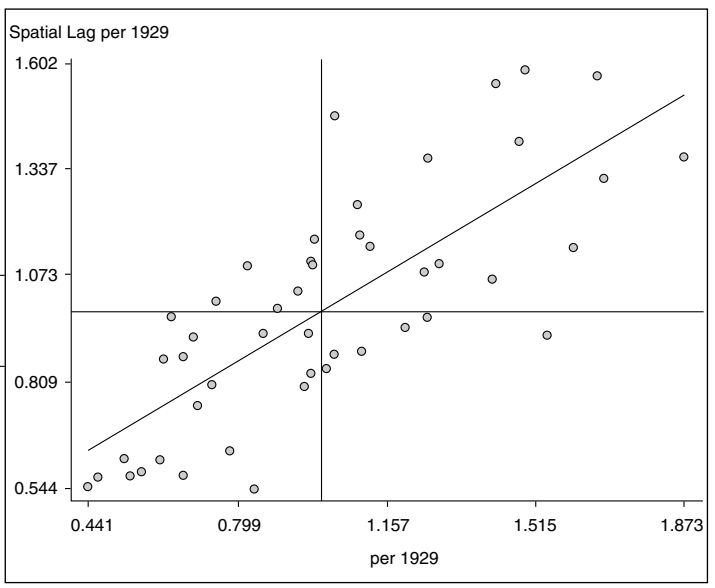


Figure 27.4 LISA Markov transitions

LISA Markov transitions there is evidence of spatial polarization in the transitions, as the "extreme" classes are the HH and LL cells, which have higher retention probabilities than is the case for the HL and LH cells.

27.3.2.2 Space-time paths

In addition to characterizing the global spatial dynamics, a more meso-level scale of analysis is possible. Figure 27.5 contains the *space-time* paths for Virginia (left) and New Jersey (right). The paths are generated interactively by selecting the states in the Moran scatterplot (lower left) and clicking. The time paths trace out the local statistic for each state, showing how its income and that of its neighbors co-evolve over the sample period. At first glance, a comparative view of the two states suggests that the same general trend is evident, reflecting strong spatial cohesion between each state and its neighbors. Closer inspection reveals, however, that

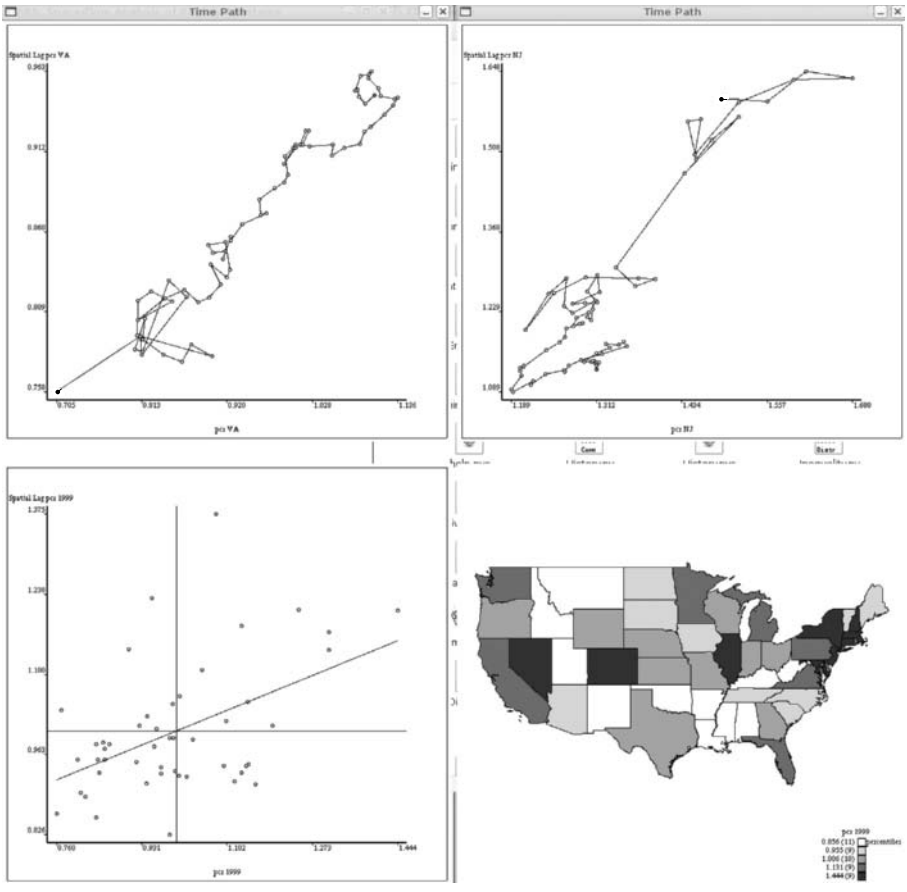


Figure 27.5 Space-time paths

the directionality between the two paths is distinct, with New Jersey's evolution moving downwards towards the mean of the distribution, while for Virginia the dynamics are upward.[7] Rey and Ye (2008) develop approaches for summarizing the properties of these time paths and their use in comparative analysis.

27.3.3 Stochastic kernels

A very active area of exploratory analysis has been the application of stochastic kernels to regional growth series (Magrini, 1999; López-Bazo *et al.*, 1999; Bulli, 2001; Basile, 2006). Here the focus is on the evolution of the distribution itself and a number of novel approaches towards the estimation of densities and their interpretation have been suggested.

27.3.3.1 Estimation

Letting $f_{x(t)}$ represent the regional income density for n economies in period t, the evolution of the cross-sectional distributions is modeled through the use of a stochastic kernel:

$$f_{x(t+s)} = \int_{-\infty}^{\infty} M_{t,s} f_{x(t)} dx, \tag{27.24}$$

where $M_{t,s}$ is the stochastic kernel which traces where points in $f_{x(t)}$ move to in $f_{x(t+s)}$. The kernel can be viewed as a continuous analog of the Markov transition matrix that we examined in the previous section. As such, the stochastic kernel contains important information regarding the distributional dynamics and thus the question of its estimation becomes important. One approach relies on an estimate of the conditional distribution:

$$\hat{M}_{t,s} = \hat{f}_{x(t+s)|x(t)} = \frac{\hat{f}_{x(t+s),x(t)}}{\hat{f}_{x(t)}}. \tag{27.25}$$

Estimates of the joint or marginal densities themselves rely on, somewhat confusingly, kernel density estimates. For example:

$$\hat{f}_x(t) = \hat{f}(x, t; h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_{i,t})/h\}, \tag{27.26}$$

where K is a function such that $\int K(x)dx = 1$, referred to as the kernel, h is the bandwidth and $X_{i,t}$ is the income for economy i for a given time period.

Based on the estimate of the stochastic kernel, alternative visualizations can be generated to explore the implied transitional dynamics. These include three-dimensional representations and the analogous two-dimensional contour plot. Evidence of polarization in the income distributions would be reflected in peaks in the 3D kernel or concentrated values in the contour. The sphericality of either of the graphs would be indicative of heightened income mobility and leap-frogging, while elongated ellipses along the 45-degree line would suggest a lack of mobility and convergence.

Other types of kernel visualizations are shown in Figure 27.6. The conditional densities (Figure 27.6(a)) have been estimated for a 71-year transition period from 1929 to 2000 for the US. By stacking the conditional densities, the figure shows how the regional income distribution from 1929 evolves into that in 2000. The highest-density region (HDR) plot in Figure 27.6(b) identifies the smallest region in the sample space that covers a given probability. For each conditional distribution, the darker shaded area is the 50% HDR, while the lighter shade represents the 99% HDR. The HDR plot reflects strong convergence over the 71-year period, as each of the conditional modes (dots) falls away from the diagonal, indicating that states tend to move towards the overall mean of the distribution.[8]

Alternatively, estimates of the marginal densities for the two periods could be examined. Bianchi (1997) has suggested that the multimodality of a density could serve as an indication of polarization. In this way, movement to a single mode would be reflective of convergence. It should be noted that this marginal approach risks loss of information on the transitions and can mask internal mixing.

While kernels provide novel visualizations of growth dynamics, their use with spatially referenced data is not without problems. To gain some understanding of the potential implications for spatial dependence in the estimation, rewrite the kernel density estimator for a given time period as:

$$\hat{f}(x;h) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{27.27}$$

where the incomes $X_1, \ldots, X_n$ are no longer independent but are identically distributed with a common density, so that $Cov(X_j, X_{j+k})$ depends only on k. In such a setting the bias of $\hat{f}(x;h)$ is robust to the dependence, but the variance becomes:

$$\begin{aligned} V\left[\hat{f}(x;h)\right] =& n^{-1} V\left[K_h(x - X_1)\right] \\ &+ 2n^{-1} \sum_{j=1}^{n-1} (1 - j/n) COV\left[K_h(x - X_j), K_h(x - X_{j+1})\right]. \end{aligned} \tag{27.28}$$

Applications of kernel density estimators to regional income datasets have implicitly assumed that the observations were pairwise independent. With spatially referenced data, however, such an assumption can be questionable. The implications for the properties of kernel density estimators and the related methods of stochastic dominance (Carrington, 2006) and relative distribution approaches (Janikas, 2007) have been largely ignored to date in the growth literature.

27.3.3.2 Regional conditioning and spatial filtering

Canova and Marcet (1995) suggest that income cross-correlations among countries need to be treated prior to constructing kernels. Their approach is to base the kernel on:

$$x^*_{i,t} = x_{i,t} / \sum_i x_{i,t}, \tag{27.29}$$

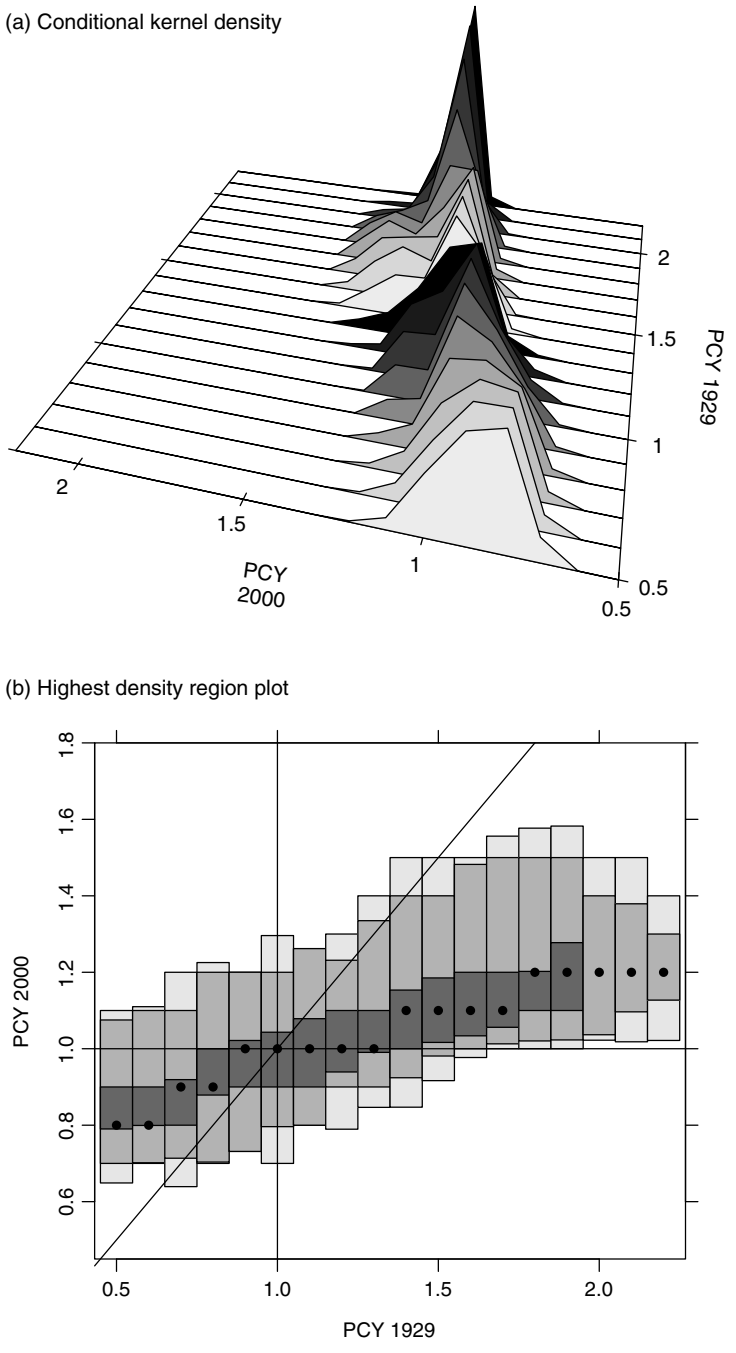


Figure 27.6 Kernel density visualization

where $x_{i,t}$ is per capita income for a given country. Canova and Marcet (1995, p. 9) argue that this alleviates potential problems of income cross-correlations among countries since recessions and expansions which affect the whole aggregate of regions would be factored out. This is not spatial autocorrelation, however, but rather a shift in the mean of the series over time.

In the same spirit, a number of approaches to regional conditioning have been suggested as ways to mitigate the impact of spatial autocorrelation in income series. Quah (1996b) and Arbia *et al*. (2003) rely on the following transformation:

$$\tilde{x}_{i,t} = \frac{x_{i,t}}{\sum_j w_{i,j} x_{j,t}}, \tag{27.30}$$

where a region's income is now expressed relative to its geographical neighbors rather than the national average. Applying this transformation to European data results in the removal of the bimodal characteristic of the income distribution.

Fischer and Stumpner (2008) apply the spatial filtering approach of Getis and Ord (1992) to European data using the following filter:

$$\tilde{x}_{i,t} = \frac{x_{i,t}\left[\frac{1}{n-1} W_i\right]}{G_i(\delta)}, \tag{27.31}$$

where $W_i = \sum_j w_{i,j}(\delta)$ and $w_{i,j}(\delta)$ is the (i, j)th element of the binary spatial weights matrix such that $w_{i,j}(\delta) = 1$ if a region's i and j are separated by a distance of less than the critical distance band δ, and $w_{i,j}(\delta) = 0$ otherwise. $G_i(\delta)$ is the local statistic defined as:

$$G_i(\delta) = \frac{\sum_j w_{i,j}(\delta) x_{j,t}}{\sum_j x_{j,t}}. \tag{27.32}$$

Comparing the unfiltered, $f(x)$, and the spatially filtered, $f(\tilde{x})$, distributions, Fischer and Stumpner (2008) find that the latter displays much less dispersion than the former, yet over the 1995–2003 period, the level of dispersion in the filtered series increases by 15%, while the unfiltered series experiences an actual decline in dispersion of 3.3%. Because the unfiltered series is highly spatially autocorrelated, the overall finding of σ-convergence is attributed to the role of the spatial dependence.

While the spatial filtering and regional conditioning approaches provide avenues to explore the impact of spatial dependence on the evolution of regional income densities, several questions remain. First, as Fischer and Stumpner (2008) note, there is currently a lack of a formal inferential framework to test hypotheses about distribution dynamics in the presence of spatial effects. Second, the filters applied in (27.30) and (27.32) keep $w_{i,j}$ fixed over time. In other words, the spatial structure is specified as time-invariant. Finally, the interpretation of just what the identified spatial components represent is not at all clear.

27.3.4 Space-time kernels

Rather than filtering out the spatial component, kernels that explicitly incorporate space can be developed. Figure 27.7 contains examples of how this can be done.

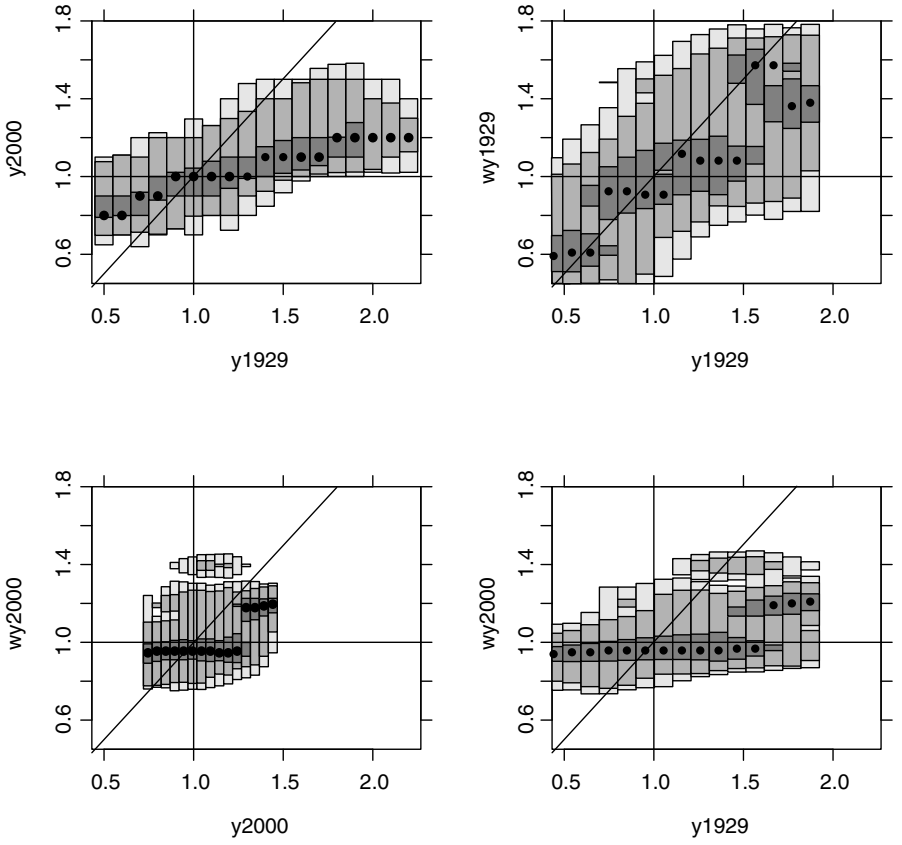


Figure 27.7 Space-time kernels US
Note: y1929 is relative income in 1929 (per capita income for a state relative to national per capita income), wy 1929 is the spatial lag of relative income for 1929 (see section 27.2.2).

The conventional HDR plot for US incomes from 1929 and 2000 is contained in the upper left figure. Next to it is the Moran HDR plot, which extends the Moran scatterplot to a conditional kernel density for the year 1929. The strong spatial autocorrelation is clearly evident as the modes for the conditional distributions for the spatial lag of high-income states in 1929 are all above the average, while for lower-income states, the mode of the conditional distributions are all below average.

The Moran HDR for 2000 in the lower left of the figure is radically different from the HDR for 1929. First, all the conditional spatial lag distributions display much less dispersion in 2000 than in 1929, while the range of the marginal distribution for relative incomes in 2000 is also much narrower than in 1929. Second, the strength of the spatial autocorrelation has weakened over the period, as reflected in the number of conditional modes located in the HH or LL quadrants of the HDR scatter dropping to 10 in 2000 from 15 in 1929.

The space-time Moran HDR in the lower right panel of Figure 27.7 provides a composite view of the spatial dynamics over this period. Here the spatial lag for incomes in 2000 is conditioned on state income in 1929. In contrast to the Moran HDR for 1929, the mode of the conditional lag distributions in 2000 display linear independence for low and moderate values of income in 1929. Moreover, in all cases, the respective modes of the conditional lag distributions have moved towards the overall average income value, with the convergence being upward for the conditional distributions below one, and downward for conditional distributions for $y_{1929} > 1.0$.

27.4 Conclusion

The previous two sections have reviewed a rapidly evolving body of literature concerned with spatial analysis of economic convergence. In both the formal confirmatory work and more recent exploratory data analysis, the unique challenges that spatially organized data pose continue to attract much attention. In this section we offer some final thoughts and identify some remaining outstanding issues for future research.

In the study of regional economic convergence it has become abundantly clear that spatial dependence tends to be the rule rather than the exception in regional income and product series. As a result, there is a growing recognition of the importance of treating this form of dependence in both formal confirmatory econometric models as well as in newer exploratory methods of data analysis.

While spatial autocorrelation and spatial dependence have attracted the majority of the attention in the literature, spatial heterogeneity has also been recognized as an important dimension of many regional series. In general terms, however, the treatment of spatial heterogeneity is more easily done using traditional (that is, a-spatial) econometric methods, while spatial dependence has necessitated a new body of models and methods.

Despite the growing awareness of spatial dependence and heterogeneity in empirical work, most studies tend to focus on only one type of spatial effect. Work on developing specification strategies within spatial econometrics (Anselin and Rey, 1991; Anselin and Florax, 1995; Florax *et al.*, 2003) has provided guidance to practitioners on how to detect and discriminate between different forms of spatial dependence, yet similar approaches to deal with alternative forms of geographical heterogeneity are lacking. Durlauf and Johnson (1995) use regression trees to examine whether international growth data obey a single Solow-type growth model or if there are specific sub-groups or regimes with distinct parameter values. They find evidence of substantial geographic homogeneity within the subgroups. That is, model parameters are found to vary across the regional regimes but are assumed to hold for all economies within a given regime. As acknowledged by Durlauf and Johnson, there is no asymptotic theory available to access the statistical significance of the identified regimes. An additional qualification offered is that there is the possibility that the parameter heterogeneity could hold within each regime as well as across the regimes. Moreover, the question of how to deal with the joint presence of spatial dependence and spatial heterogeneity remains unanswered.

Spatial dependence and heterogeneity do not exhaust the class of spatial effects that confront the regional convergence literature. The issue of the appropriate spatial scale has largely been ignored as the specification of the geographical unit of analysis has tended to be based on data availability rather than theoretical concerns. In the context of convergence research, the geographical unit has varied from countries, to regions, states and metropolitan areas (Drennan, 2005). As a vast literature in spatial analysis has demonstrated (cf. Wong and Amrhein, 1996), the modifiable area unit problem can give rise to inferences that are not robust to changes in the spatial scale and aggregation of the data.

The aggregation issue can also be confounded with the regional heterogeneity question. As work by Miller and Genc (2005) demonstrates, alternative definitions of regional groupings can lead to different inferences regarding the speed of convergence. Similarly, Rey (2004a) shows that changes in regional definitions can have similar impacts on measuring regional inequality dynamics. In these studies the groups are taken exogenously with regard to administrative boundaries, yet the possibility of endogenously determining these groupings was touched on in section 27.2.3. Regionally constrained clustering algorithms (Rey and Anselin, 2007) could be used to determine spatially explicit convergence clubs.

In existing studies of regional convergence, the underlying spatial structure has been assumed time invariant. Over the short run, the assumptions that the boundaries of economies, spatial weights matrices, and the composition of convergence clubs are unchanging, are likely to be plausible. However, as the period under consideration lengthens, such assumptions become increasingly untenable. A critical area for future research will be the integration of evolving spatial structure into formal growth models and distribution dynamics approaches.

Closely related to the issue of spatial scale is the relationship between regional inequality dynamics and personal income distribution dynamics. In the case of the US we have witnessed an apparent paradox of convergence over the last 150 years at the state and regional scales, yet strong evidence of growing polarization in personal and household income distributions (Jones and Weinberg, 2000). Regional patterns in personal income inequality, that is, the spatial distribution of personal income distributions, has attracted some attention (Bishop *et al.*, 1994; Levernier *et al.*, 1995; Partridge *et al.*, 1996; Morrill, 2000). However, the link between convergence in the mean of these distributions (that is, regional convergence) and the evolution of increasing inequality between individuals within these distributions remains largely unexplored and is an important avenue for future research. In particular, the relationship between spatial clustering at one scale and the pace of convergence at a higher spatial scale has received only limited attention (Janikas and Rey, 2008).

Finally, although we organized our review along the dimensions of confirmatory and exploratory analysis, we see these approaches as complements rather than substitutes. We can identify several areas where cross-fertilization between these literatures is likely to lead to new advances. The first is the use of ESDA methods applied to regional data series to identify interesting new patterns from which suggestions for new types of theories and hypotheses about the spatial nature of economic convergence may emerge. Rather than seeing ESDA as a case of

"measurement without theory," such applications may actually lead to theoretical advances.

The gap between the restrictive nature of most growth theory on the one hand and empirical complexity of regional datasets on the other has clearly been a motivating factor in the turn to EDA methods in convergence analysis. But rather than abandoning the classical regression based approach, we suggest that using ESDA and EDA methods to refine model specifications could work to relax some of the restrictions. In this regard, there is a rich conceptual literature on spatial poverty traps (cf. Bowles *et al*., 2006) that could, in our view, be tied to recent work in ESDA to develop operational measures for these theoretical constructs. A final area is the use of ESDA methods in enhanced diagnostic methods for spatial econometric modeling (Ord, 2008).

