Modelling Individual Choice
The Econometrics of Corners, Kinks and Holes

Stephen Pudney

Basil Blackwell
Further reading

The subject matter of section 1.1 is covered by most advanced microeconomics textbooks, although indivisibility and preferences defined over characteristics are topics which usually receive only cursory treatment. Deaton and Muellbauer (1980a) and Simons (1974) are excellent general references. On the existence of utility functions, Debreu (1954) is the seminal paper, and his work is taken much further in the various contributions to Chipman et al. (1971). Kagel et al. (1975) describe an ingenious attempt to test the implications of the theory of rational choice by means of experiments with laboratory animals.

Deaton and Muellbauer (1980a, chapter 1) provide a useful discussion of many of the awkward opportunity sets discussed in section 1.2 and later chapters.

Sections 1.3–1.6 are covered by Deaton and Muellbauer (1980a, chapter 2); Simons (1974) is also useful, although it does not deal with the distance function. Gorman (1976), Blackorby, Primont and Russell (1978) and Diewert (1974) are valuable references on the use of duality theory in conventional optimizing models. Deaton (1986) contains an excellent survey of alternative functional forms for the various representations of utility and their role in applied demand analysis.

Deaton and Muellbauer (1980, chapter 8) and Deaton (1986) discuss the methods that are commonly used to incorporate the effects of household composition and other demographic variables in demand analysis.

Inherent randomness of preferences has been analysed by Block and Marschak (1960), Marschak (1960) and McFadden and Richter (1971), who give conditions for the existence of a stochastic utility indicator. See also McFadden (1982) for a survey of stochastic choice theory. However, the applied literature contains very little discussion of the choice of a stochastic specification; analytical convenience usually has to be the overriding consideration.

2.

Survey methods and cross-section econometrics

Let observation with extensive view
Survey mankind from China to Peru.

Samuel Johnson

Every day, thousands of people are asked thousands of questions about their behaviour, attitudes, habits, circumstances and opinions. Regular sample surveys are conducted by most governments to satisfy various national accounting needs: the construction of weights for the retail price index, the index of industrial production etc. and for a host of other policy-related purposes. Commercial market research surveys are conducted regularly to provide information on demand conditions for producers and to assess the effectiveness of advertising campaigns. Many research bodies and pressure groups carry out surveys to meet the needs of their research or supply evidence to support their case. A particularly interesting form of cross-section enquiry, carried out in several parts of the USA during the 1970s, is the social experiment, in which families' economic circumstances are not only observed but are also modified by subjecting them to some controlled change – an experimental form of negative income tax, for instance.

With the exception of these experimental enquiries, economists have played a surprisingly small part in this survey activity. It is still a comparatively rare for economists to commission a survey enquiry dedicated to the objectives of a specific piece of econometric research; it remains normal practice to make use of survey data which is collected for entirely different purposes, by methods over which the econometrician has no control. Although this is to be deplored, since it imposes severe constraints on the kind of research that is possible, we shall neglect the topic of optimal sample survey design and instead discuss the theory of sample surveys from the more realistic point of view of the econometrician who must use survey data gathered by a sampling technique which he or she is powerless to influence.

In section 2.1 we briefly survey the theory of common sampling methods.
Survey methods and cross-section econometrics

Section 2.2 deals with some of the difficulties associated with the practical form of the enquiry. Section 2.3 discusses a specific example: the UK Family Expenditure Survey. Sections 2.4 and 2.5 are concerned with the econometric implications of exogenous and endogenous biased sampling methods, and section 2.6 deals with the problem of non-response, attrition and other forms of extraneous sample truncation.

2.1 Sampling techniques

We begin with some definitions. The sampling unit is the individual subject of our analysis; sampling units may be single persons, households, firms etc. Sometimes there may be some problems of definition here: if an aunt comes to visit in the week of the survey, is she counted as a member of the household? These definitional problems must be resolved by the survey designer, often arbitrarily.

The relevant attributes of these sampling units we represent by two (vector) variables: \( x \) is the subject of our economic model, for instance expenditure on food, hours of labour supplied etc.; \( x \) is a vector containing all attributes which are to be used as explanatory variables - income, prices, wage rates, family size and structure, age etc. Thus \( x \) is endogenous and \( x \) exogenous to our model.

2.1.1 The population

The actual population is simply the full set of sampling units in existence: the group from which the sample is chosen. For example, the population of individual persons in the UK comprises some 56 million sampling units. In virtually all practical cases, this actual population has a finite number of members, and therefore the distribution of the observable variables \( x \) and \( x \) is discrete - there are at most 56 million distinct possible values for any set of individual attributes observed at any one time. This is so even for quantities such as income that are potentially continuously variable. We denote the number of units in the actual population by \( N \) and the empirical c.d.f. of \( x \) and \( x \) by \( G(x, Z) \), where \( G(x, Z) \) is the probability of observing \( x \) \( x \) \( Z \) in a unit drawn at random from the actual population.

Note that, for it to be possible to sample members of this population, it is necessary that there exist some practical means of identifying sampling units. This list of sampling units is known as the sample frame. Ideally, one would like the sample frame to cover the whole of the population that the enquiry attempts to survey. However, this is rarely achievable in practice; sample frames are generally based on lists assembled for other purposes - the electoral register, the telephone directory etc. These are often highly unsatisfactory, omitting certain types of individual in a systematic way, and the possible consequences of a poor sampling frame should always be borne in mind.

Most statistical sampling theory (see, for instance, Cochran, 1963; Kish, 1965) is concerned with the problem of estimating characteristics of the actual population: quantities such as the average age, the quantities of the income distribution, and so on. Our perspective is rather different, since we are interested in the construction of structural models. Such models are not designed to describe the actual population but rather to answer hypothetical questions: what would happen to the demand for bread if its price were to double? How would labour supply change in response to an increase in the tax rate? If such changes were actually to occur, there would be no reason why the stochastic factors affecting behaviour should remain fixed at the values they happen to take in the actual population. Thus the hypothetical worlds described by an econometric model have no particular relationship with the actual population that supplies the data from which it is estimated. To reflect this, we introduce the idea of a super-population consisting of a continuum of potential individuals. The actual sample is regarded as a set of \( N \) individuals drawn by (nature) at random from this super-population, which is characterized by a c.d.f. \( G(x, Z) \) \( x \leq \) \( Z \) or, equivalently, by a probability density/mass function \( g(x, Z) \), which is the function \( G(x, Z) \) partially differentiated with respect to all elements of \( x \) and \( x \) that are (locally) continuously variable and differentiated with respect to any elements (such as dummy variables) that are discrete.

2.1.2 Exogeneity and the statistical model

The econometric model that we seek to estimate from sample information attempts to 'explain' \( x \) in terms of \( x \), and thus amounts to an assumed parametric form for the distribution of \( x \) conditional on \( x \), in the super-population. Let the vector of unknown parameters \( \theta \) define the precise form of this distribution be \( \theta \), and write the model as \( g(x; \theta) \). For this to be an autonomous model (or, equivalently, for \( x \) to be weakly exogenous or ancillary to \( \theta \), the marginal distribution of \( x \) must be independent of \( x \), in which case it is uninformative for \( \theta \). Hence:

\[
g(x; \theta) = g(x; \theta) g(x; \theta) \quad (2.1)
\]

where we are using \( g(x; \theta) \) as generic notation for any population density/mass function.

The exogeneity assumption (2.1) is often questionable. A particular problem concerns the use of prices as explanatory variables, since prices are determined simultaneously with quantities within the market. Few applied workers give any justification for the assumption of exogenous prices, but there is an implicit assumption that, in a large market with a uniform average price, the price variable is approximately exogenous.

To illustrate the issues involved, consider a very simple market model with an actual population of \( N \) consumers (drawn at random from the super-
Survey methods and cross-section econometrics

(2.2)

\[ x_t = \gamma_0 + \gamma_1 p_t + \varepsilon_t \]

where \( p_t \) (the sole element of \( f \)) is the price faced by individual \( i \), and \( \varepsilon_t \) is a behavioural disturbance term, with an \( N(0, \sigma) \) distribution (say). The typical cross-section demand analysis proceeds by translating (2.2) into the assumption that \( x_t | p_t \) has an \( N(\gamma_0 + \gamma_1 p_t, \sigma) \) in the population. The parameters of interest are thus \( \theta = (\gamma_0, \gamma_1, \sigma) \).

Assume now that there is a single producer, operating under constant costs and having perfect knowledge of all aspects of the market. The retail system is such that individual prices \( p_t \) deviate randomly and uncontrollably from the mean price \( P \) fixed by the producer:

\[ p_t = P + \varepsilon_t \]

(2.3)

where \( \varepsilon_t \) is normally distributed with zero mean and variance \( \sigma^2 \). Assume that \( p_t \) and \( \varepsilon_t \) are independent. The producer’s profit function is

\[ \Pi(P) = \sum P - \sum (p_t + \varepsilon_t - c) \left( \gamma_0 + \gamma_1 (P + \varepsilon_t) + \varepsilon_t \right) \]

(2.4)

where \( c \) is unit cost, which we assume to be known. The profit-maximizing price is \( P = c - \bar{z}/\gamma_1 - \bar{z} \), where bars denote averages over all \( N \) consumers. Thus the actual price is

\[ p_t = c - \bar{z}/\gamma_1 - \bar{z} + \varepsilon_t \]

(2.5)

Equations (2.2) and (2.5) have the reduced form

\[ p_t = c - \bar{z}/\gamma_1 - \bar{z} + \varepsilon_t - \gamma_1 \]

(2.6)

\[ x_t = \bar{z}/\gamma_1 - \bar{z} + \varepsilon_t - \gamma_1 \bar{z} + \varepsilon_t \]

(2.7)

and these expressions imply that \( p_t \) and \( x_t \) have a bivariate normal distribution with means

\[ \nu_p = (c - \bar{z}/\gamma_1 - \bar{z})/\gamma_1 \]

(2.8)

\[ \nu_x = (\bar{z}/\gamma_1 - \bar{z})/\gamma_1 \]

(2.9)

and covariances

\[ \sigma_{p\varepsilon} = \sigma_p + \frac{1}{\gamma_1} \left( \frac{\sigma_p}{\gamma_1} - \gamma_1 \sigma_x \right) \]

(2.10)

\[ \sigma_{p\varepsilon} = \frac{\gamma_1^2 \sigma_p + \sigma_x}{\gamma_1} - \frac{3}{\gamma_1^2} \left( \gamma_1 \sigma_x + \sigma_x \right) \]

(2.11)

The marginal distribution of \( p_t \) is thus

\[ p_t \sim N(\mu_p, \sigma^2) \]

(2.12)

Since \( \sigma_p \) is a known function of \( \theta \), this distribution is informative for \( \theta \), and price is not exogenous in the sense of (2.1). We therefore lose information if we choose to work only with the conditional distribution of \( x_t \). More seriously, the conditional distribution of \( x_t \) has the following form:

\[ x_t | p_t \sim N(\mu_p, \sigma^2)(\sigma_p/\gamma_1, \sigma_x) \]

(2.13)

For finite \( N \), this is not the \( N(\gamma_0, \gamma_1, \sigma) \) distribution that equation (2.2) suggests, and simple estimation of (2.2) by regression methods leads to biased estimates of \( \gamma_0 \) and \( \gamma_1 \). However, this problem disappears as \( N \) becomes large. If the limiting forms of \( \sigma_p \) and \( \sigma_x \) are substituted into (2.14), the result is the \( N(\gamma_0, \gamma_1, \sigma) \) distribution that we would find if \( p_t \) were known to be exogenous.

