PART ONE

Preliminaries

Part 1 covers the essential components of microeconometric analysis – an economic specification, a statistical model and a data set.

Chapter 1 discusses the distinctive aspects of microeconometrics, and provides an outline of the book. It emphasizes that discreteness of data, and nonlinearity and heterogeneity of behavioral relationships are key aspects of individual-level microeconometric models. It concludes by presenting the notation and conventions used throughout the book.

Chapters 2 and 3 set the scene for the remainder of the book by introducing the reader to key model and data concepts that shape the analyses of later chapters.

A key distinction in econometrics is between essentially descriptive models and data summaries at various levels of statistical sophistication and models that go beyond associations and attempt to estimate causal parameters. The classic definitions of causality in econometrics derive from the Cowles Commission simultaneous equations models that draw sharp distinctions between exogenous and endogenous variables, and between structural and reduced form parameters. Although reduced form models are very useful for some purposes, knowledge of structural or causal parameters is essential for policy analyses. Identification of structural parameters within the simultaneous equations framework poses numerous conceptual and practical difficulties. An increasingly-used alternative approach based on the potential outcome model, also attempts to identify causal parameters but it does so by posing limited questions within a more manageable framework. Chapter 2 attempts to provide an overview of the fundamental issues that arise in these and other alternative frameworks. Readers who initially find this material challenging should return to this chapter after gaining greater familiarity with specific models covered later in the book.

The empirical researcher's ability to identify causal parameters depends not only on the statistical tools and models but also on the type of data available. An experimental framework provides a standard for establishing causal connections. However, observational, not experimental, data form the basis of much of econometric inference. Chapter 3 surveys the pros and cons of three main types of data: observational data, data from social experiments, and data from natural experiments. The strengths and weaknesses of conducting causal inference based on each type of data are reviewed.

CHAPTER 1

Overview

1.1. Introduction

This book provides a detailed treatment of **microeconometric analysis**, the analysis of individual-level data on the economic behavior of individuals or firms. A broader definition would also include grouped data. Usually regression methods are applied to cross-section or panel data.

Analysis of individual data has a long history. Ernst Engel (1857) was among the earliest quantitative investigators of household budgets. Allen and Bowley (1935), Houthakker (1957), and Prais and Houthakker (1955) made important contributions following the same research and modeling tradition. Other landmark studies that were also influential in stimulating the development of microeconometrics, even though they did not always use individual-level information, include those by Marschak and Andrews (1944) in production theory and by Wold and Jureen (1953), Stone (1953), and Tobin (1958) in consumer demand.

As important as the above earlier cited work is on household budgets and demand analysis, the material covered in this book has stronger connections with the work on discrete choice analysis and censored and truncated variable models that saw their first serious econometric applications in the work of McFadden (1973, 1984) and Heckman (1974, 1979), respectively. These works involved a major departure from the overwhelming reliance on linear models that characterized earlier work. Subsequently, they have led to significant methodological innovations in econometrics. Among the earlier textbook-level treatments of this material (and more) are the works of Maddala (1983) and Amemiya (1985). As emphasized by Heckman (2001), McFadden (2001), and others, many of the fundamental issues that dominated earlier work based on market data remain important, especially concerning the conditions necessary for identifiability of causal economic relations. Nonetheless, the style of microeconometrics is sufficiently distinct to justify writing a text that is exclusively devoted to it.

Modern microeconometrics based on individual-, household-, and establishment-level data owes a great deal to the greater availability of data from cross-section and longitudinal sample surveys and census data. In the past two decades, with the

expansion of electronic recording and collection of data at the individual level, data volume has grown explosively. So too has the available computing power for analyzing large and complex data sets. In many cases event-level data are available; for example, marketing science often deals with purchase data collected by electronic scanners in supermarkets, and industrial organization literature contains econometric analyses of airline travel data collected by online booking systems. There are now new branches of economics, such as social experimentation and experimental economics, that generate "experimental" data. These developments have created many new modeling opportunities that are absent when only aggregated market-level data are available. Meanwhile the explosive growth in the volume and types of data has also given rise to numerous methodological issues. Processing and econometric analysis of such large microdatabases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeconometrics. Econometric analysis of such data is the subject matter of this book.

Key precursors of this book are the books by Maddala (1983) and Amemiya (1985). Like them it covers topics that are presented only briefly, or not at all, in undergraduate and first-year graduate econometrics courses. Especially compared to Amemiya (1985) this book is more oriented to the practitioner. The level of presentation is nonetheless advanced in places, especially for applied researchers in disciplines that are less mathematically oriented than economics.