Notes

1. We focus here on the essential properties of these models, as they have been applied in the growth econometrics literature. For extensive technical discussion on these models, see Anselin and Bera (1998), Anselin (2003, 2006).
2. In assessing the effect of structural funds on the European regional convergence process, Dall'erba and Le Gallo (2008) use a spatial lag model and estimate it with IV by considering that both the spatial lag variable and the structural funds variable are endogenous.
3. See Fingleton and López-Bazo (2006) for a more complete description of papers including spatial effects in growth and Verdoorn regressions.
4. This assumption corresponds to one of Kaldor's stylized facts.
5. For a recent overview of software for ESDA, see Rey and Anselin (2006).
6. Since the probability of moving out of the current income class is 1 in each period and class.
7. The solid black circle indicates the first year in the time path for each state.
8. The conditional kernel and HDR plot were generated using the HDRCDE package (Hyndman, 1996; Hyndman *et al*., 1996).

References

Abreu, M., H. de Groot and R. Florax (2005) Space and growth: a survey of empirical evidence and methods. *Région et Développement* **21**, 13–44.

Angeriz, A., J. McCombie and M. Roberts (2008) New estimates of returns to scale and spatial spillovers for EU regional manufacturing, 1986–2002. *International Regional Science Review* **31**, 62–87.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. (1995) Local indicators of spatial association: LISA. *Geographical Analysis* **27**, 93–115.

Anselin, L. (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H. Scholten and D. Unwin (eds.), *Spatial Analytical Perspectives on GIS*. London: Taylor and Francis.

Anselin, L. (2003) Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review* **26**, 153–6.

Anselin, L. (2006) Spatial econometrics. In T.C. Mills, A. and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*. Basingstoke: Palgrave Macmillan.

Anselin, L. and A. Bera (1998) Spatial dependence in linear regression models with an application to spatial econometrics. In A. Ullah and D. Giles (eds.), *Handbook of Applied Economics Statistics*. New York. Marcel Dekker.

Anselin, L. and R. Florax (1995) Small sample properties of tests for spatial dependence in regression models. In L. Anselin and R. Florax (eds.), *New Directions in Spatial Econometrics*, Advances in Spatial Science. Berlin: Springer-Verlag.

Anselin, L., R.J. Florax and S. Rey (2004) *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Berlin: Springer-Verlag.

Anselin, L., J. Le Gallo and H. Jayet (2008) Spatial panel econometrics. In L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data*. Dordrecht: Kluwer Academic.

Anselin, L. and S. Rey (1991) Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* **23**, 112–31.

Arbia, G. (2006) *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Berlin: Springer.

Arbia, G., R. Basile and M. Salvatore (2003) Measuring spatial effects in parametric and non-parametric modelling of regional growth and convergence. In *UNU/WIDER Project Meeting on Spatial Inequality in Development*. Available at http://website1.wider.unu.edu/conference/conference-2003-2/conference%202003-2-papers/inequality%20project%20meeting%20papers/Basile.pdf.

Arbia, G. and G. Piras (2005) Convergence in per-capita GDP across EU-NUTS2 regions using panel data models extended to spatial autocorrelation effects. REAL Working Paper No. 05-T-3.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277–97.

Azariadis, C. and A. Drazen (1990) Threshold externalities in economic development. *Quarterly Journal of Economics* **105**, 501–26.

Azzoni, C.R. and R. Silveira-Neto (2006) Location and regional income disparity dynamics: The Brazilian case. *Papers in Regional Science* **85**(4), 599–613.

Badinger, H., W. Muller and G. Tondl (2004) Regional convergence in the European Union, 1985–1999: a spatial dynamic panel analysis. *Regional Studies* **38**, 241–54.

Barro, R.J. and X. Sala-i-Martin (1992) Convergence. *Journal of Political Economy* **100**, 223–51.

Basile, R. (2006) Intra-distribution dynamics of regional per capita income in Europe: evidence from alternative conditional density estimators. In *Papers Presented at the 53rd North American Meetings of the RSAI*. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=931106.

Basile, R. and B. Gress (2005) Semi-parametric spatial auto-covariance models of regional growth in Europe. *Région et Développement* **21**, 93–118.

Bernat, G.A. (1996) Does manufacturing matter? A spatial econometric view of Kaldor's laws. *Journal of Regional Science* **36**, 463–77.

Bianchi, M. (1997) Testing for convergence: evidence from non-parametric multimodality tests. *Journal of Applied Econometrics* **12**, 393–408.

Bickenbach, F. and E. Bode (2003) Evaluating the Markov property in studies of economic convergence. *International Regional Science Review* **26**, 363–92.

Bishop, J., J. Formby and P. Thistle (1994) Convergence and divergence of regional income distributions and welfare. *Review of Economics and Statistics* **76**(2), 228–35.

Bivand, R.S. and R.J. Brunstad (2003) Regional growth in Western Europe: an empirical exploration of interactions with agriculture and agricultural policy. In B. Fingleton (ed.), *European Regional Growth*. Berlin: Springer-Verlag.

Bowles, S., S.N. Durlauf and K. Hoff (2006) *Poverty Traps*. Princeton: Russell Sage Foundation.

Boyle, G. and T. McCarthy (1997) Simple measures of convergence in per capita GDP: a note on some further international evidence. Economics, Finance and Accounting Department Working Paper Series No. 751197, Department of Economics, Finance and Accounting, National University of Ireland, Maynooth.

Brock, W. and S. Durlauf (2001) Growth economics and reality. *World Bank Economic Review* **15**, 229–72.
Bulli, S. (2001) Distribution dynamics and cross-country convergence: a new approach. *Scottish Journal of Political Economy* **48**, 226–43.
Burridge, P. (1981) Testing for a common factor in a spatial autoregressive model. *Environment and Planning A* **13**, 795–800.
Canova, F. (2004) Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review* **45**, 49–77.
Canova, F. and A. Marcet (1995) The poor stay poor: non-convergence across countries and regions. Universitat Pompeu Fabre Working Paper.
Carlino, G.A. and L.O. Mills (1993) Are U.S. regional incomes converging? A time series analysis. *Journal of Monetary Economics* **32**, 335–46.
Carrington, A. (2003) A divided Europe? Regional convergence and neighborhood spillover effects. *Kyklos* **3**, 381–93.
Carrington, A. (2006) Regional convergence in the European Union: a stochastic dominance approach. *International Regional Science Review* **29**(1), 64.
Caselli, F., G. Esquivel and F. Lefort (1996) Reopening the convergence debate: a new look at cross-country growth empirics. *Journal of Economic Growth* **1**(3), 363–90.
Cressie, N. (1993) *Statistics for Spatial Data*. New York: Wiley.
Dall'erba, S. and J. Le Gallo (2008) Regional convergence and the impact of European structural funds over 1989–1999: a spatial econometric analysis. *Papers in Regional Science* **87**, 219–44.
Dall'erba, S., M. Percoco and G. Piras (2008) The European regional growth process revisited: increasing returns and spatial dynamic setting. *Spatial Economic Analysis* **3**, 7–25.
Desdoigts, A. (1999) Patterns of economic development and the formation of clubs. *Journal of Economic Growth* **4**, 305–30.
Dixit, A. and J. Stiglitz (1977) Monopolistic competition and optimum product diversity. *American Economic Review* **67**, 297–308.
Doppelhofer, G., R. Miller and X. Sala-i-Martin (2004) Determinants of long-term growth: A Bayesian averaging of classical estimated (BACE) approach. *American Economic Review* **94**(4), 813–35.
Drennan, M. (2005) Possible sources of wage divergence among metropolitan areas of the United States. *Urban Studies* **42**(9), 1609.
Durlauf, S.N. and P.A. Johnson (1995) Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* **10**, 365–84.
Durlauf, S., P. Johnson and J. Temple (2005) Growth econometrics. In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*. Amsterdam: North-Holland.
Durlauf, S., A. Kourtellos and A. Minkin (2001) The local Solow growth model. *European Economic Review* **45**, 928–40.
Durlauf, S.N. and D.T. Quah (1999) The new empirics of economic growth. In J.B. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics: Volume 1A*. Amsterdam: North-Holland.
Eckey, H., R. Kosfeld and M. Türck (2007) Regional convergence in Germany: a geographically weighted regression approach. *Spatial Economic Analysis* **2**(1), 45–64.
Egger, P. and M. Pfaffermayr (2006) Spatial convergence. *Papers in Regional Science* **85**(2), 199–215.
Ertur, C. and W. Koch (2007) Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of Applied Econometrics* **22**, 1023–62.
Ertur, C., J. Le Gallo and C. Baumont (2006) The European regional convergence process, 1980–1995: do spatial dependence and spatial heterogeneity matter? *International Regional Science Review* **29**(1), 2–34.
Ertur, C., J. Le Gallo and J. LeSage (2007) The European regional convergence process, 1980–1995: do spatial dependence and spatial heterogeneity matter? *Review of Regional Studies* **37**, 82–108.

Fernandez, C., E. Ley and M. Steel (2001) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**, 563–76.
Fingleton, B. (1999) Estimates of time to economic convergence: an analysis of regions of the European Union. *International Regional Science Review* **22**, 5–35.
Fingleton, B. (2000) Spatial econometrics, economic geography, dynamics and equilibrium: a "third way"? *Environment and Planning A* **32**, 1481–98.
Fingleton, B. (2001a) Equilibrium and economic growth: spatial econometric models and simulation. *Journal of Regional Science* **41**, 117–47.
Fingleton, B. (2001b) Theoretical economic geography and spatial econometrics: dynamic perspectives. *Journal of Economic Geography* **1**(2), 201–25.
Fingleton, B. (2004) Regional economic growth and convergence: insights from a spatial econometric perspective. In L. Anselin, R. Florax and S.J. Rey (eds.), *Advances in Spatial Econometrics*. Berlin: Springer-Verlag.
Fingleton, B. (2008) A generalized method of moments estimator for a spatial model with moving average errors, with application to real estate prices. *Empirical Economics* **34**(1), 35–57.
Fingleton, B. and J. Le Gallo (2008) Finite sample properties of estimators of spatial models with autoregressive, or moving average, disturbances and system feedback. *Annals of Economics and Statistics*. Forthcoming.
Fingleton, B. and E. López-Bazo (2006) Empirical growth models with spatial effects. *Papers in Regional Science* **85**(2), 177–98.
Fingleton, B. and J. McCombie (1998) Increasing returns and economic growth: some evidence for manufacturing from the European Union regions. *Oxford Economic Papers* **50**, 89–105.
Fischer, M. and C. Stirböck (2006) Pan-European regional growth and club-convergence; insights from a spatial econometric perspective. *Annals of Regional Science* **40**(4), 693–721.
Fischer, M.M. and P. Stumpner (2008) Income distribution dynamics and cross-region convergence in Europe. *Journal of Geographical Systems* **10**, 109–39.
Florax, R., H. Folmer and R. Rey (2003) Specification searches in spatial econometrics: the relevance of hendry's methodology. *Regional Science and Urban Economics* **33**, 557–79.
Fotheringham, A., C. Brundson and M. Charlton (2004) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York: Wiley.
Garrett, T., G. Wagner and D. Wheelock (2007) Regional disparities in the spatial correlation of state income growth, 1977–2002. *Annals of Regional Science* **41**(3), 601–18.
Getis, A. and D. Griffith (2002) Comparative spatial filtering in regression analysis. *Geographical Analysis* **34**(2), 130–40.
Getis, A. and J. Ord (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**(3), 189–206.
Gezici, F. and G. Hewings (2004) Regional convergence and the economic performance of peripheral areas in Turkey. *Review of Urban and Regional Development Studies* **16**(2), 113–32.
Hammond, G. (2004) Metropolitan/non-metropolitan divergence: a spatial Markov chains approach. *Papers in Regional Science* **83**, 543–63.
Hastie, T. and R. Tibshirani (1993) Varying coefficient models. *Journal of the Royal Statistical Society* **B55**, 757–96.
Hyndman, R. (1996) Computing and graphing highest density regions. *American Statistician* **50**, 120–6.
Hyndman, R., D. Bashtannyk and G. Grunwald (1996) Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statisitics* **5**, 313–36.
Islam, N. (1995) Growth empirics: a panel data approach. *Quarterly Journal of Economics* **110**, 1127–70.
Islam, N. (2003) What have we learnt from the convergence debate? *Journal of Economic Surveys* **17**(3), 309–62.
Janikas, M.V. (2007) Comparative regional income dynamics: clustering, scale and geocomputation. PhD thesis, San Diego State University.

Janikas, M.V. and S.J. Rey (2008) On the relationship between spatial clustering, inequality, and economic growth in the United States: 1969–2000. *Region et Developpement* **27**, 13–34.
Jones, A.F. and D.H. Weinberg (2000) The changing shape of the nation's income distribution. Current Population Report, U.S. Census Bureau.
Kaldor, N. (1975) Economic growth and Verdoorn's law: a comment on Mr Rowthorn's article. *Economic Journal* **85**, 891–6.
Kelejian, H. and I. Prucha (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model. *Journal of Real Estate Finance and Economics* **17**, 99–121.
Kelejian, H. and I. Prucha (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* **40**, 509–34.
Kelejian, H. and I. Prucha (2007) HAC estimation in a spatial framework. *Journal of Econometrics* **140**, 131–54.
Kosfeld, R., H. Eckey and C. Dreger (2006) Regional productivity and income convergence in the Unified Germany, 1992–2000. *Regional Studies* **40**(7), 755–67.
Lall, S. and Z. Shalizi (2003) Location and growth in the Brazilian Northeast. *Journal of Regional Science* **43**(6), 663–82.
Le Gallo, J. (2004) Space-time analysis of GDP disparities across European regions: a Markov chains approach. *International Regional Science Review* **27**(2), 138–63.
Le Gallo, J., C. Baumont, S. Dallerba and C. Ertur (2005) On the property of diffusion in the spatial error model. *Applied Economics Letters* **12**(9), 533–6.
Le Gallo, J. and S. Dall'erba (2006) Evaluating the temporal and spatial heterogeneity of the European convergence process, 1980–1999. *Journal of Regional Science* **46**(2), 269–88.
Le Gallo, J. and C. Ertur (2003) Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe. *Papers in Regional Science* **82**, 175–201.
Le Gallo, J., C. Ertur and C. Baumont (2003) A spatial econometric analysis of convergence across European regions, 1980–1995. In B. Fingleton (ed.), *European Regional Growth*. Advances in Spatial Science. Berlin: Springer-Verlag.
Lee, S. (2004) Spatial data analysis for the U.S. regional income convergence, 1969–1999: a critical appraisal of β-convergence. *Journal of the Korean Geographical Society* **39**(2), 212–28.
LeSage, J. and M. Fischer (2007) Spatial growth regressions: model specification, estimation and interpretation. Mimeo.
LeSage, J. and O. Parent (2007) Bayesian model averaging for spatial econometric models. *Geographical Analysis* **39**(3), 241–67.
Levernier, W., D. Rickman and M. Partridge (1995) Variation in US state income inequality: 1960–1990. *International Regional Science Review* **18**(3), 355–78.
Liu, Z. and T. Stengos (1999) Non-linearities in cross-country growth regressions: a semiparametric approach. *Journal of Applied Econometrics* **14**, 527–38.
López-Bazo, E., E. Vaya and M. Artis (2004) Regional externalities and growth: evidence from European regions. *Journal of Regional Science* **44**, 43–73.
López-Bazo, E., E. Vaya, A.J. Mora and J. Suriñach (1999) Regional economic dynamics and convergence in the European Union. *Annals of Regional Science* **33**, 343–70.
Magalhães, A., G.J. Hewings and C.R. Azzoni (2005) Spatial dependence and regional convergence in Brazil. *Investigaciones Regionales* **6**, 5–20.
Magrini, S. (1999) The evolution of income dynamics among regions of the European Union. *Regional Science and Urban Economics* **29**, 257–81.
Magrini, S. (2004) Regional (di)convergence. In V. Henderson and J. Thisse (eds.), *Handbook of Regional and Urban Economics*. New York: Elsevier.
Mankiw, N., D. Romer and D. Weil (1992) A contribution to the empirics of economic growth. *Quarterly Journal of Economics* **107**, 407–37.
Maurseth, P. (2001) Convergence, geography and technology. *Structural Change and Economic Dynamics* **12**, 247–76.
Miller, J. and I. Genc (2005) Alternative regional specification and convergence of US regional growth rates. *Annals of Regional Science* **39**(2), 241–52.

Morrill, R. (2000) Geographic variation in change in income inequality among US states, 1970–1990. *Annals of Regional Science* **34**(1), 109–30.
Mossi, M., P. Aroca, I. Fernandez and C. Azzoni (2004) Regional disparities in the EU: mobility and polarization. *Applied Economics Letters* **11**, 517–22.
Neven, D. and C. Gouyette (1995) Regional convergence in the European Community. *Journal of Common Market Studies* **33**, 47–65.
Ord, J.K. (2008) Spatial autocorrelation: a statistician's view. In L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*. Berlin: Springer.
Ord, J. and A. Getis (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* **27**, 286–305.
Pace, R. Kelley and J. LeSage (2004) Spatial autoregressive local estimation. In A. Getis, J. Mur and H. Zoller (eds.), *Spatial Econometrics and Spatial Statistics*. Basingstoke: Palgrave Macmillan.
Pace, R. Kelley and J. LeSage (2007) Interpreting spatial econometric models. Mimeo.
Paez, A., T. Uchida and K. Miyamoto (2002) A general framework for estimation and inference of geographically weighted regression models: 2. spatial association and model specification tests. *Environment and Planning A* **34**, 883–904.
Partridge, M., D. Rickman and W. Levernier (1996) Trends in US income inequality: evidence from a panel of states. *Quarterly Review of Economics and Finance* **36**(1), 17–37.
Patacchini, E. and P. Rice (2007) Geography and economic performance: exploratory spatial data analysis for Great Britain. *Regional Studies* **41**, 489–508.
Pons-Novell, J. and E. Viladecans-Marsal (1999) Kaldor's laws and spatial dependence: evidence for the European regions. *Regional Studies* **35**, 443–51.
Quah, D. (1993a). Empirical cross-section dynamics in economic growth. *European Economic Review* **37**, 426–37.
Quah, D. (1993b). Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics* **95**, 427–43.
Quah, D. (1996a). Empirics for economic growth and convergence. *European Economic Review* **40**, 1353–75.
Quah, D. (1996b). Regional convergence clusters across Europe. *European Economic Review* **40**(3-5), 951–8.
Ramajo, J., M. Marquez, G. Hewings and M. Salinas (2008) Spatial heterogeneity and interregional spillovers in the European Union: do cohesion policies encourage convergence across regions? *European Economic Review* **52**(3), 551–67.
Rey, S.J. (2001) Spatial empirics for economic growth and convergence. *Geographical Analysis* **33**(3), 195–214.
Rey, S.J. (2004a) Spatial analysis of regional income inequality. In M. Goodchild and D. Janelle (eds.), *Spatially Integrated Social Science: Examples in Best Practice*. Oxford: Oxford University Press.
Rey, S.J. (2004b) Spatial dependence in the evolution of regional income distributions. In A. Getis, J. Múr and H. Zoeller (eds.), *Spatial Econometrics and Spatial Statistics*. Basingstoke: Palgrave Macmillan.
Rey, S.J. and L. Anselin (2006) Recent advances in software for spatial analysis in the social sciences. *Geographical Analysis* **38**, 1–4.
Rey, S.J. and L. Anselin (2007) PySAL: a Python library of spatial analytical methods. *Review of Regional Studies* **37**(1), 5–27.
Rey, S.J. and M.G. Boarnet (2004) A taxonomy of spatial econometric models for systems of simultaneous equations. In L. Anselin, R. Florax and S.J. Rey (eds.), *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Berlin: Springer-Verlag.
Rey, S.J. and B. Dev (2006) σ-convergence in the presence of spatial effects. *Papers in Regional Science* **85**(2), 217–34.
Rey, S.J. and M.V. Janikas (2005) Regional convergence, inequality and space. *Journal of Economic Geography* **5**(2), 155–76.

Rey, S.J. and M.V. Janikas (2006) STARS: space-time analysis of regional systems. *Geographical Analysis* **38**, 67–84.
Rey, S.J. and B.D Montouri (1999) U.S. regional income convergence: a spatial econometric perspective. *Regional Studies* **33**, 143–56.
Rey, S. J. and X. Ye (2008) Comparative spatial dynamics of regional systems. Unpublished manuscript.
Roberts, M. (2004) The growth performance of the GB counties: some empirical evidence for 1977–1993. *Regional Studies* **38**, 149–65.
Rowthorn, R. (1975a) A reply to Lord Kaldor's law? *Economic Journal* **85**, 897–901.
Rowthorn, R. (1975b) What remains of Kaldor's law? *Economic Journal* **85**, 10–19.
Sala-i-Martin, X. (1996) Regional cohesion: evidence and theories of regional growth and convergence. *European Economic Review* **40**(6), 1325–52.
Sala-i-Martin, X. (1997) I just ran two million regressions. *American Economic Review* **87**(2), 178–83.
Savin, N. (1984) Multiple hypothesis testing. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics, Volume II*. Elsevier Science.
Shorrocks, A. (1978) Income inequality and income mobility. *Journal of Economic Theory* **19**, 376–93.
Temple, J. (1999) The new growth evidence. *Journal of Economic Literature* **37**, 112–56.
Tukey, J. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.
Villaverde, J. (2005) Provincial convergence in Spain: a spatial econometric approach. *Applied Economics Letters* **12**, 697–700.
Villaverde, J. (2006) A new look to convergence in Spain: a spatial econometric approach. *European Urban and Regional Studies* **13**, 131–41.
Webber, D.J., P. White and D.O. Allen (2005) Income convergence across U.S. states: an analysis using measures of concordance and discordance. *Journal of Regional Science* **4**, 565–89.
Wheeler, D. and M. Tiefelsdorf (2005) Multicollinearity and correlation among local regression coefficients in geographical weighted regression. *Journal of Geographical Systems* **7**, 161–87.
Wong, D. and C. Amrhein (1996) Research on the MAUP: old wine in a new bottle or real breakthrough. *Geographical Systems*, **3**(2–3), 73–76.
Yildirim, J. and N. Öcal (2006) Income inequality and economic convergence in Turkey. *Transition Studies Review* **13**(3), 559–68.
Ying, L. (2000) Measuring the spillover effects: Some Chinese evidence. *Papers in Regional Science* **79**(1), 75–89.

Part X
Applied Econometrics and Computing

This page intentionally left blank

28
Testing Econometric Software

B.D. McCullough

Abstract

The first part of this chapter is a non-technical survey of the relatively sparse literature on testing the accuracy of econometric software. Accuracy is primarily assessed by taking a test problem, with known inputs and outputs, giving it to the software, and comparing the software's output with the output of the test problem. We discuss the various types of tests (introductory, intermediate, and advanced) and the types of errors that these tests have uncovered. The reader is directed to specific resources for further information. The second part, which is technical, constructs a test problem (i.e., benchmark) for autoregressive moving average (ARMA) estimation. In 1994 it was reported in the literature that different packages give different answers to the same ARMA estimation problem. To date, this open problem has been unresolved. We provide benchmarks for conditional least squares and unconditional least squares ARMA estimation.

28.1 Introduction

McCullough (2000a) rhetorically entitled an article, "Is it safe to assume that software is accurate?" The answer, of course, is "No." Testing econometric software

is important because econometric software does not always produce accurate answers. Sometimes the inaccuracy can be traced to a bug, an incorrectly written line of code; at other times the algorithm is simply not up to the task of calculating the correct answer (e.g., as will be discussed in section 28.2, do not use the "calculator formula" to compute the sample variance, and, as will be discussed in section 28.3, do not invert the $(X'X)$ matrix when calculating the least squares regression coefficients $\hat{\beta}$; neither of these approaches is numerically sound for economic data). Unless the software is tested, the user has nothing more than mere hope or the software developer's blandishments on which to base his or her confidence that the computer program's output is correct. As we shall see, such blind confidence is misplaced.

Forty years ago, back when mainframe computers took their input from lines of code on punchcards, and a program consisted of a stack of punchcards, Longley (1967) worked out by hand the solution to a linear regression problem with a constant and six independent variables, for 16 observations, and he did so to several digits of accuracy. The dependent variable was "total employment" and the independent variables were: implicit price deflator, gross national product, unemployment, size of armed forces, population, and time. When he compared his hand-calculated results to those from the computer programs, he found wildly disparate results. As he put it (*ibid.*, p. 822):

> Test problems were run on available programs for use on the following electronic computers: IBM 1401, 7070/7074/ 360 model 50, 7090/7094, and General Electric 235. With identical inputs, all except four programs produced outputs which differed from each other in every digit.

To convey the essence of what Longley found, below we reproduce (to only three decimals) some of the regression coefficients produced by four computers/software packages from his Table 10 in our Table 28.1.

