CHAPTER 3

Microeconomic Data Structures

3.1. Introduction

This chapter surveys issues concerning the potential usefulness and limitations of different types of microeconomic data. By far the most common data structure used in microeconometrics is survey or census data. These data are usually called **observational data** to distinguish them from **experimental data**.

This chapter discusses the potential limitation of the aforementioned data structures. The inherent limitations of observational data may be further compounded by the manner in which the data are collected, that is, by the sample frame (the way the sample is generated), sample design (simple random sample versus stratified random sample), and sample scope (cross-section versus longitudinal data). Hence we also discuss sampling issues in connection with the use of observational data. Some of this terminology is new at this stage but will be explained later in this chapter.

Microeconometrics goes beyond the analysis of survey data under the assumptions of simple random sampling. This chapter considers extensions. Section 3.2 outlines the structure of multistage sample surveys and some common forms of departure from random sampling; a more detailed analysis of their statistical implications is provided in later chapters. It also considers some commonly occurring complications that result in the data not being necessarily representative of the population. Given the deficiencies of observational data in estimating causal parameters, there has been an increased attempt at exploiting experimental and quasi-experimental data and frameworks. Section 3.3 examines the potential of data from social experiments. Section 3.4 considers the modeling opportunities arising from a special type of observational data, generated under quasi-experimental conditions, that naturally provide treated and untreated subjects and hence are called natural experiments. Section 3.5 covers practical issues of microdata management.

3.2. Observational Data

The major source of microeconomic observational data is surveys of households, firms, and government administrative data. Census data can also be used to generate samples. Many other samples are often generated at points of contact between transacting parties. For example, marketing data may be generated at the point of sale and/or surveys among (actual or potential) purchasers. The Internet (e.g., online auctions) is also a source of data.

There is a huge literature on sample surveys from the viewpoint of both survey statisticians and users of survey data. The first discusses how to sample from the population and the results from different sampling designs, and the second deals with the issues of estimation and inference that arise when survey data are collected using different sampling designs. A key issue is how well the sample represents the population. This chapter deals with both strands of the literature in an introductory fashion. Many additional details are given in Chapter 24.

3.2.1. Nature of Survey Data

The term observational data usually refers to survey data collected by sampling the relevant population of subjects without any attempt to control the characteristics of the sampled data. Let t denote the time subscript, let $\mathbf{w}$ denote a set of variables of interest. In the present context t can be a point in time or time interval. Let $\mathcal{S}_t$ denote a sample from population probability distribution $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$; $\mathcal{S}_t$ is a draw from $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}$ is a parameter vector. The population should be thought of as a set of points with characteristics of interest, and for simplicity we assume that the form of the probability distribution F is known. A simple random sampling scheme allows every element of the population to have an equal probability of being included in the sample. More complex sampling schemes will be considered later.

The abstract concept of a **stationary population** provides a useful benchmark. If the moments of the characteristics of the population are constant, then we can write $\boldsymbol{\theta}_t = \boldsymbol{\theta}$, for all t. This is a strong assumption because it implies that the moments of the characteristics of the population are time-invariant. For example, the age–sex distribution should be constant. More realistically, some population characteristics would not be constant. To handle such a possibility, (the parameters of) each population may be regarded as a draw from a **superpopulation** with constant characteristics. Specifically, we think of each $\boldsymbol{\theta}_t$ as a draw from a probability distribution with constant (hyper)parameter $\boldsymbol{\theta}$. The terms superpopulation and hyperparameters occur frequently in the literature on hierarchical models discussed in Chapter 24. Additional complications arise if $\boldsymbol{\theta}_t$ has an evolutionary component, for example through dependence on t, or if successive values are interdependent. Using hierarchical models, discussed in Chapters 13 and 26, provides one approach for modeling the relation between hyperparameters and subpopulation characteristics.

3.2.2. Simple Random Samples

As a benchmark for subsequent discussion, consider simple random sampling in which the probability of sampling unit i from a population of size N, with N large, is $1/N$ for all i. Partition $\mathbf{w}$ as $[y : \mathbf{x}]$. Suppose our interest is in modeling y, a possibly vector-valued outcome variable, conditional on the exogenous covariate vector $\mathbf{x}$, whose joint distribution is denoted $f_J(y, \mathbf{x})$. This can be always be factored as the product of the conditional distribution $f_C(y|\mathbf{x}, \boldsymbol{\theta})$ and the marginal distribution $f_M(\mathbf{x})$:

$$f_J(y, \mathbf{x}) = f_C(y|\mathbf{x}, \boldsymbol{\theta}) f_M(\mathbf{x}). \tag{3.1}$$

Simple random sampling involves drawing the $(y, \mathbf{x})$ combinations uniformly from the entire population.

3.2.3. Multistage Surveys

One alternative is a **stratified multistage cluster sampling**, also referred to as a **complex survey** method. Large-scale surveys like the Current Population Survey (CPS) and the Panel Survey of Income Dynamics (PSID) take this approach. Section 24.2 provides additional detail on the structure of the CPS.

The complex survey design has advantages. It is more cost effective because it reduces geographical dispersion, and it becomes possible to sample certain subpopulations more intensively. For example, "oversampling" of small subpopulations exhibiting some relevant characteristic becomes feasible whereas a random sample of the population would produce too few observations to support reliable results. A disadvantage is that stratified sampling will reduce interindividual variation, which is essential for greater precision.

The sample survey literature focuses on **multistage surveys** that sequentially partition the population into the following categories:

1. **Strata**: Nonoverlapping subpopulations that exhaust the population.
2. **Primary sampling units** (PSUs): Nonoverlapping subsets of the strata.
3. **Secondary sampling units** (SSUs): Sub-units of the PSU, which may in turn be partitioned, and so on.
4. **Ultimate sampling unit** (USU): The final unit chosen for interview, which could be a household or a collection of households (a segment).

As an example, the strata may be the various states or provinces in a country, the PSU may be regions within the state or province, and the USU may be a small cluster of households in the same neighborhood.

Usually all strata are surveyed so that, for example, all states will be included in the sample with certainty. But not all of the PSUs and their subdivisions are surveyed, and they may be sampled at different rates. In **two-stage sampling** the surveyed PSUs are drawn at random and the USU is then drawn at random from the selected PSUs. In **multistage sampling** intermediate sampling units such as SSUs also appear.

A consequence of these sampling methods is that different households will have different probabilities of being sampled. The sample is then unrepresentative of the population. Many surveys provide **sampling weights** that are intended to be inversely proportional to the probability of being sampled, in which case these weights can be used to obtain unbiased estimators of population characteristics.

Survey data may be clustered due to, for example, sampling of many households in the same small neighborhood. Observations in the same cluster are likely to be dependent or correlated because they may depend on some observable or unobservable factor that could affect all observations in a stratum. For example, a suburb may be dominated by high-income households or by households that are relatively homogeneous in some dimension of their preferences. Data from these households will tend to be correlated, at least unconditionally, though it is possible that such correlation is negligible after conditioning on observable characteristics of the households. Statistical inference ignoring correlation between sampled observations yields erroneous estimates of variances that are smaller than those from the correct formula. These issues are covered in greater depth in Section 24.5. Two-stage and multistage samples potentially further complicate the computation of standard errors.