The conclusion is therefore that, provided demand and supply side disturbances are independent and that the number of consumers is large,

i. prices can be regarded as exogenous for the purpose of constructing the conditional distribution associated with the econometric model, but

ii. the marginal distribution of price may contain information on the demand parameters - so statistical methods based only on the conditional distribution may be inefficient.

Note that these conclusions are quite general and do not depend critically on normality or on the details of our model of supply.

In the remainder of the book we shall usually ignore the endogeneity of prices, since the cost is only a theoretical efficiency loss. However, in a few cases this endogeneity problem cannot be ignored. If, for instance, the consumer can choose between different qualities of the good, so that price is a choice variable, the price disturbance \( \varepsilon_t \) is partly demand determined, and unlikely to be independent of \( x_t \). A second example concerns labour supply, where each individual commands a different specific wage, determined in part by his or her unobserved characteristics, which may also partly influence labour supply. For instance, an ambitious person is likely to supply more labour and be paid better than others, implying some correlation between the two random disturbances. In such cases, the price or wage must be regarded as endogenous, even if there are a large number of individuals in the market, and we are forced to estimate simultaneous models of demand and price. With these caveats in mind, we now return to the assumption that our explanatory variables are exogenous in the population (in the strong sense
of (2.1) and consider the methods of sampling that underlie cross-section econometrics.

Sampling is a process by which we draw a subset of \( N \) individuals from the actual population so that their observed \( x \) values can be used as the basis for inferences about the nature of the distribution \( g(x; \theta) \), which is formally a feature of the super-population. There are several common methods of drawing a sample from a given sampling frame. In discussing these, we shall consider only the case where the size of sample to be drawn, \( N \), is fixed. This is somewhat restrictive, since it excludes sequential sampling, in which successive observations are selected until some fixed criterion is reached, yielding a random sample of random size. However, such sampling schemes are unusual in economic and social surveys. We shall adhere to the notation already established: distributions relating to the super-population have c.d.f. and p.d.f. written \( G(x) \) and \( g(x) \); the symbols \( F(x) \) and \( f(x) \) are used for distributions relating to the sample rather than the population.

### Simple Random Sampling

Much of statistical theory is orientated by the concept of the random sample. A random sample of size \( N \) is a set of observations \( (x_1, \ldots, x_N) \) chosen in such a way that their joint distribution, which has probability density/mass function \( f(x_1, \ldots, x_N; \theta) \), displays two properties:

1. The \( N \) pairs \( (x_i, \xi_i) \) are independently distributed:

   \[
   f(x_1, \ldots, x_N; \xi_1, \ldots, \xi_N) = \prod_{i=1}^{N} f_i(x_i, \xi_i) \tag{2.15}
   \]

   where \( f_i(\cdot, \cdot) \) is the marginal density/mass function of \( (x_i, \xi_i) \).

2. Sampling is unbiased, in the sense that each observation has the same distribution as the super-population itself:

   \[
   \begin{align*}
   f(x_1, \ldots, x_N; \xi_1, \ldots, \xi_N) &= \prod_{i=1}^{N} g(x_i, \xi_i) \\
   &= \prod_{i=1}^{N} g_{\xi_i}(x_i; \theta) g(\xi_i) \tag{2.16}
   \end{align*}
   \]

The term random sample is a very misleading one, since there are many possible methods of sampling which are random in character but do not possess these two properties; a better term would be unbiased independent sampling.

A random sample has tremendous analytical advantages. Because its distribution perfectly mirrors that of the population, statistical inference from a random sample is particularly simple and direct. However, despite this simplicity, and despite its prominent place in statistical theory, random samples are almost never encountered in practice. There are two main reasons for this.

Random samples are often a very inefficient way of collecting information which is representative of (as opposed to having the same theoretical distribution as) the population. For instance, if the survey designer is required to produce a sample with adequate representation of a number of subgroups in the population (e.g. workers in different industries, or consumers in different countries), it may be necessary to draw a very large and therefore expensive sample. Other sampling techniques can usually provide a much cheaper means of meeting the objectives of the body commissioning the survey.

A second reason is that the selection of a simple random sample requires the sample frame to cover the whole population and the method of choosing units from the frame to be by repeated random drawing. It is very unusual for a survey to be based on a perfect sampling frame. For instance, the electoral register omits the young and the transient, and therefore even if individuals were drawn at random from the electoral register the resulting sample would not be random in the present sense. In any case, many practical sampling schemes do not make repeated random drawings from the frame but instead use simpler methods such as interval sampling, which is discussed below.

### Exhaustive Sampling

Since the actual population is finite, it is a principle possible to adopt a policy of complete enumeration or exhaustive sampling, where the sample and actual population are identical. Many countries attempt to do this for the population of individual people by conducting a census every ten years, covering mainly demographic attributes. However, smaller-scale exhaustive samples are often conducted when the target population is small or when the required information is already gathered for some other administrative purpose. Unemployment statistics, for example, are usually compiled from an exhaustive sample (of the population of unemployed individuals), which is constructed for the purposes of paying unemployment benefits.

Exhaustive samples present no statistical problems, since they can be treated as simple random samples from the super-population. However, there are severe practical problems in most cases. Complete enumeration is usually very expensive, since it requires an enormous number of individual enquiries, and it therefore rather rare. Moreover, census data usually only become available after a long preparation delay, so that it is impossible to produce up-to-date econometric results. The latter problem is often avoided by the early publication of data for a small random sample of the full census, thus using census returns as a sampling frame for a smaller random sample. Perhaps the most serious problem with large exhaustive samples is that they simply generate too much data — the analysis of detailed statistical models
Survey methods and cross-section econometrics

using millions of observations is a practical impossibility, and one is forced either to use subsamples or to work with averaged data.

2.1.5 Interval sampling

Interval sampling is a particularly simple and widely used method of selecting a sample of units from a given sampling frame. Denote by \( i \in \{1, \ldots, N^*\} \) the position of the typical sampling unit in the frame. In order to establish a relationship between the ordering of units in the frame and the distribution of individuals in the super-population, define also a continuous random variable \( u \in (-\infty, \infty) \) which underlies this ordering. Thus we suppose that 'nature' selects \( N^* \) units from the super-population and arranges these within the frame in ascending \( u \) order. Write the probability density/mass function of \( x \), \( \xi \) and \( u \) in the super-population as

\[
g(x, \xi, u) = g(x, \xi|u)g(u).
\]

(2.17)

Interval sampling proceeds as follows: define \( k \) as the largest integer such that \( k \leq N^*/N \); draw the first observation at random from the first \( k \) units in the sampling frame (suppose that this turns out to be the \( i_k \); then draw the remaining observations from positions \( i_k = i_k + 1, i_k = i_k + 2k, \ldots, i_k = i_k + (N^*/N - 1)k \).

If viewed as a subsample of the actual population, this procedure clearly does not satisfy the requirements of random sampling. In particular, the observations are not independently distributed, conditional on the sampling frame, since \( (x_k, \xi_k), \ldots, (x_{N^*}, \xi_{N^*}) \) are completely determined once the first observation is selected. However, this is necessarily so only if the frame is taken as given, whereas we are treating the units in the frame as a random sample from a super-population.

The distribution of the variables \( x, \xi, u \) and \( s \) for the first sample observation has probability density/mass function

\[
f(x_k, \xi_k, u_k, s_k) = \frac{1}{k} \left[ \prod_{i=1}^{k} g(x_i, \xi_i|u_i) \right] g^{\xi}(u_k)\]

(2.18)

where \( g^{\xi}(u) \) is the p.d.f. of the \( i \)th smallest \( u \) value in a random sample of size \( N^* \) from the super-population. The joint distribution of the first two observations is similarly characterized

\[
f(x_k, \xi_k, u_k, s_k, x_{k+1}, \xi_{k+1}, u_{k+1}) = \frac{1}{k^2} \left[ \prod_{i=1}^{k} g(x_i, \xi_i|u_i) \right] g^{\xi}(u_k)g^{\xi}(u_{k+1}) f(x_{k+1}, \xi_{k+1}, u_{k+1}, s_{k+1})
\]

(2.19)

where \( g^{\xi}(\ldots) \) is the p.d.f. of the \( i \)th and \( i+1 \)th smallest \( u \) values in a random sample of size \( N^* \) from the super-population. Continuing in this way for the full sample

\[
f(x_k, \xi_k, u_k, s_k, \ldots, x_{N^*}, \xi_{N^*}, u_{N^*}) = \frac{1}{k^{N^*/N}} \left[ \prod_{i=1}^{N^*/N} g(x_i, \xi_i|u_i) \right] g^{\xi}(u_{N^*/N})
\]

Then, summing over \( i_k \) and integrating with respect to \( u_k \ldots u_{N^*} \)

\[
f(x_k, \xi_k, \ldots, x_{N^*}, \xi_{N^*}) = \frac{1}{k^{N^*/N}} \left[ \prod_{i=1}^{N^*/N} g(x_i, \xi_i|u_i) \right] g^{\xi}(u_{N^*/N})
\]

(2.21)

In general, this joint distribution for the observables is very complicated and not the same as that which is yielded by the random sample.

However, provided that the mechanism for ordering the units within the sampling frame is independent of the mechanism generating \( x \) and \( \xi \), the conditional distribution \( g(x, \xi|u) \) does not depend on \( u \) and is identical with the marginal distribution \( g(x, \xi) = g(x|\xi)g(\xi) \). In that case, (2.21) reduces to the simple form (2.1) since \( g^{\xi}(u_k) \ldots g^{\xi}(u_{N^*/N}) \) integrates to unity. Therefore, if the organization of the sample frame is independent of \( x \) and \( \xi \), interval sampling can be treated as random sampling for the purposes of drawing inferences about the super-population. Usually it is safe to make this independence assumption.

2.1.6 Stratified sampling

For stratified sampling, the population is divided into \( m \) sub-populations, or strata. Suppose that these strata are defined in terms of a variable (or variables) denoted \( s \), and that the joint density/mass function of \( x, \xi \) and \( s \) in the super-population is \( g(x, \xi|s) \), which can be decomposed as follows:

\[
g(x, \xi, s) = g(x|s, \xi|s)g(\xi|s)
\]

(2.22)

where \( g(x|s, \xi|s) \) is the essence of our (now extended) econometric model. The typical stratum \( j \) is defined as that part of the population for which \( s = s_j \), where \( s_j, s_{j+1}, \ldots, s_{j+m-1} \) are non-overlapping regions in the domain of variation of \( s \). In practice, these strata represent particular groups of individuals that the survey design wishes to see adequately represented in the sample. Often some of these groups are quite small and therefore unlikely to be well represented in a moderately sized random sample but are deemed to be of particular interest—single parent families, old people, the poor, etc. Note that some of the criteria used for defining these strata may appear in the \( x \) or \( \xi \) vectors; for instance, the poor are a group defined in terms of income, which is a quantity that is also likely to be used as an endogenous or exogenous variable in an econometric study. In practice, the sub-populations are often
Survey methods and cross-section econometrics

is independent of $\theta$. This is an example of exogenous sampling, discussed further in section 2.3.