Within several years, most (if not all) linear regression programs could meet the so-called "Longley benchmark." In the present day, it is almost unheard of for any program not to meet the Longley benchmark. One may think that several different packages giving several different answers to the same problem is a relic of the "old" days, but it is very much a problem that still persists. In 1999, McCullough and Vinod (1999, p. 534) published three widely divergent sets of full information

Table 28.1 Some of Longley's regression results

	Defl.	*GNP*	*Unem.*	*Mil.*	*Pop.*	*Time*	*Constant*
Correct	15.026	−0.036	−2.020	−1.033	−0.051	1829	−3482258
IBM 7074	−36.187	0.059	−0.593	−0.607	−0.344	183	−269126
G.E. 235	13.944	−0.346	−2.00	−1.028	−0.055	1809	−344297
BMD	27.082	−0.032	−1.946	−0.987	−0.013	1653	146264
NIPD	−0.039	0.536	−0.068	−0.064	−3.44	31.5	−5.22

Table 28.2 Multivariate GARCH results

Package	μ_c	μ_f	c_1	a_1	b_1	c_2	a_2	b_2	c_3	a_3	b_3
Package A	0.064	0.064	0.377	0.128	0.411	0.566	0.145	0.365	0.474	0.128	0.348
Package B	0.062	0.069	0.012	0.041	0.946	0.012	0.034	0.956	0.011	0.035	0.953
Package C	0.061	0.037	0.010	0.037	0.952	0.010	0.031	0.961	0.009	0.032	0.959
Package D	0.073	0.082	0.076	0.112	0.798	0.125	0.134	0.762	0.099	0.120	0.773

maximum likelihood (FIML) estimates for Klein's Model I. Clearly, not all three sets of answers can be correct. McCullough and Renfro (1999) published seven different estimates for the parameters of the same generalized autoregressive conditional heteroskedasticity (GARCH) model. What does this make one think of the GARCH results published in the literature? As far as GARCH estimation is concerned, recall that any nonlinear estimation procedure requires starting values which need to be carefully chosen if the solver is to have much of a chance to find an extremum. Some GARCH procedures in some packages do not allow the user to set the starting values! What does one think of the developers of such packages? (What do the developers of such packages think of their users?) Brooks, Burke and Persand (2003) did the same thing for multivariate GARCH models, and it is instructive to present the model they estimated; the results from their Table II are given in our Table 28.2.

The multivariate GARCH model estimated by Brooks, Burke and Persand was:

$$
\begin{aligned}
s_t &= \mu_s + \epsilon_{s,t} \\
f_t &= \mu_f + \epsilon_{f,t} \\
h_{s,t} &= c_1 + a_1 \epsilon^2_{s,t-1} + b_1 h_{s,t-1} \\
h_{f,t} &= c_2 + a_2 \epsilon^2_{f,t-1} + b_2 h_{f,t-1} \\
h_{s,f,t} &= c_3 + a_3 \epsilon_{s,t-1} \epsilon_{f,t-1} + b_3 h_{s,f,t-1}.
\end{aligned}
$$

Study the coefficients in Table 28.2. Each package estimates completely different parameters for the same model. What does this say about multivariate GARCH results published in the literature? Perhaps the multivariate GARCH likelihood is complicated and difficult to optimize, and may have multiple optima. Not that these numerical difficulties are an excuse for inaccuracy, but maybe a user could have more faith in a simple nonlinear estimation problem for a convex likelihood like the probit model? Surely nothing could go wrong with that? Stokes (2004) gave the same probit problem to six packages and got six different answers. What is particularly interesting is that, for this particular probit problem, no solution exists. The problem, as posed, exhibited what is called "complete separability" and so there is no set of parameters that maximizes the likelihood. This phenomenon is ignored by most econometrics texts that present the probit model (but see Davidson and Mackinnon, 2004, pp. 458–9, for an exception). This minor impediment did not stop the six packages from reporting that they had found solutions. A seventh

package correctly reported that the problem has no solution. If you were to estimate a probit model, which package would you prefer to use?

Do not think that in the present day this lack of accuracy affects only nonlinear problems. McCullough (2004a) reported on econometrics packages that produced correlation coefficients larger than unity. Further evidence of inaccuracy of statistical and econometric software will be presented in subsequent sections. For the present, what is of interest is this: "How can a user protect him/herself against inaccurate econometric software?" The answer is, "Test the software." The natural follow-up question, "How?," is the subject of this chapter.

In its purest form, to test a software package is to give it some input for which the correct output is known. In the aforementioned case of the Longley benchmark, the input is the dependent and independent variables, and the correct output is the coefficients that he calculated by hand. For some problems it is too difficult to work out the correct answer, and an alternative is to have two (or more) independent programming efforts reach the same conclusion. This is the method used by Drukker and Guan (2003) to produce a benchmark for a particular panel data estimator. This chapter will concern itself with the former method – known inputs and known outputs. There are three classes of tests: introductory, intermediate, and advanced. We provide an overview of each class, including a description of inaccuracies uncovered, with directions to further reading for those who wish to read up on the subject. Section 28.2 briefly introduces a few necessary ideas about the numerical limitations of computational software. Section 28.3 discusses a set of introductory tests, known as "Wilkinson tests." Section 28.4 discusses a set of intermediate tests. Section 28.5 discusses a pair of advanced tests, one for FIML estimation and one for GARCH estimation. To illuminate the numerical issues involved, Section 28.6 creates a benchmark for ARMA estimation. Section 28.7 offers some conclusions.

In what follows we typically do not mention specific software packages, even when they were identified in the article or review that is cited. The reason for this is that many of the errors we recount have long since been corrected. A notable exception is Microsoft Excel, because Microsoft has a track record of not fixing errors in Excel. Should Excel even be mentioned in this chapter? Yes. One need only surf the web to find that many professors teach introductory econometrics with Excel.

28.2 Computer arithmetic

Generally, computers do not store numbers that are perfectly accurate, nor are the calculations performed on those numbers carried out with perfect accuracy. A computer stores numbers in bits (a bit is simply a single binary digit that holds either a zero or a one) and bytes (eight bits to a byte). Currently, most desktop computers have a 32-bit word length, but there do exist 64-bit computers, and soon there will be 256-bit computers.

To make ideas concrete, suppose that we have a 4-bit word, with each bit being a zero or a one. All counting must be in base-2 with up to four places. Zero is

represented as 0000, one is represented as 0001, two is represented as 0010, and three is represented as 0011. The biggest number that can be represented is 15, which is represented as 1111 = 1000 (8) + 0100 (4) + 0010 (2) + 0001 (1).

With a 32-bit word, rather than devote all 32 bits to representing powers of two, we use some for a *mantissa* and others for an exponent (which can be either positive or negative) to which the number two can be raised. This two raised to the exponent is then multiplied by the mantissa. Such a scheme provides for a wider range of representable numbers. This single-precision scheme has one sign bit, eight bits for the exponent (which can range from −126 to 127) and 23 bits for the mantissa. The smallest mantissa is 22 zeros followed by a one, which equals one, and the smallest exponent is −126, so the smallest number that can be represented is $1 \times 2^{-126} \approx 1.2 \times 10^{-38}$. Similarly, the largest number that can be represented is $(2 - 2^{-23}) \times 2^{127} \approx 3.4 \times 10^{38}$. If two words are chained together to permit a larger exponent and larger mantissa, then this is called "double precision."

Because a computer represents numbers in base-2, it cannot accurately represent all the real numbers. For example, the number 0.5 can be represented exactly, since it equals 1/2, which is 2^{-1}. The number 0.1 cannot be represented exactly. The binary representation of the real number 0.1 is given by 0.0001100110011... where the 0011 repeats infinitely. With a finite word length, this infinite sequence must be truncated, and when it is truncated and converted back to base-10, we get 0.099999994. For a quick overview of this topic, see McCullough and Vinod (1999, sec. 2.1). For a much more detailed, yet still very accessible discussion of computer arithmetic, see Goldberg (1991).

This small difference between 0.1 and 0.099999994 is an example of rounding error. A similar type of inaccuracy is called *truncation error*, an example of which is the calculation of $\sin(x)$ by infinite series. In a computer, the series cannot be cumulated infinitely, and must be terminated at some point. The difference between termininating and continuing forever is truncation error. Like a rounding error, it can be very small. Some calculations, e.g., matrix inversion, can require millions of operations, and these small rounding and/or truncation errors can add up. Eventually, they can swamp all the accurate digits, producing a final answer that is completely inaccurate. This is very probably what happened in Longley's paper.

28.3 Introductory tests

In 1985, Leland Wilkinson, developer of the SYSTAT statistical software package, produced a pamphlet describing some simple tests of software accuracy. The primary documents for understanding and applying Wilkinson tests are Wilkinson (1985) and McCullough (2004a). The tests are all based on a dataset that he called the "Nasty" dataset, which is reproduced in Table 28.3.

The Wilkinson tests are fully described in Wilkinson (1985). They have been applied by Sawitzki (1994a, 1994b), Bankhofer and Hilbert (1997a, 1997b), McCullough (2004a) and Choi and Kiefer (2005). These papers usually report errors in packages. For example, some packages could not accurately compute the sample standard deviation for either BIG or LITTLE. What this reveals is that the packages

Table 28.3 Dataset NASTY.DAT

LABEL$	*X*	*ZERO*	*MISS*	*BIG*	*LITTLE*	*HUGE*	*TINY*	*ROUND*
ONE	1	0	.	99999991	0.99999991	1.0E12	1.0E-12	0.5
TWO	2	0	.	99999992	0.99999992	2.0E12	2.0E-12	1.5
THREE	3	0	.	99999993	0.99999993	3.0E12	3.0E-12	2.5
FOUR	4	0	.	99999994	0.99999994	4.0E12	4.0E-12	3.5
FIVE	5	0	.	99999995	0.99999995	5.0E12	5.0E-12	4.5
SIX	6	0	.	99999996	0.99999996	6.0E12	6.0E-12	5.5
SEVEN	7	0	.	99999997	0.99999997	7.0E12	7.0E-12	6.5
EIGHT	8	0	.	99999998	0.99999998	8.0E12	8.0E-12	7.5
NINE	9	0	.	99999999	0.99999999	9.0E12	9.0E-12	8.5

in question employed the "calculator formula":

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}{n-1}. \tag{28.1}$$

This formula makes one pass through the data and squares the observations. Consider squaring the first two observations on the variable BIG: $99999991^2 = 9999998200000081$ and $99999992^2 = 9999998400000064$. Now subtract the former from the latter: $9999998400000064 - 9999998200000081 = 199999983$. On a typical desktop econometric package, executing the command $99999992^2 - 99999991^2$ produces 199999984. The squaring of these large numbers has just barely used up the computer's finite precision. To see this, simply drop the leading nines and perform the following subtraction on your desktop computer: $8400000064 - 8200000081 = 199999983$, which does not exhaust a desktop computer's precision. Think of what would happen if nine-digit numbers were used instead of eight-digit numbers! Using the calculator formula to compute the sample variance of the variable BIG yields 2.424 instead of the correct 2.738. This is an example of what was described in the first section as an algorithm simply not being up to the task. The calculator formula contrasts sharply with the usual formula:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \tag{28.2}$$

which squares much smaller numbers, and so it is more difficult for this formula to exhaust the computer's finite precision.

The calculator formula, often presented in textbooks in the pre-computer era, was designed for use on toy problems found in textbooks, not for real-life problems. This fact was known in the statistical computing literature at least as far back as Ling (1974), and in one of the two classic texts on statistical computing that every competent statistical programmer would have read, i.e., Thisted (1988) (the other classic text is Kennedy and Gentle, 1980). Nonetheless, the calculator formula appeared in many statistical and econometric software packages. What this tells us is that the people who designed these programs were not versed in the basics of

statistical computing, one of the primary tenets of which is this: do not program the formula from statistics texts. The fact that allegedly competent programmers would implement the calculator formula underscores the need for users to test their software.

Calculating the correlation coefficient is conceptually simple, especially since it is bounded between +1 and −1. Yet some packages returned a correlation between BIG and HUGE, or LITTLE and ROUND, or X and BIG, that was bigger than unity! The developers of these packages never revealed why their packages were unable to compute correctly a correlation coefficient. Also of interest, the correlation between ZERO and any other variable should be undefined by definition because ZERO is a constant so its standard deviation $\sigma_Z = 0$ and:

$$\rho_{ZX} = \frac{cov(z, x)}{\sigma_Z \sigma_X}. \tag{28.3}$$

The 0 in the denominator means that the value of the ratio is undefined, yet some packages computed $\rho_{ZX} = 0$, and one package even managed to compute $\rho_{ZZ} = 1$.

Plotting BIG against LITTLE obviously should produce a straight line. Some packages were unable to do this. In one case, the software produced a single point in the middle of the graph, dropping all the other points. Again, the developers did not reveal the reasons for the failure.

Performing operations that involved the MISS variable revealed that not all packages correctly handled missing values.

Regressing X on a constant, BIG and LITTLE should produce an error, since BIG and LITTLE are linear transforms of each other, i.e., the matrix of independent variables is singular. If a package does not recognize the singularity, then it can grind through all the calculations and produce an answer – an incorrect answer, but an answer nonetheless. Not all packages passed this test.

Wilkinson tests have been applied to many statistic and econometric software packages, almost invariably revealing flaws of one sort or another. The only mystery is why software developers don't apply these tests themselves and fix the errors before someone writes an article or software review about it. These tests are quick and easy, taking less than an hour, and every user should make sure his package passes all these tests – or if it doesn't, have the developer fix the problem. If he or she doesn't fix it fast enough, a well-placed software review will convince him or her to do so.

28.4 Intermediate tests

McCullough (1998a) proposed an intermediate set of tests covering three areas: coefficient estimation based on the then recently released National Institute of Standards and Technology's (NIST) Statistical Reference Datasets (StRD), random number generation, and statistical distributions (e.g., the functions used to determine critical values for various distributions). This methodology has been applied by McCullough (1999a, 1999b), Vinod (2000), Sall (2002), Altman and McDonald

(2002), Kitchen, Drachenberg and Symanzik (2003), Teyssiere (2005), Keeling and Pavur (2004, 2007), Yalta and Yalta (2007), and Yalta (2008), among others.[1]

28.4.1 StRD

The StRD presents estimation problems in four suites: univariate summary statistics, one-way analysis of variance, linear regression, and nonlinear least squares. Each suite contains test problems at three levels of difficulty: lower, average, and higher. The primary discussions for understanding and applying the StRD are found in McCullough (1998a, 2000b). For the linear problems, NIST computed accurate solutions by carrying 500 digits through all calculations, effectively eliminating rounding error, and then rounding the final answer to 15 digits. Nonlinear problems were solved using different algorithms in quadruple precision, and rounding the final solution to 11 digits. For each nonlinear problem there are two sets of starting values: Start I and Start II. The former is far from the solution, which makes it harder for a solver to find the solution. The latter is closer to the solution, making it easier for a solver to find the solution.

One of the nine univariate problems (six lower, two average, one higher difficulty) require the calculation of three summary statistics: mean, standard deviation, and first-order autocorrelation coefficient. One of the lower-difficulty problems, NumAcc1, consists of just three observations: 10000001, 10000003, 10000002. A program that employs the "calculator formula" will fail this test starkly, even in double precision. It is a bit easier to diagnose a failure with these data than with the similar data in Wilkinson tests. Many packages fail to get good accuracy when calculating the first-order autocorrelation coefficient because they use bad algorithms.

The ANOVA suite has four lower-, four average-, and three higher-difficulty problems. Even a good ANOVA algorithm that does not recenter the data will return a completely inaccurate answer for the most difficult problem. If a package returns four digits of accuracy on this problem, then it is safe to conclude that the package recenters the data (to reduce the effect of squaring) before calculating all the relevant sums of squares. It is not uncommon to see packages, especially those that have been around for a long time, fail even on the average difficulty tests. The reason is that the packages use legacy code left over from the days of single-precision computers. After the StRD is applied to such packages, the developers usually update the code.

The linear regression suite has 11 test problems, two of lower difficulty, two of average, and seven of higher difficulty. One of the higher-difficulty test problems is the Longley benchmark, which most packages can handle easily these days. A problem that many packages cannot handle is the Filip dataset, which is a tenth-order polynomial that is nearly singular. A good package will either produce accurate coefficients or detect the singularity and refuse to produce a solution. This latter result is *not* a failure, because the user has not been misled. A package that fails this test, e.g., Excel 2000 and earlier (McCullough and Wilson, 2005), is capable of producing completely inaccurate coefficients when confronted with collinear data. A package that directly solves the normal equations, e.g., that calculates

$\hat{\beta} = (X'X)^{-1}X'y$ will fail many of the linear regression tests. The reason is that computing the inverse of $(X'X)$ is like a squaring operation that produces a great deal of cumulated rounding error, sometimes enough cumulated rounding error to completely ruin the estimate $\hat{\beta}$. A better method is to use what is called the QR decomposition, which avoids the squaring in the computation of $\hat{\beta}$. If a package fails many of these linear problems, it is safe to say that the package does not use a good QR decomposition, and almost certainly calculates the least squares coefficients by inverting $(X'X)$.

Most nonlinear solvers offer various options to the user: different algorithms, convergence tolerance, etc. Two major lessons have been learned from the early application of the StRD nonlinear suite. First, the default options for most nonlinear solvers will not find the correct answer. One primary example is the convergence tolerance. Suppose the solver is set, at default, to stop when the difference between the sum of squares on two successive iterations changes by less than 0.001. This might not be tight enough and the solver might give a "failure to converge" message; it might be necessary to set the tolerance at 0.000001. For a nonlinear least squares problem, changing the tolerance by this amount may well dramatically alter the coefficients and markedly reduce the sum of squares. Do not rely on the default options for nonlinear solvers!

Normally, this first problem would not be of much consequence; simply change the options until convergence is declared. However, there is a common second problem that greatly complicates the matter: many packages have a tendency to stop at a point that is not a solution and claim that they have found a solution. (How is a user to know whether his/her solver has this second problem? See whether anyone has applied the StRD nonlinear suite to his/her software package!) If a user happens to have such a defective solver, then a reasonable strategy is to keep decreasing the convergence tolerance – getting a "convergence achieved message" every time – until the user verifies that the coefficients have stopped changing, that the solver really has achieved a stationarity point. For a detailed example of applying this strategy, see McCullough (2004b). For extended discussion of using nonlinear solvers to find solutions to nonlinear problems, see McCullough and Renfro (2000) and McCullough and Vinod (2003a, 2004).

To belabor this important point, almost any nonlinear problem solved using a package that has this second problem will produce an incorrect answer. The user simply accepts the default options, the solver produces an incorrect answer, and the user accepts the incorrect results. The literature is filled with results from such packages, and because journals typically do not even compel authors to identify their software packages, let alone make their data and code available, there is little hope of purging these incorrect results from the literature. As will be discussed in the conclusions, replication of published results and software accuracy are intimately connected.

One interesting aspect of the nonlinear StRD suite that has yet to be published in the literature is the following. The StRD nonlinear suite is for nonlinear least squares problems. Nonlinear least squares solvers make use of the special structure of nonlinear least squares problems. Reformulating these problems as maximum

likelihood problems and giving them not to the package's nonlinear least squares solver, but instead to its maximum likelihood solver, would produce a method for testing the efficacy of maximum likelihood solvers. To date, there has been no such systematic testing of the maximum likelihood solver of any econometrics package. Preliminary examination of this topic by the author suggests that many packages have maximum likelihood solvers that tend to stop at points that are not solutions but nonetheless declare that a solution has been found. The implications of this line of research for applied econometrics would be enormous.

28.4.2 Random number generators

Random numbers form the basis of many econometric estimators, e.g., simulation and bootstrapping. Yet the efficacy of these methods depends on the random numbers being truly random. In fact, the random number generator (RNG) in an econometrics package does not produce random numbers, but deterministic numbers. However, they are nonlinearly deterministic and can give the appearance of being truly random. For a short primer see McCullough (2001) or McCullough and Vinod (1999, sec. 5). The physics literature is filled with stories of bad simulation results due to bad RNGs (see, e.g., Coddington, 1994, 1996). The economics literature has no such examples because there is practically no replication in economics. When was the last time one economist took another economist's simulation code and swapped out one RNG for another RNG to see if the results would change? Nonetheless, the literature on testing econometric software has found bad RNGs, and the usual response by a developer is to put in a good RNG.

The primary method for determining whether an RNG produces "random" numbers is to take the numbers and apply a statistical test of some sort. As a simple example, take 1,000 random numbers, compute the first-order autocorrelation coefficient, and test the null that the coefficient equals zero. Rejecting the null constitutes evidence that the numbers are not random; they are correlated. There are many, many types of tests because there are many, many ways for a sequence of numbers not to be random. One of the first standard collections of such tests was given by the eminent computer scientist Donald Knuth in the 1981 first edition of his classic text (Knuth, 1998). Coding these many tests and applying them was time-consuming and tedious. The random-number specialist George Marsaglia (1996) collected several tests and coded them as "DIEHARD: a battery of tests of randomness." DIEHARD has been used to uncover bad RNGs in many statistical and econometric software packages.

As the scale of computing has grown since then, the DIEHARD tests are no longer up to the task of validating RNGs for the huge simulations that are run today. The RNG testing program TESTU01 has been the new standard since its first release in 2002. The program is reviewed by McCullough (2006) and described in more detail in L'Ecuyer and Simard (2007). Using TESTU01 requires a passing knowledge of C, and is available freely. It comes with three batteries of tests pre-programmed, Small Crush (which can take a few minutes), Crush (which can take a few hours), and Big Crush (which can take a day or more). By contrast, on a modern PC, DIEHARD takes several seconds. If an RNG passes Small Crush, then give it to Crush. If it

passes that, then give it to Big Crush. If it passes that, then it is fit to use. There are too many good RNGs to use one that fails any reasonable test of randomness. There are many bad RNGs, even ones published in journals, so testing is imperative.

It is not enough that the underlying uniform RNG passes tests. Often the need for random normal variates, or random variables with other distributions, arises. These are usually obtained from the uniform RNG via transformation. However, the process of turning seemingly good uniforms into normals, e.g., requires not just good uniforms, but also a good transformation. The transformations from uniformity to another distribution, e.g., standard normal, can be flawed. The Marsaglia Multicarry RNG, which passed all the DIEHARD tests and was popular for some years, does not play well in the tails with the Kinderman–Ramage transform to normality. So if your econometric software package uses the Multicarry, and it produces random normals via the Kinderman–Ramage, then the tails of the so-called random normals actually deviate substantially from true random normals. For a graphical depiction of this phenomenon, see Figure 28.1. Note the slight gap at about 3.4 and a pronounced gap at about 3.6. These gaps would wreak havoc upon any simulation that focused on the tails of the standard normal.[2]

Further, the Multicarry fails both Small Crush and Crush batteries in TESUT01 (there was no need to apply Big Crush). Testing random normals, random-t's, and random chi-squares needs to be done. As this time, such testing is in its infancy. A common approach is to back transform the random normals (say) to uniformity, and then apply tests for uniformity to the backtransformed uniforms. This approach was used successfully by Tirler *et al.* (2004). Do you think it safe to trust the random normals in your software package?

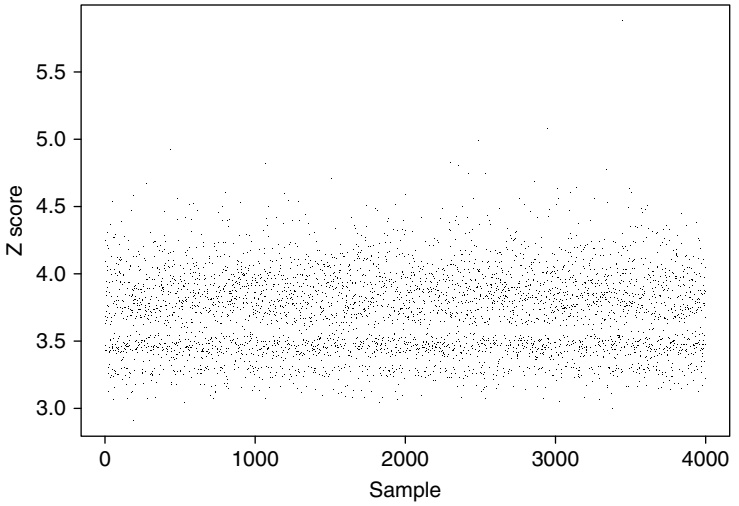


Figure 28.1 The extreme tails of 4,000 samples of random normals from Marsaglia Multicarry and Kinderman–Ramage

Generally speaking, econometric discussions of simulation simply assume a high-quality RNG the same way that econometrics texts discussing nonlinear estimation assume a high-quality solver. For example, the text on simulation in econometrics by Gouieroux and Montfort (1997) makes no mention of testing RNGs.

28.4.3 Statistical Distributions

Bai and Perron (2003) (henceforth BP) published an article in the *Journal of Applied Econometrics* about structural change models, and used an econometric software package to write software for their method. Zeileis and Kleiber (2005) (henceforth ZK) attempted to port the code to R. We first note that the *Journal of Applied Econometrics* has a data-only mandatory archive, i.e., BP were under no obligation to supply their code to would-be replicators, but they did so. While ZK were able to reproduce much of what BP did, they were unable to match some confidence intervals for break-points. As one example, both packages estimated a break-point at the third quarter of 1972, 1972:3, but the BP interval was [1970:3, 1972:4] while the ZK interval was [1969:1, 1972:4]. After much numerical detective work, ZK finally ascertained that the BP software package had a function for the Normal CDF (cumulative distribution function) that was inaccurate in the tails, while the R package had a Normal CDF that was accurate in the tails. The accuracy of statistical distributions matters.