In summary, (1) stratification with different sampling rates within strata means that the sample is unrepresentative of the population; (2) sampling weights inversely proportional to the probability of being sampled can be used to obtain unbiased estimation of population characteristics; and (3) clustering may lead to correlation of observations and understatement of the true standard errors of estimators unless appropriate adjustments are made.

3.2.4. Biased Samples

If a random sample is drawn then the probability distribution for the data is the same as the population distribution. Certain departures from random sampling cause a divergence between the two; this is referred to as **biased sampling**. The data distribution differs from the population distribution in a manner that depends on the nature of the deviation from random sampling. Deviation from random sampling occurs because it is sometimes more convenient or cost effective to obtain the data from a subpopulation even though it is not representative of the entire population. We now consider several examples of such departures, beginning with a case in which there is no departure from randomness.

Exogenous Sampling

Exogenous sampling from survey data occurs if the analyst segments the available sample into subsamples based only on a set of exogenous variables $\mathbf{x}$, but not on the response variable. For example, in a study of hospitalizations in Germany, Geil et al. (1997) segmented the data into two categories, those with and without chronic conditions. Classification by income categories is also common. Perhaps it is more accurate to depict this type of sampling as exogenous subsampling because it is done by reference to an existing sample that has already been collected. Segmenting an existing

sample by gender, health, or socioeconomic status is very common. Under the assumptions of exogenous sampling the probability distribution of the exogenous variables is independent of y and contains no information about the population parameters of interest, $\boldsymbol{\theta}$. Therefore, one may ignore the marginal distribution of the exogenous variables and simply base estimation on the conditional distribution $f(y|\mathbf{x}, \boldsymbol{\theta})$. Of course, the assumption may be wrong and the observed distribution of the outcome variable may depend on the selected segmenting variable, which may be correlated with the outcome, thus causing departure from exogenous sampling.

Response-Based Sampling

Response-based sampling occurs if the probability of an individual being included in the sample depends on the responses or choices made by that individual. In this case sample selection proceeds in terms of rules defined in terms of the endogenous variable under study.

Three examples are as follows: (1) In a study of the effect of negative income tax or Aid to Families with Dependent Children (AFDC) on labor supply only those below the poverty line are surveyed. (2) In a study of determinants of public transport modal choice, only users of public transport (a subpopulation) are surveyed. (3) In a study of the determinants of number of visits to a recreational site, only those with at least one visit are included.

Lower survey costs provide an important motivation for using choice-based samples in preference to simple random samples. It would require a very large random sample to generate enough observations (information) about a relatively infrequent outcome or choice, and hence it is cheaper to collect a sample from those who have actually made the choice.

The practical significance of this is that consistent estimation of population parameters $\boldsymbol{\theta}$ can no longer be carried out using the conditional population density $f(y|\mathbf{x})$ alone. The effect of the sampling scheme must also be taken into account. This topic is discussed further in Section 24.4.

Length-Biased Sampling

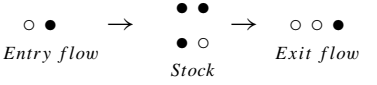
Length-biased sampling illustrates how biases may result from sampling one population to make inferences about a different population. Strictly speaking, it is not so much an example of departure from randomness in sampling as one of sampling the "wrong" population.

Econometric studies of transitions model the time spent in origin state j by individual i before transiting to another destination state s. An example is when j corresponds to unemployment and s to employment. The data used in such studies can come from one of several possible sources. One source is sampling individuals who are unemployed on a particular date, another is to sample those who are in the labor force regardless of their current state, and a third is to sample individuals who are either entering or leaving unemployment during a specified period of time. Each type of sampling scheme is based on a different concept of the relevant population. In the

first case the relevant population is the stock of unemployed individuals, in the second the labor force, and in the third individuals with transitioning employment status. This topic is discussed further in Section 18.6.

Suppose that the purpose of the survey is to calculate a measure of the average duration of unemployment. This is the average length of time a randomly chosen individual will spend in unemployment if he or she becomes unemployed. The answer to this apparently straightforward question may vary depending on how the sample data are obtained. The flow distribution of completed durations is in general quite different from the stock distribution. When we sample the stock, the probability of being in the sample is higher for individuals with longer durations. When we sample the flow out of the state, the probability does not depend on the time spent in the state. This is the well-known example of length-biased sampling in which the estimate obtained by sampling the stock is a biased estimate of the average length of an unemployment spell of a random entrant to unemployment.

The following simple schematic diagram may clarify the point:



Here we use the symbol • to denote slow movers and the symbol ○ to denote fast movers. Suppose the two types are equally represented in the flow, but the slow movers stay in the stock longer than the fast movers. Then the stock population has a higher proportion of slow movers. Finally, the exit population has a higher proportion of fast movers. The argument will generalize to other types of heterogeneity.

The point of this example is not that flow sampling is a better thing to do than stock sampling. Rather, it is that, depending on what the question is, stock sampling may not yield a random sample of the relevant population.

3.2.5. Bias due to Sample Selection

Consider the following problem. A researcher is interested in measuring the effect of training, denoted z (treatment), on posttraining wages, denoted y (outcome), given the worker's characteristics, denoted x. The variable z takes the value 1 if the worker has received training and is 0 otherwise. Observations are available on (x, D) for all workers but on y only for those who received training ($D = 1$). One would like to make inferences about the average impact of training on the posttraining wage of a randomly chosen worker with known characteristics who is currently untrained ($D = 0$). The problem of **sample selection** concerns the difficulty of making such an inference.

Manski (1995), who views this as a problem of identification, defines the selection problem formally as follows:

> This is the problem of identifying conditional probability distributions from random sample data in which the realizations of the conditioning variables are always observed but realizations of the outcomes are censored.

Suppose y is the outcome to be predicted, and the conditioning variables are denoted by x. The variable z is a censoring indicator that takes the value 1 if the outcome y is observed and 0 otherwise. Because the variables (D, x) are always observed, but y is observed only when $D = 1$, Manski views this as a *censored sampling process*. The censored sampling process does not identify $\Pr[y|x]$, as can be seen from

$$\Pr[y|x] = \Pr[y|x, D = 1]\Pr[D = 1|x] + \Pr[y|x, D = 0]\Pr[D = 0|x]. \tag{3.2}$$

The sampling process can identify three of the four terms on the right-hand side, but provides no information about the term $\Pr[y|x, D = 0]$. Because

$$\mathrm{E}[y|x] = \mathrm{E}[y|x, D = 1] \cdot \Pr[D = 1|x] + \mathrm{E}[y|x, D = 0] \cdot \Pr[D = 0|x],$$

whenever the censoring probability $\Pr[D = 0|x]$ is positive, the available empirical evidence places no restrictions on $\mathrm{E}[y|x]$. Consequently, the censored-sampling process can identify $\Pr[y|x]$ only for some unknown value of $\Pr[y|x, D = 0]$. To learn anything about the $\mathrm{E}[y|x]$, restrictions will need to be placed on $\Pr[y|x]$.