Since the sample and population densities are not identical under stratified sampling, inference is sometimes less straightforward than for random sampling. For instance, consider the problem of estimating the superpopulation mean $E(x)$. The ordinary sample mean has expected value

$$E(\bar{x}) = \frac{1}{N} \sum_{j=1}^{J} N_j E(x_j)$$

$$= \frac{1}{N} \sum_{j=1}^{J} N_j \sum_{k=1}^{N_j} E(x_{jk} | s_j)$$

$$= \frac{1}{N} \sum_{j=1}^{J} N_j \sum_{k=1}^{N_j} E(x_{jk} | s_j).$$

(2.29)

However, since $E(x) = \sum_{j=1}^{J} P_j E(x_{j})$, $\bar{x}$ is generally biased unless the ratios $N_j/N$ coincide with the population proportions $P_j$. This suggests the use of reweighted data, $x^* = P_j N_j x_{j} / N$, where $j$ is the stratum to which observation $x$ belongs. The weighted mean $x^* = N^{-1} \sum_{j=1}^{J} N_j x^*_j$ is unbiased, and for this reason survey data are often made available in weighted form.

2.1.7 Multistage sampling

Ordinary stratification as described above involves the drawing of observations from all the strata. This can be thought of as exhaustive sampling of the population of strata, followed by non-exhaustive sampling within strata. In practice, this first stage is often not exhaustive, and instead samples are drawn only from a subset of the available strata (which are often called primary sampling units or PSUs in this context). For the simple two-stage case, a typical sampling process might be as follows.

i. Draw a sample of $M < m$ strata from the set $\{S_1, \ldots, S_p\}$. Assume that this is done by making repeated independent drawings with probabilities $\Pi_1, \ldots, \Pi_p$ of selection for the strata, where the $\Pi_i$ are chosen as per the design of the sample.

ii. From each stratum that is selected at the first stage, draw $k = N/M$ sampling units by random sampling.

This process yields observations on $x$ and $y$ that are not generally independent (unless the stratification variable $s$ is independent of both $x$ and $y$). The marginal distribution of any single observation is

$$f(x, y) = \sum_{j=1}^{J} P(\text{stratum } j \text{ selected}) g(x, y | s_j).$$

(2.28)
Survey methods and cross-section econometrics

\[
\sum_{j} \prod_{i} g(x_i, t_j) = S_{j}
\]

\[
\sum_{j} \prod_{P_j} \prod_{t_j} g(x_i, t_j) g(z_i) dG(s).
\]

However, the joint distribution of the full sample is more complicated. Define \(N_j\) as the index set of observations drawn from the \(r\) th stratum that is sampled (thus, if the observations are ordered by strata, \(N_j\) is the set \(\{k(r-1)+1, \ldots, kr\}\)). Then

\[
f(x_1, \ldots, x_n, z_0) = \prod_{i} \text{Joint p.d.f. of observations in the set } N_j\]

\[
= \prod_{i} \left( \sum_{j} \prod_{P_j} \prod_{t_j} g(x_i, t_j; z_0; \theta) g(z_0) dG(s) \right).
\]

Note that, in general, the inner summation and product operations cannot be interchanged, and so the joint distribution cannot be expressed in the product form \(\prod f_i(x_i, z_0)\). Therefore the observations are not independently distributed.

In the special case where the strata are defined independently of \(x\), the conditional distribution \(g(x; \theta)\) does not depend on \(s\), and (2.31) simplifies to

\[
f(x_1, \ldots, x_n, z_0) = \prod_{i} g(x_i, z_0; \theta) f(z_0, \ldots, z_0).
\]

where

\[
f(z_0, \ldots, z_0) = \prod_{i} \left( \sum_{j} \prod_{P_j} \prod_{t_j} g(z_i) dG(s) \right).
\]

Thus, in the case of exogenously defined strata, the observations on the endogenous variables are independent conditional on \(z_0, \ldots, z_0\), but the sample as a whole is not fully independent, unless \(z\) and \(s\) are independently distributed in the population. If, as is often the case in practice, the strata are sampled without replacement, then the marginal distribution of \(z_0, \ldots, z_0\) becomes more complicated still, but the general form (2.32) still applies. As with ordinary stratification, problems arise in the estimation of population means, and to achieve unbiasedness observations must be reweighted by the factors \(P_j/1\).

Large-scale surveys generally make use of rather complex schemes involving both multiple levels of ordinary stratification and multistage sampling, sometimes with interval sampling used at the lowest level. The extension of our theory to such cases is straightforward in principle.

2.2 The survey enquiry

The way in which one asks a question often has a marked effect on the answer one receives. This is a fundamental truth in survey analysis, and the design of the survey questionnaire, the phrasing of questions, the duration of the period of observation etc., are at least as important to the user of survey data as the formal sampling technique used. Since these are essentially practical matters, there is only a limited amount that can be said that is of very wide validity; every field poses its own special problems for the survey designer.

One of the potential difficulties of the survey method is that the very act of observation may change the behaviour that is under study. Someone who is asked detailed questions about his or her current economic decisions may be forced into a re-evaluation of the way those decisions are made, and may become more cautious or moderate as a result. Alternatively, the attention of an external observer may cause the survey to be conducted in a way that is different from a normal survey. There is obviously little that can be done to investigate this suspicion: if the results of observation are suspect, the 'true' behaviour which is operative only in the absence of observation is unknowable, and there is no direct means of checking the validity of any given set of survey data. Thus most of the evidence on either side of this question is indirect and often anecdotal. However, one of the Family Budget Survey experiments carried out by the UK Office of Population Censuses and Surveys (see Kemsley, Redpath and Holmes, 1980) in 1976 did attempt to gain some information on this by asking households who had been interviewed for the Family Budget Survey (see section 2.6) whether they thought that keeping detailed expenditure records had affected their normal spending in any way. Of those questioned in this follow-up survey, 12.7 per cent believed that there had been some change, and this does not suggest that observation-induced bias is an especially serious problem.

Even if behaviour is itself unaffected by the survey, errors may still arise through various forms of misreporting. People may forget to record all the relevant events, or they may deliberately give inaccurate responses in order to conceal aspects of their behaviour that might seem shameful or that are felt to be too personal. This is particularly true of surveys covering sensitive areas such as drinking habits, sexual behaviour, participation in the black economy etc. Some individuals express the fear that complete information might be
Survey methods and cross-section econometrics

day than respondents claimed would be read on a typical day. Any econometrician seeking to model these frequency data would be led seriously astray by a repeating bias of this magnitude.

Survey enquiries are of two possible types: those that attempt to record activity as it occurs, usually by requiring that subjects keep continuous records of their behaviour for some period (the diary method): those that obtain information by retrospective interview, where the subject is asked to produce a record of past behaviour from memory (the recall method). These both have advantages and disadvantages and are best suited to different fields of enquiry. Some surveys use a combination of the two methods.

The diary method relies little on memory and can therefore be used to elicit information in much more detail. Its disadvantages are that it imposes a much greater burden on the respondent, and it is consequently hard to achieve a good response rate, particularly if the observation period is longer than one or two weeks. Moreover, because it is so intrusive, it risks causing some change in behaviour, which a retrospective enquiry cannot do.

The recall method usually produces much better co-operation, is cheaper to administer (and can therefore produce larger samples) but is highly suspect when the behaviour involved comprises numerous small-scale activities such as the purchase of household goods. In these circumstances, retrospective enquiry tends to produce underestimates of the level of activity, since minor events tend to be forgotten. The recall method is more better suited to surveys of infrequent events of a substantial magnitude, such as the purchase of major durables or the timing and duration of spells of unemployment.

In these areas, the period of time covered by the retrospective interview can be much longer than would be feasible for a continuous diary-based enquiry.

The duration of the period over which the individual is observed is of major importance, and practice varies widely. For example, one study conducted their household expenditure surveys with observation periods of one or two weeks, some a month, some (particularly in Eastern Europe) as long as a year. Economic theory is generally couched in terms of flows of time, but this is an abstraction: in reality goods are purchased in discrete amounts at irregular intervals, for instance. If we view the theoretical area of flow as a hypothetical average (holding external conditions constant) of this irregular sequence of actions, then there is some difficulty in interpreting the observed demand over, say, a two-week period as the true rate of demand. The two will differ by possibly large amounts which are entirely fortuitous - if someone happens to be interviewed in the week in which he or she buys a refrigerator, a naive interpretation of this as an observation of a rate of flow of one refrigerator a week is obviously wildly mistaken. As the length of the observation period is increased, the importance of these fortuitous errors diminishes, although a long observation period brings its own problems since external conditions (prices etc.) may change within the period. The length of the observation period increases...
Survey methods and cross-section econometrics
ists were above a critical level; various time-of-day electricity pricing experiments have changed the cost of electricity supply to some households (see Aigner, 1985).

There are two separate issues here. Since these programmes are of limited duration, households may not respond as fully as they might to the introduction of a similar permanent programme. Thus, quite apart from any dynamic adjustment problems, analysis of these experiments might give an unrepresentative picture of long-term behaviour, simply as a consequence of rational intertemporal choice. Responses to transitory hours are less than responses to permanent factors. This problem has been investigated by Burdless and Greenberg (1982) in the context of a labour supply equation estimated from data on subjects subjected to a negative income tax programme in the Seattle-Denver experiment. Simple linear equations were estimated separately for each of the programmes involving schemes with three and five years' duration, and large differences were found in the coefficients. Thus the transient nature of the schemes does seem to be a significant problem in the use of econometric methods to predict the consequences of any proposed permanent income support system.

The second difficulty that arises when the survey makes changes to individual circumstances is that it may take a long time before adjustment to the change is completed. For instance, to alter one's hours of work may necessitate finding a new job; to alter one's spending on rent will mean moving house. These responses are costly and time consuming, and will not be observed immediately. Depending on the timing of the observation, fitting an ordinary econometric model to a single cross-section generated by one of these experiments may yield very misleading results. In these circumstances, panel data allow the explicit modelling of the process of adjustment (possibly as a discrete change: moving house or changing job) and is therefore clearly preferable in principle, although it often raises almost as many econometric problems as it solves. The analysis of panel data is beyond the scope of this book.

2.3 An example: the UK Family Expenditure Survey

Many countries carry out regular surveys of household income and expenditure. In most cases, these surveys have their origins in the need to produce up-to-date weights for use in the construction of an aggregate retail price index, although now they are generally expected to meet many other needs, such as the measurement of welfare needs. The annual UK Family Expenditure Survey (FES) is fairly typical of these, and is used as the basis for a wide variety of econometric work.
2.3.1 The sample design

The basic sampling unit of the FES is the household. A household is defined as a group of one or more individuals (not necessarily related), who have the same address, whose meals are prepared together and who have exclusive use of at least one room. The sample design used in the FES is considerably more complicated than any of the simple techniques discussed in section 2.1 above; this is true of a vast majority of the sample surveys that are used in applied economic and social research.

There are four stages in the FES procedure; these are described in considerably greater detail in Kemsley, Redpath and Holmes (1980), on which this section is based. (Note that the design used since 1986 differs slightly from this.)

Stage 1 At the first stage, the whole of Great Britain (with the exclusion of some remote areas) is divided into 168 strata, which are the outcome of the application of three stratification factors: initially by region, then within region according to the degree of urbanization and the relative value of the private housing stock. Each of the strata is then divided into geographical areas belonging to different local administrative districts. Such an area is divided at random from each stratum, with each area’s selection probability equal to its share of the stratum’s total of registered voters.