On a related note, there is a definite need for the profession to find every article using the Normal CDF function in the package used by BP, and check to see if the results are wrong. Of course, since journals largely do not require authors even to identify what package they use, let alone supply the code, there is no hope of ever accomplishing this task. This example shows how replication can uncover errors not only in published articles, but in software, too.

To test statistical distributions, one needs an accurate source for the desired quantities – not just any software package will do. For years the primary sources were Knüsel's (1989) ELV package and Brown's DCDFLIB (available in C and Fortran). McCullough (2000c) showed that the program Mathematica was at least as good as ELV and, lately, Yalta (2008) showed that Mathematica produces results more accurate than ELV. One simply compares the output from one of these three programs to the output produced by the econometric software in question. Naturally, one cannot assess all possible outputs, so one examines a carefully chosen subset. For the normal distribution, one might first test the following percentiles: $\{0.0001, 0.001, 0.01, 0.1, 0.2, \ldots 0.9, 0.99, 0.999, 0.9999\}$ and check the extreme tails to find out where the algorithm in the econometric software breaks down; it might be completely inaccurate for the 0.9999999999 percentile. A similar approach might be undertaken for the t-distribution, except it has to be done for different degrees of freedom. The process gets more complex for distributions with two or more parameters, e.g., the F-distribution. Complete details for testing can be found in McCullough (1998a, sec. 6). Most packages are able to compute these distributions for simple hypothesis testing, but methods that require distributions to be evaluated in the extreme tails, e.g., value-at-risk testing or saddle-point approximation, should only be undertaken with packages that have very accurate distributions.

28.5 Advanced tests

Look at the table of contents of an advanced econometrics text, or the list of functions in any econometrics software package, and you will see many familiar names: Kalman filtering, multinomial logit, ARMA, etc. Just as in the case of multivariate GARCH, with which this chapter opened, different packages will give different answers to the same problem, and no one has any idea which package, if any, is correct. This is because there are no benchmarks for any of these procedures.

We have already alluded to different packages giving different answers to the same FIML problem. In fact, a benchmark was worked out for this problem by Calzolari and Panattoni (1988). Not many software developers were aware of this, because it was not advertised as a benchmark. Silk (1996) recognized that it was benchmark-quality work, and used it as a benchmark for his comparison of software packages.

When Bollerslev (1986) published the first GARCH article, he did not completely describe his method, leaving developers to guess at important details. Consequently, while every package soon had a GARCH command, they all gave different answers. McCullough and Renfro (1999) (henceforth MR) documented this fact, but also sought to alleviate the problem. Fiorentini, Calzolari and Panattoni (1996) published an article entitled "Analytic derivatives and the computation of GARCH estimates." Recognizing that two of the remaining three authors had already written benchmark-quality code, MR suspected that this might be code of a similar quality. Indeed, it was, and MR offered the "FCP GARCH benchmark" on which many developers quickly converged. MR analyzed seven packages, and found that four of them could not estimate the FCP GARCH model, and only one of the remaining three was reasonably accurate. When Brooks, Burke and Persand (2001) reassessed the situation only two years later, they found that most packages could estimate the model and do so with reasonable accuracy (but see McCullough and Vinod, 2003b, for another example of many packages giving different answers to the same GARCH problem). Developers will converge on a benchmark *if* one is available.

Bruno and De Bonis (2004) wrote a benchmark for a garden-variety panel data estimator, for both fixed and random effects. They then gave the data to three software packages. All packages agreed on the fixed effects estimation, but disagreed on the random effects. Investigation of the matter required correspondence with the developers because the user guide and reference manuals did not provide much information on the algorithms employed (which is typical of econometric software packages). As Bruno and De Bonis (2004, p. 281) discovered, "it is clear that all the numerical differences produced by the random-effects estimates are caused by the differences in the small-sample formulas for the computation of the between-regression variance." All three packages used consistent estimators, so there was no theoretical reason to prefer one over the other. The literature provided no guidance, so Bruno and De Bonis conducted a Monte Carlo study to determine which of the three estimators had the better finite-sample properties.

The Yule–Walker equations are often used to compute partial autocorrelation coefficients; the method is presented in most econometrics texts, and most econometric software packages offer the method. What is not presented is that it is the

least reliable method for such computation. The classic time series text by Priestley (1981) presents four methods for computing partial autocorrelation coefficients, and he presents them in *decreasing* order of reliability: the Yule–Walker method is presented fourth. The Yule–Walker equations are, from a computational standpoint, easy to implement, and they might have been justified back in the days when computing power was expensive; in the present day, they cannot be justified. Very few econometric packages offer better methods. One such method is the Burg algorithm. However, there was no benchmark for the method, so McCullough (1998b) computed one. Now some econometrics time series packages offer the Burg method as an improvement over the Yule–Walker equations.

There are not many advanced benchmarks for econometric software, and there is precious little tangible evidence that any econometric software package is giving the correct answer for any moderately complicated problem.

28.6 Benchmarks for ARMA models

In an important article, Newbold, Agiakloglou and Miller (1994) (henceforth NAM) observed: "fitting the same model to the same data will yield more or less identical results whatever software is used for multiple regression. That is not the case for the estimation of the parameters of an ARIMA model." In part, this may be due to the fact that NAM placed themselves in the position of a novice user, i.e., "though many programs allow the user a range of optional modifications, we generally ran them in default mode." If one thing has been learned from the literature on statistical and econometric software accuracy, it is that default options for nonlinear estimation procedures typically do not produce accurate answers. Such matters as choice of algorithm, convergence criterion, convergence tolerance, and initial conditions, can all greatly affect the quality of the answer produced by a nonlinear estimation procedure. For example, in the case of autoregressive integrated moving average (ARIMA) procedures, some packages conduct preliminary estimations to determine starting values, while others simply use zeros. This fact alone could account for much variation between packages. Therefore, it may seem entirely possible that the packages examined by NAM would have exhibited little variation in the range of results produced if only they had adopted the posture of an experienced user. Such, however, turns out not to be the case, as will be shown.

Given that the differences are not due to the use of default options, the notion that algorithmic differences may be responsible comes to mind. In the case of unconditional least squares (ULS) with backcasting, there is no one preferred method of backcasting, so perhaps this may account for the differences. NAM (1994, p. 580) pointedly address this notion in the discussion of their conditional least squares (CLS) results, for which no such difference is possible. Even in the cases when point estimates agree, NAM note substantial variation in the estimates of standard errors.

Thus, the only means to resolve the discrepancies between packages is the production of a benchmark. The production of a benchmark typically requires the use of extended precision computation, i.e., more than double precision. One

common method of achieving this level of accuracy is to use FORTRAN with a multiple precision pre-processor (e.g., Bailey, 1993). An alternative is to employ the software package Mathematica, which can combine symbolic calculation with extended precision computation to produce benchmark-quality results. For example, on the NIST StRD ANOVA tests, 32-bit word double precision can do no better than four or five digits of accuracy. Mathematica can, on these problems, return a full 15 digits of accuracy (McCullough, 2000c). For this analysis, we therefore employ Mathematica.

Here we consider only Box and Jenkins (1976) (henceforth BJ) methods based on least squares for two reasons. First, these are by far the most widely-used methods; only two of the packages considered by NAM offered exact maximum likelihood. Second, there is a definitive reference for the procedures – BJ. There exist many alternative methods for computing maximum likelihood methods, and computing a maximum likelihood benchmark for ARIMA estimation constitutes a project in itself.

This is not an unimportant topic. ARMA estimation and forecasting is a mainstay of applied time series analysis. Note that forecasts have not been benchmarked: this needs to be done, too. Recently, Yalta and Jenal (2009) attempted to double-check, by hand, the forecasts coming from an ARMA procedure. They could not reproduce the program's results. On further investigation, they found that not only were the forecasts incorrect, even the ARMA coefficients were incorrect. (For a benchmark, they used the approach of having several packages give the same answer to the same problem – the package in question did not agree with all the other packages.)

Sub-section 28.6.2 describes definitions and notations. Sub-section 28.6.3 presents CLS results; analytic derivatives are employed for the ARMA(1,1) case, and comparison with numerical derivatives sheds light on the choice of the differencing interval for the computation of numerical derivatives. Then, using extended precision computation, a benchmark for CLS is presented. Sub-section 28.6.4 presents a benchmark for ULS.

28.6.1 Definitions and notation

The ARMA(p,q) model can be written as:

$$a_t = \tilde{w}_t - \phi_1 \tilde{w}_{t-1} - \phi_2 \tilde{w}_{t-2} - \ldots - \phi_p \tilde{w}_{t-p} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots \theta_q a_{t-q}, \quad (28.4)$$

where $w_t = \nabla^d z_t$ and $\tilde{w}_t = w_t - \mu$ with $E[w_t] = \mu$. In general, when $d > 0$ it is assumed that $\mu = 0$. However, we take $\mu \neq 0$ as an additional parameter to be estimated. The form in which the equation is written affects the intercept term in an ARMA model. For example, BJ write $(1 - \Phi(B))y_t = c + (1 - \Theta(B))a_t$, while an alternate formulation, adopted here, takes $(1 - \Phi(B))(y_t - \mu) = (1 - \Theta(B))a_t$. Comparing the two formulations, it is obvious that $\mu = c/(1 - \sum_{j=1}^{p} \phi_p)$. In the sequel, only the case $p = 1, q = 1$ is treated.

The variation on the Marquardt algorithm proposed by BJ, and which is implemented in several packages, is extremely simple, and focuses on minimizing a sum

of squares. In the case of CLS, the relevant quantity is:

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta} | \boldsymbol{w}_*, \boldsymbol{a}_*, \boldsymbol{w}), \tag{28.5}$$

where the subscript asterisks emphasize conditioning on the choice of starting values. The problem, obviously, is in the choice of pre-sample values for w and a, i.e., how to define $\boldsymbol{w}_*$ and $\boldsymbol{a}_*$.

BJ (1976, p. 211) consider two approaches. The first involves setting $\boldsymbol{w}_*$ and $\boldsymbol{a}_*$ equal to their unconditional expectations, which are μ and 0, respectively. In the event that $\mu \neq 0$, $\bar{w}$ can be substituted for each element of $\boldsymbol{w}_*$. This preserves the full sample. However, due to potential instabilities when the roots of $\phi(B)$ are near the unit circle, this method is not much used. A more reliable method is to discard observations so that actual values of w are used for all calculations. This implies that the sum of squares can be taken only over observations $p+1$ through n. (Clearly, when $p = 0$ this method is equivalent to the previous method.) Thus, the second approach sets $a_t = 0$ for $t = 1, \ldots, p$ and calculates the a's from a_{p+1} onward; observations must be dropped from the sample. It is the second approach that BJ adopt for CLS, and which is considered here.

Setting to zero values that might vary substantially from zero may induce a transient effect that can adversely affect the quality of the final estimates. BJ (1976, p. 211) notes that CLS is "not satisfactory" for seasonal models.

To derive the ULS method, BJ introduce the unconditional sum of squares:

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-Q}^{n} [a_t | (\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{w}]^2, \tag{28.6}$$

where the $a_t, t \leq 0$ are computed recursively by taking expectations of equation (28.4). The values necessary to compute these $a_t, t \leq 0$, are the $[w_t], t \leq 0$, which can be calculated via the procedure known as *backcasting*.

In the usual fashion, the backward and forward representations are given by:

$$e_t = (w_t - \mu) - \phi(w_{t+1} - \mu) + \theta e_{t+1} \tag{28.7}$$

$$a_t = (w_t - \mu) - \phi(w_{t-1} - \mu) + \theta a_{t-1}. \tag{28.8}$$

Given initial estimates of the parameters, μ_0, ϕ_0 and θ_0, setting $e_{n+1} = 0$ allows equation (28.7) to be executed from $t = n$ to $t = 1$. At $t = 0$ the value $e_t = 0$, and so the equation can be rewritten as an expression for w_t, from which $[w_t], t \leq 0$ can be calculated back to some $t = -Q$, the quantity $(w_t - \mu)$ being neglible for $t < -Q$. These backcasted values of w_t can then be used to start the recursion equation (28.8) from $-Q$ upon setting $a_{-Q-1} = 0$.

Exactly when the quantity $w_t - \mu$ becomes neglible is not explicitly stated in BJ. Their example in their Table 7.4 (1976, p. 218) suggests that $w_t - \mu < 0.01$ is an appropriate stopping rule. Other stopping rules for backcasting have been described in the literature. For example, Granger and Newbold (1977, p. 88) suggest stopping when the magnitude of three successive values of $E_c(Y_t)$ is less than 1% of the standard deviation of y_t.

In whatever way it is determined, the choice of Q, the number of observations to backcast, will affect the final estimation results. What is surprising is that many packages that offer the ULS method do not mention the stopping rule employed. Even more surprising, when this researcher contacted many such developers, they refused to reveal their stopping rules. Not only is this tantamount to refusing to reveal the method of computation, but it also makes provision of a benchmark impossible for these packages. Thus, users of such packages are in the unenviable position of relying on unproven and unverifiable code.

The variation of the Marquardt algorithm proposed by BJ for estimating CLS and ULS models is the essence of simplicity. Given initial estimates of the coefficients, μ_0, ϕ_0 and θ_0, compute the vector of residuals a, which will have length $n - p$ in the case of CLS and length $n + Q + 1$ in the ULS case. Compute the derivative of a with respect to the parameters, denoted $a^{(\mu)}$, $a^{(\phi)}$ and $a^{(\theta)}$, and run the regression:

$$a = b_\mu a^{(\mu)} + b_\phi a^{(\phi)} + b_\theta a^{(\theta)}, \tag{28.9}$$

to obtain estimates $\hat{b}_\mu, \hat{b}_\phi$ and $\hat{b}_\theta$. Compute the coefficient estimates at the end of the first iteration as $\mu_1 = \mu_0 + \hat{b}_\mu$, $\phi_1 = \phi_0 + \hat{b}_\phi$ and $\theta_1 = \theta_0 + \hat{b}_\theta$. To commence the second iteration, based on μ_1, ϕ_1 and θ_1, compute a (note that the value of Q on this iteration may well not be equal to the value of Q on the previous iteration) and repeat the process until a termination criterion is achieved (e.g., successive sum of squared residuals is less than ϵ, etc.). Suppose the termination test is successful at the end of the cth iteration. Then the procedure is said to have terminated after c iterations. Note, though, that the gradients currently in the computer's memory are computed based on μ_{c-1}, ϕ_{c-1} and θ_{c-1}.

Computation of the standard errors is effected in the usual fashion. First, the gradients must be recomputed using μ_c, ϕ_c and θ_c, and each gradient will have length $n-p$ (CLS) or $Q+1+n$ (ULS). In the latter case, drop the first $Q+1$ elements of each vector. Form the matrix with three columns and either $(n - p)$ (CLS) or n (ULS) rows: $g = [a^{(\mu)}\ a^{(\phi)}\ a^{(\theta)}]$. The covariance matrix is given by $(g'g)^{-1}$ which, when multiplied by $\sum a_i^2/n$, has as its trace the variances of the coefficients.

28.6.2 Calculation of derivatives

Given the general superiority of analytical derivatives over numerical derivatives, no benchmark for a nonlinear procedure should be attempted on the basis of numerical derivatives alone, except in exceptional circumstances (e.g., when calculation of the derivatives is nearly impossible). Comparing the performance of numerical and analytic derivatives in a benchmark setting can determine whether it is safe for a user to rely on numerical derivatives or whether, as Fiorentini, Calzolari and Panattoni (1996) found in the case of GARCH models, analytic derivatives are necessary to achieve decent accuracy.

Computation of numerical derivatives is easy. Consider computing $a^{(\phi)}$ on the ith iteration. Compute the residuals based on μ_i, ϕ_i and θ_i, and call this vector a_i. Now for some differencing interval h, compute a_i^h based on $\mu_i, \phi_i + h$ and θ_i. Then the numerical estimate of $a^{(\phi)}$ is given by $a_i - a_i^h$. The choice of differencing interval

can greatly affect the quality of the numerical derivatives. For example, in the days when single precision was common, $h = 0.01$ or $h = 0.001$ were common choices. In the present day, when double precision is standard, smaller values of h are used.

It is well known that analytic derivatives are generally more accurate than numerical derivatives. It is also well known that there exist some problems for which implementing analytic derivatives is too difficult, and there is no recourse to numerical differentiation. ARIMA estimation is just such a case. For CLS, analytic derivatives are not difficult to implement. Backcasting, however, is another case. BJ (1976, p. 235) provide the analytic derivatives for an ARMA(1,1) model, where $a_t^{(\phi)}$ denotes $\partial[a_t]/\partial\phi$:

$$e_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t+1}^{(\phi)} + \theta e_{t+1}^{(\phi)} - [w_{t+1}] \tag{28.10}$$

$$a_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t-1}^{(\phi)} + \theta a_{t-1}^{(\phi)} - [w_{t-1}] \tag{28.11}$$

$$e_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t+1}^{(\theta)} + \theta e_{t+1}^{(\theta)} + [e_{t+1}] \tag{28.12}$$

$$a_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t-1}^{(\theta)} + \theta a_{t-1}^{(\theta)} + [a_{t-1}], \tag{28.13}$$

where:

$$[w_t] = w_t, \ t > 0 \tag{28.14}$$

$$w_t^{(\phi)} = w_t^{(\theta)} = 0, \ t > 0 \tag{28.15}$$

$$[e_{-j}] = 0, \ j \geq 0. \tag{28.16}$$

These derivatives are tedious but straightforward to implement, and are similar to the previous use of backward and forward equations. Setting $e_{n+1}^{(\phi)} = 0$, equation (28.10) can be solved from n to 1. For $t = 0$, $e_t^{(\phi)} = 0$ and so this equation can be re-expressed to solve for values of $w_t^{(\phi)}, t \leq 0$. Then, setting $a_{-Q-1}^{(\phi)} = 0$ and using the backcasted values of $w_t^{(\phi)}, t \leq 0$, equation (28.11) can be solved from $t = -Q$ to $t = n$, and similarly for equations (28.12) and (28.13).

By comparison, the analytic derivatives for an ARIMA(1,0,1) are much more difficult to implement. Then differentiating equations (28.7) and (28.8) yields:

$$e_t^{(\phi)} = w_t^{(\phi)} - (w_{t+1} - \mu) - \phi w_{t+1}^{(\phi)} + \theta e_{t+1}^{(\phi)} \tag{28.17}$$

$$a_t^{(\phi)} = w_t^{(\phi)} - (w_{t-1} - \mu) - \phi w_{t-1}^{(\phi)} + \theta a_{t-1}^{(\phi)} \tag{28.18}$$

$$e_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t+1}^{(\theta)} + e_{t+1} + \theta e_{t+1}^{(\theta)} \tag{28.19}$$

$$a_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t-1}^{(\theta)} + a_{t-1} + \theta a_{t-1}^{(\theta)} \tag{28.20}$$

$$e_t^{(\mu)} = (w_t^{(\mu)} - 1) - \phi(w_{t+1}^{(\mu)} - 1) + \theta e_{t+1}^{(\mu)} \tag{28.21}$$

$$a_t^{(\mu)} = (w_t^{(\mu)} - 1) - \phi(w_{t-1}^{(\mu)} - 1) + \theta a_{t-1}^{(\mu)}, \tag{28.22}$$

and programming these, of course, is more difficult than the ARMA(1,1) case.

More sophisticated BJ models are much more difficult. Were it the case that the most estimated ARIMA model was an ARIMA(1,0,1) then it would be worthwhile for a developer to implement analytic derivatives. For example, the FCP GARCH benchmark (Fiorentini, Calzaroli and Panattoni 1996; McCullough and Renfro, 1999; Brooks, Burke and Persand, 2001) is based on a GARCH(1,1) model, with analytic first and second derivatives because most applications of GARCH involve the GARCH(1,1). The same is not true of ARIMA models.

Nonetheless, it is of interest to know what price is paid for the use of numerical derivatives instead of analytic derivatives. Moreover, the use of analytic derivatives can shed light on the appropriate choice of h, the differencing interval. Since analytic derivatives are easily implemented in CLS, this will be done in the next section.

28.6.3 Conditional least squares

We are now in a position to create a CLS benchmark. We use the 197 observations from Box–Jenkins Series A. In the CLS case, the analytic derivatives given by equations (28.17)–(28.22) reduce to:

$$a_t^{(\phi)} = (w_{t-1} - \mu) + \theta a_{t-1}^{(\phi)} \tag{28.23}$$

$$a_t^{(\theta)} = a_{t-1} + \theta a_{t-1}^{(\theta)} \tag{28.24}$$

$$a_t^{(\mu)} = -1 - \phi + \theta a_{t-1}^{(\mu)} \tag{28.25}$$

which can be calculated recursively after setting $a_0^{(\phi)} = a_0^{(\theta)} = a_0^{(\mu)} = 0$.

Carrying 50 digits through all calculations,[3] using the maximum of the relative change in the coefficients as the convergence criterion, setting the convergence tolerance to 1E-13, and rounding to 11 digits produced the benchmark. To determine what we can expect from ordinary double-precision calculation, we also re-ran the program using ordinary double precision. As can be seen in Table 28.4, double precision delivers the benchmark answer for the first seven digits of the constant, five digits of ϕ, and four digits for θ. The standard errors are computed on division by n, not $n - k$.

Recall that the benchmark was produced with analytic derivatives. We can use some capabilities of Mathematica to obtain this level of accuracy without

Table 28.4 CLS benchmark with analytic first derivatives

Parameter	μ	ϕ	θ
Double precision	17.093753099	0.90658876305	0.56881361427
Benchmark	17.093752390	0.90658703600	0.56880910281
Standard error (MLE)	0.10520938686	0.045388753586	0.086811221485

Note: MLE = maximum likelihood estimation.

the trouble of implementing analytic derivatives. Issuing the commands: "$MinPrecision=50," rationalizing the input data, setting h = 1E-12 and using the "Rationalize" command on the input data series enabled numerical differentiation to recover the full eleven digits of the benchmark. Thus, without using analytic derivatives, Mathematica is capable of producing the same accuracy as if analytic derivatives were employed. This information will be useful in producing the benchmark for ULS – it will save us the trouble of implementing analytic derivatives.

Next we investigate the accuracy that can be attained with numerical derivatives with varying sizes of the forward differencing parameter, h. The forward difference derivative is computed by $f'(x) = (f(x+h) - f(x))/h$. We do not consider numerical derivatives via central differences, because the software packages in question do not offer such an option for estimation.

Though obviously the desired degree of accuracy depends on the particular problem at hand, based on the limits of achievable accuracy in this situation as discussed above, we take as our desideratum that the ARMA estimation procedure should produce at least three accurate digits for the coefficients and standard errors. Table 28.5 shows the effect of various choices of h on the accuracy of the estimates.

Recall that we ran analytic derivatives two ways: at regular double precision, and also carrying 50 digits; the former agreed with the latter to about five digits. Clearly, we cannot expect more than five digits of agreement when using numerical derivatives. We see that this level of accuracy is attained in the bottom row, and almost in the penultimate row. Hence, we see that, for best accuracy, the differencing parameter h should be set to at most 0.00001. (Very few packages even permit users to control this feature, so this information is mostly for the benefit of software developers rather than users.)

Programming analytic derivatives, especially recursive ones, can leave the programmer wondering whether he/she did it correctly. A useful device in this

Table 28.5 Effect of differencing interval on accuracy (MLE standard errors)

Derivative	μ	ϕ	θ
Analytic	17.0938	0.906587	0.568809
	(0.105209)	(0.0453888)	0.0868112)
$h = 0.01$	17.0945	0.908317	0.573343
	(0.106161)	(0.0454922)	(0.0866366)
$h = 0.001$	17.0938	0.906760	0.569260
	(0.105303)	(0.0453983)	(0.0867941)
$h = 0.0001$	17.0938	0.906604	0.568854
	(0.105219)	(0.0453897)	(0.0868095)
$h = 0.00001$	17.0938	0.906589	0.568814
	(0.105210)	(0.0453888)	(0.0868111)
$h = 0.000001$	17.0938	0.906587	0.568810
	(0.105209)	(0.0453888)	(0.0868112)

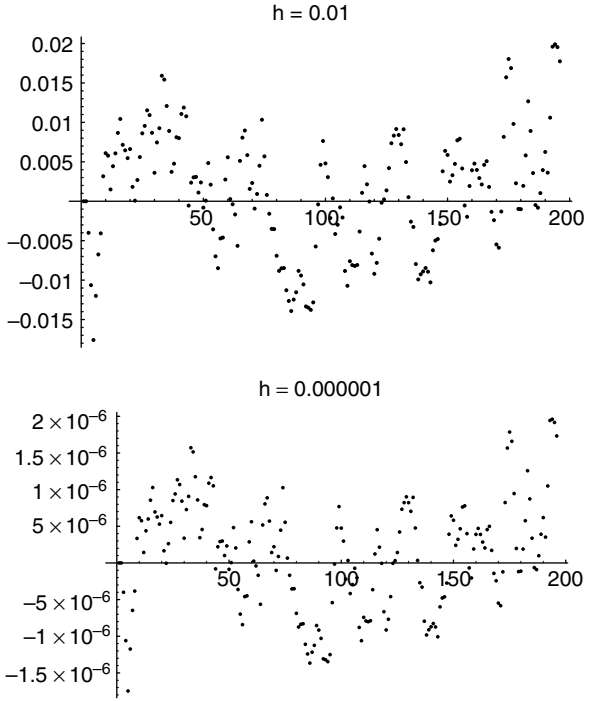


Figure 28.2 Analytic minus numerical derivative for $a_t^{(\theta)}$ at various differencing intervals h for all 197 observations on Series A

Table 28.6 CLS benchmark – parameters

Package	μ	ϕ	θ
Benchmark	17.093752390	0.90658703600	0.56880910281
Mathematica default	17.093753099	0.90658876305	0.56881361427
Package X	17.09375129	0.90658475	0.56880310
Package Y	17.09366504	0.90660389	0.56883103
Package Z	17.0937032318	0.90648328387	0.568536846

situation is to compare the difference between the numeric and analytic derivatives (which should be close to zero) for different values of h (it should get smaller as h gets smaller). Both these conditions can be seen in Figure 28.2.