The alternative approaches for solving this problem are discussed in Section 16.5.

3.2.6. Quality of Survey Data

The quality of sample data depends not only on the sample design and the survey instrument but also on the survey responses. This observation applies especially to observational data. We consider several ways in which the quality of the sample data may be compromised. Some of the problems (e.g., attrition) can also occur with other types of data. This topic overlaps with that of biased sampling.

Problem of Survey Nonresponse

Surveys are normally voluntary, and incentive to participate may vary systematically according to household characteristics and type of question asked. Individuals may refuse to answer some questions. If there is a systematic relationship between refusal to answer a question and the characteristics of the individual, then the issue of the representativeness of a survey after allowing for **nonresponse** arises. If nonresponse is ignored, and if the analysis is carried out using the data from respondents only, how will the estimation of parameters of interest be affected?

Survey nonresponse is a special case of the selection problem mentioned in the preceding section. Both involve biased samples. To illustrate how it leads to distorted inference consider the following model:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \Bigg| \mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}'\boldsymbol{\beta} \\ \mathbf{z}'\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right), \tag{3.3}$$

where y_1 is a continuous random variable of interest (e.g., expenditure) that depends on $\mathbf{x}$, and y_2 is a latent variable that measures the "propensity to participate" in a survey

and depends on $\mathbf{z}$. The individual participates if $y_2 > 0$; otherwise the individual does not. The variables $\mathbf{x}$ and $\mathbf{z}$ are assumed to be exogenous. The formulation allows y_1 and y_2 to be correlated.

Suppose we estimate $\boldsymbol{\beta}$ from the data supplied by participants by least squares. Is this estimator unbiased in the presence of nonparticipation? The answer is that if nonparticipation is random and independent of y_1, the variable of interest, then there is no bias, but otherwise there will be.

The argument is as follows:

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}_1,$$
$$\mathrm{E}[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathrm{E}\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathrm{E}[\mathbf{y}_1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}, \mathbf{y}_2 > 0\right],$$

where the first line gives the least-squares formula for the estimates of $\boldsymbol{\beta}$ and the second line gives its bias. If $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent, conditional on $\mathbf{X}$ and $\mathbf{Z}$, $\sigma_{12} = 0$, then

$$\mathrm{E}[\mathbf{y}_1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}, \mathbf{y}_2 > 0] = \mathrm{E}[\mathbf{y}_1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}] = \mathbf{0},$$

and there is no bias.

Missing and Mismeasured Data

Survey respondents dealing with an extensive questionnaire will not necessarily answer every question and even if they do, the answers may be deliberately or fortuitously false. Suppose that the sample survey attempts to obtain a vector of responses denoted as $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})$ from N individuals, $i = 1, \ldots, N$. Suppose now that if an individual fails to provide information on any one or more elements of $\mathbf{x}_i$, then the entire vector is discarded. The first problem resulting from **missing data** is that the sample size is reduced. The second potentially more serious problem is that missing data can potentially lead to biases similar to the selection bias. If the data are missing in a systematic manner, then the sample that is left to analyze may not be representative of the population. A form of selection bias may be induced by any systematic pattern of nonresponse. For example, high-income respondents may systematically not respond to questions about income. Conversely, if the data are missing completely at random then discarding incomplete observations will reduce precision but not generate biases. Chapter 27 discusses the missing-data problem and solutions in greater depth.

Measurement errors in survey responses are a pervasive problem. They can arise from a variety of causes, including incorrect responses arising from carelessness, deliberate misreporting, faulty recall of past events, incorrect interpretation of questions, and data-processing errors. A deeper source of measurement error is due to the measured variable being at best an imperfect **proxy** for the relevant theoretical concept. The consequences of such measurement errors is a major topic and is discussed in Chapter 26.

Sample Attrition

In panel data situations the survey involves repeated observations on a set of individuals. In this case we can have

- full response in all periods (full participation),
- nonresponse in the first period and in all subsequent periods (nonparticipation), or
- partial response in the sense of response in the initial periods but nonresponse in later periods (incomplete participation) – a situation referred to as **sample attrition**.

Sample attrition leads to missing data, and the presence of any nonrandom pattern of "missingness" will lead to the sample selection type problems already mentioned. This can be interpreted as a special case of the sample selection problem. Sample attrition is discussed briefly in Sections 21.8.5 and 23.5.2.

3.2.7. Types of Observational Data

Cross-section data are obtained by observing $\mathbf{w}$, for the sample $\mathcal{S}_t$ for some t. Although it is usually impractical to sample all households at the same point of time, cross-section data are still a snapshot of characteristics of each element of a subset of the population that will be used to make inferences about the population. If the population is stationary, then inferences made about $\boldsymbol{\theta}_t$ using $\mathcal{S}_t$ may be valid also for $t' \neq t$. If there is significant dependence between past and current behavior, then longitudinal data are required to identify the relationship of interest. For example, past decisions may affect current outcomes; inertia or habit persistence may account for current purchases, but such dependence cannot be modeled if the history of purchases is not available. This is one of the limitations imposed by cross-section data.

Repeated cross-section data are obtained by a sequence of independent samples $\mathcal{S}_t$ taken from $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$, $t = 1, \ldots, T$. Because the sample design does not attempt to retain the same units in the sample, information about dynamic dependence in behavior is lost. If the population is stationary then repeated cross-section data are obtained by a sampling process somewhat akin to sampling with replacement from the constant population. If the population is nonstationary, repeated cross sections are related in a manner that depends on how the population is changing over time. In such a case the objective is to make inferences about the underlying constant (hyper)parameters. The analysis of repeated cross sections is discussed in Section 22.7.

Panel or longitudinal data are obtained by initially selecting a sample $\mathcal{S}$ and then collecting observations for a sequence of time periods, $t = 1, \ldots, T$. This can be achieved by interviewing subjects and collecting both present and past data at the same time, or by tracking the subjects once they have been inducted into the survey. This produces a sequence of data vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_T\}$ that are used to make inferences about either the behavior of the population or that of the particular sample of individuals. The appropriate methodology in each case may not be the same. If the data are drawn from a nonstationary population, the appropriate objective should be inference on (hyper)parameters of the superpopulation.

Some limitations of these types of data are immediately obvious. Cross-section samples and repeated cross-sections do not in general provide suitable data for modeling intertemporal dependence in outcomes. Such data are only suitable for modeling static relationships. In contrast, longitudinal data, especially if they span a sufficiently long time period, are suitable for modeling both static and dynamic relationships.

Longitudinal data are not free from problems. The first issue is representativeness of the panel. Problems of inference regarding population behavior using longitudinal data become more difficult if the population is not stationary. For analyzing dynamics of behavior, retaining original households in the panel for as long as possible is an attractive option. In practice, longitudinal data sets suffer from the problem of "sample attrition," perhaps due to "sample fatigue." This simply means that survey respondents do not continue to provide responses to questionnaires. This creates two problems: (1) The panel becomes unbalanced and (2) there is the danger that the retained household may not be "typical" and that the sample becomes unrepresentative of the population. When the available sample data are not a random draw from the population, results based on different types of data will be susceptible to biases to different degrees. The problem of "sample fatigue" arises because over time it becomes more difficult to retain individuals within the panel or they may be "lost" (censored) for some other reason, such as a change of location. These issues are dealt with later in the book. Analysis of longitudinal data may nevertheless provide information about some aspects of the behavior of the sampled units, although extrapolation to population behavior may not be straightforward.