Stage 2 Each area is composed of wards (the basic geographical unit of the electoral process), each ward being small enough to be covered by a single interviewer. Four of these wards are then drawn from each area (one for survey in each quarter of the year), using an adaptation of the interval sampling technique designed to give each ward a probability of selection proportional to the number of electors in that ward.

Stage 3 From each ward, sixteen addresses are selected by interval sampling from the electoral register. Thus (with some minor exceptions), an address has a probability of selection proportional to the number of electors resident there. If an address is ineligible because it is that of a hotel, public house, institution, vacant property etc., it is discarded and not replaced.

Stage 4 In 97 per cent of cases, an address contains only one household, and that household is selected for interview. For multi-household addresses, a further random sampling procedure is used to produce up to three households for interview.

2.3.2 The form of the enquiry

The FES is designed to gather information on general household characteris-
studies using data on respondents. It is possible to record some information for all non-respondents: the location of their dwelling and (from local authority records) its rateable value. This non-response is not a complete lack of information. Response tends to be relatively poor for households living in properties with high rateable values or situated in densely populated areas, and also varies seasonally, with a slight decline in co-operation in the fourth quarter of each year.

In census years, it is possible to go further than this: since participation in the census is compulsory, census information should be available for all households, including those who refused to cooperate with the FES. In 1971 and 1981, attempts were made to identify the census returns of all households that had been approached in the course of the 1971 and 1981 FES enquiries (see Kemmley, 1971, and Redpath, 1986); this proved possible in over 90 per cent of cases. For these, the responding and non-responding households were compared in terms of a number of general household characteristics recorded in the census. Significant differences were found for the age of the head of household (younger households being more likely to co-operate), for the number of children (the more children, the greater the likelihood of response), for employment status (much lower response for the self-employed) and for household size (better response for large households).

Unfortunately, census returns do not give information on income or expenditure, so it is not possible to find any direct evidence of a correlation between response and the variables which are most likely to be endogenous to an econometric study of individual behaviour. However, there are indications that wealthier and higher-spending households are less likely to respond: for example, multi-car households tend to have low response rates. Thus endogenous self-selection is definitely a potential difficulty. The econometric problems raised by this are examined in section 2.6 below.

2.3.4 Reliability of survey data

It is only possible to check the accuracy of survey data by indirect means, since any independent enquiry conducted for the purposes of checking is likely to be subject to the same inaccuracies as the FES itself. The major source of evidence about FES reliability is the comparison of grossed-up income and expenditure averages with the corresponding figures in the National Accounts. The latter figures are constructed from a wide variety of sources and are therefore a largely independent check. There are inevitable difficulties in putting FES and National Accounts data on a comparable basis, but attempts have been made. Table 2.2, containing comparisons for the year 1976, has been constructed from information given by Kemmley, Redpath and Holmes (1980) and Atkinson and Micklewright (1983). The correspondence appears to be reasonably close except for the problem areas of expenditure on alcohol, tobacco, durables and catering, and income from self-employment, investment income and receipts of sickness benefit.

Much depends on the reasons for these apparent errors. If they are inaccuracies inherent in the survey method (for instance, if heavy drinkers feel guilty about their rates of consumption and either temporally change their behaviour or lie about it) then little can be done - the data are simply
wrong. On the other hand, if the discrepancies are due to differential response, for instance if heavy drinkers are reluctant to co-operate (or incapable), then an attempt can be made to adjust for this endogenous sample selection.

Kemlesy, Redpath and Holmes report as experiment which attempted to prevent the concealment or misrepresentation of expenditures by requiring households to keep much more detailed records, producing an income-expenditure account balanced to the last penny. This failed to reveal any hidden expenditures. There is some evidence, however, of survey-induced change in behaviour: the first few days’ records for a household often show expenditures which are rather higher than is typical of the remainder of the 14 days. Expenditure tends to be 2 or 3 per cent higher in the first week of the survey than in the second. However, this effect is very small and in the wrong direction to explain FES understatement of national rates of expenditure, and it is necessary to look elsewhere for the explanation.

Since many consumer durables are purchased by instalment, and records of these committed expenditures are gathered at the initial interview rather than through a diary, the understatement of durables expenditure seems likely to be due mainly to memory bias. However, durable goods are relative luxuries, so the lower response rate of wealthier households is presumably also a factor. For alcohol and tobacco, differential response and limitations of the coverage of the survey seem to be mainly responsible. Consumption of alcohol is highly skewed in the cross-section, and many of the groups who are identified as heavy drinkers (publicans, hoteliers, seafarers etc.) are underrepresented because they do not live in eligible households or because they are away from home for extended periods. The response rate is also lowest during December when consumption of alcohol is at its greatest. Since tobacco and alcohol consumption are highly correlated, the same response deficiencies would appear to account for the underestimation of tobacco expenditures.

Stimson and Micklewright (1983) give a very detailed analysis of the FES sample figures. They conclude that the understatement of income from self-employment is roughly even mixture of three factors: differences in the time period to which the income figures refer; the relatively low response of the unemployed; and under-reporting. This under-reporting seems to be as much a consequence of a lack of knowledge of true income and difficulty in estimating net profits as a deliberate attempt to conceal income. The measurement of individuals’ investment incomes is also notoriously difficult, with many surveys encountering substantial underestimation. There is little reliable evidence to suggest the reason for this, although both under-reporting and the relatively low response of the wealthy seem to be involved. The substantial understatement of income from state sickness benefit seems capable of being explained mainly in terms of differential response, since people who are ill are likely to be away from home or unwilling to co-operate at such a difficult time.

2.4 Estimation under exogenous sampling

The term exogenous sampling refers to the case where the mechanism used to collect the sample is independent of the behaviour that the economic model attempts to explain. In formal terms, this means that it is possible to write the probability density/mass function for the sample in the form

\[ f(g_1, \ldots, g_n) = \prod_{i=1}^{n} g_i(g_{1i}, \ldots, g_{ni}) f(g_{1i}, \ldots, g_{ni}) \]  

(2.34)

where the marginal distribution of the \( g_i \) is independent of the parameter vector \( \theta \). If the sample is fully independent, then (2.34) can be written

\[ f(g_1, \ldots, g_n) = \prod_{i=1}^{n} g_i(g_{1i}, \ldots, g_{ni}) f_i(g_{1i}, \ldots, g_{ni}) \]  

(2.35)

Random sampling (and, under our assumptions about the relationship of the actual population to a super-population, exhaustive sampling and interval sampling also) leads to the form (2.35) with

\[ f_i(g_{1i}, \ldots, g_{ni}) = g_i(g_{1i}, \ldots, g_{ni}). \]  

(2.36)

Under stratified random sampling, with the stratification variables \( s \) independent of \( s \), we also have (2.35) with

\[ f_i(g_{1i}, \ldots, g_{ni}) = \int g_i(g_{1i}, s) dG_i(s) / P_i. \]  

(2.37)

where \( j \) is the index of the stratum to which observation \( n \) belongs and \( P_i \) is the population frequency of stratum \( j \). This is an instance of a sample of independent but not identically distributed (NID) observations, since the different strata use different forms for \( f_i(g_{1i}, \ldots, g_{ni}) \). For two-stage random sampling, with the strata exogenously defined, the observations are not fully independent, but the sample distribution is of the form (2.34) with

\[ f_i(g_{1i}, \ldots, g_{ni}) \]  

given by expression (2.33) of section 2.1.

2.4.1 Maximum likelihood

There are three common estimation principles that we shall consider here: maximum likelihood, least-squares regression and instrumental variables. The maximum likelihood (ML) estimator is the most widely used in this field and is the technique we shall automatically consider whenever estimation is discussed in the remainder of this book. This merely reflects its dominance in applied work. The ML estimator \( \hat{\theta} \) is the value of \( \theta \) which maximizes the likelihood function \( L(\theta) \) or, more conveniently, its logarithm:

\[ L(\theta) = \log \prod_{i=1}^{n} g_i(g_{1i}, \ldots, g_{ni}) \]

(2.38)
Survey methods and cross-section econometrics

\[ \log L(\theta) = \sum_{i} \log f(z_i, \xi_i) \]

\[ = \sum_{i} \log g(x_i|z_i; \theta) + \log f(z_i; \xi_i) \] (2.38)

Under our assumption that both \( L \) and the sampling process are exogenous, in the sense that the marginal density/mass function \( f(z_i; \xi_i) \) does not depend on \( \theta \), the ML estimator is also identical with the value of \( \theta \) that maximizes the conditional log-likelihood function:

\[ \log L(\theta) = \sum_{i} \log f(z_i|\xi_i) \]

\[ = \sum_{i} \log g(x_i|z_i; \theta) \] (2.39)

In later chapters, we make extensive use of ML estimators, and generally we shall assume that the sampling process is exogenous. Thus the log-likelihood functions we derive will usually be conditional on the exogenous variables, although for economy of notation we shall not make this conditioning explicit, as we have here.

The important consequence of the equivalence of (2.38) and (2.40) is that it is possible to compute the ML estimator without knowing the sampling distribution of the exogenous variables. This is an enormous practical advantage. Of course, the statistical properties of the ML estimator will depend on the nature of the sampling scheme: for example, a sampling process which collects observations from a very limited range of values for \( z \) will generate relatively imprecise estimates of \( \theta \). Moreover, the ease with which one can derive the asymptotic properties of the ML estimator certainly depends on the nature of the sampling process. Standard proofs (see, for instance, Le Cam, 1953) are based on the assumption that the sample observations are independently and identically distributed, and much less attention has been devoted to the general case. However, Bradley and Gart (1962) and Hoadley (1971) give regularity conditions under which standard asymptotic results remain valid.

Under these conditions, the ML estimator is consistent, asymptotically efficient and asymptotically normal. The usual asymptotic approximation to the covariance matrix of \( \hat{\theta} \) is

\[ \hat{\Sigma} = \left( \sum_{i} \frac{\partial \log g(x_i|z_i; \hat{\theta})}{\partial \theta} \frac{\partial \log g(x_i|z_i; \hat{\theta})}{\partial \theta'} \right)^{-1} \] (2.40)

or the asymptotically equivalent expression

\[ \hat{\Sigma} = \left( \sum_{i} \frac{\partial g(x_i|z_i; \theta)}{\partial \theta} \frac{\partial g(x_i|z_i; \theta)}{\partial \theta'} \right)^{-1} \]

White (1982a) points out that the alternative formula

\[ \hat{\Sigma} = \left( \frac{1}{n} \sum_{i} \frac{\partial g(x_i|z_i; \theta)}{\partial \theta} \frac{\partial g(x_i|z_i; \theta)}{\partial \theta'} \right)^{-1} \] (2.42)

is a more robust estimate, in the sense that it remains valid even in cases where the model is mis-specified. We shall not normally trouble to give explicit forms for these expressions.

### 2.4.2 Least squares

Except in special cases, non-linear least-squares estimators are inefficient, since they are based only on the mean rather than on the whole of the population distribution of \( x \) conditional on \( z \). This mean is known as the regression function \( \mu(z; \theta) \),

\[ \mu(z; \theta) = E(x|z; \theta) \] (2.43)

and is the basis for the structural form

\[ x = \mu(z; \theta) + e \] (2.44)

where \( e = x - \mu(z; \theta) \) is a disturbance term with mean zero conditional on \( z \).