A relatively advanced presentation is needed for several reasons. First, the data are often discrete or censored, in which case **nonlinear methods** such as logit, probit, and Tobit models are used. This leads to statistical inference based on more difficult asymptotic theory.

Second, **distributional assumptions** for such data become critically important. One response is to develop highly parametric models that are sufficiently detailed to capture the complexities of data, but these models can be challenging to estimate. A more common response is to minimize parametric assumptions and perform statistical inference based on standard errors that are "robust" to complications such as heteroskedasticity and clustering. In such cases considerable knowledge can be needed to ensure valid statistical inference even if a standard regression package is used.

Third, economic studies often aim to determine **causation** rather than merely measure correlation, despite access to observational rather than experimental data. This leads to methods to isolate causation such as instrumental variables, simultaneous equations, measurement error correction, selection bias correction, panel data fixed effects, and differences-in-differences.

Fourth, microeconomic data are typically collected using cross-section and panel surveys, censuses, or social experiments. **Survey data** collected using these methods are subject to problems of complex survey methodology, departures from simple random sampling assumptions, and problems of sample selection, measurement errors, and incomplete, and/or missing data. Dealing with such issues in a way that can support valid population inferences from the estimated econometric models population requires use of advanced methods.

Finally, it is not unusual that two or more **complications occur simultaneously**, such as endogeneity in a logit model with panel data. Then a cookbook approach

becomes very difficult to implement. Instead, considerable understanding of the theory underlying the methods is needed, as the researcher may need to read econometrics journal articles and adapt standard econometrics software.

1.2. Distinctive Aspects of Microeconometrics

We now consider several advantages of microeconometrics that derive from its distinctive features.

1.2.1. Discreteness and Nonlinearity

The first and most obvious point is that microeconometric data are usually at a low level of aggregation. This has a major consequence for the functional forms used to analyze the variables of interest. In many, if not most, cases linear functional forms turn out to be simply inappropriate. More fundamentally, disaggregation brings to the forefront **heterogeneity** of individuals, firms, and organizations that should be properly controlled (modeled) if one wants to make valid inferences about the underlying relationships. We discuss these issues in greater detail in the following sections.

Although aggregation is not entirely absent in microdata, as for example when household- or establishment-level data are collected, the level of aggregation is usually orders of magnitude lower than is common in macro analyses. In the latter case the process of aggregation leads to smoothing, with many of the movements in opposite directions canceling in the course of summation. The aggregated variables often show smoother behavior than their components, and the relationships between the aggregates frequently show greater smoothness than the components. For example, a relation between two variables at a micro level may be piecewise linear with many nodes. After aggregation the relationship is likely to be well approximated by a smooth function. Hence an immediate consequence of disaggregation is the absence of features of continuity and smoothness both of the variables themselves and of the relationships between them.

Usually individual- and firm-level data cover a huge range of variation, both in the cross-section and time-series dimensions. For example, average weekly consumption of (say) beef is highly likely to be positive and smoothly varying, whereas that of an individual household in a given week may be frequently zero and may also switch to positive values from time to time. The average number of hours worked by female workers is unlikely to be zero, but many individual females have zero market hours of work (corner solutions), switching to positive values at other times in the course of their labor market history. Average household expenditure on vacations is usually positive, but many individual households may have zero expenditure on vacations in any given year. Average per capita consumption of tobacco products will usually be positive, but many individuals in the population have never consumed these products and never will, irrespective of price and income considerations. As Pudney (1989) has observed, microdata exhibit "holes, kinks and corners." The holes correspond to nonparticipation in the activity of interest, kinks correspond to the switching behavior, and corners correspond

to the incidence of nonconsumption or nonparticipation at specific points of time. That is, discreteness and nonlinearity of response are intrinsic to microeconometrics.

An important class of nonlinear models in microeconometrics deals with **limited dependent variables** (Maddala, 1983). This class includes many models that provide suitable frameworks for analyzing discrete responses and responses with limited range of variation. Such tools of analyses are of course also available for analyzing macrodata, if required. The point is that they are indispensable in microeconometrics and give it its distinctive feature.

1.2.2. Greater Realism

Macroeconometrics is sometimes based on strong assumptions; the representative agent assumption is a leading example. A frequent appeal is made to microeconomic reasoning to justify certain specifications and interpretations of empirical results. However, it is rarely possible to say explicitly how these are affected by aggregation over time and micro units. Alternatively, very extreme aggregation assumptions are made. For example, aggregates are said to reflect the behavior of a hypothetical representative agent. Such assumptions also are not credible.