3.3. Data from Social Experiments

Observational and experimental data are distinct because an experimental environment can in principle be closely monitored and controlled. This makes it possible to vary a causal variable of interest, holding other covariates at controlled settings. In contrast, observational data are generated in an uncontrolled environment, leaving open the possibility that the presence of confounding factors will make it more difficult to identify the causal relationship of interest. For example, when one attempts to study the earnings–schooling relationship using observational data, one must accept that the years of schooling of an individual is itself an outcome of an individual's decision-making process, and hence one cannot regard the level of schooling as if it had been set by a hypothetical experimenter.

In social sciences, data analogous to experimental data come from either **social experiments**, defined and described in greater detail in the following, or from "laboratory" experiments on small groups of voluntary participants that mimic the behavior of economic agents in the real-life counterpart of the experiment. Social experiments are relatively uncommon, and yet experimental concepts, methods, and data serve as a benchmark for evaluating econometric studies based on observational data.

This section provides a brief account of the methodology of social experiments, the nature of the data emanating from them, and some problems and issues of econometric methodology that they generate.

The central feature of the experimental methodology involves a comparison between the outcomes of the randomly selected experimental group that is subjected to a "**treatment**"with those of a **control** (comparison) group. In a good experiment considerable care is exercised in matching the control and experimental ("treated") groups, and in avoiding potential biases in outcomes. Such conditions may not be realized in observational environments, thereby leading to a possible lack of identification of causal parameters of interest. Sometimes, however, experimental conditions may be approximately replicated in observational data. Consider, for example, two contiguous regions or states, one of which pursues a different minimum-wage policy from the other, creating the conditions of a **natural experiment** in which observations from the "treated" state can be compared with those from the "control" state. The data structure of a natural experiment has also attracted attention in econometrics.

A social experiment involves exogenous variations in the economic environment facing the set of experimental subjects, which is partitioned into one subset that receives the experimental treatment and another that serves as a control group. In contrast to observational studies in which changes in exogenous and endogenous factors are often confounded, a well-designed social experiment aims to isolate the role of treatment variables. In some experimental designs there may be no explicit **control group**, but varying levels of the treatment are applied, in which case it becomes possible in principle to estimate the entire **response surface** of experimental outcomes.

The primary object of a social experiment is to estimate the impact of an actual or potential social program. The potential outcome model of Section 2.7 provides a relevant background for modeling the impact of social experiments. Several alternative measures of impact have been proposed and these will be discussed in the chapter on program evaluation (Chapter 25).

Burtless (1995) summarizes the case for social experiments, while noting some potential limitations. In a companion article Heckman and Smith (1995) focus on limitations of actual social experiments that have been implemented. The remaining discussion in this section borrows significantly from these papers.

3.3.1. Leading Features of Social Experiments

Social experiments are motivated by policy issues about how subjects would react to a type of policy that has never been tried and hence one for which no observed response data exist. The idea of a social experiment is to enlist a group of willing participants, some of whom are randomly assigned to a treatment group and the rest to a control group. The difference between the responses of those in the treatment group, subjected to the policy change, and those in the control group, who are not, is the estimated effect of the policy. Schematically the standard experimental design is as depicted in Figure 3.1.

The term "experimentals" refers to the group receiving treatments, "controls" to the group not receiving treatment, and "**random assignment**" to the process of assigning individuals to the two groups.

Randomized trials were introduced in statistics by R. A. Fisher (1928) and his co-workers. A typical agricultural experiment would consist of a trial in which a new

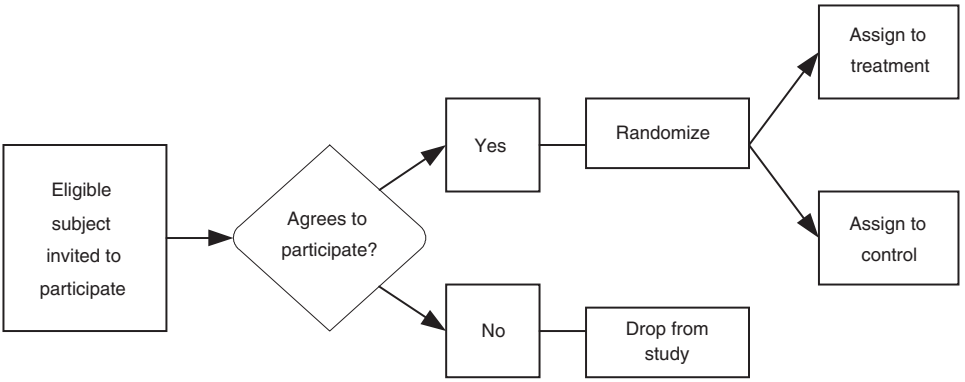


Figure 3.1: Social experiment with random assignment.

treatment such as fertilizer application would be applied to plants growing on randomly chosen blocks of land and then the responses would be compared with those of a control group of plants, similar to the experimentals in all relevant respects but not given experimental treatment. If the effect of all other differences between the experimental and control groups can be eliminated, the estimated difference between the two sets of responses can be attributed to the treatment. In the simplest situation one can concentrate on a comparison of the mean outcome of the treated group and of the untreated group.

Although in agricultural and biomedical sciences, the randomized experiments methodology has been long established, in economics and social sciences it is new. It is attractive for studying responses to policy changes for which no observational data exist, perhaps because the policy changes of interest have never occurred. Randomized experiments also permit a greater variation in policy variables and parameters than are present in observational data, thereby making it easier to identify and study responses to policy changes. In many cases the social experiment may try out a policy that has never been tried, so the observational data remain completely silent on its potential impact.

Social experiments are still rather rare outside the United States, partly because they are expensive to run. In the United States a number of such experiments have taken place since the early 1970s. Table 3.1 summarizes features of some relatively well-known examples; for a more extensive coverage see Burtless (1995).

An experiment may produce either cross-section or longitudinal data, although cost considerations will usually limit the time dimension well below what is typical in observational data. When an experiment lasts several years and has multiple stages and/or geographical locations, as in the case of RHIE, interim analyses based on "incomplete" data are not uncommon (Newhouse et al., 1993).

3.3.2. Advantages of Social Experiments

Burtless (1995) surveys the advantages of social experiments with great clarity. The key advantage stems from randomized trials that remove any correlation between the observed and unobserved characteristics of program participants. Hence the

Table 3.1. *Features of Some Selected Social Experiments*

Experiment	Tested Treatments	Target Population
Rand Health Insurance Experiment (RHIE), 1974–1982	Health insurance plans with varying copayment rate and differing levels of maximum out-of-pocket expenses	Low- and moderate-level income persons and families
Negative Income Tax (NIT), 1968–1978	NIT plans with alternative income guarantees and tax rates	Low- and moderate-level income persons and families with nonaged head of household
Job Training Partnership Act (JTPA), (1986–1994)	Job search assistance, on-the-job training, classroom training financed under JTPA	Out-of-school youths and disadvantaged adults

contribution of the treatment to the outcome difference between the treated and control groups can be estimated without confounding bias even if one cannot control for the confounding variables. The presence of correlation between treatment and confounding variables often plagues observational studies and complicates causal inference. By contrast, an experimental study conducted under ideal circumstances can produce a consistent estimate of the average difference in outcomes of the treated and nontreated groups without much computational complexity.