For some models, the density/mass function \( g(x|z; \theta) \) is very awkward from the analytical or computational point of view, whereas the regression function is relatively tractable. In these cases, non-linear least-squares estimation has considerable practical advantages over the more efficient ML estimator. Consider the case where \( x \) is a scalar variable. The unweighted least-squares estimator \( \hat{\theta} \) is the value of \( \theta \) that minimizes the following criterion:

\[ s(\theta) = \sum_{i} (x_i - \mu(z_i; \theta))^2 \] (2.45)

In the general non-linear case, \( s(\theta) \) must usually be minimized by an iterative numerical algorithm. In the familiar linear case, where \( \mu(z; \theta) = z^T \theta \), the solution is the usual linear regression:

\[ \hat{\theta} = \left( \sum_{i} z_i z_i^T \right)^{-1} \left( \sum_{i} z_i x_i \right). \] (2.46)

The asymptotic properties of \( \hat{\theta} \) have been studied by White (1980), who has shown that under weak conditions which also cover the NID case \( \hat{\theta} \) is strongly consistent and asymptotically normal, with covariance matrix approximated by...
Survey methods and cross-section econometrics

\[ x(\theta) = \{x(\theta), x(\theta)|\xi| \} \]  

(2.53)

The objective function \( x(\theta) \) can be thought of as a generalized vector norm \( \| x(\theta) \| \), and so, in a sense, the IV estimator minimizes the covariance between the instruments and the residual. In the special case that \( \phi(\theta) \) is the linear function, \( \theta(\theta) \), the IV estimator takes the form

\[ \theta = \left( \Sigma_{\iota,\iota}^{-1} \Sigma_{\iota,\epsilon}^{-1} \right) \Sigma_{\iota,\epsilon}^{-1} \epsilon \]  

(2.54)

which specializes to the following form when \( \epsilon \) is of the same dimension as \( \xi \),

\[ \theta = \left( \Sigma_{\iota,\iota}^{-1} \right) \epsilon \]  

(2.55)

Under suitable regularity conditions, it can be shown (see White, 1982b) that \( \hat{\theta} \) is consistent and asymptotically normal, with approximate covariance matrix

\[ \Sigma_{\hat{\theta}} = A^{-1} \Sigma A^{-1} \]  

(2.56)

where

\[ A = \sum \frac{\partial \phi(\theta)}{\partial \theta} \Sigma_{\iota,\iota}^{-1} \Sigma_{\iota,\epsilon}^{-1} \frac{\partial \phi(\theta)}{\partial \theta} \]  

(2.57)

and

\[ B = \sum \frac{\partial \phi(\theta)}{\partial \theta} \Sigma_{\iota,\iota}^{-1} \Sigma_{\iota,\epsilon}^{-1} \frac{\partial \phi(\theta)}{\partial \theta} \]  

(2.58)

Formula (2.56) remains valid when \( \epsilon \) is heteroscedastic.

The best choice of instruments is usually not obvious. Under homoscedasticity the optimal choice is \( \eta = \Sigma_{\iota,\iota}^{-1} \epsilon \), which is generally dependent on \( \theta \) and thus unobservable. Moreover, if \( \phi(\theta) \) is tractable, it is often difficult to construct a very close approximation to this ideal.

2.5 Estimation under endogenous sampling

By the term endogenous sampling, we mean the case where sample selection proceeds according to a rule defined in terms of the endogenous variables \( x \) (and possibly also \( \xi \)). Thus the variables defining an individual's likelihood of being included in the sample are both endogenous and observable. We leave until the next section the case where sample selection is based on variables that are unobservable. The formal distinction between these two cases is not very significant, but their implications for applied work are very different. Both cases can be characterized in the same general way. If \( \xi \) is the vector of variables on which sample selection is based
and \( g(x_i; \theta) \) is the conditional probability density/mass function in the population, for the \( n \)th observation; we have
\[
J(\theta) = \sum_j \frac{g(x_i; \theta)}{g(x_i; \theta)} dF_i(s)
\]
(2.59)
where \( F_i(s) \) is the (possibly discrete) c.d.f. reflecting the sampling process by which observation \( n \) is drawn. Integration is over the full domain of variation of \( s \).

If \( s \) is observable, comprising a subset of the variables in \( x \) and \( \gamma \), the distribution of \( s \) and \( \gamma \) conditional on \( s \) is degenerate with respect to those variables, and the integral reduces to
\[
J(\theta) = \sum \left[ \log g(s; \gamma; \theta) f_i(s) \right]
\]
(2.60)
where \( x^* \) and \( \gamma^* \) represent the elements of \( x \) and \( \gamma \) that are not part of \( s \) and \( f_i(s) \) is the density/mass function corresponding to \( F_i(s) \). Under endogenous sampling, because \( s_n \) contains part of \( x_n \), the last two terms of (2.60) are generally dependent on \( \theta \). Rewriting their ratio as \( \omega_i(s; \theta) = f_i(s)/g(s) \), ML estimation requires the maximization of
\[
\log L(\theta) = \sum \left[ \log g(s; \gamma; \theta) + \log g(s) + \log \omega_i(s; \theta) \right]
\]
(2.61)
Although the second term in brackets is as usual independent of \( \theta \), the third is not, and therefore the simple procedure of maximizing the conventional objective function \( \Sigma \log g(s; \gamma; \theta) \) is not equivalent to ML estimation and generally yields inconsistent estimates. This is in sharp contrast with the case of exogenous sampling.

The true log-likelihood (2.61) is usually very awkward, and most applied work based on this type of sample uses instead a conventional log-likelihood function
\[
\log L(\theta) = \sum \log J(\theta)
\]
(2.62)
which is constructed from derived forms for the conditional sample density/mass functions \( f_i(s; \gamma; \theta) \). Econometricians often treat this objective function as if it were the true log-likelihood, assuming that the estimates possess the properties of consistency and asymptotic normality, with asymptotic standard errors computable from the information matrix. This is rather dangerous, since there do not appear to be any results of general validity on the asymptotic properties of conditional ML estimators. They certainly cannot be assumed to be asymptotically efficient, for instance.

Some specific examples will make all this rather more clear. The area where the theory has been most fully developed and applied is in discrete choice models; this will be discussed in the next chapter. For the present, we shall assume that \( x \) is a continuously variable scalar with unlimited range and concentrate on the normal regression model
\[
x_i = N(\gamma_i; \sigma^2)
\]
(2.63)
which implies the following p.d.f. and c.d.f.:
\[
g(x_i | \gamma_i; \theta) = \exp \left( -\frac{(x_i - \gamma_i)^2}{2\sigma^2} \right)
\]
(2.64)
\[
Pr(x_i < X | \gamma_i) = G(X | \gamma_i; \theta) = \Phi \left( \frac{X - \gamma_i}{\sigma} \right)
\]
(2.65)
where \( G(\cdot \cdot) \) and \( \Phi(\cdot \cdot) \) are the p.d.f. and c.d.f. of the \( N(0, 1) \) distribution, and \( \gamma = (\gamma \gamma \cdot) \).

The techniques discussed here and in the next section can readily be extended to more complicated models and to distributions other than the normal. Since the appropriateness of these techniques is sensitive to distributional form, tests of the normality hypothesis are particularly important; Bera, Jarque and Lee (1984) and Lee (1984) discuss tests applicable to many of the cases we consider here.

We shall concentrate on two particular sampling schemes that have appeared in the applied literature: endogenous truncation and stratification.

2.5.1 Simple truncation

We shall consider two forms of truncation, both of which are widely encountered in applied work. These differ in the way in which observations are gathered. In both cases, we shall assume a convenient and very simple truncation rule; extension of this to more complicated truncation rules is straightforward. Suppose that sampling is to be done only from the part of the population for which \( x > c \), where \( c \) is a known constant. For example, if \( x \) is hours of labour supplied, we might have a sample that excludes part-time and non-workers, where these are defined as people working fewer than \( c \) hours a week.

Fixed \( N \)

For the first form of truncated sampling, a sample of fixed size \( N \) is drawn. In practice, this might be done either by making repeated drawings from a sampling frame covering the whole population but retaining only the first \( N \) for which \( x > c \), or alternatively by using a frame that covers only the relevant sub-population. Whichever is the case, assume for simplicity that these drawings are purely random, so that sample and sub-population distributions are identical. The distribution of \( s \) and \( \gamma \) in the sample is therefore the population distribution of \( x \) and \( \gamma \) conditional on \( x > c \).
Survey methods and cross-section econometrics

\[ f(x_i, \gamma) = g(x_i, \gamma_i | x_i > c) = g(x_i, \gamma_i) / \Pr(x_i > c) \]

\[ = \exp \left( \frac{-X - \gamma \gamma_i}{\sigma} \right) \frac{g(\gamma_i)}{\Pr(y, \sigma)} \]

(2.66)

where

\[ \Pr(y, \sigma) = \int_{\gamma} \int \phi(x; \gamma, \sigma) \, dx \, G(\gamma) \]

\[ = 1 - \int_{\gamma} \phi \left( \frac{c - \gamma Y}{\sigma} \right) \, dG(\gamma) \]

(2.67)

and \( R_i \) is the domain of variation of \( \gamma \). This is a special case of the general formulation (2.59) and (2.60) with \( x_i = x \) and the ratio \( f(x_i) / g(x_i) \) being the constant \( \omega(x_i, \gamma) = 1 / \Pr(y, \sigma) \).

Thus, the true log-likelihood is

\[ \log L(y, \sigma) = \sum_{i=1}^{n} \log g(\gamma_i) - \frac{N}{2} \log 2 \pi - \frac{N}{2} \log \sigma^2 \]

(2.69)

This is very awkward and never used in practice. In order to evaluate \( P(y, \sigma) \) one must know the distributional form \( G(\gamma) \) and also be able to perform the integration in (2.67), and this is too demanding for practical use. The difficulty cannot be circumvented by treating \( P(y, \sigma) \) as an unknown parameter to be estimated separately from \( \gamma \) and \( \sigma \), for this would yield an ML estimator identical with inconsistent least-squares regression.

In practice, a conditional ML estimator is used. In the sample, the distribution of \( x_i \) conditional on \( \gamma_i \) is

\[ f(x_i | \gamma_i) = g(x_i, \gamma_i | x_i > c) = g(x_i, \gamma_i) / \Pr(x_i > c | \gamma_i) \]

\[ = \exp \left( \frac{-X - \gamma \gamma_i}{\sigma} \right) \frac{1}{1 - \phi \left( \frac{c - \gamma \gamma_i}{\sigma} \right) \sigma^2} \]

(2.70)

Thus, conditionally on \( \gamma_i \), \( x_i \) has an LN(\( \gamma_i \), \( \sigma^2 \)) distribution, which is discussed in detail in appendix 2. This leads to the following conditional log-likelihood:

\[ \log L^T(y, \sigma) = -\frac{N}{2} \log 2 \pi - \frac{N}{2} \log \sigma^2 - \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (X_i - \gamma_i \gamma_i)^2 \]

\[ - \frac{N}{2} \sum_{i=1}^{n} \log \left( 1 - \phi \left( \frac{c - \gamma_i \gamma_i}{\sigma} \right) \right) \]

(2.71)

The first-order conditions for a maximum are

\[ \frac{\partial \log L}{\partial \gamma} = -\frac{1}{2 \sigma^2} \sum_{i=1}^{n} (X_i - \gamma \gamma_i - \sigma \lambda(\gamma^2)) \gamma_i = 0 \]

(2.72)

\[ \frac{\partial \log L}{\partial \sigma} = \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (X_i - \gamma \gamma_i)^2 - \sigma^2 \lambda(\gamma^2) = \frac{-N}{2 \sigma^2} = 0 \]

(2.73)

where \( \lambda(\cdot) = \phi(\cdot) [1 - \Phi(\cdot)] \) is the inverse of Mills’ ratio and \( \gamma^2 \) is the normalized truncation threshold \( c - \gamma, \sigma, \sigma \).