From the viewpoint of microeconomic theory, quantitative analysis founded on microdata may be regarded as more realistic than that based on aggregated data. There are three justifications for this claim. First, the measurement of the variables involved in such hypotheses is often more direct (though not necessarily free from measurement error) and has greater correspondence to the theory being tested. Second, hypotheses about economic behavior are usually developed from theories of individual behavior. If these hypotheses are tested using aggregated data, then many approximations and simplifying assumptions have to be made. The simplifying assumption of a representative agent causes a great loss of information and severely limits the scope of an empirical investigation. Because such assumptions can be avoided in microeconometrics, and usually are, in principle the microdata provide a more realistic framework for testing microeconomic hypotheses. This is not a claim that the promise of microdata is necessarily achieved in empirical work. Such a claim must be assessed on a case-by-case basis. Finally, a realistic portrayal of economic activity should accommodate a broad range of outcomes and responses that are a consequence of individual heterogeneity and that are predicted by underlying theory. In this sense microeconomic data sets can support more realistic models.

Microeconometric data are often derived from household or firm surveys, typically encompassing a wide range of behavior, with many of the behavioral outcomes taking the form of discrete or categorical responses. Such data sets have many awkward features that call for special tools in the formulation and analysis that, although not entirely absent from macroeconometric work, nevertheless are less widely used.

1.2.3. Greater Information Content

The potential advantages of microdata sets can be realized if such data are informative. Because sample surveys often provide independent observations on thousands of

cross-sectional units, such data are thought to be more informative than the standard, usually highly serially correlated, macro time series typically consisting of at most a few hundred observations.

As will be explained in the next chapter, in practice the situation is not so clear-cut because the microdata may be quite noisy. At the individual level many (idiosyncratic) factors may play a large role in determining responses. Often these cannot be observed, leading one to treat them under the heading of a random component, which can be a very large part of observed variation. In this sense randomness plays a larger role in microdata than in macrodata. Of course, this affects measures of goodness of fit of the regressions. Students whose initial exposure to econometrics comes through aggregate time-series analysis are often conditioned to see large R^2 values. When encountering cross-section regressions for the first time, they express disappointment or even alarm at the "low explanatory power" of the regression equation. Nevertheless, there remains a strong presumption that, at least in certain dimensions, large microdata sets are highly informative.

Another qualification is that when one is dealing with purely cross-section data, very little can be said about the intertemporal aspects of relationships under study. This particular aspect of behavior can be studied using panel and transition data.

In many cases one is interested in the behavioral responses of a specific group of economic agents under some specified economic environment. One example is the impact of unemployment insurance on the job search behavior of young unemployed persons. Another example is the labor supply responses of low-income individuals receiving income support. Unless microdata are used such issues cannot be addressed directly in empirical work.

1.2.4. Microeconomic Foundations

Econometric models vary in the explicit role given to economic theory. At one end of the spectrum there are models in which the a priori theorizing may play a dominant role in the specification of the model and in the choice of an estimation procedure. At the other end of the spectrum are empirical investigations that make much less use of economic theory.

The goal of the analysis in the first case is to identify and estimate fundamental parameters, sometimes called deep parameters, that characterize individual taste and preferences and/or technological relationships. As a shorthand designation, we call this the **structural approach**. Its hallmark is a heavy dependence on economic theory and emphasis on causal inference. Such models may require many assumptions, such as the precise specification of a cost or production function or specification of the distribution of error terms. The empirical conclusions of such an exercise may not be robust with respect to the departures from the assumptions. In Section 2.4.4 we shall say more about this approach. At the present stage we simply emphasize that if the structural approach is implemented with aggregated data, it will yield estimates of the fundamental parameters only under very stringent (and possibly unrealistic) conditions. Microdata sets provide a more promising environment for the structural approach, essentially because they permit greater flexibility in model specification.

The goal of the analysis in the second case is to model relationship(s) between response variables of interest conditionally on variables the researcher takes as given, or exogenous. More formal definitions of **endogeneity** and **exogeneity** are given in Chapter 2. As a shorthand designation, we call this a **reduced form approach**. The essential point is that reduced form analysis does not always take into account all causal interdependencies. A regression model in which the focus is on the prediction of y given regressors $\mathbf{x}$, and not on the causal interpretation of the regression parameters, is often referred to as a reduced form regression. As will be explained in Chapter 2, in general the parameters of the reduced form model are functions of structural parameters. They may not be interpretable without some information about the structural parameters.