If, however, an outcome depends on treatment as well as other observable factors, then controlling for the latter will in general improve the precision of the impact estimate.

Even if observational data are available, the generation and use of experimental data has great appeal because it offers the possibility of **exogenizing** a policy variable, and randomization of treatments can potentially lead to great simplification of statistical analysis. Conclusions based on observational data often lack generality because they are based on a nonrandom sample from the population – the problem of selection bias. An example is the aforementioned RHIE study whose major focus is on the price responsiveness of the demand for health services. Availability of health insurance affects the user price of health services and thereby its use. An important policy issue is the extent to which "overutilization" of health services would result from subsidized health insurance. One can, of course, use observational data to model the relation between the demand for health services and the level of insurance. However, such analyses are subject to the criticism that the level of health insurance should not be treated as exogenous. Theoretical analyses show that the demand for health insurance and health care are jointly determined, so causation is not unidirectional. This fact can potentially make it difficult to identify the role of health insurance. Treating health insurance as exogenous biases the estimate of price responsiveness. However, in an experimental setup the participating households could be assigned an insurance policy, making it an exogenous variable. The role of insurance is then identifiable. Once the key variable of interest is exogenized, the direction of causation becomes clear and the impact of

the treatment can be studied unambiguously. Furthermore, if the experiment is free from some of the problems that we mention in the following, this greatly simplifies statistical analysis relative to what is often necessary in survey data.

3.3.3. Limitations of Social Experiments

The application of a nonhuman methodology, initially that is, one developed for and applied to nonhuman subjects, to human subjects has generated a lively debate in the literature. See especially Heckman and Smith (1995), who argue that many social experiments may suffer from limitations that apply to observational studies. These issues concern general points such as the merits of experimental versus observational methodology, as well as specific issues concerning the biases and problems inherent in the use of human subjects. Several of the issues are covered in more detail in later chapters but a brief overview follows.

Social experiments are very costly to run. Sometimes, perhaps often, they do not correspond to "clean" randomized trials. Hence the results from such experiments are not always unambiguous and easily interpretable, or free from biases. If the treatment variable has many alternative settings of interest, or if extrapolation is an important objective, then a very large sample must be collected to ensure sufficient data variation and to precisely gauge the effect of treatment variation. In that case the cost of the experiment will also increase. If the cost factor prevents a large enough experiment, its utility relative to observational studies may be questionable; see the papers by Rosen and Stafford in Hausman and Wise (1985).

Unfortunately the design of some social experiments is flawed. Hausman and Wise (1985) argue that the data from the New Jersey negative income tax experiment was subject to endogenous stratification, which they describe as follows:

> . . . [T]he reason for an experiment is, by randomization, to eliminate correlation between the treatment variable and other determinants of the response variable that is under study. In each of the income-maintenance experiments, however, the experimental sample was selected in part on the basis of the dependent variable, and the assignment to treatment versus control group was based in part on the dependent variable as well. In general, the group eligible for selection – based on family status, race, age of family head, etc. – was stratified on the basis of income (and other variables) and persons were selected from within the strata. (Hausman and Wise, 1985, pp. 190–191)

The authors conclude that, in the presence of endogenous stratification, unbiased estimation of treatment effects is not straightforward. Unfortunately, a fully randomized trial in which treatment assignment within a randomly selected experimental group from the population is independent of income would be much more costly and may not be feasible.

There are several other issues that detract from the ideal simplicity of a randomized experiment. First, if experimental sites are selected randomly, cooperation of administrators and potential participants at that site would be required. If this is not forthcoming, then alternative treatment sites where such cooperation is obtainable

will be substituted, thereby compromising the random assignment principle; see Hotz (1992).

A second problem is that of sample selection, which is relevant because participation is voluntary. For ethical reasons there are many experiments that simply cannot be done (e.g., random assignment of students to years of education). Unlike medical experiments that can achieve the gold standard of a double-blind protocol, in social experiments experimenters and subjects know whether they are in treatment or control groups. Furthermore, those in control groups may obtain treatment, (e.g., training) from alternative sources. If the decision to participate is uncorrelated with either $\mathbf{x}$ or ε, the analysis of the experimental data is simplified.

A third problem is sample attrition caused by subjects dropping out of the experiment after it has started. Even if the initial sample was random the effect of nonrandom attrition may well lead to a problem similar to the attrition bias in panels. Finally, there is the problem of **Hawthorne effect**. The term originates in social psychology research conducted jointly by the Harvard Graduate School of Business Administration and the management of the Western Electric Company at the latter's Hawthorne works in Chicago from 1926 to 1932. Human subjects, unlike inanimate objects, may change or adapt their behavior while participating in the experiment. In this case the variation in the response observed under experimental conditions cannot be attributed solely to treatment.

Heckman and Smith (1995) mention several other difficulties in implementing a randomized treatment. Because the administration of a social experiment involves a bureaucracy, there is a potential for biases. **Randomization bias** occurs if the assignment introduces a systematic difference between the experimental participant and the participant during its normal operation. Heckman and Smith document the possibilities of such bias in actual experiments. Another type of bias, called **substitution bias**, is introduced when the controls may be receiving some form of treatment that substitutes for the experimental treatment. Finally, analysis of social experiments is inevitably of a partial equilibrium nature. One cannot reliably extrapolate the treatment effects to the entire population because the *ceteris paribus* assumption will not hold when the entire population is involved.

Specifically, the key issue is whether one can extrapolate the results from the experiment to the population at large. If the experiment is conducted as a pilot program on a small scale, but the intention is to predict the impact of policies that are more broadly applied, then the obvious limitation is that the pilot program cannot incorporate the broader impact of the treatment. A broadly applied treatment may change the economic environment sufficiently to invalidate the predictions from a partial equilibrium setup. So the treatment will not be like the actual policy that it mimics.

In summary, social experiments, in principle, could yield data that are easier to analyze and to understand in terms of cause and effect than observational data. Whether this promise is realized depends on the experimental design. A poor experimental design generates its own statistical complications, which affect the precision of the conclusions. Social experiments differ fundamentally from those in biology and agriculture because human subjects and treatment administrators tend to be both active and forward-looking individuals with personal preferences, rather than

Table 3.2. *Features of Some Selected Natural Experiments*

Experiment	Treatments Studied	Reference
Outcomes for identical twins with different schooling levels	Differences in returns to schooling through correlation between schooling and wages	Ashenfelter and Krueger (1994)
Transition to National Health Insurance in Canada as Sasketchwan moves to NHI and other states follow several years later	Labor market effects of NHI based on comparison of provinces with and without NHI	Gruber and Hanratty (1995)
New Jersey increases minimum wage while neighboring Pennsylvania does not	Minimum wage effects on employment	Card and Krueger (1994)

passive administrators of a standard protocol or willing recipients of randomly assigned treatment.