Having determined the benchmark coefficients and standard errors, we are now in a position to see whether our packages can hit the benchmark. In Table 28.6 we present default estimation for three packages that offer CLS.

Package X is the most accurate, followed by Package Y, followed by Package Z. The first two seem to provide about four accurate digits, and the third about three

Table 28.7 How close to the benchmark?

Package	*Tolerance*	μ	ϕ	θ
Benchmark	–	17.093752390	0.90658703600	0.56880910281
Package X	Default	17.09375129	0.90658475	0.56880310
Package X	1E-6	17.09375234	0.90658690	0.56880876
Package X	1E-8	17.09375234	0.90658690	0.56880876
Package X	1E-10	17.09375234	0.90658690	0.56880876

Table 28.8 CLS benchmark – standard errors (constant omitted)

Package	ϕ	θ
Benchmark	0.045388753586	0.086811221485
Package X	0.0457	0.0874
Package Y	0.0547	0.1186
Package Z	0.0433	0.0894

digits. How much more accuracy can be squeezed out of a package by varying the default options? We consider only the case of Package X.

The primary option in this case is the convergence tolerance which, by default, is set to 0.00001, i.e., 1E-5. What this tolerance controls is unknown, because the package's extensive documentation does not say! (This is typical of econometric software packages.) Let us vary this convergence tolerance nonetheless.

As can be seen in Table 28.7, tightening up to 1E-6 yields a small improvement in about the sixth digit (which is negligible since the difference between the benchmark with 50 digits and the benchmark with double-precision occurs in about the same place), and no further improvement occurs with more tightening.

We now turn to the question of standard errors. Since all packages do not use the same parameterization for the constant term, the standard errors thereof are not directly comparable and are omitted. NAM noted that even when the parameters were the same, different standard errors could be observed, and we find the same with our three packages, as seen in Table 28.8. There are many possible sources for this (see McCullough and Renfro, 2000, for a discussion); here we mention just one. There is no single method for computing standard errors for nonlinear estimators; the product of the gradient (recommended by BJ, as discussed at the very end of sub-section 28.2.2), which is used in our benchmark program, the inverse of the Hessian, and the information matrix are all prime candidates.

Though the documentation for Package X gives no indication of how its standard errors are computed, we can see that Package X probably uses the product of the gradient method. Similarly, the documentation for both Packages Y and Z are silent on this important point. We have no idea whether these standard errors are incorrect or based on some other approach, e.g., inverse of the Hessian or the

information matrix. The users of these packages will just have to trust that these standard errors – whatever type they are – are correctly programmed.

28.6.4 Unconditional least squares

We have seen that an important part of ULS is the determination of Q, the number of observations to backcast. Many packages claim to use backcasting and offer only BJ as a reference. Given the atrocious state of software documentation, it is not surprising that most packages that offer backcasting do not give the rule used to determine the number of backcasts. What is perhaps surprising is that, when this author contacted the developers who did not mention the rule in their documentation, only one of them would reveal its rule.

This is either pathetic or amusing, depending on your point of view, because the text does not give the rule used to determine the number of backcasts. The backcasts in an example in BJ (1976, pp. 217–18) "die out quickly and to the accuracy with which we are working are equal to zero" for backcasts number 4 and greater. Confusingly, BJ do not declare the accuracy to which they are working! In their code in the back of the book, BJ (*ibid*., p. 502) recommend stopping backcasts when $w_t - \hat{\mu}$ becomes negligibly small, but they do not state what constitutes "negligibly small": is it 0.1, 0.01, 0.001, or even smaller? For this benchmark we backcast until $w_t - \hat{\mu} < 0.01$. Decreasing this tolerance to 0.001 changes the estimates of ϕ and θ in the third decimal.

Analytic derivatives are not used here, as mentioned in the previous section; yet we can be confident that we are achieving the same level of accuracy as if we were implementing analytic derivatives. The benchmark is presented in Table 28.9.

Note that this result comports with BJ's Table 7.13, the BJ results for estimation of Series A. BJ report (with standard errors in parentheses) a constant of 1.45, a ϕ of 0.92(0.04) and a θ of 0.58(0.08). The BJ constant is consistent with our benchmark constant because $17.066 \approx 1.45 \times 1/(1 - 0.9149...)$. Note that our standard errors agree with those of BJ. Further, BJ give the residual variance as 0.097 while that from the benchmark, to four decimals, is 0.0974. So it seems that we are estimating the same model for which BJ present results.

Of interest is whether any of the three packages that offer backcasting come close to the benchmark. These results are presented in Table 28.10.

Overall, no package comes close to the benchmark. As mentioned, this is because they use different, secret methods for determining Q that, for some reason, they will not reveal to their users. There is no point in checking the standard errors.

Table 28.9 ULS benchmark

Parameter	μ	ϕ	θ
Coefficient	17.065547663	0.91494836959	0.58268097638
Standard error (MLE)	0.10808561791	0.042209513625	0.083811338527

Table 28.10 Packages that backcast

Package	μ	ϕ	θ
Benchmark	17.065503687	0.91492053108	0.58262567074
Package U	17.11392	0.917600	0.607982
Package V	17.08580	0.919264	0.595042
Package W	17.06372	0.884730	0.530030

28.6.5 Final thoughts on ARMA benchmarking

We have used analytic derivatives and extended precision calculation in Mathematica to produce much-needed benchmarks for ARMA least squares models. Generally, packages can hit the point estimates for the CLS benchmark, but the method of standard error calculation for most packages is unknown. Packages offer disparate answers for the ULS benchmark because each uses its own specialized, undocumented algorithm for backcasting.

While this is a damning indictment of customary practice in econometric and statistical software practice, what can we hope for the future? Should we expect developers to fix these problems? Generally, no. To rewrite the CLS and ULS code would be very time-consuming and of little benefit. CLS is an approximation to ULS which, in turn, is an approximation to maximum likelihood. Everybody should be using exact maximum likelihood instead of CLS or ULS (Choudhury, Hubata and St. Louis, 1999). There is currently no exact maximum likelihood benchmark. Someone should develop it, and all packages should converge on it. A package that does not offer exact maximum likelihood should either implement it, or better document its existing CLS/ULS code, and ensure that it hits the benchmark presented in this section.

28.7 Conclusions

We have seen that it is not safe to assume that econometric software is accurate, and we have reviewed methods of testing econometric software, of which there are far too few. While this chapter has primarily concerned itself with the "known inputs – known outputs" approach to testing, it was mentioned that there is another approach: two independently developed methods producing the same answer. The former approach is very time-consuming and requires some knowledge of numerical methods. The latter approach simply requires two (or more) software packages, and the ability to use them correctly. This latter method has not been much employed simply because of the dearth of replication in economics. However, it is reasonable to expect that there will be much more replication in economics in the future, and the relation between the accuracy of econometric software and replication merits exposition here in the concluding section.

Over 20 years ago, Dewald, Thursby and Anderson (1986) attempted to replicate many articles from the *Journal of Money, Credit and Banking*. Dewald, Thursby and Anderson advised against the adoption of an honor system, whereby publishing

authors pledge to provide their data and code to researchers wishing to replicate the published results, due to the obvious incentive-compatibility problems. They recommended a mandatory data/code archive, whereby authors would have to deposit their data and code prior to publication. The *American Economic Review*, which published the article, nevertheless adopted an honor system. McCullough and Vinod (2003a) attempted to replicate every article in a single issue of the *American Economic Review* and discovered that half the authors would not honor their pledges; the percentage of compliant authors at other journals with honor systems was much less.

Under the then editor, Ben Bernanke, in direct response to McCullough and Vinod (2003a), the *American Economic Review* adopted a mandatory data/code archive (Bernanke, 2004). Many journals followed suit: *Econometrica*, 2005; *Review of Economic Studies*, 2005; *Journal of Political Economy*, 2006; *Spanish Economic Review*, 2007; *Canadian Journal of Economics*, 2008; and the *Review of Economics and Statistics*, 2009. More can be expected to follow. It should be noted that simply having a mandatory data/code archive is no guarantee of replicable research being published (see McCullough, McGeary and Harrison, 2006, 2008, for details). The topic of replication in economics, including its relation to software, is covered in great detail in Anderson *et al*. (2008).

As more data code and published results (read "inputs and alleged outputs") are available, more and more code will be run on more than one econometric software package, both uncovering discrepancies that need to be resolved as well as verifying that two different programs give the same answer to the same problem (increasing our confidence that both programs are correct). In the case of uncovered discrepancies, software developers are generally willing to fix these problems. The net result will be more accurate software, as evidenced by some cases we have discussed here: Drukker and Guan (2003), Zeileis and Kleiber (2005) and Bruno and De Bonis (2004).

While some journals have always been willing to publish software reviews that address accuracy issues, software reviews carry little professional credit, and until recently practically no journal would publish articles on accuracy. *Computational Statistics and Data Analysis*, *Computational Statistics*, the *International Journal of Forecasting* and the *Journal of Statistical Software* have all published such articles in recent years. So there are outlets for persons willing to do the computational work of creating benchmarks. And there will be a much greater need for it as progress on the replication of economic research leads to the identification of fruitful areas for developing such benchmarks.

Acknowledgments

The author would like to thank Houston Stokes for useful comments on this chapter.

Notes

1. Because using the StRD is much more interesting than, and not nearly so tedious as testing random number generators or statistical distributions, some of these authors only apply the StRD.

2. The R code to produce this graph is two lines: f <- function(n)max(rnorm(n)), and plot(sapply(rep(5000,4000),f)).
3. In Mathematica, this is effected by issuing the command "$MinPrecision=50."

References

Altman, M. and M. McDonald (2002) Choosing reliable statistical software. *PS: Political Science and Politics* **24**(3), 681–8.

Anderson, R., W.H. Greene, B.D. McCullough and H.D. Vinod (2008) The role of data/code archives in the future of economic research. *Journal of Economic Methodology* **15**(1), 99–119.

Bai, J. and P. Perron (2003) Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* **18**, 1–22.

Bailey, D.H. (1993) Algorithm 719: multiprecision translation and execution of FORTRAN programs. *ACM TOMS* **19**(3), 288–319.

Bankhofer, U. and A. Hilbert (1997a) An application of two-mode classification to analyze the statistical software market. In R. Klar and O. Opitz (eds.), *Classification and Knowledge Organisation*, pp. 567–72. Heidelberg: Springer.

Bankhofer, U. and A. Hilbert (1997b) Statistical software packages for windows: a market survey. *Statistical Papers* **38**, 393–407.

Bernanke, B. (2004) Editorial statement. *American Economic Review* **94**(1), 404.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–27.

Box, G.E.P. and G. Jenkins (1976) *Time Series Analysis: Forecasting and Control*. Oakland, Calif.: Holden-Day.

Brooks, C., S.P. Burke and G. Persand (2001) Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting* **17**(1), 45–56.

Brooks, C., S.P. Burke and G. Persand (2003) Multivariate GARCH models: software choice and estimation issues. *Journal of Applied Econometrics* **18**(6), 725–34.

Bruno, G. and R. De Bonis (2004) A comparative study of alternative econometric packages with an application to Italian deposit interest rates. *Journal of Economic and Social Measurement* **29**, 271–95.

Calzolari, G. and L. Panattoni (1988) Alternative estimators of FIML covariance matrix: a Monte Carlo study. *Econometrica* **56**(3), 701–14.

Choi, H.-S. and N.M. Kiefer (2005) Software evaluation: EasyReg International. *International Journal of Forecasting* **21**(3), 609–16.

Choudhury, A.-H., R. Hubata and R.D. St. Louis (1999) Understanding time-series regression estimators. *American Statistician* **53**(4), 342–8.

Coddington, P.D. (1994) Analysis of random number generators using Monte Carlo simulation. *International Journal of Modern Physics C* **3**, 547–60.

Coddington, P.D. (1996) Tests of random number generators using Ising model simulations. *International Journal of Modern Physics* **7**(3), 295–303.

Davidson, R. and J.G. MacKinnon (2004) *Econometric Theory and Methods*. New York: Oxford University Press.

Dewald, W.C., J. Thursby and R. Anderson (1986) Replication in empirical economics: the *Journal of Money, Credit and Banking* project. *American Economic Review* **76**(4), 587–603.

Drukker, D.M. and W. Guan (2003) Replicating the results in "On efficient estimation with panel data: an empirical comparison and instrumental variables estimators." *Journal of Applied Econometrics* **18**(1), 119–20.

Fiorentini G., G. Calzolari and L. Panattoni (1996) Analytic derivatives and the computation of GARCH estimates. *Journal of Applied Econometrics* **11**(4), 399–417.

Goldberg, D. (1991) What every computer scientist should know about floating point arithmetic. *ACM Computing Surveys* **23**(1), 5–48.

Gourieroux, C. and A. Montfort (1997) *Simulation-based Econometric Methods*. New York: Oxford University Press.

Granger, C.W.J. and P. Newbold (1977) *Forecasting Economic Time Series*. Orlando, Flor.: Academic Press.

Keeling, K.B. and R.J. Pavur (2004) Numerical accuracy issues in using Excel for simulation studies. In R.G. Ingalls, M.D. Rossetti, J.S. Smith and B.A. Peters (eds.), *Proceedings of the 2004 Winter Simulation Conference, Volume 2*, pp. 1513–18. Piscataway, NJ: IEEE.

Keeling, K.B. and R.J. Pavur (2007) A comparative study of the reliability of nine statistical software packages. *Computational Statistics and Data Analysis* **51**(8), 3811–31.

Kennedy, W. and J. Gentle (1980) *Statistical Computing*. New York: Marcel-Dekker.

Kitchen, A.M., R. Drachenerg and J. Symanzik (2003) Assessing the reliability of web-based statistical software. *Computational Statistics* **18**(1), 107–22.

Knüsel, L. (1989) Computergestützte Berechnung statistischer Verteilungen. München-Wien: Oldenbourg (English version available at http://www.stat.uni-muenchen.de/ knuesel/elv).

Knuth, D. (1998) *The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3e*. Reading, Mass.: Addison-Wesley.

L'Ecuyer, P. and R. Simard (2007) TestU01: a C library for empirical testing of random number generators. *ACM TOMS* **33**(4), Article 22.

Ling, R.F. (1974) Comparison of several algorithms for computing sample means and variances. *Journal of the American Statistical Association* **69**, 859–66.

Longley, J.W. (1967) An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association* **62**(319), 819–41.

Marsaglia, G. (1996) DIEHARD: a battery of tests of randomness, http://stat.fsu.edu/pub/diehard.

McCullough, B.D. (1998a) Assessing the reliability of statistical software: part I. *American Statistician* **52**(4), 358–66.

McCullough, B.D. (1998b) Algorithm choice for (partial) autocorrelation functions. *Journal of Economic and Social Measurement* **24**(3/4), 265–78.

McCullough, B.D. (1999a) Assessing the reliability of statistical software: part II. *The American Statistician* **53**(2), 149–59.

McCullough, B.D. (1999b) Econometric software reliability: E-Views, LIMDEP, SHAZAM, and TSP. *Journal of Applied Econometrics* **14**(2), 191–202.

McCullough, B.D. (2000a) Is it safe to assume that software is accurate? *International Journal of Forecasting* **16**(3), 349–57.

McCullough, B.D. (2000b) Experience with the StRD: application and interpretation. *Computing Science and Statistics* **31**, 16–21.

McCullough, B.D. (2000c) The accuracy of Mathematica 4 as a statistical package. *Computational Statistics* **15**(2), 279–99.

McCullough, B.D. (2001) Random number generators. In A. El-Shaarawi and W. Piegorsch (eds.), *Encyclopedia of Environmetrics, Volume 3*, pp. 1678–81. New York: Wiley.

McCullough, B.D. (2004a) Wilkinson's tests and econometric software. *Journal of Economic and Social Measurement* **29**(1–3), 261–70.

McCullough, B.D. (2004b) Some details of nonlinear estimation. In M. Altman, J. Gill and M.P. McDonald (eds.), *Numerical Methods in Statistical Computing for the Social Sciences*, pp. 199–218. New York: Wiley.

McCullough, B.D. (2006) A review of TESTU01. *Journal of Applied Econometrics* **21**(5), 677–82.

McCullough, B.D., K.A. McGeary and T.D. Harrison (2006) Lessons from the JMCB archive. *Journal of Money, Credit and Banking* **38**(4), 1093–107.

McCullough, B.D., K.A. McGeary and T.D. Harrison (2008) Do economics journal archives promote replicable research? *Canadian Journal of Economics*. Forthcoming.

McCullough, B.D. and C.G. Renfro (1999) Benchmarks and software standards: A case study of GARCH procedures. *Journal of Economic and Social Measurement* **25**(2), 59–71.

McCullough, B.D. and C.G. Renfro (2000) Some numerical aspects of nonlinear estimation. *Journal of Economic and Social Measurement* **26**(1), 63–77.

McCullough, B.D. and H.D. Vinod (1999) The numerical reliability of econometric software. *Journal of Economic Literature* **37**(2), 633–55.
McCullough, B.D. and H.D. Vinod (2003a) Verifying the solution from a nonlinear solver: A case study. *American Economic Review* **93**(3), 873–92.
McCullough, B.D. and H.D. Vinod (2003b) Comment: econometrics and software. *Journal of Economic Perspectives* **17**(1), 223–4.
McCullough, B.D. and H.D. Vinod (2004) Verifying the solution from a nonlinear solver: reply. *American Economic Review* **94**(1), 400–3.
McCullough, B.D. and B. Wilson (2005) On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis* **49**(4), 1244–52.
Newbold, P., C. Agiakloglou and J. Miller (1994) Adventures with ARIMA software. *International Journal of Forecasting* **10**(4), 573–81.
Priestley, M.B. (1981) *Spectral Analysis and Time Series*. New York: Academic Press.
Sall, J. (2002) Comment. *American Statistician* **56**, 160–1.
Sawitzki, G. (1994a) Testing numerical reliability in data analysis systems. *Computational Statistics and Data Analysis* **18**(2), 269–85.
Sawitzki, G. (1994b) Report on the numerical reliability of data analysis systems. *Computational Statistics and Data Analysis (SSN)* **18**(2), 289–301.
Silk, J. (1996) Systems estimation: a comparison of SAS, SHAZAM and TSP. *Journal of Applied Econometrics* **11**(4), 437–50.
Stokes, H.H. (2004) On the advantage of using two or more econometric software systems to solve the same problem. *Journal of Economic and Social Measurement* **29**, 307–320.
Teyssiere, G. (2005) Structural time series modelling with STAMP 6.02. *Journal of Applied Econometrics* **20**(4), 571–7.
Thisted, R.A. (1988) *Elements of Statistical Computing*. New York: Chapman and Hall.
Tirler, G., P. Dalgaard, W.Hörmann and J. Leydold (2004) An error in the Kinderman-Ramage method and how to fix it. *Computational Statistics and Data Analysis* **47**(3), 433–40.
Vinod, H.D. (2000). Review of GAUSS for windows, including its numerical accuracy. *Journal of Applied Econometrics* **15**, 211–22.
Wilkinson, L. (1985) *Statistics Quiz*. Evanston, Ill.: SYSTAT.
Yalta, A.T. (2007) The numerical reliability of GAUSS 8.0. *American Statistician* **61**, 262–8.
Yalta, A.T. (2008) The reliability of statistical distributions in Microsoft Excel 2007. *Computational Statistics and Data Analysis* **52**(10), 4579–86.
Yalta, A.T. and A.Y. Yalta (2007) GRETL 1.6.0 and its numerical accuracy. *Journal of Applied Econometrics* **22**(4), 849–54.
Yalta, A.T. and O. Jenal (2009) On the importance of verifying forecasting results. *International Journal of Forecasting*. Forthcoming.
Zeileis, A. and C. Kleiber (2005) Validating multiple structural change models – a case study. *Journal of Applied Econometrics* **20**, 685–90.

29

Trends in Applied Econometrics Software Development 1985–2008: An Analysis of *Journal of Applied Econometrics* Research Articles, Software Reviews, Data and Code[1]

Marius Ooms

Abstract

Trends in software development for applied econometrics emerge from an analysis of the research articles and software reviews of the *Journal of Applied Econometrics* (*JAE*) appearing since 1986. The data and code archive of the journal provides more specific information on software use for applied econometrics since 1995. GAUSS, Stata, MATLAB and Ox have been the most important software since 2001. I compare these higher-level programming languages and R in somewhat more detail. An increasing number of packages are being used. A surprisingly low number of products have been discontinued since 1987. I put the time series count data on the number of articles using different software and on the number of reviews discussing different products in a historical perspective, where I distinguish several software types. Two waves of new products showed up in the period under study, the first associated with the introduction of the personal computer and new graphical interfaces, the second with the appearance of the internet. The *JAE* has reviewed 77 packages. In this chapter I discuss 13 other relevant packages. A table with all mentioned packages, their authors and latest versions provides a comprehensive overview of the relevant software as in June 2008.

29.1 Introduction

In this chapter I provide an overview of academic applied econometrics software development, deriving time series count data from the *JAE* software reviews (1987–2008), *JAE* research articles and the *JAE* data archive (1995–2008). The *JAE* has promoted documentation and indexing of softwares and codes for applied econometrics by publishing software reviews and replication studies. Most importantly, James MacKinnon has patiently, successfully and consistently added the software codes of *JAE* authors to the *JAE* data archive.

I first provide a contingency table of used data type versus year of publication. The types of data used indicate a gradual shift from traditional macroeconometrics and time series analysis to microeconometric applications and panel data research. Second, I present the distribution of reviews per software category per two years, and I check which software still existed in June 2008. Third, I present the yearly distribution of software use over the 25 specifically mentioned software packages.

During the observation period, the *JAE* reviewed the usefulness of 77 different packages for applied econometrics research and education. Surprisingly, only a handful of these products have been discontinued before June 2008 and a large majority have recently been updated. Trends in general and individual applied econometric software development emerge from the corresponding tables. In recent years the range of effective specific softwares in applied econometric research has increased. GAUSS, Stata and MATLAB dominate. Freely downloadable alternatives like R and Ox have not had a similar impact as yet.

Econometric programs such as LIMDEP, SHAZAM, TSP, RATS and Ox are also used for scientific research outside applied econometrics, not only in the traditionally related areas of econometric theory, applied statistics and applied economics, but also in marketing, finance, management science, accounting, regional science and transportation science. For example, Altman and McDonald (2001) survey the use of software in Political Science, including many econometrics packages. My analysis is therefore admittedly very focused. Many interesting applied econometrics articles have been published outside the *JAE*, but data on software use and development for other journals are not easy to obtain and results are therefore difficult to check.

This chapter implicitly defines applied econometrics as the econometrics that leads to publication in the *JAE*. Cleaning and preparing complicated empirical datasets, writing code for advanced estimation procedures or new types of inference, and presenting and interpreting results for *JAE* articles involves expert knowledge that distinguishes applied econometrics from both applied economics and econometric theory.

The remainder of this chapter is organized as follows. The *JAE* research articles, software reviews, data archive and software use are discussed in sections 29.2–29.5, respectively. The most intensively used high-level programming languages are treated in more detail in section 29.6.

A deeper understanding of the tables is obtained by a selective description of the history and characteristics of the packages, given in section 29.7. This section

draws heavily on Ooms and Doornik (2006) and on the extensive account of Renfro (2004b), who corresponded with many econometric software developers in preparing his article and in editing Renfro (2004a). It also reflects my experience as editor of the Econometric Software Links of the *Econometrics Journal* at http://www.econometriclinks.com. Section 29.8 discusses the combination of software and the concluding section 29.9 looks into future aspects of econometric modeling software.

