
1 Introduction

1.1 About Econometrics

Economists are frequently interested in relationships between different quantities, for example between individual wages and the level of schooling. The most important job of econometrics is to quantify these relationships on the basis of available data and using statistical techniques, and to interpret, use or exploit the resulting outcomes appropriately. Consequently, econometrics is the interaction of economic theory, observed data and statistical methods. It is the interaction of these three that makes econometrics interesting, challenging and, perhaps, difficult. In the words of a seminar speaker, several years ago: ‘Econometrics is much easier without data’.

Traditionally econometrics has focused upon aggregate economic relationships. Macro-economic models consisting of several up to many hundreds of equations were specified, estimated and used for policy evaluation and forecasting. The recent theoretical developments in this area, most importantly the concept of cointegration, have generated increased attention to the modelling of macro-economic relationships and their dynamics, although typically focusing on particular aspects of the economy. Since the 1970s econometric methods have increasingly been employed in micro-economic models describing individual, household or firm behaviour, stimulated by the development of appropriate econometric models and estimators that take into account problems like discrete dependent variables and sample selection, by the availability of large survey data sets and by the increasing computational possibilities. More recently, the empirical analysis of financial markets has required and stimulated many theoretical developments in econometrics. Currently econometrics plays a major role in empirical work in all fields of economics, almost without exception, and in most cases it is no longer sufficient to be able to run a few regressions and interpret the results. As a result, introductory econometrics textbooks usually provide insufficient coverage for applied researchers. On the other hand, the more advanced econometrics textbooks are often too technical or too detailed for the average economist to grasp the essential ideas and to extract the information that is needed. Thus there is a need for an accessible textbook that discusses the recent and relatively more advanced developments.

The relationships that economists are interested in are formally specified in mathematical terms, which lead to econometric or statistical models. In such models there is room for deviations from the strict theoretical relationships owing to, for example, measurement errors, unpredictable behaviour, optimization errors or unexpected events. Broadly, econometric models can be classified in a number of categories.

A first class of models describes relationships between present and past. For example, how does the short-term interest rate depend on its own history? This type of model, typically referred to as a time series model, usually lacks any economic theory and is mainly built to get forecasts for future values and the corresponding uncertainty or volatility.

A second type of model considers relationships between economic quantities over a certain time period. These relationships give us information on how (aggregate) economic quantities fluctuate over time in relation to other quantities. For example, what happens to the long-term interest rate if the monetary authority adjusts the short-term one? These models often give insight into the economic processes that are operating.

Thirdly, there are models that describe relationships between different variables measured at a given point in time for different units (for example households or firms). Most of the time, this type of relationship is meant to explain why these units are different or behave differently. For example, one can analyse to what extent differences in household savings can be attributed to differences in household income. Under particular conditions, these cross-sectional relationships can be used to analyse 'what if' questions. For example, how much more would a given household, or the average household, save if income were to increase by 1%?

Finally, one can consider relationships between different variables measured for different units over a longer time span (at least two periods). These relationships simultaneously describe differences between different individuals (why does person 1 save much more than person 2?), and differences in behaviour of a given individual over time (why does person 1 save more in 1992 than in 1990?). This type of model usually requires panel data, repeated observations over the same units. They are ideally suited for analysing policy changes on an individual level, provided that it can be assumed that the structure of the model is constant into the (near) future.

The job of econometrics is to specify and quantify these relationships. That is, econometricians formulate a statistical model, usually based on economic theory, confront it with the data and try to come up with a specification that meets the required goals. The unknown elements in the specification, the parameters, are *estimated* from a sample of available data. Another job of the econometrician is to judge whether the resulting model is 'appropriate'. That is, to check whether the assumptions made to motivate the estimators (and their properties) are correct, and to check whether the model can be used for its intended purpose. For example, can it be used for prediction or analysing policy changes? Often, economic theory implies that certain restrictions apply to the model that is estimated. For example, the efficient market hypothesis implies that stock market returns are not predictable from their own past. An important goal of econometrics is to formulate such hypotheses in terms of the parameters in the model and to test their validity.