3.4. Data from Natural Experiments

Sometimes, however, a researcher may have available data from a "**natural experiment**." A natural experiment occurs when a subset of the population is subjected to an exogenous variation in a variable, perhaps as a result of a policy shift, that would ordinarily be subject to endogenous variation. Ideally, the source of the variation is well understood.

In microeconometrics there are broadly two ways in which the idea of a natural experiment is exploited. For concreteness consider the simple regression model

$$y = \beta_1 + \beta_2 x + u, \tag{3.4}$$

where x is an endogenous treatment variable correlated with u.

Suppose that there is an exogenous intervention that changes x. Examples of such external intervention are administrative rules, unanticipated legislation, natural events such as twin births, weather-related shocks, and geographical variation; see Table 3.2 for examples. Exogenous intervention creates an opportunity for evaluating its impact by comparing the behavior of the impacted group both pre- and postintervention, or with that of a nonimpacted group postintervention. That is, "natural" comparison groups are generated by the event that facilitates estimation of the β_2. Estimation is simplified because x can be treated as exogenous.

The second way in which a natural experiment can assist inference is by generating natural instrumental variables. Suppose z is a variable that is correlated with x, or perhaps causally related to x, and uncorrelated with u. Then an **instrumental variable** estimator of β_2, expressed in terms of sample covariances, is

$$\widehat{\beta}_2 = \frac{\text{Cov}[z, y]}{\text{Cov}[z, x]} \tag{3.5}$$

(see Section 4.8.5). In an observational data setup an instrumental variable with the right properties may be difficult to find, but it could arise naturally in a favorable natural experiment. Then estimation would be simplified. We consider the first case in the next section; the topic of naturally generated instruments will be covered in Chapter 25.

3.4.1. Natural Exogenous Interventions

Such data are less expensive to collect and they also allow the researcher to evaluate the role of some specific factor in isolation, as in a controlled experiment, because "nature" holds constant variations attributed to other factors that are not of direct interest. Such natural experiments are attractive because they generate treatment and control groups inexpensively and in a real-world setting. Whether a natural experiment can support convincing inference depends, in part, on whether the supposed natural intervention is genuinely exogenous, whether its impact is sufficiently large to be measurable, and whether there are good treatment and control groups. Just because a change is legislated, for example, does not mean that it is an exogenous intervention. However, in appropriate cases, opportunistic exploitation of such data sets can yield valuable empirical insights.

Investigations based on natural experiments have several potential limitations whose importance in any given study can only be assessed through a careful consideration of the relevant theory, facts, and institutional setting. Following Campbell (1969) and Meyer (1995), these are grouped into limitations that affect a study's internal validity (i.e., the inferences about policy impact drawn from the study) and those that affect a study's external validity (i.e., the generalization of the conclusions to other members of the population).

Consider an investigation of a policy change in which conclusions are drawn from a comparison of pre- and postintervention data, using the regression method briefly described in the following and in greater detail in Chapter 25. In any study there will be omitted variables that may have also changed in the time interval between policy change and its impact. The characteristics of sampled individuals such as age, health status, and their actual or anticipated economic environment may also change. These omitted factors will directly affect the measured impact of the policy change. Whether the results can be generalized to other members of the population will depend on the absence of bias due to nonrandom sampling, existence of significant interaction effects between the policy change and its setting, and an absence of the role of historical factors that would cause the impact to vary from one situation to another. Of course, these considerations are not unique to data from natural experiments; rather, the point is that the latter are not necessarily free from these problems.

3.4.2. Differences in Differences

One simple regression method is based on a comparison of outcomes in one group before and after a policy intervention. For example, consider

$$y_{it} = \alpha + \beta D_t + \varepsilon_{it}, \quad i = 1, \ldots, N, \; t = 0, 1,$$

where $D_t = 1$ in period 1 (postintervention), $D_t = 0$ in period 0 (preintervention), and y_{it} measures the outcome. The regression estimated from the pooled data will yield an estimate of policy impact parameter β. This is easily shown to be equal to the average difference in the pre- and postintervention outcome,

$$\begin{aligned}\widehat{\beta} &= N^{-1}\sum\nolimits_i (y_{i1} - y_{i0}) \\ &= \overline{y}_1 - \overline{y}_0.\end{aligned}$$

The one-group before and after design makes the strong assumption that the group remains comparable over time. This is required for identifiability of β. If, for example, we allowed α to vary between the two periods, β would no longer be identified. Changes in α are confounded with the policy impact.

One way to improve on the previous design is to include an additional untreated comparison group, that is, one not impacted by policy, and for which the data are available in both periods. Using Meyer's (1995) notation, the relevant regression now is

$$y_{it}^j = \alpha + \alpha_1 D_t + \alpha^1 D^j + \beta D_t^j + \varepsilon_{it}^j, \quad i = 1, \ldots, N, \; t = 0, 1,$$

where j is the group superscript, $D^j = 1$ if j equals 1 and $D^j = 0$ otherwise, $D_t^j = 1$ if both j and t equal 1 and $D_t^j = 0$ otherwise, and ε is a zero-mean constant-variance error term. The equation does not include covariates but they can be added, and those that do not vary are already subsumed under α. This relation implies that, for the treated group, we have preintervention

$$y_{i0}^1 = \alpha + \alpha^1 D^1 + \varepsilon_{i0}^1$$

and postintervention

$$y_{i1}^1 = \alpha + \alpha_1 + \alpha^1 D^1 + \beta + \varepsilon_{i1}^1.$$

The impact is therefore

$$y_{i1}^1 - y_{i0}^1 = \alpha_1 + \beta + \varepsilon_{i1}^1 - \varepsilon_{i0}^1. \tag{3.6}$$

The corresponding equations for the untreated group are

$$y_{i0}^0 = \alpha + \varepsilon_{i0}^0,$$

$$y_{i1}^0 = \alpha + \alpha_1 + \varepsilon_{i1}^0,$$

and hence the difference is

$$y_{i1}^1 - y_{i0}^0 = \alpha_1 + \varepsilon_{i1}^0 - \varepsilon_{i0}^0. \tag{3.7}$$

Both the first-difference equations include the period-1 specific effect α_1, which can be eliminated by taking the difference between Equations (3.6) and (3.7):

$$\left(y_{i1}^1 - y_{i0}^1\right) - \left(y_{i1}^0 - y_{i0}^0\right) = \beta + \left(\varepsilon_{i1}^1 - \varepsilon_{i0}^1\right) - \left(\varepsilon_{i1}^0 - \varepsilon_{i0}^0\right). \tag{3.8}$$

Assuming that $\mathrm{E}[(\varepsilon_{i1}^1 - \varepsilon_{i0}^1) - (\varepsilon_{i1}^0 - \varepsilon_{i0}^0)]$ equals zero, we can obtain an unbiased estimate of β by the sample average of $(y_{i1}^1 - y_{i0}^1) - (y_{i1}^0 - y_{i0}^0)$. This method uses

differences in differences. If time-varying covariates are present, they can be included in the relevant equations and their differences will appear in the regression equation (3.8).