These conditions can be manipulated to the more revealing forms

\[ \hat{\gamma} = \left( \sum_{i=1}^{n} (X_i \gamma_i)^{-1} \right)^{-1} \left( \sum_{i=1}^{n} (X_i \gamma_i - \delta \lambda(\gamma_i) \gamma_i) \right) \]

(2.74)

\[ \hat{\sigma} = \sum_{i=1}^{n} (X_i - \hat{\gamma} \gamma_i)^2 \left[ N + \sum_{i=1}^{n} \lambda(\gamma_i)^2 \right] \]

(2.75)

where \( \lambda(\cdot) \) is evaluated at \( \hat{\gamma} \) and \( \hat{\sigma} \). Although (2.74) and (2.75) appear to be very simple modifications of the analogous regression formulae, they are non-linear in \( \gamma \) and \( \delta \) and must be solved numerically by means of an iterative algorithm. Since the resulting estimator is not a true ML estimator, standard results on the asymptotic properties of ML estimators cannot be relied upon. However, the work of Amemiya (1973) can be used to show strong consistency and asymptotic normality, at least for the case where \( \lambda(\cdot) \) is a non-stochastic sequence with convenient limiting properties. Amemiya also shows that the second derivative matrix of log \( L^T \) can be used for the calculation of asymptotic standard errors.

Alternative regression-based estimators have been proposed for this truncated model; these are based on the moments of \( X_i | x_i \) rather than the full distribution. From equations (A2.5) and (A2.7) of appendix 2, the conditional mean of \( x_i \) is

\[ E(x_i | x_i > c) = \gamma_i \gamma_i + \sigma \lambda(\gamma_i)^2 \]

(2.76)

Thus, when the sample is gathered by means of a truncated sampling mechanism, the regression function of \( x_i \) is no longer the simple linear form \( \gamma_i \gamma_i \) but rather the non-linear form (2.76). Note, however, that the
Survey methods and cross-section econometrics

conditional mean of \( x_i - \phi(c) \) is equal to \( \gamma' \xi \), and so the ML estimator (2.79) can be interpreted as a regression of this corrected variable on \( \xi \).

The conditional variance of \( x_i \) is

\[
\text{var}(x_i | \xi, x_i > c) = \sigma^2 \left[ 1 + \lambda(c \xi) - \lambda(c) \xi \right].
\]

(2.77)

Therefore, although the full population regression has constant variance \( \sigma^2 \), this non-linear regression displays a complicated pattern of heteroscedasticity.

The expressions are specific to our assumption of a conditional normal distribution for \( x_i \), and other distributional forms will give rise to different expressions; appendix 2 gives the truncated moments of some other common forms. However, whatever the true distribution, the linear regression model will always be mis-specified: truncation from below has the effect of increasing the mean of \( y_i \) by an amount that depends on the difference between \( \gamma' \xi \) and the threshold \( c \). Thus there will be some covariance in the sample between the truncation effect which is omitted from the regression model and the explanatory variable \( \xi \), and this will impart some omitted variable bias to the regression coefficients. The same will apply to the case of truncation from above. Unbiased estimation of \( \gamma \) by multiple regression is only possible if there happens to be exactly offsetting truncation both from above and below, and this is a case of no practical importance. Regression is an inappropriate technique in truncated samples, and can lead to seriously flawed empirical results.

For example, Hausman and Wise (1977) compare regression estimates with more appropriate ML estimates for a model of individual earnings fitted to a sample deriving from the New Jersey negative income tax experiment. This sample is quite severely truncated (from above, rather than below), with the truncation threshold equal to 150 per cent of the poverty level defined for each family (thus \( c \) is a function of \( \beta \) and varies from family to family). They find evidence of very serious biases, with the ML coefficients being two or three times as large as the corresponding regression coefficients.

One way of avoiding these biases is to treat (2.74) as the basis of a non-linear regression model, with \( \gamma \) and \( \delta \) found by numerically minimizing the residual sum of squares:

\[
S(y, \gamma, \delta) = \sum_{x_i > c} [y_i - \gamma' \xi_i - \delta \lambda(c \xi_i)]^2.
\]

(2.78)

Since \( x_i \) is heteroscedastic, it is important that suitable expressions are used for computing asymptotic standard errors (for example, those of White, equations (2.69)-(2.71)). Although this least-squares estimator has seen some use (see Wales and Woodland, 1980), it has no obvious advantages over conditional ML estimation since it is inefficient but still requires an iterative computational algorithm.

In a slightly different context, Amemiya (1973) has proposed a simple estimator which, like non-linear least squares, is inefficient, but has the advantage that iterative computations are not required. Combining equations (A2.5) and (A2.6) from appendix 2, we can show that

\[
E(x_i (x_i - c) | \xi_i, x_i > c) = \sigma^2 + \gamma' \xi_i E(x_i | \xi_i, x_i > c).
\]

(2.79)

This suggests the regression model

\[
x_i (x_i - c) = \sigma^2 + \gamma' \xi_i + \nu_i,
\]

(2.80)

where \( \nu_i \) is a random disturbance with \( E(\nu_i | \xi) = 0 \). However, a regression of \( x_i (x_i - c) \) on \( \xi_i \) will not provide a consistent estimate of \( \gamma \) and \( \delta \), since \( x_i \) is correlated with \( \nu_i \). Amemiya proposes a two-step procedure. First regress \( x_i \) on \( \xi_i \) and higher powers of the variables in \( \xi_i \), and construct the fitted value \( \hat{x}_i \). Then use the elements of the vector \( x_i \hat{x}_i \) as instrumental variables in estimating the regression of \( x_i (x_i - c) \) on \( \xi_i \). This procedure yields consistent estimates of \( \gamma \) and \( \delta \); Amemiya gives an expression for their asymptotic covariance matrix.

Random N

It is quite common for samples to be truncated in such a way that the final sample size is random rather than fixed. This is sometimes done as part of a two-step estimation technique (see, for example, chapter 4, section 4.1) but is also sometimes a feature of the survey design (for instance, the New Jersey negative income tax experiment). In such cases, a sample of fixed size \( N \) is drawn from the whole population (assume again that this is a random sample), and the final sample is arrived at by deleting all observations which fail to satisfy the acceptance criterion \( x > c \). This sample then contains a random number \( N \) of observations.

Consider the marginal distribution of \( N \). Since a randomly drawn observation satisfies \( x > c \) with probability \( P(y, \beta) \), \( N \) has a binomial distribution with frequency function

\[
P(N) = \frac{N!}{N! (N^*-N)!} P(y, \beta)^N (1 - P(y, \beta))^{N^*-N},
\]

(2.81)

Conditional on a specific positive value for \( N \), the observations have a joint density/mass function

\[
\begin{align*}
S(y, \xi, \ldots, x_i | N) &= \prod_{i=1}^{N} \gamma(x_i | \xi_i), x_i > c \\
&= \prod_{i=1}^{N} \exp \left[ \frac{(x_i - \gamma' \xi_i)}{\sigma} \right] P(y, \beta) \\
&= \exp \left[ \frac{-N^* P(y, \beta)}{\sigma} \right] \prod_{i=1}^{N} \exp \left[ \frac{x_i - \gamma' \xi_i}{\sigma} \right] P(y, \beta).
\end{align*}
\]

(2.82)
Survey methods and cross-section econometrics

Thus the distribution of the full sample is

\[ f(x_1, y_1, \ldots, x_n, y_n) = \frac{N!}{N^n} \prod_{i=1}^{n} \frac{(1 - P(x_i, y_i))^{n-i} x_i^{y_i}}{N} \frac{1}{(N^m - N)!} \times \prod_{i=1}^{n} \frac{1}{y_i} \left( \frac{x_i - y_i}{\sigma} \right)^{y_i} g(z_i) \]  

(2.83)

where the product term is to be taken as unity if \( N = 0 \).

Ignoring incidental constants, the true log-likelihood is therefore

\[ \log L(y, \theta) = -\frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - y_i)^2 \]

\[ + \frac{N^m - N}{N} \log (1 - P(y, \theta)) \]  

(2.84)

This is again intractable for most purposes, and in practice the conditional log-likelihood (2.71) is used, although it now must be regarded as conditional on both \( N \) and \( y_1, \ldots, y_n \). The asymptotic properties of the conditional ML estimator have not been derived for this case; Amemiya's (1973) results do not apply directly, since they are asymptotic in \( N \), whereas here \( N \) is random so the relevant theory must be asymptotic in \( N^m \).

The previously discussed regression-based methods are also widely used, and there is no difficulty in showing that these have standard asymptotic properties in this case.

### 2.5.2 Endogenous stratification

Sample truncation can be thought of as an example of stratification of the population into two sub-populations, defined by \( x < c \) and \( x \geq c \), the latter having a sample of size zero drawn from it. Occasionally, more complicated forms of endogenous stratification are encountered. An example is the Gary income maintenance experiment, discussed by Hausman and Wise (1982), which used the following sampling technique. A pool of households was drawn at random from the relevant population and then divided into income classes, with limits defined as fixed multiples of a prespecified poverty level. These endogenously defined income classes were sampled at different rates until a final sample of fixed size \( N \) was compiled. This is a somewhat simplified picture of the procedure that was followed in the Gary experiment, but it suffices to illustrate the ideas involved.

Define the \( m \) strata, \( S_1, \ldots, S_m \), as follows:

\[ S_i = \{ x \mid x \leq c_i < x \leq c_{i+1} \} \quad i = 1, \ldots, m \]  

(2.85)

where \( c_0 = -\infty \) and \( c_m = \infty \). Note that, for the Gary experiment, \( m = 5 \) strata were used. In practice, the \( c_i \) depend on factors such as family size; however, we shall simplify matters further by assuming them to be constant. We consider two different interpretations of the Gary sampling procedure.

**Fixed stratum sample sizes** Suppose that subsamples of fixed sizes \( N_1, \ldots, N_m \) are drawn from the \( m \) strata. The result is an NID sample with the typical observation having the following density/mass function:

\[ f(x_i, y_i) = \frac{g(x_i, y_i; c_i, \theta) \chi_{x_i < c_i}}{\int g(x_i, y_i; c_i, \theta) \chi_{x_i < c_i} \, dy_i} \]

\[ = \frac{g(x_i; y_i; \theta) \chi_{x_i < c_i}}{\int g(x_i; y_i; \theta) \chi_{x_i < c_i} \, dy_i} \]

\[ = \frac{1}{\int g(x_i; y_i; \theta) \, dy_i} \left[ (x_i - y_i)^2 \right] \]

(2.86)

where

\[ P(\gamma; \theta) = \left[ \Phi \left( \frac{c_i - y_i}{\sigma} \right) - \Phi \left( \frac{c_{i-1} - y_i}{\sigma} \right) \right] g(\gamma) \]  

(2.87)

This is a special case of our general formulation of endogenous sampling, equation (2.60), with

\[ \omega(x_i; \theta) = \frac{1}{P(\gamma_i; y_i, \sigma)} \]  

(2.88)

and presents familiar problems: since the terms \( P(\gamma; \theta) \) require both knowledge of the form of \( G(\gamma) \) and the evaluation of the integral (2.87), the true log-likelihood will be too awkward for applied work. Moreover, if, instead of being evaluated, they are treated as unknown parameters, ML estimation again reduces to the regression of \( x_i \) on \( y_i \) and yields inconsistent estimates of \( \gamma \) and \( \sigma \) (the \( P(\gamma; \theta) \) are also inconsistently estimated as \( N_i/N \)).