The number of econometric techniques that can be used is numerous and their validity often depends crucially upon the validity of the underlying assumptions. This book attempts to guide the reader through this forest of estimation and testing procedures,

not by describing the beauty of all possible trees, but by walking through this forest in a structured way, skipping unnecessary side-paths, stressing the similarity of the different species that are encountered and pointing out dangerous pitfalls. The resulting walk is hopefully enjoyable and prevents the reader from getting lost in the econometric forest.

1.2 The Structure of this Book

The first part of this book consists of Chapters 2, 3 and 4. Like most textbooks, it starts with discussing the linear regression model and the OLS estimation method. Chapter 2 presents the basics of this important estimation method, with some emphasis on its validity under fairly weak conditions, while Chapter 3 focuses on the interpretation of the models and the comparison of alternative specifications. Chapter 4 considers two particular deviations from the standard assumptions of the linear model: autocorrelation and heteroskedasticity of the error terms. It is discussed how one can test for these phenomena, how they affect the validity of the OLS estimator and how this can be corrected. This includes a critical inspection of the model specification, the use of adjusted standard errors for the OLS estimator and the use of alternative (GLS) estimators. These three chapters are essential for the remaining part of this book and should be the starting point in any course.

In Chapter 5 another deviation from the standard assumptions of the linear model is discussed that is, however, fatal for the OLS estimator. As soon as the error term in the model is correlated with one or more of the explanatory variables, all good properties of the OLS estimator disappear and we necessarily have to use alternative estimators. The chapter discusses instrumental variable (IV) estimators and, more generally, the generalized method of moments (GMM). This chapter, at least its earlier sections, is also recommended as an essential part of any econometrics course.

Chapter 6 is mainly theoretical and discusses maximum likelihood (ML) estimation. Because in empirical work maximum likelihood is often criticized for its dependence upon distributional assumptions, it is not discussed in the earlier chapters where alternatives are readily available that are either more robust than maximum likelihood or (asymptotically) equivalent to it. Particular emphasis in Chapter 6 is on misspecification tests based upon the Lagrange multiplier principle. While many empirical studies tend to take the distributional assumptions for granted, their validity is crucial for consistency of the estimators that are employed and should therefore be tested. Often these tests are relatively easy to perform, although most software does not routinely provide them (yet). Chapter 6 is crucial for understanding Chapter 7 on limited dependent variable models and for a small number of sections in Chapters 8 to 10.

The last part of this book contains four chapters. Chapter 7 presents models that are typically (though not exclusively) used in micro-economics, where the dependent variable is discrete (e.g. zero or one), partly discrete (e.g. zero or positive) or a duration. It also includes discussions of the sample selection problem and the estimation of treatment effects that go further than their typical textbook treatment.

Chapters 8 and 9 discuss time series modelling including unit roots, cointegration and error-correction models. These chapters can be read immediately after Chapter 4 or 5, with the exception of a few parts that relate to maximum likelihood estimation.

The theoretical developments in this area over the last 25 years have been substantial, and many recent textbooks seem to focus upon it almost exclusively. Univariate time series models are covered in Chapter 8. In this case, models are developed that explain an economic variable from its own past. This includes ARIMA models, as well as GARCH models for the conditional variance of a series. Multivariate time series models that consider several variables simultaneously are discussed in Chapter 9. This includes vector autoregressive models, cointegration and error-correction models.

Finally, Chapter 10 covers models based on panel data. Panel data are available if we have repeated observations of the same units (for example households, firms or countries). In the last decade the use of panel data has become important in many areas of economics. Micro-economic panels of households and firms are readily available and, given the increase in computing resources, more manageable than in the past. In addition, it has become increasingly common to pool time series of several countries. One of the reasons for this may be that researchers believe that a cross-sectional comparison of countries provides interesting information, in addition to a historical comparison of a country with its own past. This chapter also discusses the recent developments on unit roots and cointegration in a panel data setting. Furthermore, a separate section is devoted to repeated cross-sections and pseudo panel data.