For simplicity our analysis ignored the possibility that there remain observable differences in the distribution of characteristics between the treatment and control groups. If so, then such differences must be controlled for. The standard solution is to include such controlling variables in the regression.

An example of a study based on a natural experiment is that of Ashenfelter and Krueger (1994). They estimate the returns to schooling by contrasting the wage rates of identical twins with different schooling levels. In this case running a regular experiment in which individuals are exogenously assigned different levels of schooling is simply not feasible. Nonetheless, some experimental-type controls are needed. As the authors explain:

> Our goal is to ensure that the correlation we observe between schooling and wage rates is not due to a correlation between schooling and a worker's ability or other characteristics. We do this by taking advantage of the fact that monozygotic twins are genetically identical and have similar family backgrounds.

Data on twins have served as a basis for a number of other econometric studies (Rosenzweig and Wolpin, 1980; Bronars and Grogger, 1994). Since the twinning probability in the population is not high, an important issue is generating a sufficiently large representative sample, allowing for some nonresponse. One source of such data is the census. Another source is the "twins festivals" that are held in the United States. Ashenfelter and Krueger (1994, p. 1158) report that their data were obtained from interviews conducted at the 16th Annual Twins Day Festival, Twinsburg, Ohio, August 1991, which is the largest gathering of twins, triplets, and quadruplets in the world.

The attraction of using the twins data is that the presence of common effects from both observable and unobservable factors can be eliminated by modeling the *differences* between the outcomes of the twins. For example, Ashenfelter and Krueger estimate a regression model of the difference in the log of wage rates between the first and the second twin. The first differencing operation eliminates the effects of age, gender, ethnicity, and so forth. The remaining explanatory variables are differences between schooling levels, which is the variable of main interest, and variables such as differences in years of tenure and marital status.

3.4.3. Identification through Natural Experiments

The natural experiments school has had a useful impact on econometric practice. By encouraging the opportunistic exploitation of quasi-experimental data, and by using modeling frameworks such as the POM of Chapter 2, econometric practice bridges the gap between observational and experimental data. The notions of parameter identification rooted in the SEM framework are broadened to include identification of measures that are interesting from a policy viewpoint. The main advantage of using data from a natural experiment is that a policy variable of interest might be validly treated as exogenous. However, in using data from natural experiments, as in the case of social

experiments, the choice of control groups plays a critical role in determining the reliability of the conclusions. Several potential problems that affect a social experiment, such as selectivity and attrition bias, will also remain potential problems in the case of natural experiments. Only a subset of interesting policy problems may lend themselves to analysis within the natural experiment framework. The experiment may apply only to a small part of the population, and the conditions under which it occurs may not replicate themselves easily. An example given in Section 22.6 illustrates this point in the context of difference in differences.

3.5. Practical Considerations

Although there has been an explosion in the number and type of microdata sets that are available, certain well-established databases have supported numerous studies. We provide a very partial list of some of very well known U.S. micro databases. For further details, see the respective Web sites for these data sets or the data clearinghouses mentioned in the following. Many of these allow you to download the data directly.

3.5.1. Some Sources of Microdata

Panel Study in Income Dynamics (PSID): Based at the Survey Research Center at the University of Michigan, PSID is a national survey that has been running since 1968. Today it covers over 40,000 individuals and collects economic and demographic data. These data have been used to support a wide variety of microeconometric analyses. Brown, Duncan and Stafford (1996) summarize recent developments in PSID data.

Current Population Survey (CPS): This is a monthly national survey of about 50,000 households that provides information on labor force characteristics. The survey has been conducted for more than 50 years. Major revisions in the sample have followed each of the decennial censuses. For additional details about this survey see Section 24.2. It is the basis of many federal government statistics on earnings and unemployment. It is also an important source of microdata that have supported numerous studies especially of labor markets. The survey was redesigned in 1994 (Polivka, 1996).

National Longitudinal Survey (NLS): The NLS has four original cohorts: NLS Older Men, NLS Young Men, NLS Mature Women, and NLS Young Women. Each of the original cohorts is a national yearly survey of over 5,000 individuals who have been repeatedly interviewed since the mid-1960s. Surveys collect information on each respondent's work experiences, education, training, family income, household composition, marital status, and health. Supplementary data on age, sex, etc. are available.

National Longitudinal Surveys of Youth (NLSY): The NLSY is a national annual survey of 12,686 young men and young women who where 14 to 22 years of age when they were first surveyed in 1979. It contains three subsamples. The data

provide a unique opportunity to study the life-course experiences of a large sample of young adults who are representative of American men and women born in the late 1950s and early 1960s. A second NLSY began in 1997.

Survey of Income and Program Participation (SIPP): SIPP is a longitudinal survey of around 8,000 housing units per month. It covers income sources, participation in entitlement programs, correlation between these items, and individual attachments to the job market over time. It is a multipanel survey with a new panel being introduced at the beginning of each calendar year. The first panel of SIPP was initiated in October 1983. Compared with CPS, SIPP has fewer employed and more unemployed persons.

Health and Retirement Study (HRS): The HRS is a longitudinal national study. The baseline consists of interviews with members of 7,600 households in 1992 (respondents aged from 51 to 61) with follow-ups every two years for 12 years. The data contain a wealth of economic, demographic, and health information.

World Bank's Living Standards Measurement Study (LSMS): The World Bank's LSMS household surveys collect data "on many dimensions of household well-being that can be used to assess household welfare, understand household behavior, and evaluate the effects of various government policies on the living conditions of the population" in many developing countries. Many examples of the use of these data can be found in Deaton (1997) and in the economic development literature. Grosh and Glewwe (1998) outline the nature of the data and provide references to research studies that have used them.

Data clearinghouses: The Interuniversity Consortium for Political and Social Research (ICPSR) provides access to many data sets, including the PSID, CPS, NLS, SIPP, National Medical Expenditure Survey (NMES), and many others. The U.S. Bureau of Labor Statistics handles the CPS and NLS surveys. The U.S. Bureau of Census handles the SIPP. The U.S. National Center for Health Statistics provides access to many health data sets. A useful gateway to European data archives is the Council of European Social Science Data Archives (CESSDA), which provides links to several European national data archives.

Journal data archives: For some purposes, such as replication of published results for classroom work, you can get the data from journal archives. Two archives in particular have well-established procedures for data uploads and downloads using an Internet browser. The *Journal of Business and Economic Statistics* archives data used in most but not all articles published in that journal. The *Journal of Applied Econometrics* data archive is also organized along similar lines and contains data pertaining to most articles published since 1994.

3.5.2. Handling Microdata

Microeconomic data sets tend to be quite large. Samples of several hundreds or thousands are common and even those of tens of thousands are not unusual. The distributions of outcomes of interest are often nonnormal, in part because one is often dealing

with discrete data such as binary outcomes, or with data that have limited variation such as proportions or shares, or with truncated or censored continuous outcomes. Handling large nonnormal data sets poses some problems of summarizing and reporting the important features of data. Often it is useful to use one computing environment (program) for data extraction, reduction, and preparation and a different one for model estimation.