As usual, the conditional density \( f(x_i, y_i; \theta) \) leads to a much simpler conditional log-likelihood:

\[ \log L(y, \theta) = -\frac{N}{2} \log \sigma^2 - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - y_i)^2 \]

\[ - \sum_{i=1}^{N} \log \left[ \Phi \left( \frac{c_i - y_i}{\sigma} \right) - \Phi \left( \frac{c_{i-1} - y_i}{\sigma} \right) \right] \]  

(2.89)

A generalized version of Amemiya's estimator can also be constructed for this case.

**Probabilistic stratum sampling** In their applied work, Hausman and Wise (1982) favour a rather different interpretation of the Gary sampling procedure. They assume that the final sample is taken from the initial pool of households by repeated random drawings, with each draw being made from stratum \( i \) with probability \( \Pi_i, i = 1, \ldots, m \). This leads to sample observations having a common density/mass function with discontinuities at the
Survey methods and cross-section econometrics

Log-likelihoods (2.89) and (2.90) appear to be very different: evaluation of \( P(\gamma; \eta, \sigma) \) is necessary for the former but not for the latter. However, this is misleading. The maximization of (2.92) with respect to \( \gamma \) leads to a simple regression of \( x_0 \) on \( \gamma \), which is known to be inconsistent. To ensure consistency, it is necessary to maximize \( L \) subject to the constraint

\[
\sum_{i \in s} \rho_i P(\gamma_i; \eta, \sigma) = 1. \tag{2.93}
\]

Again, this is intractable.

The natural alternative to these unhelpful expressions is to use the conditional density function \( f(x_0|\gamma_i) = f(x_0, \gamma_i)/f(\gamma_i) \). The marginal density of \( \gamma \) is found by integrating and summing the sums (2.90):

\[
f(\gamma_i) = g(\gamma_i) \sum_{s \in m} P(\eta_i - \Phi_\gamma - \Phi_\eta_i)/P(\gamma_i; \eta, \sigma) \tag{2.94}
\]

where

\[
\Phi_\gamma = \phi \left( \frac{x_0 - \gamma}{\sigma} \right). \tag{2.95}
\]

The conditional density function therefore has \( \gamma \) segment

\[
f(\gamma_i|\xi_i) = \sum_{s \in m} P(\eta_i - \Phi_\gamma - \Phi_\eta_i)/P(\gamma_i; \eta, \sigma) \tag{2.96}
\]

defined for \( x_0 \in (c_i, c_{i+1}), i = 1, m \).

Although this conditional density also depends on the terms \( P(\gamma_i; \eta, \sigma) \), conditional likelihood methods are feasible. The sampling rates \( \rho_i = P_i/P(\gamma_i; \sigma) \) can be treated as parameters and estimated independently of \( \gamma \) and \( \sigma \) by maximizing the unconstrained conditional log-likelihood

\[
\log L(\rho_1, \ldots, \rho_m; \gamma, \sigma) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{x \in s} (x - \gamma)^2 2N_i \log P(\sigma_i; \gamma) \tag{2.97}
\]

Note that, since (2.96) is nonhomogenous of degree zero in the \( \rho_i \), only the \( m-1 \) ratios \( \rho_2/\rho_1, \ldots, \rho_m/\rho_{m-1} \) (say) can be identified. Although no formal derivation of the asymptotic properties of this conditional ML estimator is available, there appears to be little difficulty in
Survey methods and cross-section econometrics

extending Amemiya's (1973) results to this case, particularly if the $u_i$ are known. Thus it is not necessary to impose restriction (2.93) to achieve consistency.

If the $u_i$ are known or can be reliably estimated, an obvious weighted least-squares estimator is available. This weights each observation in inverse proportion to the square root of the degree of oversampling in its stratum, and requires only a regression of $x_i/\sqrt{u_i}$ on $\sqrt{u_i} x_i$:

$$
\overline{y} = \left( \sum \frac{1}{u_i} \sqrt{u_i} x_i \right)^{-1} \sum \frac{1}{u_i} \sqrt{u_i} x_i y_i.
$$

This estimator is easily shown to be consistent under weak conditions. However, Hausman and Wise present some calculations which suggest that it may be rather inefficient relative to conditional ML if the sampling scheme is very unbalanced.

2.4 Non-response and extreme sample selection

With rare exceptions, participation in a sample survey is voluntary. Thus samples are in part self-selected, and any econometric analysis is conditional on individuals' decisions to co-operate with the survey. If this decision to co-operate is exogenous, in the sense that it is related to the behaviour under study, then the implicit conditioning will impart some bias to econometric results unless corrective action is taken.

2.6.1 Consequences of differential response

To illustrate the difficulties here, we shall continue to work with the simple normal regression model (2.63) and extend it by adding a related regression-like equation governing survey participation. Thus we have the following conditional bivariate normal model:

$$
x_i, y_i, x_{i0}, \xi_i \sim N \left( \begin{array}{c}
\gamma x_i \xi_i, \\
h \\
\delta \\
\end{array} \right), 
\begin{array}{c}
\sigma_x^2 & \sigma_{xy} & 0 \\
\sigma_{xy} & \sigma_y^2 & 0 \\
0 & 0 & 1
\end{array} \right)
$$

where $y_i$ is an unobservable variable representing the individual's 'willingness to participate in the survey', $\xi_i$ is a vector of exogenous variables explaining the participation decision and $\delta$ is the corresponding coefficient vector. The variance of $y_i$ is arbitrary at zero: since $y_i$ is unobserved, its scale is arbitrary.

Let us suppose (without loss of generality, if $\xi_i$, includes an intercept term) that an individual decides to participate $\xi_i$ 'willingness' is positive. Furthermore, assume for simplicity (but the original sample of individuals (both respondents and non-respondents) is a random sample. Then the distribution of $x$ and $\xi_i$ in the sample is the population distribution of $x$ and $\xi_i$ conditional on the event $\gamma \ast > 0$. Now consider the regression estimator

\begin{align*}
\overline{y} = \left( \sum \frac{1}{u_i} \sqrt{u_i} x_i \right)^{-1} \sum \frac{1}{u_i} \sqrt{u_i} x_i y_i & \\
\text{This has expected value} & \\
E(\overline{y}) = E \left( \sum \frac{1}{u_i} \sqrt{u_i} x_i \right)^{-1} \sum \frac{1}{u_i} \sqrt{u_i} x_i E(x_{i0}|x_i, \xi_i, \nu > 0) & \\
\text{and bias} & \\
E(\overline{y} - \gamma) = E \left( \sum \frac{1}{u_i} \sqrt{u_i} x_i \right)^{-1} \sum \frac{1}{u_i} \sqrt{u_i} x_i E(x_{i0} - \gamma|\xi_i, \nu > 0) & 
\end{align*}

There are two principal circumstances under which this bias will be zero. If $x$ and $\nu$ are independent conditional on $\xi_i$ and $\xi_i$, then $E(x_{i0} - \gamma|\xi_i, \nu > 0) = E(x_{i0} - \gamma|\xi_i, \xi_i) = 0$.

Alternatively, consider the case where $L_i$ and $L_i$ are mutually independent, with distributions that do not vary with $\nu$. The regression model implies that $x_{i0} - \gamma \ast \xi_i$ is independent of $L_i$, and therefore $E(x_{i0} - \gamma \ast L_i|\xi_i, \nu > 0)$ is a function of $L_i$ only, say $\mu(L_i)$. Therefore (2.101) can be factored as follows:

\begin{align*}
E(\overline{y} - \gamma) & = \sum E(A^{-1} l_{i0}) L_i \mu(L_i) \\
& - E \left( A^{-1} \sum l_{i0} \mu \right)
\end{align*}

where $A = \sum l_{i0}$, and $\mu = E[\mu(L_i)]$. This is the regression of a constant, $\mu_i$, on $\xi_i$ if, say, the first element of $\xi_i$ is a dummy variable so that $\xi_i$ is the intercept term, then the bias vector is $E(\overline{y} - \gamma) = (a_0 \ldots 0)$, and all slope coefficients are unbiased. This is because the conditional mean of $\mu_i$ is greater than $\gamma \ast \xi_i$ by a truncation factor that fluctuates randomly about the constant $\mu_i$, independently of $\xi_i$.

It is often very difficult to justify making either of these two independence assumptions. Our discussion of the FES in section 2.3, for instance, suggested that response rates are likely to be related to many variables such as income, employment status, location, wealth and some types of expenditure. It is very unlikely that such variables are independent of the explanatory variables $\mu_i$, or, indeed, that the unobservable factors determining response are independent of the unobservable influences on $\mu_i$. Thus, least-squares regression and most other statistical techniques must be regarded as biased when applied to economic data originating from a sample survey in which response is voluntary. If response is low and non-uniform, then this bias may be substantial.
Survey methods and cross-section econometrics

For the special case of normality, relation (2.99), the conditional expectation of \( x_i \) given \( x_i \) is given by equation (A2.36) of appendix 2,

\[
E(x_i | \epsilon_i, x_i > 0) = \phi(x_i) \Lambda^*(0, \Lambda) \tag{2.104}
\]

and the least-squares bias is

\[
E(\tilde{y} - y) = \alpha E \left[ \left( \sum \xi_i \epsilon_i \right) \left( \sum \epsilon_i \Lambda^*(0, \Lambda) \right) \right] \tag{2.105}
\]

where \( \Lambda^* = \lambda - \lambda = \phi(0)/\phi(\lambda) \) is the complement of the inverse Mills ratio. Thus, in the case of a normal bivariate regression structure, the simple parameter restriction \( \alpha = 0 \) is a sufficient condition for consistency of least squares in the presence of differential response.

Although we have considered the effects of differential response on a very simple regression model, the general considerations involved here are applicable to any statistical model and any estimation technique - the neglect of difficulties caused by non-response can have serious consequences.

2.6.2 Estimation without data on non-respondents

For obvious reasons it is extremely difficult to obtain worthwhile information about non-respondents. Most surveys yield no information at all, and those that do usually provide very limited and unreliable data on the characteristics of non-respondents. This is the fundamental problem faced by an applied worker attempting to correct for differential response.