1.2.5. Disaggregation and Heterogeneity

It is sometimes said that many problems and issues of macroeconometrics arise from serial correlation of macro time series, and those of microeconometrics arise from heteroskedasticity of individual-level data. Although this is a useful characterization of the modeling effort in many microeconometric analyses, it needs amplification and is subject to important qualifications. In a range of microeconometric models, modeling of dynamic dependence may be an important issue.

The benefits of disaggregation, which were emphasized earlier in this section, come at a cost: As the data become more disaggregated the importance of controlling for interindividual heterogeneity increases. Heterogeneity, or more precisely unobserved heterogeneity, plays a very important role in microeconometrics. Obviously, many variables that reflect interindividual heterogeneity, such as gender, race, educational background, and social and demographic factors, are directly observed and hence can be controlled for. In contrast, differences in individual motivation, ability, intelligence, and so forth are either not observed or, at best, imperfectly observed.

The simplest response is to ignore such heterogeneity, that is, to absorb it into the regression disturbance. After all this is how one treats the myriad small unobserved factors. This step of course increases the unexplained part of the variation. More seriously, ignoring persistent interindividual differences leads to **confounding** with other factors that are also sources of persistent interindividual differences. Confounding is said to occur when the individual contributions of different regressors (predictor variables) to the variation in the variable of interest cannot be statistically separated. Suppose, for example, that the factor x_1 (schooling) is said to be the source of variation in y (earnings), when another variable x_2 (ability), which is another source of variation, does not appear in the model. Then that part of total variation that is attributable to the second variable may be incorrectly attributed to the first variable. Intuitively, their relative importances are confounded. A leading source of confounding bias is the incorrect omission of regressors from the model and the inclusion of other variables that are proxies for the omitted variable.

Consider, for example, the case in which a program participation (0/1 dummy) variable D is included in the regression mean function with a vector of regressors $\mathbf{x}$,

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha D + u, \tag{1.1}$$

where u is an error term. The term "treatment" is used in biological and experimental sciences to refer to an administered regimen involving participants in some trial. In econometrics it commonly refers to participation in some activity that may impact an outcome of interest. This activity may be randomly assigned to the participants or may be self-selected by the participant. Thus, although it is acknowledged that individuals choose their years of schooling, one still thinks of years of schooling as a "treatment" variable. Suppose that program participation is taken to be a discrete variable. The coefficient α of the "treatment variable" measures the average impact of the program participation ($D = 1$), conditional on covariates. If one does not control for unobserved heterogeneity, then a potential ambiguity affects the interpretation of the results. If d is found to have a significant impact, then the following question arises: Is α significantly different from zero because D is correlated with some unobserved variable that affects y or because there is a causal relationship between D and y? For example, if the program considered is university education, and the covariates do not include a measure of ability, giving a fully causal interpretation becomes questionable. Because the issue is important, more attention should be given to how to control for unobserved heterogeneity.

In some cases where dynamic considerations are involved the type of data available may put restrictions on how one can control for heterogeneity. Consider the example of two households, identical in all relevant respects except that one exhibits a systematically higher preference for consuming good A. One could control for this by allowing individual utility functions to include a heterogeneity parameter that reflects their different preferences. Suppose now that there is a theory of consumer behavior that claims that consumers become addicted to good A, in the sense that the more they consume of it in one period, the greater is the probability that they will consume more of it in the future. This theory provides another explanation of persistent interindividual differences in the consumption of good A. By controlling for heterogeneous preferences it becomes possible to test which source of persistence in consumption – preference heterogeneity or addiction – accounts for different consumption patterns. This type of problem arises whenever some dynamic element generates persistence in the observed outcomes. Several examples of this type of problem arise in various places in the book.

A variety of approaches for modeling heterogeneity coexist in microeconometrics. A brief mention of some of these follows, with details postponed until later.

An extreme solution is to ignore all unobserved interindividual differences. If unobserved heterogeneity is uncorrelated with observed heterogeneity, and if the outcome being studied has no intertemporal dependence, then the aforementioned problems will not arise. Of course, these are strong assumptions and even with these assumptions not all econometric difficulties disappear.

One approach for handling heterogeneity is to treat it as a **fixed effect** and to estimate it as a coefficient of an individual specific $0/1$ dummy variable. For example, in a cross-section regression, each micro unit is allowed its own dummy variable (intercept). This leads to an extreme proliferation of parameters because when a new individual is added to the sample, a new intercept parameter is also added. Thus this approach will not work if our data are cross sectional. The availability of multiple observations

per individual unit, most commonly in the form of panel data with T time-series observations for each of the N cross-section units, makes it possible to either estimate or eliminate the fixed effect, for example by first differencing if the model is linear and the fixed effect is additive. If the model is nonlinear, as is often the case, the fixed effect will usually not be additive and other approaches will need to be considered.