29.2 *JAE* research articles

The *JAE* is an important source of information on trends in software development. The founding editor, Hashem Pesaran, has been based at the Cambridge (UK) Department of Applied Economics (DAE) for most of the time since 1986. Richard Stone, the founder of the DAE, wrote the first *JAE* article, this being his Nobel Prize lecture on national accounts (Stone, 1986). Stone's methods still underpin the basic data source for applied macroeconometric research today. Whereas Stone pioneered mainframe econometric software development in Cambridge, Pesaran was one of the first to produce user-friendly software for the PC, Data-Fit and Microfit, as reviewed in Ericsson (1988). He initiated the software review section and a replication section for the *JAE*. He has written influential publications in theoretical and applied time series econometrics, and in theoretical and applied microeconometrics for cross-section and panel data.

The *JAE* publishes applied econometric research in all important areas in the field. Special issues of the journal indicate the wide range of topics and methods: time series and cross-section model specification as in McAleer (1989) and Magnus and Morgan (1997); event counts as in Trivedi (1997); nonlinear dynamics as in Pesaran and Potter (1992); simulation-based inference (frequentist and Bayesian) as in Brown *et al*. (1993), macro time series as in Pagan (1994), Diebold and Watson (1996), Hendry and Pesaran (2001) and Franses *et al*. (2005); microeconometric structural dynamics as in Kapteyn *et al*. (1995) and Christensen *et al*. (2004); semiparametric microeconometrics as in Horowitz *et al*. (1998); statistical decision making (Bayesian and frequentist, macro, micro and finance) as in Geweke *et al*. (2000); financial time series analysis as in Franses and McAleer (2002); social and spatial interactions as in Durlauf and Moffitt (2003); and, finally, empirical industrial organization as in Bauwens *et al*. (2007).

The *JAE* co-editors have worked on both sides of the Atlantic and the Pacific and represent the major fields and schools of applied econometrics. Table 29.1 also illustrates this point. It shows the frequency distributions of the dataset types over three main categories, panel data, time series data and cross-section data, for the years 1995–2008, although with only four issues of 2008 covered. The basic source of these counts were the *JAE* authors' readme files on the *JAE* data archive. If these were unclear I checked the corresponding articles on the JSTOR archive and on Wiley Interscience. The gradual shift from traditional macroeconometrics and time series analysis to microeconometric applications and panel data research emerges. Time series articles are overrepresented in the years with corresponding

Table 29.1 Research articles in *JAE* per data type per year

	95	*96*	*97*	*98*	*99*	*00*	*01*	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*Total*
Panel data	9	5	7	8	5	8	4	4	15	21	9	17	21	7	140
Time series	16	22	13	14	18	15	24	18	14	16	30	27	18	8	253
Cross-section	8	4	10	8	4	7	5	5	3	9	6	14	16	6	105
Simulated	.	1	.	.	.	.	1	.	1	.	.	1	.	.	4
Experiment					1	.	.	1	.	.	.	.	.	.	2
Meta-data					1	.	.	.	.	.	.	.	1	.	2
Auction					1	.	.	.	1	.	.	2	.	.	4
Scanner						1	.	.	.	.	.	.	1	.	2
Algebra								1	.	.	.	.	.	.	1
Total	33	32	30	30	30	31	34	29	34	46	45	61	57	21	513

Notes: panel data: data with a small time series dimension and a large cross-section dimension; time series: data with large time series dimension, larger than cross-section dimension; cross-section: cross-section data without time series dimension; experiment: data from experimental economics; simulated: data from random number generator (RNG) and known data-generating process (DGP); meta-data: data summarizing results from other articles; auction: empirical data from auctions.
Sources: JAE data archive, http://www.econ.queensu.ca/jae/; JSTOR, http://www.jstor.org; http://www3.interscience.wiley.com; ISSN code JAE: 08837252.

special issues: 1996, 2001 and 2005. Four articles are based on simulated (Monte Carlo) data, reflecting the research interest of James MacKinnon. Two articles use data from economic experiments and four from auctions, new fields for serious applied econometrics. One article uses cross-section meta-data in a traditional way, and Baltagi (1999) uses bibliographical panel meta-data to construct rankings of authors and departments in applied econometrics. Finally, Meddahi (2002) is the only pure econometric theory *JAE* article I have come across. In computing the total number of research articles, I have included articles from the *JAE*'s replication section, edited by Badi Baltagi.

29.3 *JAE* software reviews

The *JAE* software reviews have been edited by Pravin Trivedi (1988–92) and James MacKinnon. The reviews vary greatly in length. Most reviews concentrate on one package, others compare up to six different packages on many features (data management, model formulation, simulation, availability of procedures, speed, help functions and documentation), as in Brillet (1989) and Cribari-Neto (1997). Other reviews compare specific functions like survival modeling, as in Goldstein *et al*. (1989); GARCH modeling, as in Brooks *et al*. (2003); or properties like numerical reliability, as in McCullough (1999).

Many packages have been reviewed only once, but dedicated widely used (inside and outside the *JAE*) econometric packages show up several times in these 20 years. Table 29.2 details the reviews of dedicated econometric software since 1987, split into two-year periods to show the distributions over time for each package. Repeated reviews of the same product occur because the package receives a major

Table 29.2 Software package and *JAE* reviews per software per two years, part 1: econometric software

Software (old) package	*Type*	*88*	*90*	*92*	*94*	*96*	*98*	*00*	*02*	*04*	*06*	*08*	*Tot.*	*V.08*	*Y.*	*Author*	*Country*
AREMOS	E	.	1	.	.	.	.	.	.	.	.	.	1	5.3	03	Global Insight	US
EasyReg	E	.	.	.	.	.	1	.	.	.	.	.	1	2007	07	H. Bierens	US
ESP	E	.	1	.	.	.	.	.	.	.	.	.	1	†	92	J.P. Cooper, O.A. Curtis	US
Eviews (MicroTSP)	E	.	1	1	.	1	.	3	1	1	1	1	10	6	07	QMS, D. Lilien	US
GAUSS - GAUSSX	E	.	.	.	.	1	.	.	.	.	.	.	1	9	08	Econotron Soft, J. Breslaw	CA
Gretl	E	.	.	.	.	.	.	.	.	1	1	1	3	1.7.5	08	A. Cottrell, R. Lucchetti	US,IT
LIMDEP	E	1	1	1	1	.	.	2	.	1	.	.	7	9	07	W. Greene	US
Microfit (DataFit)	E	2	.	.	1	.	1	.	.	.	.	.	3	4	98	B. Pesaran, M.H. Pesaran	UK
MODLER	E	.	1	.	.	.	.	.	.	.	.	.	1	10.7	08	C. Renfro	US
PCBRAP	E	.	.	1	.	.	.	.	.	.	.	.	1		02	A. Zellner	US
OxMetrics-PcGive(PcFiml)	E	1	.	1	.	2	1	.	.	1	1	.	7	12	07	J.A. Doornik, D.F. Hendry	UK
PERM	E	.	.	1	.	.	.	.	.	.	.	.	1	†	94		US
SHAZAM	E	2	.	1	.	1	.	1	.	.	.	.	5	10	08	D. Whistler, K. White	US
SORITEC	E	.	1	.	.	.	.	.	1	.	.	.	2		98	J. Sneed	US
TSP	E	.	.	.	.	1	1	1	.	.	.	.	3	5	08	B. Hall, C. Cummins	US
Autobox	ETS	1	.	.	.	.	.	.	.	.	.	.	1	6	07	Automatic Forecasting Sys.	US
Dynare	ETS	.	.	.	.	.	.	.	.	.	.	.	.	4	08	M. Juillard	FR
Forecast Master	ETS	.	1	.	.	.	.	.	.	.	.	.	1		98	Scientific Systems Company	US
GAUSS - COINT	ETS	.	.	.	.	1	.	.	.	.	.	.	1	2	94	S. Ouliaris, P.C.B. Phillips	US
GAUSS - FANPAC	ETS	.	.	.	.	.	.	1	.	1	.	.	2	2	02	R. Schoenberg	US
GAUSS - GRTE - JMulti	ETS	.	.	.	.	.	.	.	.	.	.	.	.	4.2.1	08	M. Krätzig, H. Lütkepohl	DE,IT
GAUSS - TSM	ETS	.	.	.	.	.	.	.	.	.	.	.	.	1	07	Aptech	US
MATLAB - BDS	ETS	.	.	.	.	.	.	.	1	.	.	.	1		99	L. Kanzler	DE
MTS	ETS	.	1	.	.	.	.	.	.	.	.	.	1	1	96	Automatic Forecasting Sys.	US
Ox - TSMod	ETS	.	.	.	.	.	.	.	.	.	1	.	1	4.26	08	J. Davidson	UK
RATS	ETS	.	1	.	.	.	1	.	.	1	.	.	3	7	07	Estima, T. Doan	US
RATS - CATS	ETS	.	.	.	.	.	1	.	.	.	.	.	1	2	06	H. Hansen, K. Juselius	DK
SAS - ETS	ETS	.	1	.	.	1	.	.	.	1	.	.	3	9.1	07	SAS Institute	US
SIMPC	ETS	.	.	.	.	1	.	.	.	.	.	.	1	†	94	H. Don	NL
S-PLUS - FinMetrics	ETS	.	.	.	.	.	.	.	.	2	.	.	2	3	07	E. Zivot, J. Wang	US
STAMP	ETS	.	1	.	.	.	.	.	.	.	1	.	2	8	07	S.J. Koopman, A.C. Harvey	NL,UK
TSW (TRAMO/SEATS)	ETS	.	.	.	.	.	.	.	1	.	.	.	1	1	08	G. Caporello, A. Maravall	ES
X-12 ARIMA (X-11)	ETS	.	.	.	.	.	.	.	.	.	.	.	.	0.3	07	U.S. Census, B. Monsell	US

Notes: See also Tables 29.3 and 29.4. Software in alphabetical order within type categories; review counts per two-year periods, 88: 1987–88,. . ., 08: 2007–08; old names in parentheses; Tot: total number of reviews per software; . for unreviewed software; †: discontinued; V.08: last version in June 2008; Y.: year of last update; Author: name of producing company/author(s) (not all authors mentioned); Type descriptions as E: econometrics package; ETS: econometrics time series package.

Table 29.3 Software and *JAE* reviews per software package per two years, part 2: various packages

Software package	*Type*	*88*	*90*	*92*	*94*	*96*	*98*	*00*	*02*	*04*	*06*	*08*	*Tot.*	*V.08*	*Y.*	*Author*	*Country*
GAUSS - Micro-EBA	ECS	.	.	1	.	.	.	.	.	.	.	.	1	†		J. Fowles	US
PcGive PcNaive	EMC	.	.	.	1	.	.	.	.	.	.	.	1	5	08	J.A. Doornik, D.F. Hendry	UK
BACC	EMC2	.	.	.	.	.	.	1	.	.	.	.	1	2003	03	J. Geweke, W. McCausland	US
Ox	EMPL	.	.	.	.	.	1	.	.	.	.	.	1	5.1	08	OxMetrics, J.A. Doornik	UK
OxGauss-M@ximize	EMPL	.	.	.	.	.	.	.	.	.	1	.	1	1.0	03	S. Laurent, J.P. Urbain	BE,NL
Fortran - FRONTIER	EPD	.	.	.	.	.	.	1	.	.	.	.	1	4.1	03	T. Coelli	AU
GAUSS - DPD	EPD	.	.	.	.	.	.	.	.	.	.	.	.	98	98	M. Arellano, S. Bond	ES,UK
GAUSS - ExpEnd	EPD	.	.	.	.	.	.	.	.	1	.	.	1	1	02	F. Windmeijer	UK
Maple	MCA	.	.	.	.	1	.	.	.	.	.	.	1	12	08	Maplesoft	CA
Mathematica	MCA	.	.	.	.	.	.	.	.	1	.	.	1	5.2	08	Wolfram Research	US
GNUPlot	G	.	.	.	.	.	.	.	.	.	1	.	1	4.3	08	GNUplot team	
JStatCom	GUI	.	.	.	.	.	.	.	.	.	.	.	.	2.4	08	M. Krätzig	DE
wxWidgets	GUI	.	.	.	.	.	.	.	.	.	.	.	.	2.8	08	wxWidgets project, J. Smart	
GAUSS	MPL	.	.	1	1	.	1	1	.	.	.	.	4	9	07	Aptech, N. Lohonen	US
MATLAB	MPL	.	.	.	1	.	1	.	1	.	.	.	3	7.6	08	Mathworks	US
Maxima	MPL	.	.	.	.	.	.	.	.	.	.	1	1	5.15	08	Maxima team	US
NAG	MPL	.	1	.	.	.	.	.	.	.	.	.	1	Mark 21	07	NAG Group Ltd.	UK
C++ - Newmat	MPL	.	.	.	.	1	.	.	.	.	.	.	1	10	06	R. Davies	US
Octave	MPL	.	.	.	.	.	.	2	1	.	.	.	3	3.01	08	J. Eaton	US
Scilab	MPL	.	.	.	.	.	.	1	1	.	.	.	2	4.1	07	Scilab Consortium	FR
Yorick	MPL	.	.	.	.	.	.	1	.	.	.	.	1	2.1	08	D. Munro	US
LaTeX	MWPL	.	.	.	.	.	.	.	1	.	.	.	1	2.7	07	C. Schenk	DE
MPI LAM	NMT	.	.	.	.	.	.	.	1	.	.	.	1	7.3	07	LAM/MPI team	
ParallelKnoppix	NMT	.	.	.	.	.	.	.	.	.	.	1	1	2.9	08	M. Creel	ES
C++	NPL	.	.	.	.	1	.	.	.	.	.	.	1	4.2	08	i.a. GNUcc team	
Fortran	NPL	.	.	.	.	.	.	.	.	.	.	.	.	95/2003	07	i.a. NAG fortran	UK
Debian-GNU/Linux	OS+	.	.	.	.	.	.	1	.	.	.	1	2	4.0	08	Debian team	
Cygwin	OST	.	.	.	.	.	.	1	.	.	.	.	1	1.5	08	Cygwin team	
Perl	TNPL	.	.	.	.	.	.	.	.	1	.	.	1	5.10	08	Perl team	
Python	TNPL	.	.	.	.	.	.	1	.	.	.	.	1	2.5	08	Python Team	

Notes: See also Tables 29.2 and 29.4. Type descriptions as ECS: econometrics cross-section package; EMC2: econometrics Bayesian Markov chain Monte Carlo package; EMPL: econometrics matrix programming language; EPD: econometrics panel data package; MCA: mathematics computer algebra package; G: graphics package; GUI: graphical user interface; MPL: mathematical matrix programming language; MWPL: mathematical word processing language; NMT: numerical tool (parallel computing); OS+: operating system plus applications; OST: operating system cross-over package; TNPL: text processing and numerical programming language.

Table 29.4 Software and *JAE* reviews per software per two years, part 3: statistical packages

Software package	*Type*	*88*	*90*	*92*	*94*	*96*	*98*	*00*	*02*	*04*	*06*	*08*	*Tot.*	*V.08*	*Y.*	*Author*	*Country*
SYSTAT	S	.	1	.	.	.	.	.	.	.	.	.	1	12	08	Systat Soft, L. Wilkinson	US
Math.- MathStatica	SCA	.	.	.	.	.	.	.	.	1	.	.	1	1.5	06	C. Rose, M.D. Smith	AU
BMDP	SCS	.	1	.	.	.	.	.	.	.	.	.	1	2007	07	W. Dixon, M.B. Brown	US
GAIM	SCS	.	.	.	1	.	.	.	.	.	.	.	1	†	.	T. Almudevar, R. Tibshirani	US
NCSS	SCS	.	1	.	.	.	.	.	.	.	.	.	1	2007	07	NCSS	US
N-KERNEL	SCS	.	1	.	.	.	.	.	.	.	.	.	1	†			US
Stat/Transfer	SDT	.	.	.	.	.	.	.	.	.	.	.	.	9	07	Circle Systems Inc.	US
Excel	SG	.	.	.	.	.	.	.	.	.	.	.	.	2007	07	Microsoft	US
STATGRAPHICS	SG	.	.	.	1	.	.	.	.	.	.	.	1	XV.II	07	StatPoint Inc.	US
ViSta	SG	.	.	.	.	.	.	.	1	.	.	.	1	7.9	07	P.M. Valero-Mora, M. Friendly	CH,CA
TESTU01	SMC	.	.	.	.	.	.	.	.	.	1	.	1	1.2.1	08	R. Simard	CA
BUGS -Open BUGS	SMC2	.	.	.	.	.	.	.	.	.	.	.	.	1.4.3	07	D. Lunn, A. Thomas	UK, FI
C++ -BIOGEME	SPD	.	.	.	.	.	.	.	.	.	.	.	.	1.6	08	M. Bierlaire	CH
R	SPL	.	.	.	.	.	.	2	1	.	1	.	4	2.7	08	R team	
SC	SPL	.	.	.	1	.	.	.	.	.	.	.	1	2.03	05	T. Dusoir	FR
S-PLUS	SPL	.	.	.	1	.	1	.	.	.	.	.	2	8	07	Insightful Corp.	US
Stata	SPL	.	1	1	1	.	.	.	1	.	.	.	4	10	07	StataCorp, W. Gould	US
SST	SPL	1	.	1	.	.	.	.	.	.	.	.	2	3	04	J. Dubin	US
Xplore	SPL	.	1	.	.	.	1	.	.	.	.	.	2	4.7	07	MD*Tech, W. Härdle	DE
LISREL	SSS	.	.	.	.	.	.	.	.	1	.	.	1	8.8	06	SSI International	US
SPSS	SSS	.	.	.	.	.	.	.	.	.	.	.	.	16	07	SPSS Inc.	CA
R ts	STS	.	.	.	.	.	.	.	1	.	.	.	1	0.15	08	A. Trapletti	AT
Total reviews		7	18	10	10	11	11	20	13	14	9	5	128				
Range reviews		6	18	10	10	11	11	15	13	13	9	5	77				
Total articles		7	9	10	9	8	9	11	9	9	6	5	92				
Not reviewed													13				

Notes: Software: in alphabetical order within type categories; review counts per two-year periods, 88: 1987–88,. . .; 08: 2007–08; old names in parentheses; Tot.: total number of reviews per software package; . for unreviewed software; V.08: last version in June 2008; Y.: year of last update; Author: name of producing company/author(s) (not all authors mentioned); Type descriptions as S: statistics package; SCA: statistics computer algebra package; SCS: statistical cross-section package; SG: statistical graphics package; SDT: statistical data transfer package; SPD: statistical panel data package; SMC: statistical simulation (random number generator testing) package; SMC2: econometrics Bayesian Markov chain Monte Carlo package; SPL: statistical programming language; SSS: Social Sciences Statistics package; STS: Statistical Time Series package. Reviews in *JAE* often discuss multiple packages, so that total reviews (Tables 29.2–29.4) exceeds total articles.

Sources: JAE in JSTOR: http://www.jstor.org, and in http://www3.interscience.wiley.com; ISSN code JAE: 08837252; URLs for softwares available via the econometric links of the *Econometrics Journal*, http://www.econometriclinks.com.

update (in the beginning of its life) or because it is interesting, important and accessible enough to include in a comparison. JSTOR provides extensive bibliographical information on archived *JAE* articles in database entries like "Reviewed work(s)," but so far, this information is inaccurate and incomplete for the *JAE* reviews, so the numbers in Table 29.2 are based on the full text of the 92 articles.

I also checked the latest update and version number to make the table interesting as a reference for the state of relevant software in June 2008. I was first surprised to find recent updates for most of the packages. This may have been caused by the introduction of Windows Vista and Excel 2007, which made updating necessary for users who are not able to choose between operating systems. Table 29.2 also shows the current software companies and main author names. This entry is not relevant for the modern freely downloadable "team" software and these are therefore missing. The last column gives the country code of the workplace of the company and main authors. Most companies and software developers work in the US; some are in the UK, and nearly all others are in Canada and mainland Europe. None are in South America and Asia, although econometrics is now a well established field of (social) science in those continents. Irregular updates of internet links to the packages will be provided on the Econometric Software links of the *Econometrics Journal*.

The popular econometrics package Eviews (formerly Micro-TSP) has been discussed most often. LIMDEP, SHAZAM, PcGive and Microfit received most attention in the twentieth century. Gretl is the latest general econometrics package to appear on the *JAE* pages, and S-PLUS-FinMetrics is the latest time series econometrics package that has been reviewed. Three reviewed econometrics packages have been discontinued, or at least I could no longer trace them on the internet: ESP, PERM and SIMPC. All other packages have been updated since the first review. I have included three unreviewed packages in the list. TSM for GAUSS was mentioned in the code archive. Dynare, by Michel Juillard, is widely used in modern applied macroeconomics. Juillard (1996) is often cited. The 2008 version is available as a stand-alone program, but also in the form of GAUSS, MATLAB and Scilab packages. JMulti is a teaching package for multivariate time series analysis (see Lütkepohl and Krätzig, 2004); it previously required GAUSS to run. Markus Krätzig developed a graphical user interface (GUI), JStatCom (see also Table 29.3, and Krätzig, 2006). Using this GUI and GRTE (the GAUSS RunTime Engine), JMulti is now also available as a free stand-alone package.

Table 29.3 shows the corresponding review counts of programs and two packages for Bayesian econometrics (Micro-EBA and BACC), specific panel data econometrics (Frontier, DPD and ExPEnd), the econometric programming language Ox, and other packages used for scientific word processing, mathematics and computer science. I added the DPD package for dynamic panel data analysis. This code, by Manuel Arellano and Stephen Bond, has been instrumental for the breakthrough of dynamic panel data econometrics, catering for large unbalanced panels as encountered in practical applications. The fundamental article, Arellano and Bond (1991), has the exceptional econometrics citation scores of 900+ in the ISI Web of Knowledge and 4,000+ in Google Scholar. Their procedures have now been implemented in most econometric packages, both in the original time series-oriented packages

(PcGive) and in the original cross-section packages (LIMDEP). FORTRAN in Table 29.3 and BIOGEME, Excel and SPSS in Table 29.4 have been included to keep consistency with Table 29.5 below. BIOGEME and SPSS are discussed later in this chapter. Stat/Transfer has not been reviewed, but it is referred to on the *JAE* data archive. It allows for user-friendly transfer of datasets between statistics packages and LIMDEP, GAUSS, MATLAB and Excel.

Finally, Table 29.4 considers the statistical software reviews and provides summary statistics. Here I also added BUGS by Lunn *et al*. (2000), because it is widely used in Bayesian econometrics teaching; WinBUGS is a popular version for Windows. The preferred version is called OpenBUGS. The summary shows that 77 different packages have been reviewed 128 times in 92 articles. Thirteen packages have not been reviewed. The number of "reviews" in the table equals or exceeds the number of "articles" by definition. As explained above, a large difference between these two numbers indicates the discussion of several packages in single articles. This phenomenon occurred in 1989–90 when many PC packages for econometrics became fit for review, and in 1999–2000 when the first "GNUwares" came into use among econometricians.

29.4 The *JAE* data and code archive and reproducibility

The data (and code) archive of *JAE*, http://www.econ.queensu.ca/jae/, consistently coordinated by James MacKinnon, contains detailed references of all articles published since 1995. Most authors (85%) have complied with the policy to provide their data in a well documented human-readable format, fit for different operating systems and econometric software, and usable for many years to come. This is a high success rate, compared to journals in economics or statistics who intend to have a similar policy. Authors who do not provide data for no reason whatsoever receive the remark: "Contrary to the policy of the Journal, the author has failed to submit the data used in this paper."

In recent years a growing number of microeconometric datasets and even some software codes are confidential for reasons of privacy, so the overall coverage of the data archive will go down in the coming years. On the other hand, the number of articles providing details on used software and codes has been high and increasing. This is the main motivation for choosing *JAE* articles, data and code as the main sources of information for this chapter.

The existence of a carefully managed and indexed data and code archive is an essential prerequisite for the scientific ideal of effortless reproducibility of key results in applied econometrics. Anderson *et al*. (2008) set the *JAE* data and code archive as an example. William Greene is a leading econometric software developer, and Bruce McCullough and Hrishikesh Vinod are influential reviewers. They discussed the disappointing compliance rates for leading American economics journals for a recent American Economic Association meeting. The situation is hardly better for leading statistics journals, like the *Journal of Business and Economic Statistics*, where the latest instructions for the FTP (file transfer protocol) data archive are now eight years old.

Of course, thanks to automatic indexing by Google and free specific internet aggregators of economic and econometric research (papers, articles, books, citations, data and software) like RePEc (http://www.repec.org), it is relatively easy to find properly documented econometric source code outside official peer-reviewed archives. Unsurprisingly, given the working environment of most econometricians, robust, high-quality econometric procedures seldom come free. Here, the situation in computer science and statistics seems to be much better as the (much larger) programmer communities are funded in a different way.