3.5.3. Data Preparation

The most basic feature of microeconometric analysis is that the process of arriving at the sample finally used in the econometric investigation is likely to be a long one. It is important to accurately document decisions and choices made by the investigator in the process of "cleaning up" the data. Let us consider some specific examples.

One of the most common features of sample survey data is **nonresponse** or partial response. The problems of nonresponse have already been discussed. Partial response usually means that some parts of survey questionnaires were not answered. If this means that some of the required information is not available, the observations in question are deleted. This is called **listwise deletion**. If this problem occurs in a significant number of cases, it should be properly analyzed and reported because it could lead to an unrepresentative sample and biases in estimation. The issue is analyzed in Chapter 27. For example, consider a question in a household survey to which high-income households do not respond, leading to a sample in which these households are underrepresented. Hence the end effect is no different from one in which there is a full response but the sample is not representative.

A second problem is *measurement error* in reported data. Microeconomic data are typically noisy. The extent, type, and seriousness of measurement error depends on the type of survey cross section or panel, the individual who responds to the survey, and the variable about which information is sought. For example, self-reported income data from panel surveys are strongly suspected to have serially correlated measurement error. In contrast, reported expenditure magnitudes are usually thought to have a smaller measurement error. Deaton (1997) surveys some of the sources of measurement error with special reference to the World Bank's *Living Standards Measurement Survey*, although several of the issues raised have wider relevance. The biases from measurement error depend on what is done to the data in terms of transformations (e.g., first differencing) and the estimator used. Hence to make informative statements about the seriousness of biases from measurement error, one must analyze well-defined models. Later chapters will give examples of the impact of measurement error in specific contexts.

3.5.4. Checking Data

In large data sets it is easy to have erroneous data resulting from keyboard and coding errors. One should therefore apply some elementary checks that would reveal the existence of problems. One can check the data before analyzing it by examining some

descriptive statistics. The following techniques are useful. First, use summary statistics (min, max, mean, and median) to make sure that the data are in the proper interval and on the proper scale. For instance, categorical variables should be between zero and one, counts should be greater than or equal to zero. Sometimes missing data are coded as −999, or some other integer, so take care not to treat these entries as data. Second, one should know whether changes are fractional or on a percentage scale. Third, use box and whisker plots to identify problematic observations. For instance, using box and whisker plots one researcher found a country that had negative population growth (owing to a war) and another country that had recorded investment as more than GDP (because foreign aid had been excluded from the GDP calculation). Checking observations before proceeding with estimation may also suggest normalizing transformations and/or distributional assumptions with features appropriate for modeling a particular data set. Third, screening data may suggest appropriate data transforms. For example, box and whisker plots and histograms could suggest which variables might be better modeled via a log or power transform. Finally, it may be important to check the scales of measurement. For some purposes, such as the use of nonlinear estimators, it may be desirable to scale variables so that they have roughly similar scale. Summary statistics can be used to check that the means, variances, and covariances of the variables indicate proper scaling.

3.5.5. Presenting Descriptive Statistics

Because microdata sets are usually large, it is essential to provide the reader with an initial table of descriptive statistics, usually mean, standard deviation, minimum, and maximum for every variable. In some cases unexpectedly large or small values may reveal the presence of a gross recording error or erroneous inclusion of an incorrect data point. Two-way scatter diagrams are usually not helpful, but tabulation of categorical variables (contingency tables) can be. For discrete variables histograms can be useful and for continuous variables density plots can be informative.

3.6. Bibliographic Notes

3.2 Deaton (1997) provides an introduction to sample surveys especially for developing economies. Several specific references to complex surveys are provided in Chapter 24. Becketti et al. (1988) investigate the importance of the issue of representativeness of the PSID.

3.3 The collective volume edited by Hausman and Wise (1985) contains several papers on individual social experiments including the RHIE, NIT, and Time-of-Use pricing experiments. Several studies question the usefulness of the experimental data and there is extensive discussion of the flaws in experimental designs that preclude clear conclusions. Pros and cons of social experiments versus observational data are discussed in an excellent pair of papers by Burtless (1995) and Heckman and Smith (1995).

3.4 A special issue of the *Journal of Business and Economic Statistics* (1995) carries a number of articles that use the methodology of quasi- or natural experiments. The collection includes an article by Meyer who surveys the issues in and the methodology of econometric

studies that use data from natural experiments. He also provides a valuable set of guidelines on the credible use of natural variation in making inferences about the impact of economic policies, partly based on the work of Campbell (1969). Kim and Singal (1993) study the impact of changes in market concentration on price using the data generated by a airline mergers. Rosenzweig and Wolpin (2000) review an extensive literature based on natural experiments such as identical twins. Isacsson (1999) uses the twins approach to study returns to schooling using Swedish data. Angrist and Lavy (1999) study the impact of class size on test scores using data from schools that are subject to "Maimonides' Rule" (briefly reviewed in Section 25.6), which states that class size should not exceed 40. The rule generates an instrument.

PART TWO
Core Methods

Part 2 presents the core estimation methods – least squares, maximum likelihood and method of moments – and associated methods of inference for nonlinear regression models that are central in microeconometrics. The material also includes modern topics such as quantile regression, sequential estimation, empirical likelihood, semiparametric and nonparametric regression, and statistical inference based on the bootstrap. In general the discussion is at a level intended to provide enough background and detail to enable the practitioner to read and comprehend articles in the leading econometrics journals and, where needed, subsequent chapters of this book. We presume prior familiarity with linear regression analysis.

The essential estimation theory is presented in three chapters. Chapter 4 begins with the linear regression model. It then covers at an introductory level quantile regression, which models distributional features other than the conditional mean. It provides a lengthy expository treatment of instrumental variables estimation, a major method of causal inference. Chapter 5 presents the most commonly-used estimation methods for nonlinear models, beginning with the topic of m-estimation, before specialization to maximum likelihood and nonlinear least squares regression. Chapter 6 provides a comprehensive treatment of generalized method of moments, which is a quite general estimation framework that is applicable for linear and nonlinear models in single-equation and multi-equation settings. The chapter emphasizes the special case of instrumental variables estimation.

We then turn to model testing. Chapter 7 covers both the classical and bootstrap approaches to hypothesis testing, while Chapter 8 presents relatively more modern methods of model selection and specification analysis. Because of their importance the computationally-intensive bootstrap methods are also the subject of a more detailed chapter, Chapter 11 in Part 3. A distinctive feature of this book is that, as much as possible, testing procedures are presented in a unified manner in just these three chapters. The procedures are then illustrated in specific applications throughout the book.

Chapter 9 is a stand-alone chapter that presents nonparametric and semiparametric estimation methods that place a flexible structure on the econometric model.

Chapter 10 presents the computational methods used to compute the nonlinear estimators presented in chapters 5 and 6. This material becomes especially relevant to the practitioner if an estimator is not automatically computed by an econometrics package, or if numerical difficulties are encountered in model estimation.