A second approach to modeling unobserved heterogeneity is through a **random effects** model. There are a number of ways in which the random effects model can be formulated. One popular formulation assumes that one or more regression parameters, often just the regression intercept, varies randomly across the cross section. In another formulation the regression error is given a component structure, with an individual specific random component. The random effects model then attempts to estimate the parameters of the distribution from which the random component is drawn. In some cases, such as demand analysis, the random term can be interpreted as random preference variation. Random effects models can be estimated using either cross-section or panel data.

1.2.6. Dynamics

A very common assumption in cross-section analysis is the absence of intertemporal dependence, that is, an absence of dynamics. Thus, implicitly it is assumed that the observations correspond to a stochastic equilibrium, with the deviation from the equilibrium being represented by serially independent random disturbances. Even in microeconometrics for some data situations such an assumption may be too strong. For example, it is inconsistent with the presence of serially correlated unobserved heterogeneity. Dependence on lagged dependent variables also violates this assumption.

The foregoing discussion illustrates some of the potential limitations of a single cross-section analysis. Some limitations may be overcome if repeated cross sections are available. However, if there is dynamic dependence, the least problematic approach might well be to use panel data.

1.3. Book Outline

The book is split into six parts. Part 1 presents the issues involved in microeconometric modeling. Parts 2 and 3 present general theory for estimation and statistical inference for nonlinear regression models. Parts 4 and 5 specialize to the core models used in applied microeconometrics for, respectively, cross-section and panel data. Part 6 covers broader topics that make considerable use of material presented in the earlier chapters.

The book outline is summarized in Table 1.1. The remainder of this section details each part in turn.

1.3.1. Part 1: Preliminaries

Chapters 2 and 3 expand on the special features of the **microeconometric** approach to modeling and **microeconomic data structures** within the more general statistical

Table 1.1. *Book Outline*

Part and Chapter	Background[a]	Example
1. Preliminaries		
1. Overview	–	
2. Causal and Noncausal Models	–	Simultaneous equations models
3. Microeconomic Data Structures	–	Observational data
2. Core Methods		
4. Linear Models	–	Ordinary least squares
5. Maximum Likelihood and Nonlinear Least-Squares Estimation	–	m-estimation or extremum estimation
6. Generalized Method of Moments and Systems Estimation	5	Instrumental variables
7. Hypothesis Tests	5	Wald, score, and likelihood ratio tests
8. Specification Tests and Model Selection	5,7	Conditional moment test
9. Semiparametric Methods	–	Kernel regression
10. Numerical Optimization	5	Newton–Raphson iterative method
3. Simulation-Based Methods		
11. Bootstrap Methods	7	Percentile t-method
12. Simulation-Based Methods	5	Maximum simulated likelihood
13. Bayesian Methods	–	Markov chain Monte Carlo
4. Models for Cross-Section Data		
14. Binary Outcome Models	5	Logit, probit for $y = (0, 1)$
15. Multinomial Models	5,14	Multinomial logit for $y = (1, . . , m)$
16. Tobit and Selection Models	5,14	Tobit for $y = \max(y^*, 0)$
17. Transition Data: Survival Analysis	5	Cox proportional hazards for $y = \min(y^*, c)$
18. Mixture Models and Unobserved Heterogeneity	5,17	Unobserved heterogeneity
19. Models for Multiple Hazards	5,17	Multiple hazards
20. Models of Count Data	5	Poisson for $y = 0, 1, 2, \ldots$
5. Models for Panel Data		
21. Linear Panel Models: Basics	–	Fixed and random effects
22. Linear Panel Models: Extensions	6,21	Dynamic and endogenous regressors
23. Nonlinear Panel Models	5,6,21,22	Panel logit, Tobit, and Poisson
6. Further Topics		
24. Stratified and Clustered Samples	5	Data $(y_{ij}, \mathbf{x}_{ij})$ correlated over j
25. Treatment Evaluation	5,21	Regressor $d = 1$ if in program
26. Measurement Error Models	5	Logit model with measurement errors
27. Missing Data and Imputation	5	Regression with missing observations

[a] The background gives the essential chapter needed in addition to the treatment of ordinary and weighted LS in Chapter 4. Note that the first panel data chapter (Chapter 21) requires only Chapter 4.