Buckheit and Donoho (1995) gave a lively discussion of the difficulties in reproducing (even) one's own computer intensive results in computer science. Koenker and Zeileis (2007) elaborate on the difficulties in reproducing exact econometric results using codes from data archives. This is a nontrivial exercise, even using the original econometric software and a similar operating system. They advocate the use of internet-based tools for subversion control (SVN) for programmer communities and recent R applications to consistently develop reproducible econometric results. Roger Koenker is the father of quantile regression in econometrics (see Koenker, 2005). Achim Zeileis is a key R developer.

The good news to derive from Tables 29.2–29.4 is that it is now unlikely that the current software and code will become completely useless because of the discontinuation of products.

29.5 Software used in *JAE* research articles

Table 29.5 details the time-varying impact of the main software in applied econometrics research since 1995. The software packages are ordered by first-mentioned use to get a clear picture of the growing range of products used. Up to three software packages were mentioned per article; for example, S-PLUS, FORTRAN and Stata for a cross-section study. The basic sources of the counts were the readme files on the data archive. If these were unclear I checked the corresponding articles on the JSTOR archive and on Wiley Interscience.

The "Range" indicates the number of different products per year, which reached a maximum of 14 in 2006. The row labeled "Missing" counts the number of articles that don't mention specific software. This number has increased in absolute terms, but it has decreased compared with the number of (research) "Articles" mentioned in the bottom row. Twenty-five packages have been used. I have distinguished seven general econometrics packages (E), four statistical programming languages (SPL), three econometric time series packages (ETS), two mathematical matrix programming languages (MPL), two third-generation numerical programming languages (NPL), Ox as an econometric matrix programming language (EMPL), BACC as an econometric MCMC (Markov chain Monte Carlo) package (EMC2), and finally, SPSS and Excel.

GAUSS is number one and consistently mentioned over time. In addition, two specific GAUSS applications figure once. Stata and MATLAB have only become attractive for applied econometrics since 2000. SAS and Ox (in later years) appear regularly. RATS has been the most important econometrics package for time series applications. FORTRAN has been consistently more important than C. Other

Table 29.5 Research articles in *JAE* with specific software package per software package per year

Software	*Type*	*Year*														
		95	*96*	*97*	*98*	*99*	*00*	*01*	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*Tot.*
TSP	E	1	1	.	.	.	1	.	1	.	.	.	.	.	.	4
GAUSS	MPL	6	1	2	3	2	3	.	2	4	5	5	8	10	7	58
RATS	ETS	2	.	.	.	.	.	1	.	.	1	.	2	.	.	6
SAS	SPL	2	.	1	3	.	1	1	.	.	1	1	1	2	.	13
Fortran	NPL	2	2	1	2	5	.	2	.	.	1	.	1	.	.	16
C	NPL			1	.	.	.	.	.	.	1	.	.	.	.	2
LIMDEP	E				1	.	1	.	.	.	.	.	1	.	.	3
S-PLUS	SPL				1	.	.	.	.	.	.	.	2	1	.	4
MATLAB	MPL					2	.	.	2	.	.	2	5	5	1	17
SHAZAM	E						1	.	.	.	.	.	.	.	.	1
Stata	SPL						1	.	.	.	7	1	4	2	6	21
SPSS	SSS						1	.	.	.	1	.	1	.	.	3
Ox	EMPL						2	.	2	2	1	2	2	1	.	12
GAUSSX	E							1	.	.	.	.	.	.	.	1
Xplore	SPL							1	.	.	.	.	.	.	.	1
PcGive	E							2	.	.	.	.	.	.	.	2
STAMP	ETS							1	.	.	.	1	.	.	.	2
Ox G@RCH	ETS									1	.	.	1	.	.	2
Excel	SG										1	.	.	.	.	1
Eviews	E											2	1	.	.	3
R	SPL											2	.	1	.	3
BACC	EMC2												1	.	.	1
GAUSS TSM	ETS												1	.	.	1
EasyReg	E													1	.	1
BIOGEME	SPD													1	.	1
Range		5	3	4	5	3	8	7	4	3	9	8	14	9	3	25
Missing		21	28	25	21	21	22	26	22	28	30	32	37	37	8	358
Articles		33	32	30	30	30	31	34	29	34	46	45	61	57	21	513

Notes: software ordered by time of first-mentioned use in *JAE* article (1995.1–2008.4). Counts of research articles using the specific software. Range: number of different softwares mentioned per year; Missing: research articles *not* mentioning specific software; Articles: number of research articles per year. Type descriptions as E: econometrics package; ECS: econometrics cross-section package; EMC2: econometrics Bayesian Markov chain Monte Carlo package; EMPL: econometrics matrix programming language; ETS: econometrics time series package; NPL: numerical programming language (third generation); SG: statistical graphics package; MPL: matrix programming language; SPL: statistical programming language; SSS: statistical package for social sciences; SPD: statistical panel data package.
Sources: data archive *JAE*, http://www.econ.queensu.ca/jae/; JSTOR: http://www.jstor.org; http://www3.interscience.wiley.com, ISSN code JAE: 08837252.

packages appear less than five times. S-PLUS appears four times and R appears three times. SHAZAM, Xplore and PcGive do not reappear after 2001. The other packages cannot be written off as tools for applied econometrics software development. They may well have been used in the preparation of the articles, but the authors did not develop new programs or procedures that they wanted to publish.

In sum, Table 29.5 shows that the programming languages GAUSS, MATLAB, Stata and Ox are the most important tools for applied econometrics software development. GAUSS, MATLAB and Stata are apparently widely available in economics and econometrics departments all over the world.

29.6 High-level programming languages in econometrics

Table 29.6 illustrates some characteristics of the dominating languages GAUSS, MATLAB, Stata and Ox. The table displays very short programs that load a simple dataset from a human-readable ASCII file, estimate regression coefficients using ordinary least squares (OLS) and show these on screen. The examples are adapted from first lessons of course notes available on the internet. The table also includes code for R (and S-PLUS) as this is an increasingly important alternative, as discussed below.

The codes for the matrix programming languages GAUSS and MATLAB are very similar. Beginning is easy, because variables don't have to be declared. The "default type" is a matrix (of double-precision floating-point numbers); and statements end with a semicolon. MATLAB uses square brackets for concatenation, GAUSS has special concatenation operators. GAUSS uses square brackets for indexing, MATLAB indexes with parentheses. Indexing in GAUSS and MATLAB starts at one. Fortunately, arguments in clear function calls are in parentheses. GAUSS provides the least squares solution for the coefficients by the "divide symbol" `/`, which looks a bit weird and mathematically incorrect, but is easy to use; MATLAB uses the more sensible \ operator instead. Neither GAUSS nor MATLAB use a formal print function to show the regression coefficients.

The Stata code is totally different and is reminiscent of many command-line-driven packages in the early 1980s. Stata is, as Baum (2002) put it, "on the middle ground" between econometric packages and matrix languages. The default regression method requires variable names (of columns of dataset, rather than a matrix) to read the data. OLS is the default estimator of the easy-to-read-and-remember regress command, which also adds a constant term and computes standard errors and *p*-values by default. The `matrix` command extracts the regression coefficients in vector format. The mathematical structure is hidden from the programmer. The standard output of regress (not shown in the table) is in the ANOVA format, rather than the standard regression output of econometrics programs. Stata recently introduced the matrix language Mata. So far, Mata has not explicitly been used for *JAE* publications.

Like Stata, R starts with a dataset rather than a matrix. In the R example we assume that the variable names are on the first line of the data file, so that "header=T(rue)". OLS is performed using a challenging call of `lm()` (linear model). This function creates a model object, and the corresponding function `coefficients` extracts the coefficient estimates from the model. The model is specified with the names of the variables and the dataset. The operator ~ separates regressand and regressor, the operator + separates the regressors. The "dollar" operator makes sure we use the coefficients from the linear model object.

Table 29.6 OLS in programming languages for applied econometrics

Software	*Type*	*How to*	*Statements/commands*
GAUSS	MPL	begin, declarations	not necessary
		read data	`load myX[4,3]=yX.asc;`
		select y	`y=myX[.,1] ;`
		add constant to X	`X=ones(rows(myX),1)~myX[.,2:3];`
		get b	`b=y/X;`
		show b	`print b;`
MATLAB	MPL	begin declarations	not necessary
		read data	`myX=load('-ascii','yX.asc');`
		select y	`y=myX(:,1);`
		add constant to X	`X=[ones(size(myX,1), myX) myX(:,2:3)];`
		get b	`b=X\y;`
		show b	`b`
Stata	SPL	begin, declarations	not necessary
		read data	`infile y x1 x2 using yX.mat`
		add constant to X	added by default
		get b	`regress y x1 x2`
		show b	`matrix list e(b)`
R and S-PLUS	SPL	begin, declarations	not necessary
		read data	`myX<-read.table("yX.dat", header=T)`
		add constant to X	added by default
		get b	`b=lm(y1~x1+x2,data=myX)$coefficients`
		show b	`b`
		extra line in `yX.dat`	`y1 x1 x2 \\ variable names`
Ox	EMPL	begin main program	`main()`
			`{`
		declarations	`decl myX, mX, vy, vb, vyhat;`
		read data	`myX=loadmat("yX.mat");`
		select y	`vy=myX[][0];`
		add constant to X	`mX=1 So just a one: mX=1~myX[]1:2];` is the total command;
		get b	`olsc(vy, mX, &vb);`
		fit y: $\widehat{y} = Xb$	`vyhat=mX*vb;`
		show b	`println("b: ",vb);`
		end	`}`
		extra line in `yX.mat`	`4 3 \\ dimensions`

Notes: MPL: matrix programming language; SPL: statistical programming language; EMPL: econometric matrix programming language. Computation 3×1 regression coefficient vector b in linear model $y = X\beta + u$, where X is a 4×3 matrix starting with a column of ones. y and X are read from file. myX: 4×3 matrix with numerical data read from human-readable data file yX.asc. For R we read yX.dat, and for Ox we read yX.mat, starting with an extra line as indicated.

Ox has a syntax similar to C++ (and Java), so all statements are executed in a `main()` function, square bracket pairs index a matrix, and indexing starts at zero. Doornik (2002) discusses differences and similarities of C++ and Ox. Variables have to be declared, but they automatically get a type (double, matrix) when they are assigned. Ox uses the same matrix concatenation operator as GAUSS. A dedicated least squares function `olsc()` is provided, but the (slower and less robust) explicit matrix formula for OLS could have been used instead. The ampersand (reference) is used to deliver the coefficients directly at the memory address of the vector variable `b`. This reduces memory use. I added a statement for computing fitted values $\widehat{y}$ to clarify the matrix programming nature of Ox. I use slightly adapted Hungarian notation, where vector names start with `v` and matrix names with `m`, but this is not required. The Object Oriented (OO) nature of Ox does not appear in this example, but a Modelbase class, derived from a Database class, is standard. Both classes are extendible. Unlike R and S-Plus, Ox does not force the OO features upon novice users.

Table 29.6 shows that Stata code is probably the most easy to use for students and researchers with limited programming experience. GAUSS and MATLAB require knowledge of matrix algebra and numerical programming, but this should not be a problem for econometricians. R is harder to get into as it requires a profound knowledge of statistical terminology and object-oriented programming. Ox is easy to learn if a basic programming language and matrix algebra are known.

Readability and complexity are not the only selection criteria for high-level programming languages. Large models require an extendible modeling language (like Stata and R), and new models require an efficient programming language in which to code new algorithms to estimate and evaluate new model types (like GAUSS and Ox). The programming language should also cater for effective data management, robust optimization methods, state-of-the-art stochastic simulation, and decent, easily adaptable, graphical and textual output facilities.

For maintenance and reproducibility one requires explicit documentation facilities that can transform comments in the code into context-sensitive, clearly structured and indexed help functions for existing and new procedures. One should be able to integrate existing numerical procedures from low-level languages. For business use, the econometric programming language should be applicable as an engine within other software, so that econometric procedures can be called by, and feed results to, programs like Excel, Access, or commercial front-office and back-office applications written in lower-level languages like C++ or Perl. In computing intensive simulation-based methods, one wants to automatically optimize code for parallel computing or for specific hardwares at a low level to increase speed. Multiprocessor computers are now standard. Efficient computing will probably return as a very important issue as electricity prices go through the roof.

The econometric language should have an interface to the Structured Query Language (SQL), a standard language that provides an interface to many relational database systems, and to specific economic, financial, and energy data management software, like FAME (http://www.fame.com), or HAVER (http://www.haver.com). For example, none of the above-mentioned languages can be used to manipulate and select data from the vast datasets on the WRDS (Wharton Research

Data Service), which is now the leading academic archive for econometric time series data. Only SAS, C and FORTRAN can be used on the WRDS server.

The next section discusses other important aspects of a large array of packages and adds a historical perspective, concentrating on the period since 1980. This discussion should help the reader in interpreting the preceding tables on the historical impact of the different products.

29.7 The historical development of econometric software

Over the last 50 years, econometric software development has developed from writing complicated sets of computer specific instructions into coding in structured purpose built programming languages and into interactive GUI-based model development. Increased backward compatibility, cross-platform and cross-operating system applicability of new software, and low cost of maintaining existing software, has increased the lifetime of packages and procedures. Less than 10% of the 77 packages reviewed in the *JAE* have been discontinued.

Econometric software development started around 55 years ago. Renfro (2004b) gives a detailed account of the history of econometric software development in the English-speaking world. Early econometric software development was labor-intensive and served only a few institutions that could manage and pay the substantial capital input for the required programmable computers. Today, this situation has completely changed. Modern econometric software is written by a few individuals and thousands of users perform econometric estimations, forecasts and tests on thousands of machines. The joint cost of standard econometric software and hardware is low and dropping. Thanks to a concentration in hardware and software development, a few developers now serve an entire community. However, expert support and tailored innovative development of user-friendly platform-independent applications is still expensive.

Three structural changes affected econometric software development in a major way in the period 1985–2008. The first was the breakthrough in hardware development: the onset and subsequent quick improvement in computer power and graphical displays of personal computers (PC or Micro computer) during the 1980s opened opportunities for new developers. Many textbook authors wrote their own packages. Cheap standard storing devices for the PC (floppy disks) made distribution (and copying) of econometric software easy. This change is reflected in the large number of different software packages reviewed in 1990, as detailed in the summary statistics of Table 29.4.

The second change was the introduction and standardization of effective GUIs for data analysis, programming and operating systems. Graphing became easy, and it was no longer necessary to memorize a list of basic commands and options.

The third change was the development and widespread use of the internet since the 1990s, more specifically the WWW standard and the later development of powerful search engines like Google. This led to the development of "free" products in mathematics, statistics and computer science. These products have now become powerful, stable and easier to use so that they are effectively applied in econometric

software development and in innovative research in econometrics, leading to *JAE* publications.

The interfaces of many computer programs for data input, programming, text processing, formula and graph editing became more and more similar, due to the worldwide concentration in operating systems and standardization of other scientific applications like LaTeX. Only three operating systems remain important: MS-Windows (Microsoft), Mac OS X (Apple) and Linux (many distributions; Ubuntu/Linux is the most popular version of late). Products developed on one platform can be ported or recompiled on other platforms, although this is far from trivial for most econometricians. Racine (2000) discusses some aspects of Cygwin ports of basic Unix tools to Windows.

Hendry and Doornik (2000) discuss and illustrate the necessary changes of the time series econometric program PcGive during the 1980s and 1990s: from command interaction to menu interaction and IDE (Integrated Development Environment); from text menus to mouse-pointer-driven drop-down menus and dialogs of a WIMP (Windows, Icons, Menus, Pointing) GUI; from black-and-white text graphs to colored bitmap, to high-quality, adjustable, publication-ready figures; from a static manual to a context-sensitive help system, from static presentation to live presentations of simulation exercises; from basically one program code in FORTRAN, and later in C++, to a modular object-oriented architecture allowing user-built extensions with an up-to-date user interface with the same look and feel as the standard applications. PcGive was extended with an independent Windows interface, GiveWin. Jurgen Doornik (1998) also developed the object-oriented econometric matrix programming language Ox, which allowed independent development of new packages and was later integrated within OxMetrics (Doornik, 2007), together with PcGive and the time series programs STAMP and G@RCH. The new interface for OxMetrics was built with the free cross-platform GUI wxWidgets. Other software packages have provided similar updates in order to keep old users and get new customers. For example, Stata introduced object-oriented features and GUI programming in Stata 8 and the matrix language Mata in Stata 9.

In the remaining sub-sections I make a distinction between five admittedly overlapping categories of software: macroeconometric software, (pure) time series econometric software, microeconometric software, statistical software for econometrics and mathematical software for econometrics. I treat each in turn.

29.7.1 Macroeconometric software

Back in the 1960s, Robert Hall laid the foundations of TSP (Time Series Processor) software. At the end of the 1970s, TSP already had many of the characteristics of a modern econometric software package: it read and wrote a variety of data formats, it included a matrix language, it made use of symbolic differentiation, it contained good nonlinear solvers, a powerful optimizer and simulation procedures. In this sense TSP can be considered as the most original econometric software on the market.

In the PC era of the 1980s, TSP was split into two separate programs, Micro-TSP, headed by David Lilien, and PC-TSP, headed by Bronwyn Hall. Micro-TSP later became the Windows program Eviews, Econometric Views, whereas PC-TSP is now

simply called TSP (see Hall and Cummins, 2005; Eviews, 2004). TSP retained the numerical and algebraic programming features. Eviews later introduced its own object-oriented programming language. One of the main attractions of Micro-TSP and Eviews was the timely interface for the first univariate econometric time series models: ARCH (autoregressive conditional heteroskedasticity), and GARCH (generalized autoregressive conditional heteroskedasticity). This user-friendly implementation of GARCH models was developed in close cooperation with Robert Engle, the father of ARCH. A special issue of the *JAE* (Franses and McAleer, 2002) was published to celebrate Engle's seminal ARCH article (Engle, 1982).

Ken White started the package SHAZAM at Wisconsin and is now at UBC in Vancouver, where SHAZAM is updated by a small team. Whistler *et al*. (2004) describe the latest version. Nobel Laureate Lawrence Klein founded the Wharton Econometric Forecasting Association (WEFA) at the University of Pennsylvania: WEFA is now part of Global Insight and markets the econometric software AREMOS, which was strongly influenced by Klein's modeling methodology. AREMOS is not frequently updated, but is still being used.

In the UK, at the Department of Applied Economics of the University of Cambridge, Hashem and Bahram Pesaran used their expertise in econometric estimation and testing for the development of Data-FIT, later called Microfit, for the PC. At the Department of Statistics at the London School of Economics, econometric software development was inspired by the hands-on tradition of Denis Sargan. David Hendry, a student and later a colleague of Sargan, developed the programs AUTOREG and GIVE. In Oxford, Hendry developed PcGive (generalized instrumental variable estimator) and PCFIML (full information maximum likelihood) on the IBM PC. Jurgen Doornik modernized and extended PcGive, as explained in the first part of this section.

More recently, Michel Juillard developed a stand-alone version of Dynare, previously only available for GAUSS and MATLAB. Dynare implements modern, small-scale, but very computer-intensive DSGE (dynamic stochastic general equilibrium) modeling. These highly nonlinear structural models are difficult to solve and estimate and require Bayesian econometric techniques to do inference. DSGE models are introduced and used at central banks throughout the world.

On the educational side of the spectrum, Gretl, by Allin Cottrell and Ricardo Lucchetti, is an international GNU (GNU's Not Unix: a free, open source Unix-like operating system) econometrics program, with menus in French, Italian, Spanish, Polish and German as well as English. It is based on code for a textbook by Ramu Ramanathan. As in other packages mentioned in this section, the traditional macroeconometric procedures are being supplemented with microeconometric functions, DPD and procedures in particular.

29.7.2 Time series econometric software

One can no longer imagine applied econometrics without implementations of ARMA (autoregressive moving average), VAR (vector autoregression) and GARCH (generalized autoregressive conditional heteroskedasticity) time series models. The Box–Jenkins methodology is a standard procedure in many fields of science. Under the direction of George Box, the first special software for ARMA analysis was

written by David Pack, and David Reilly turned this into AutoBox. He also coded the Multivariate Time Series (VARMA) program MTS. AutoBox and MTS are now marketed by Reilly's company AFS.

Chris Sims developed SPECTRE at the end of the 1970s. This was one of the first econometric programs offering spectral analysis. Subsequently, Sims' 1980 VAR modeling methodology was made available in RATS (Regression Analysis of Time Series) by Thomas Doan (see Doan, 2004). CATS in RATS was (shortly after PcGive) one of the first widely available software packages for Søren Johansen's likelihood-based analysis of the concept of cointegration, eventually published as Johansen (1991).

The Census Bureau in Washington, DC, produced the first reliable software for seasonal adjustment of economic time series, Census X-11, implementing a methodology (updated to X-12 ARIMA) that is now an international standard and available in most time series econometrics softwares (see Ladiray and Quenneville, 2001).

At the London School of Economics, Andrew Harvey initiated the development of STAMP, for structural time series modeling, implementing an econometric methodology which serves as an alternative both to Box–Jenkins forecasting models and to Census X-11 seasonal adjustment. Siem Jan Koopman now develops the (multivariate) STAMP software at the VU University in Amsterdam (see Koopman *et al.*, 2007).

At the Bank of Spain, Victor Gómez and Augusín Maravall developed the second influential alternative software for seasonal adjustment: TRAMO/SEATS. Their procedures are also available in many time series programs.

Herman Bierens is the independent author of EasyReg International, a free software package (developed in visual Basic), primarily developed for econometrics education but equipped with many advanced procedures in Bierens' area of research (nonparametric methods, first for time series and later for cross-sections), and therefore also featuring in a recent *JAE* research article.

29.7.3 Microeconometric software

This sub-section is short as there is only one surviving dedicated econometric software for nonstandard econometric models for cross-section data, LIMDEP. Microeconometricians have mainly been using lower-level programming languages and statistical packages, discussed below.

William Greene based the first versions of LIMDEP, for LIMited DEPendent variable econometrics on code for multinomial logit models by Marc Nerlove and James Press at the University of Wisconsin. Greene (2007) describes the current features of the program. Previous versions of Greene's influential and popular textbook, now in its sixth edition (Greene, 2008), contained a special student edition, EA/LIMDEP, of the software. Over the last 20 years, most standard econometric procedures (time series and panel data) have been added. Greene also authored the packages ET and NLOGIT. Greene is now at New York University.

29.7.4 Statistical software for econometrics

In the last 25 years, several statistical programs have become more geared towards econometrics and subsequently widely used by econometricians. The general

statistics package SAS (SAS, 2004) has a long tradition (starting in the 1960s) of implementing macroeconometric and microeconometric procedures for large datasets. In academic research and education in econometrics, SAS/ETS has lost ground from its strong position at the end of the 1980s, though its econometrics features are still being developed, recently in state space procedures, in generalized maximum entropy estimation and in automatic model selection for forecasting. Of course, SAS is widely used in official institutions and in business applications, but few modern econometrics textbooks continue to use SAS examples.

SPSS, dating back to the 1970s, is not particularly suited for econometrics, but it is used for handling large and complicated datasets. Interesting third party packages for SPSS exist, like Jeroen Vermunt's LATENT GOLD for Latent Class models and event history modeling in marketing and social sciences. It is also suitable for modern microeconometrics problems (as other packages which were primarily developed for the social sciences).

The beginning of the PC era saw the birth of the "Data Analysis and Statistical Software" Stata. Stata, by William Gould, was not an instant success among econometricians, whereas it was for statistics in medicine. At first, it did not have extensive programming facilities and specialized in applications for survival data (see Goldstein *et al.*, 1989). It was not suited for dynamic econometric modeling. Peterson (1991) correctly predicted: "this shortcoming could be mitigated substantially in future versions." Later Stata introduced more programming tools and eventually a matrix language and it was completed with more and more econometric models. Stata's data management features made it well suited for the econometric analysis of complicated panel data like event histories. Time series procedures have been added. Stata is now a popular package in applied economics and econometrics and a large number of introductory econometric textbooks present examples using Stata. Kit Baum maintains a large Statistical Software Components (SSC) archive within RePEc with over 1,000 free open-source Stata procedures and programs for statistics, economics and econometrics. Baum (2006) also wrote an applied econometric textbook for Stata.

S-PLUS and corresponding packages cater for financial econometrics and operations research: financial time series analysis, modeling credit risks and optimizing asset allocation. S-PLUS, originally a product of StatSci, founded by R. Douglas Martin in Seattle, Washington, is a commercial version of the object oriented statistical programming language S, which Martin learned at Bell Laboratories in Murray Hill, New Jersey, now Lucent Technologies. The software was primarily developed for statistical data analysis of many types (see Venables and Ripley, 2002), with excellent graphs. Martin added robust estimation procedures, inspired by John Tukey, inventor of the term "bit," FFT (fast Fourier transform) and EDA (exploratory data analysis). The current owner of S-PLUS, Insightful, focuses on data mining and risk management. Zivot and Wang (2005), also in Seattle, Washington, developed the S-PLUS FinMetrics software for financial econometric time series analysis. The package also includes financial engineering procedures developed by Carmona (2004) and efficient Kalman filter state-space procedures by Siem Jan Koopman (see Koopman *et al.*, 1999). The popular financial time series textbook by Tsay (2005) makes intensive use of S-PLUS FinMetrics.