At the end of the book the reader will find two short appendices discussing mathematical and statistical results that are used in several places in the book. This includes a discussion of some relevant matrix algebra and distribution theory. In particular, a discussion of properties of the (bivariate) normal distribution, including conditional expectations, variances and truncation, is provided.

In my experience the material in this book is too much to be covered in a single course. Different courses can be scheduled on the basis of the chapters that follow. For example, a typical graduate course in applied econometrics would cover Chapters 2, 3, 4 and parts of Chapter 5, and then continue with selected parts of Chapters 8 and 9 if the focus is on time series analysis, or continue with Section 6.1 and Chapter 7 if the focus is on cross-sectional models. A more advanced undergraduate or graduate course may focus attention on the time series chapters (Chapters 8 and 9), the micro-econometric chapters (Chapters 6 and 7) or panel data (Chapter 10 with some selected parts from Chapters 6 and 7).

Given the focus and length of this book, I had to make many choices concerning which material to present or not. As a general rule I did not want to bother the reader with details that I considered not essential or not to have empirical relevance. The main goal was to give a general and comprehensive overview of the different methodologies and approaches, focusing on what is relevant for doing and understanding empirical work. Some topics are only very briefly mentioned, and no attempt is made to discuss them at any length. To compensate for this I have tried to give references in appropriate places to other, often more advanced, textbooks that do cover these issues.

1.3 Illustrations and Exercises

In most chapters a variety of empirical illustrations is provided in separate sections or subsections. While it is possible to skip these illustrations essentially without losing continuity, these sections do provide important aspects concerning the implementation

of the methodology discussed in the preceding text. In addition, I have attempted to provide illustrations that are of economic interest in themselves, using data that are typical of current empirical work and cover a wide range of different areas. This means that most data sets are used in recently published empirical work and are fairly large, both in terms of number of observations and in terms of number of variables. Given the current state of computing facilities, it is usually not a problem to handle such large data sets empirically.

Learning econometrics is not just a matter of studying a textbook. Hands-on experience is crucial in the process of understanding the different methods and how and when to implement them. Therefore, readers are strongly encouraged to get their hands dirty and to estimate a number of models using appropriate or inappropriate methods, and to perform a number of alternative specification tests. With modern software becoming more and more user friendly, the actual computation of even the more complicated estimators and test statistics is often surprisingly simple, sometimes dangerously simple. That is, even with the wrong data, the wrong model and the wrong methodology, programs may come up with results that are seemingly all right. At least some expertise is required to prevent the practitioner from such situations, and this book plays an important role in this.

To stimulate the reader to use actual data and estimate some models, almost all data sets used in this text are available through the website www.wileyurope.com/college/verbeek. Readers are encouraged to re-estimate the models reported in this text and check whether their results are the same, as well as to experiment with alternative specifications or methods. Some of the exercises make use of the same or additional data sets and provide a number of specific issues to consider. It should be stressed that, for estimation methods that require numerical optimization, alternative programs, algorithms or settings may give slightly different outcomes. However, you should get results that are close to the ones reported.

I do not advocate the use of any particular software package. For the linear regression model any package will do, while for the more advanced techniques each package has its particular advantages and disadvantages. There is typically a trade-off between user-friendliness and flexibility. Menu-driven packages often do not allow you to compute anything other than what's on the menu, but, if the menu is sufficiently rich, that may not be a problem. Command-driven packages require somewhat more input from the user, but are typically quite flexible. For the illustrations in the text, I made use of Eviews 5.1, Microfit 4.0, RATS 5.1 and Stata 9.2. Several alternative econometrics programs are available, including PcGive, TSP and SHAZAM; for more advanced or tailored methods, econometricians make use of GAUSS, Matlab, Ox, S-Plus and many other programs, as well as specialized software for specific methods or types of model. Journals like the *Journal of Applied Econometrics* and the *Journal of Economic Surveys* regularly publish software reviews.

The exercises included at the end of each chapter consist of a number of questions that are primarily intended to check whether the reader has grasped the most important concepts. Therefore, they typically do not go into technical details or ask for derivations or proofs. In addition, several exercises are of an empirical nature and require the reader to use actual data, made available through the book's website.