The popularity of the internet motivated the start of the statistical software Xplore in the later 1990s. There was great optimism about online cooperative development and use of software for advanced statistical computations. Härdle and Horowitz (2000) envisaged that the establishment of well-documented method archives, central common platform independent compilers and new web user interfaces would give easy access to the most advanced nonparametric methods. One of their suggested method and data technology centers was created and a (Java-based) web interface, Xplore Quantlet Client (XQC), was realized. Online electronic books with econometric and financial time series applications were provided for educational purposes, but online web-based econometric computing has not caught on yet. Xplore is now freely downloadable from http://www.xplore-stat.de.

In recent years, Michel Bierlaire has developed BIOGEME, an open source package (in C++ and Python) for modern random coefficient (or mixed) discrete choice modeling; he cooperates with Moshe Ben-Akiva and Nobel Laureate Daniel McFadden. Train (2003) treats this important topic in a textbook.

Young and old econometricians are switching from S-PLUS and other packages to the freely-available statistical system R, an open source statistical system that was initiated by statisticians Ross Ihaka and Robert Gentleman from Auckland, New Zealand. R has the S syntax (and is also known as GNU S). Graphs in R are provided via Gnuplot (which is also used in SHAZAM, discussed above, and TSMod, discussed below). R is part of the free GNU operating system (OS) and is part of all standard installations of this OS and therefore of many Linux installations. Officially, Gnuplot does not belong to GNU. Over 1,200 packages are available for R at the CRAN (Comprehensive R Archive Network) at http://www.r-project.org. Cribari-Neto and Zarkos (1999) reviewed an early version of R from an econometric research point of view, and Racine and Hyndman (2002) took a teaching perspective. Shumway and Stoffer (2006) provided up-to-date R code for their time series textbook. Rossi *et al*. (2005) developed an R package (`bayesm`) for their marketing statistics textbook. Li and Racine (2007) wrote the `np` package for a text on nonparametric econometrics. Modern statistical methods are often made available in R. For example, Hastie *et al*. (2001) discuss their well-known automatic model selection methods for regression and classification implemented in R.

Most R developers seem to work under the Linux OS and choose short, Unix-style package names. Many R packages are not difficult to use under Windows and Mac OS. Developing R packages under MS Windows has not been too easy, though as Rossi (2006) reports in his 15-page tutorial on this topic: "There is a sense in which the Windows R environment is a house of cards that must be carefully assembled or it won't work!" A specialized archive of R for econometrics does not exist. A comprehensive package for financial engineering, http://www.rmetrics.org, which encompasses many econometric time series functions, has been built by Diethelm Würtz at the ETH in Zürich.

29.7.5 Mathematical software for econometrics

The beginning of the PC era also witnessed the start of the matrix programming language GAUSS developed by Lee Edlefsen and Sam Jones in Washington State.

GAUSS did not offer a new econometric methodology, but it did have a very appealing combination of price and features for econometricians and economists (see GAUSS, 2005). It soon became popular and has remained popular ever since. It has a simple language with short matrix expressions (as illustrated in Table 29.6), decent graphs, fast numerical algorithms, tools to handle large datasets with limited memory, and a wide range of free and powerful packages implementing econometric applications for cross-section models and time series. Ron Schoenberg (1997), affiliated with Washington University, developed early procedures for constrained maximum likelihood for GAUSS, which found widespread application in the estimation of GARCH models. Schoenberg also wrote FANPAC, a financial time series analysis package with early applications of multivariate GARCH models.

The matrix programming language and signal processing tools of MATLAB (MATLAB, 2004), of the Mathworks, founded by Clive Moler, are used by many econometricians to implement model solvers and estimation methods. Econometricians use the free and comprehensive archive of econometric tools, at http://spatial-econometrics.com, administered by James P. LeSage at the university of Toledo, Ohio. Although the archive is set up for spatial econometrics procedures (LeSage and Pace, 2004), it contains many "estimation functions that provide printed and graphical output similar to that found in RATS, SAS or TSP."

Table 29.3 lists seven other mathematical programming languages which have not been used for *JAE* research articles so far, but code for these languages is provided by prominent econometricians. For example, Scilab code can be obtained for Dynare. Christopher Sims provides recent Octave code for solving rational expectations models on his own (Ubuntu/Linux) web server: http://sims.princeton.edu. Octave is a free alternative for MATLAB, but Sims points out that procedures with the same names can have different effects in the two languages.

Computer algebra packages like Mathematica and Maple are now also used for fast numerical computations, and are therefore more suited for applied econometrics, but they haven't had a big impact yet. The recently developed package MathStatica for Mathematica, by Colin Rose and Murray Smith, can save applied econometricians work in the analytical derivations of complicated likelihoods.

29.8 Simultaneous use of different software

As the tables and the discussion in the previous sections illustrate, many econometric techniques can now be implemented using existing mathematical and statistical software packages. No single software can serve all purposes, which explains why more and more packages coexist and why many researchers use several products next to each other.

Thanks to the search engine Google and free specific internet aggregators of economic and econometric research (papers, articles, books, citations, data and software) like RePEc, it is now easy to find properly documented econometric source code written for one of the main econometric softwares on the web. However, it is still difficult to assess the quality of this code if one does have access to the software for which it was originally developed. As most of these codes for academic

research papers are available free of charge, authors cannot be expected to set up a helpdesk, and one has to resort to mailing lists and internet forums, which also may be unreliable. Unsurprisingly, given the background of most econometricians, robust, high-quality econometric procedures seldom come free.

The modular structure of econometric and statistical software makes it possible to use codes outside their original environment. This helps the reproducibility required in academic econometrics. For example, Laurent and Urbain (2003) provide an interface called `M@ximize` for Ox, based on OxGauss, so that the wide range of econometric GAUSS programs available on the internet can be run without a licence for GAUSS or constrained maximum likelihood for GAUSS. Markus Krätzig developed a GUI for econometric modeling, JStatCom (see Kräetzig, 2006), which he built on top of GAUSS code and the GRTE (Gauss run time engine) to create JMulti as a stand-alone program. JStatCom can also be used in combination with MATLAB and Ox. John Breslaw of Econotron software introduced Symbolic Tools, which extends GAUSS and the GRTE with the infinite precision computer algebra of Maple. Cameron Rookley wrote the free GTOML (GAUSStoMATLAB) scripts which translate GAUSS code into MATLAB. This requires the free powerful OO programming language Perl (see http://www.perl.com, and http://www.cameronrookley.com).

Diethelm Würtz, author of Rmetrics, provided an interface in R for the G@RCH package that Laurent and Peters (2005) developed for Ox, but this still requires the availability of Ox. Many statistical packages have been ported to R; for example, BRugs, which embeds OpenBUGS in R. Robert Henson (2004) introduced a MATLAB R-link with functions for calling R from within MATLAB; Bengtsson (2005) increased the communication possibilities between MATLAB and R.

Integrating codes from different applications can save time, but has its dangers. Evaluation and improvement of existing implementations for nontrivial procedures should be a constant concern (see, for example, the discussion of numerical precision of econometric packages by McCullough and Vinod, 1999, which generated a series of changes in testing procedures). Note also the evaluation of random number generators (RNGs) as in McCullough (2006) and Doornik (2006). Reliability of RNGs is now extremely important as simulation-based inference starts to dominate both macroeconometrics and microeconometrics. Even if the RNG is right, and expert econometric knowledge is available, there is plenty of room for undetected mistakes. The home page of the BUGS project (Bayesian inference using Gibbs sampling) phrases this as follows: "Independent corroboration of MCMC results is always valuable!"; "MCMC is inherently less robust than analytic statistical methods. There is no in-built protection against misuse." Even before econometric modeling starts, one should apply Hendry's (1980) "three golden rules of econometrics: test, test and test" to the freshly developed or imported software.

29.9 New econometric modeling features and conclusions

Pagan and Wickens (1989) surveyed applied econometric methods 20 years ago. Four estimation methods were discussed: maximum likelihood, GMM (generalized

method of moments), M-estimators and nonparametric estimation, and different types of inference: frequentist and Bayesian, large sample asymptotics and the bootstrap for tests in small samples. They concluded: "when it comes to an area such as econometrics. Gone are the days when a single individual could have a detailed knowledge of all divisions of the subject. Just twenty years ago this might have been possible"; "the years since then have witnessed a fragmentation of econometrics. The biggest division has been between micro and macro econometrics." As indicated in section 29.2, many new data types, estimators, inference methods and diagnostic procedures have been analyzed by applied econometricians since 1989. The fragmentation now also applies to software development, with dozens of procedures published on the internet for the same purpose.

Although applied nonparametric econometrics has been on the rise, model-based econometrics still dominates the field of applied econometrics. A key aspect that distinguishes model-based econometric software is the standard availability of features for the interactive modeling cycle: models not only are easily specified and estimated, but diagnostic tests, easy respecification, and re-estimation facilities are provided in order to make the interpretation of parameter estimates and forecasts as credible as possible. Today, this requires a graphical (WIMP) interface that is sufficiently intuitive and easy to learn and remember for new users.

This recursive modeling is especially relevant for the econometric analysis of time series, where new observations become available in a natural order, with associated testing possibilities and possible adaptations of existing models. In the context of dynamic linear regression models, PcGive was the first program to cater for the influential general-to-specific methodology of econometric model selection. A "Progress" menu in PcGive simplifies the interactive model selection process. Although this feature *per se* has not been copied in other packages, a wide range of standard specification tests and diagnostics for estimated models has now become a crucial ingredient of every econometric software.

The model selection process can be automated. Successful automated model selection has long been available for pure Box–Jenkins time series modeling for forecasting in the AutoBox software by David Reilly and in the Census X-11-ARIMA program for seasonal adjustment of the US Census. Automated linear dynamic model selection for economic analysis, based on a wide range of robust diagnostic tests and multiple-path general-to-specific modeling, is available in the PcGive procedure Autometrics (Doornik, 2008).

However, automated model selection methods, even if they encompass generalized linear models of "Statistical Learning," as in Hastie *et al*. (2003), or fractional instead of zero-one model weights of Bayesian model averaging (BMA), as in Raftery *et al*. (1997), still require a "most general" adequately specified model, for which extensive tests should be available.

Stochastic simulation and bootstrap analysis of econometric models should be available as a matter of course, both for the interpretation of nonlinear models and for associated statistical inference. James Davidson's (nonlinear) time series modeling package TSMod, reviewed by Fuertes *et al*. (2005), has this feature for all models in the package: "Bootstrap *p*-values for diagnostic and significance tests, using the

simulation module to generate bootstrap draws." If the inference is simulation-based, one also needs diagnostics on the efficacy and reliability of the associated simulation methods.

User interfaces will have to be updated. Following Google and Gretl, users will expect econometric software to deal with labels and numbers in their native language and application menus to use their own character sets. The graphical interface will also need reconstruction as customers adapt to modern graphical interfaces. New interfaces will help to make better use of the many options that programs and procedures have, both on the user's own computer and on internet archives. Many procedures are ineffective because they are hard to find in the current menu structures. Based on a user history, the menus will "automatically" select the best options for the user.

The market for specific econometric software is too small for one program to keep up with all recent scientific developments in econometrics, mathematics and statistics, to keep advanced knowledgeable customers interested in buying updates, and to implement lessons from human–computer interaction (HCI) research to keep attracting new customers.

The presence of trends implies some predictability of future developments. The pattern that has emerged in the last 25 years does not make it likely that new, fully-fledged, dedicated econometric software packages with high academic standards are going to be developed. Academic returns on high-quality, robust, versatile, and well-documented and supported econometric software development are low. Changing citation practices for software use, as exemplified by the *JAE* data and code archive, may increase these returns in the years ahead.

In this chapter I have discussed over 20 years of changing software use and software development for innovative applied econometrics. An increasing range of software has became relevant in this period. I also classified this large collection of programs and assessed the continuity of their use. Finally, I pointed out new direction for econometric modeling software development.

Acknowledgments

I would like to thank Christopher Baum, Jurgen Doornik, Bill Rising, Ronald Schoenberg and Christian Kleiber for helpful comments on this chapter. All errors are my own.

Note

1. Software: I used MS Excel 2000, Windows XP, 5.1 Service Pack 2, MikTeX 2.4, OxEdit 5.0, Firefox 3, OxMetrics 5.0, GAUSS 7, R 2.7, Google, Google Scholar, Google Books, JSTOR and Wiley Interscience to prepare this chapter.

References

Altman, M. and M. McDonald (2001) Choosing reliable statistical software. *PS: Political Science & Politics* **34**, 681–7.

Anderson, R., W. Greene, B.D. McCullough and H.D. Vinod (2008) The role of data/code archives in the future of economic research. *Journal of Economic Methodology* **15**(1), 99–119.

Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**(2), 277–97.

Baltagi, B.H. (1999) Applied econometrics rankings: 1989–1995. *Journal of Applied Econometrics* **14**, 423–41.

Baum, C. (2002) Facilitating applied economic research with Stata. In S.B. Nielsen (ed.), *Programming Languages and Systems in Computational Economics and Finance*, pp. 173–9. Dordrecht: Kluwer Academic Publishers.

Baum, C.F. (2006) *An Introduction to Modern Econometrics Using Stata*. College Station, Texas: Stata Press.

Bauwens, L., A. Escribano and M. Lubrano (2007) The econometrics of industrial organization. *Journal of Applied Econometrics* **22**(7), 1153–6.

Bengtsson, H. (2005) R.matlab – local and remote Matlab connectivity in R. Preprint in Mathematical Sciences (manuscript in progress), Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden.

Brillet, J.-L. (1989) Econometric modelling on microcomputers: a review of major software packages. *Journal of Applied Econometrics* **4**, 73–92.

Brooks, C., S.P. Burke and G. Persand (2003) Multivariate GARCH models: software choice and estimation issues. *Journal of Applied Econometrics* **18**, 725–34.

Brown, B.W., A. Monfort and H.K. van Dijk (1993) Introduction to special issue on econometric inference using simulation techniques. *Journal of Applied Econometrics* **8**, S1–3.

Buckheit, J. and D.L. Donoho (1995) WaveLab and reproducible research. In A. Antoniadis and G. Oppenheim (eds.), *Wavelets in Statistics*. New York: Springer-Verlag.

Carmona, R.A. (2004) *Statistical Analysis of Financial Data in S-Plus*. New York: Springer-Verlag.

Christensen, B.J., N.D. Gupta and J. Rust (2004) Introduction: special issue on the econometrics of social insurance. *Journal of Applied Econometrics* **19**(6), 647–8.

Cribari-Neto, F. (1997) Econometric programming environments: GAUSS, Ox and S-PLUS. *Journal of Applied Econometrics* **12**, 77–89.

Cribari-Neto, F. and S. Zarkos (1999) R: yet another econometric programming environment. *Journal of Applied Econometrics* **14**, 319–29.

Diebold, F.X. and M.W. Watson (1996) Introduction: econometric forecasting. *Journal of Applied Econometrics* **11**, 453–4.

Doan, T.A. (2004) *User's Manual RATS, Version 5*. Evanston, Illin., http://www.estima.com: Estima.

Doornik, J.A. (1998) *Object-oriented Matrix Programming using Ox*. London: Timberlake Consultants Ltd, http://www.oxmetrics.com.

Doornik, J.A. (2002) Object oriented programming in econometrics and statistics using Ox: a comparison with C++, Java and C#. In S.B. Nielsen (ed.), *Programming Languages and Systems in Computational Economics and Finance*, pp. 173–9. Dordrecht: Kluwer Academic Publishers.

Doornik, J.A. (2006) The role of simulation in econometrics. In T.C. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics. Volume 1: Econometric Theory*, Ch. 22, pp. 787–811. Basingstoke: Palgrave Macmillan.

Doornik, J.A. (2007) *An Introduction to OxMetrics 5*. London: Timberlake Consultants Press.

Doornik, J.A. (2008) Autometrics. In J.L. Castle and N. Shephard (eds.), *Festschrift in Honour of David F. Hendry.* Oxford: Oxford University Press.

Durlauf, S.N. and R.A. Moffitt (2003) Introduction: special issue on empirical analysis of social interactions. *Journal of Applied Econometrics* **18**(5), 499.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

Ericsson, N.R. (1988) A review of Data-FIT: an interactive econometric modelling package for IBM-compatible PCs. *Journal of Applied Econometrics* **3**, 319–32.

Eviews (2004) *Eviews 5 User's Guide*. Irvine, Calif., http://www.eviews.com: Quantitative Micro Software.

Franses, P.H. and M. McAleer (2002) Financial volatility: an introduction. *Journal of Applied Econometrics* **17**, 419–24.
Franses, P.H., H.K. van Dijk and D. van Dijk (2005) On the dynamics of business cycle analysis: editors' introduction. *Journal of Applied Econometrics* **20**(2), 147–50.
Fuertes, A.-M., M. Izzeldin and A. Murphy (2005) A guided tour of TSMod 4.03. *Journal of Applied Econometrics* **20**(5), 691–8.
GAUSS (2005) *GAUSS 7.0 User's Guide*. Maple Valley, Wash. Aptech systems Inc. /Trafford Publishing, http://www.aptech.com.
Geweke, J., J. Rust and H.K. Van Dijk (2000) Introduction: inference and decision making. *Journal of Applied Econometrics* **15**, 545–6.
Goldstein, R., J. Anderson, A. Ash, B. Craig, D. Harrington and M. Pagano (1989) Survival analysis software on MS/PC-DOS computers. *Journal of Applied Econometrics* **4**, 393–414.
Greene, W.H. (2007) *LIMDEP 9.0 Econometric Modeling Guide*. New York, http://www.limdep.com: Econometric Software Inc.
Greene, W.H. (2008) *Econometric Analysis* (sixth edition). Upper Saddle River, NJ: Prentice-Hall.
Hall, B.H. and C. Cummins (2005) *TSP 5.0 User's Guide*. Palo Alto, Calif., http://www.tspintl.com: TSP International.
Härdle, W. and J. Horowitz (2000) Internet-based econometric computing. *Journal of Econometrics* **95**, 333–45.
Hastie, T., R. Tibshirani and J. Friedman (2001) *The Elements of Statistical Learning* (second edition). New York: Springer.
Hendry, D.F. (1980) Econometrics – alchemy or science? *Economica* **47**, 386–406.
Hendry, D.F. and J.A. Doornik (2000) The impact of computational tools on time-series econometrics. In T. Coppock (ed.), *Information Technology and Scholarship Applications in the Humanities and Social Sciences*, pp. 257–69. Oxford: Oxford University Press/British Academy.
Hendry, D.F. and M.H. Pesaran (2001) Introduction to special issue in memory of John Denis Sargan, 1924–1996: studies in empirical macroeconometrics. *Journal of Applied Econometrics* **16**, 197–202.
Henson, R. (2004) *MATLAB R-Link*. In MATLAB Central, http://www.mathworks.com/matlabcentral.
Horowitz, J., M.-J. Lee, B. Melenberg and A. van Soest (1998) Introduction: application of semiparametric methods for micro-data. *Journal of Applied Econometrics* **13**, 431–3.
Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551–80.
Juillard, M. (1996) Dynare: a program for the resolution and simulation of dynamic models with forward variables through the use of a relaxation algorithm. CEPREMAP Working Papers (Couverture Orange) 9602, CEPREMAP, Paris.
Kapteyn, A., N. Kiefer and J. Rust (1995) Introduction: the microeconometrics of dynamic decision Making. *Journal of Applied Econometrics* **10**, S1–7.
Koenker, R. (2005) *Quantile Regression*. Econometric Society Monographs, Cambridge University Press.
Koenker, R. and A. Zeileis (2007) Reproducible econometric research (a critical review of the state of the art). Technical Report 60, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien.
Koopman, S.J., A.C. Harvey, J.A. Doornik and N. Shephard (2007) *STAMP 8, Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants Press.
Koopman, S.J., N. Shephard and J.A. Doornik (1999) Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal* **2**, 107–60, http://www.ssfpack.com.
Krätzig, M. (2006) A software framework for data analysis. *Computational Statistics and Data Analysis* **53**, 618–34.

Ladiray, D. and B. Quenneville (2001) *Seasonal Adjustment with the X-11 Method*. New York: Springer-Verlag.
Laurent, S. and J.-P. Peters (2005) *G@RCH 4.0, Estimating and Forecasting ARCH Models*. London: Timberlake Consultants Press.
Laurent, S. and J.-P. Urbain (2003) Bridging the gap between Ox and Gauss using OxGauss. Paper presented at the first Oxmetrics user conference, London. Technical Report, Center for Econometrics and Operations Research, Louvain-la-Neuve, Belgium.
LeSage, J.P. and R. Kelley Pace (eds.) (2004) *Advances in Econometrics, Volume 18: Spatial and Spatiotemporal Econometrics*. Oxford: Elsevier.
Li, Q. and J.S. Racine (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.
Lunn, D.J., A. Thomas, N. Best and D. Spiegelhalter (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–37.
Lütkepohl, H. and M. Krätzig (2004) The software JMulti. In H. Lütkepohl and M. Krätzig (eds.), *Applied Time Series Econometrics*, Ch. 8. Cambridge: Cambridge University Press, http://www.jmulti.de.
Magnus, J.R. and M.S. Morgan (1997) Design of the experiment. *Journal of Applied Econometrics* **12**, 459–65.
MATLAB (2004) *MATLAB 7.1*. Boston, Mass.: The Mathworks Inc., http://www.mathworks.com.
McAleer, M. (1989) Introduction to special issue on Topics in Applied Econometrics. *Journal of Applied Econometrics* **4**, S1–3.
McCullough, B.D. (1999) Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP. *Journal of Applied Econometrics* **14**, 191–202.
McCullough, B.D. (2006) A review of TESTU01. *Journal of Applied Econometrics* **21**(5), 677–82.
McCullough, B.D. and H.D. Vinod (1999) The numerical reliability of econometric software. *Journal of Economic Literature* **37**, 633–65.
Meddahi, N. (2002) A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics* **17**, 479–508.
Ooms, M. and J.A. Doornik (2006) Econometric software development: past, present and future. *Statistica Neerlandica* **60**, 206–24.
Pagan, A. (1994) Introduction: calibration and econometric research: an overview. *Journal of Applied Econometrics* **9**, S1–10.
Pagan, A.R. and M.R. Wickens (1989) A survey of some recent econometric methods. *Economic Journal* **99**, 962–1025.
Pesaran, M.H. and M.P. Potter (1992) Nonlinear dynamics and econometrics: an introduction. *Journal of Applied Econometrics* **7**, S1–7.
Peterson, S.P. (1991) A review of Stata 2.1. *Journal of Applied Econometrics* **6**, 207–12.
Racine, J. (2000) Review: the Cygwin tools: a GNU toolkit for Windows. *Journal of Applied Econometrics* **15**, 331–41.
Racine, J.S. and R. Hyndman (2002) Using R to teach econometrics. *Journal of Applied Econometrics* **17**, 175–89.
Raftery, A.E., D. Madigan and J.A. Hoeting (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–91.
Renfro, C.G. (ed.) (2004a) *Computational Econometrics: Its Impact on the Development of Quantitative Economics*. Amsterdam: IOS Press, http://www.iospress.com.
Renfro, C.G. (2004b) Econometric software: the first fifty years in perspective. *Journal of Economic and Social Measurement* **29**, 9–107.
Rossi, P.E. (2006) Making R packages under Windows, a tutorial. Technical Report, Graduate School of Business, University of Chicago, http://faculty.chicagogsb.edu/peter.rossi/.
Rossi, P.E., G. Allenby and R. McCulloch (2005) *Bayesian Statistics and Marketing*. New York: John Wiley.

SAS (2004) *SAS/ETS (Econometrics and Time Series 9.1 User's Guide*) Cary, NC: SAS Publishing, http://www.sas.com.

Schoenberg, R. (1997) Constrained maximum likelihood. *Computational Economics* **10**, 251–66.

Shumway, R.H. and D.S. Stoffer (2006) *Time Series Analysis and its Applications: With R Examples* (second edition). Springer.

Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* **48**(1), 1–48.

Stone, R. (1986) Nobel Memorial Lecture 1984: The accounts of society. *Journal of Applied Econometrics* **1**, 5–28.

Train, K.E. (2003) *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Trivedi, P.K. (1997) Introduction: econometric models of event counts. *Journal of Applied Econometrics* **12**, 199–201.

Tsay, R.S. (2005) *Analysis of Financial Time Series* (second edition). New York: John Wiley & Sons.

Venables, W. and B. Ripley (2002) *Modern Applied Statistics with S* (fourth edition). New York: Springer-Verlag.

Whistler, D.K., K.J. White, S.D. Wong and D. Bates (2004) *SHAZAM Version 10 User's Reference Manual*. Vancouver: Northwest Econometrics, http://www.econometrics.com.

Zivot, E. and J. Wang (2005) *Modeling Financial Time Series with S-PLUS* (second edition). New York: Springer-Verlag.