Even if the sample is confined to respondents, correction for the effects of differential response still requires the specification of a model of response. We shall retain the normal regression model (2.99), assuming that \( x_i, \xi_i \) and \( \epsilon_i \) (but not \( v_i \), of course) are observed. Since no information is available on the population of non-responders (those for whom \( v_i < 0 \)), the sample distribution has a joint probability density/mass function

\[
f(\alpha, \epsilon, \xi, \nu = 1, \nu > 0) = \frac{1}{\sqrt{2\pi}} \int f(\alpha, \xi, \nu = 1, \nu > 0) dz \tag{2.106}
\]

The special case of (2.106) for our normal linear model can be found by decomposing the distribution of \( x \) and \( v \) (conditional on \( f \) and \( \epsilon \)) into marginal and conditional components (see appendix 3, section A3.3).

\[
f(\alpha, \xi, \nu = 1) = \left( \frac{1}{2\pi} \right)^{\frac{1}{2}} \phi \left( \frac{x_i - \epsilon_i}{\sigma_i} \right) \Phi \left( \frac{\epsilon_i \Lambda^*(0, \Lambda)}{\sigma_i} \right) \Phi(0) \tag{2.107}
\]

where

\[
P(\theta) = P(\nu > 0) = \int \phi (\theta) d\theta(\tilde{y}). \tag{2.108}
\]

For the usual reasons, \( P(\theta) \) is an intractable expression and an operational ML estimation technique must be based instead on the conditional density \( f(\alpha, \xi_i | \nu = 1) \), which is arrived at by replacing \( P(\theta) \) in (2.107) by \( P(\nu > 0| \nu = 1) \phi (\theta) \) and deleting the term \( g(\xi_i, \nu = 1) \) (see also equation (A2.34) of appendix 2). Thus the conditional log-likelihood is

\[
\log L(\alpha, \xi, \epsilon, \omega, \nu = 1) = -N \log 2 \pi - N \log \sigma_i - \frac{N}{2} \omega \frac{\epsilon_i \Lambda^*(0, \Lambda)}{\sigma_i} \left( \frac{\epsilon_i \Lambda^*(0, \Lambda)}{\sigma_i} \right)^2 - \sum \frac{\epsilon_i \Lambda^*(0, \Lambda)}{\sigma_i} \log \phi (\theta). \tag{2.109}
\]

Note that if a fixed number of individuals are approached initially, the number of respondents \( N \) is random, and therefore (2.108) must be interpreted as conditional on \( N \) also.

Non-likelihood methods are also available. From equation (A2.36) of appendix 2, the conditional expectation of \( x_i \) is

\[
E(x_i | \nu > 0, \nu = 1) = \frac{\sigma_i}{\sigma_i} \epsilon_i \Lambda^*(0, \Lambda) \tag{2.110}
\]

which suggests the non-linear least-squares estimator

\[
\min_{\alpha, \xi} \sum (\epsilon_i - \epsilon_i \Lambda^*(0, \Lambda))^2. \tag{2.111}
\]

Observe that this has a very simple interpretation as a correction of the least-squares bias (2.105). However, it has little to recommend it as an alternative to conditional ML since it is inefficient and yet still requires the use of an iterative computational algorithm.

Both the conditional ML and least-squares estimation techniques make very great demands on the sample. Information on \( x_i \) is assumed to be adequate for the simultaneous estimation of both the primary model itself and an additional model of survey response. Bearing in mind that \( \Lambda_i \) and \( \epsilon_i \) are likely to have variables in common, or at least to be highly correlated, it seems unlikely that one would be able to obtain good estimates of \( \delta_i \), \( \beta_i \) and \( \sigma_i \). A simulation study by Mutlib and Jöreskog (1981) was used to confirm this suspicion and suggest that \( \delta_i \) in particular is likely to be unreliable. Moreover, without direct observation of the response-non-response dichotomy
and with no real behavioural theory as a guide to specification, it is very difficult to produce a convincing choice of variables for the vector $\xi$.

These considerations suggest that differential response bias is unlikely to be a curable ailment in most applied cross-section work. In the absence of reliable information on non-respondents, the best that can be done is probably to give some rough indication of the likely direction of any such bias. For instance, in the case of an analysis of the relationship between the demand for alcoholic drink and income using FES data, one could reasonably expect a downward bias in the estimated income coefficient, for the following reasons. The evidence described in section 2.3 suggests that the probability of non-response increases with alcohol consumption (implying $\sigma_n < 0$) and with income (implying that $\theta \xi_n$ decreases with income). Since $\lambda^r(\cdot)$ is a decreasing function, the term $\mathcal{E}(\xi_n - \gamma^r \xi_n > 0, \xi_n, \xi_n) = \sigma_n \lambda^r(\theta \xi_n)$ will be negatively correlated with income. Since the least-squares bias can be seen as an omitted variable bias, with $\sigma_n \lambda^r(\theta \xi_n)$ as the omitted explanatory factor, the conventional analysis of mis-specification in linear regression indicates a downward bias in the income coefficient. For models more complicated than linear regression, the nature of non-response bias may, of course, be much more difficult to establish.

2.6.3 Estimation using data on non-respondents

Occasionally, it is possible to obtain useful information on the characteristics of non-respondents. If this is felt to be adequate for the specification of a model of survey participation, then estimation methods more efficient than those described above become available. Again, suppose that the simple mode (2.99) is to be used. There are two observational regimes: for respondents, $\xi_n$ and $\xi_n$ are all observed; for non-respondents only $\xi_n$ is observable. The sample distribution then consists of two parts. For respondents

$$f(\xi_n, \xi_n, \xi_n) = \frac{1}{\mathcal{E}(\xi_n, \xi_n)} \Phi \left( \frac{\xi_n - \gamma \xi_n}{\sigma_n} \right) \sigma_n \lambda^r(\theta \xi_n) \mathcal{E}(\xi_n, \xi_n)$$

and for non-respondents

$$f(\xi_n < 0, \xi_n) = \Phi(\xi_n) \mathcal{E}(\xi_n, \xi_n)$$

Thus the log-likelihood function for a sample covering both respondents and non-respondents is

$$\log L(\gamma, \theta, \xi_n) = K + \sum \log \left[ 1 - \Phi(\theta \xi_n) \right] - \frac{N}{2} \log \sigma_n^2$$

where $K$ is an inessential constant, $\Phi$ and $R$ represent the set of observations on respondents and non-respondents, and $N$ is the number of respondents. Note that there is no difficulty in evaluating and hence maximizing this log-likelihood, and that the resulting estimator is a true ML estimator. Thus standard asymptotic results apply.

Two-step regression-based techniques can also be used. These follow a suggestion made by Heckman (1976, 1979). The first step involves the estimation of a binary probit response model in isolation from the principal regression model. The probabilities of response and non-response are

$$\Pr(\xi_n > 0, \xi_n) = \Phi(\theta \xi_n) \mathcal{E}(\xi_n)$$

and

$$\Pr(\xi_n < 0, \xi_n) = \left[ 1 - \Phi(\theta \xi_n) \right] \mathcal{E}(\xi_n)$$

Thus a log-likelihood for the response-non-response dichotomy is

$$\log L(\theta) = \sum \log \mathcal{E}(\xi_n) + \sum \log \Phi(\theta \xi_n) + \sum \log \left[ 1 - \Phi(\theta \xi_n) \right]$$

The maximizing value $\hat{\theta}$ is the widely used probit estimator, which is available in most statistical computing packages. The second stage of the Heckman procedure exploits expression (2.110). First an artificial variable $\lambda^r = \lambda^r(\theta \xi_n)$ is constructed, and then a regression of $\lambda^r$ on $\xi_n$ and $\lambda^r$ yields coefficients which are the final estimates of $\gamma$ and $\sigma_n^2$. There is a complication, arising from the fact that $\lambda^r$ depends on the estimate $\hat{\theta}$. This introduces an additional source of random variation into the final regression and invalidates the usual regression formula for the sampling variances of $\gamma$ and $\sigma_n^2$. Heckman (1979) gives a formula for the correct asymptotic covariance matrix.

Heckman's estimator has been widely applied in cases where endogenous sample selection is due to deliberate truncation of the sample rather than differential survey response. For instance, a study of individual wage rates based on a sample including both labour force participants and non-participants would be essentially equivalent to our model (2.99) with $\xi_n$ interpreted as hours of work (or, equivalently, as the difference between the actual wage and the threshold wage above which work is undertaken).
wage and a reservation wage). Early examples of this type of model are those of Gronau (1974) and Heckman (1976). However, it is usually possible in this labour supply context to observe the extraneous truncation variable \( x \) (hours of work), and then more efficient techniques are available. These are described in chapter 4, section 4.5.

### 2.6.4 Multiple truncation factors

Cross-section samples are often subject to more than one type of extraneous truncation. Many surveys display both incomplete coverage and differential response. For example, earnings might exclude non-workers as part of the survey design, and also have a poor rate of response. In such cases, each observation is conditional on two events: that the individual is not excluded by the sample design (supplies a positive amount of labour), and that he or she is willing to cooperate with the survey. Thus, the equivalent of (2.97) in this case would consist of three equations, i.e. a primary earnings model and two sample participation equations:

\[
x_n, x_{na}, x_{na1}, \xi_n, \xi_{na}, \xi_{na1} \sim N \left( \begin{array}{c} \gamma_n' \xi_n \\ \delta_{n1} \xi_{na1} \\ \delta_{n2} \xi_{na2} \\
\end{array} \right), \begin{array}{ccc}
\sigma_n^2 & \sigma_n \sigma_{na} & \sigma_n \\
\sigma_n \sigma_{na} & \sigma_{na}^2 & \sigma_{na} \\
\sigma_n & \sigma_{na} & \sigma_n^2
\end{array} \right)
\]  

(2.118)

Using the expressions derived in section A.2.4 of appendix 2, it is straightforward to extend the ML estimation techniques derived above to this case, and we do not present the details here. The log-likelihood function is more difficult in this case, however, since bivariate normal probabilities are involved (see section A.3.4 of appendix 3 for a review of computational methods). For regression-based methods, we need the appropriate conditional expectation of \( x_n \). This is found by adapting equation (A.2.52) of appendix 2 to give

\[
E(x_n | v_n = 0, v_{na} > 0, \xi_n, \xi_{na}) = \gamma' \xi_n + \sigma_n \lambda_n + \sigma_n \lambda_{na}
\]  

(2.119)

where

\[
\lambda_n = \phi(\delta_{n1} \xi_{na1} - \sigma_n \delta_{n2} \xi_{na2})
\]

(2.120)

and

\[
Pr(v_n = 0, v_{na} > 0 | \xi_n, \xi_{na}) = \frac{\delta_{n1} \xi_{na1} - \sigma_n \delta_{n2} \xi_{na2}}{1 - \sigma_n^2}
\]

(2.121)

and \( Pr(v_{na} = 0, v_{na1}, v_{na2}, v_{na2}, v_{na1}, \xi_{na}, \xi_{na1}, \xi_{na2}, \xi_{na2}) \sim N(\mu, \Sigma) \)

(2.122)

where

\[
\mu = \begin{bmatrix} \gamma' \xi_n \\ \delta_{n1} \xi_{na1} \\ \delta_{n2} \xi_{na2} \\ \end{bmatrix}
\]

(2.123)

and \( \Sigma = \{ \sigma_n \} \), with \( \sigma_n \) and \( \sigma_{na} \) normalized to be unity. Note that (2.121) allows the possibility that the experimental change in \( \xi \) alters the whole relationship between \( x \) and \( \xi \). The restriction \( \gamma_n = \gamma_n \) would be an important one to test.

The conditional distribution of \( x_n \) and \( x_{na} \) has components corresponding to three possible observational regimes: full participation; drop-out after the first interview; and refusal to participate at all. These three components are as follows.

### 2.6.5 Response in experimental surveys

Social experimentation programs typically involve repeated observations. Households are first observed before being subjected to any experimental change (in the exogenous variables \( \xi \)) and then re-interviewed at some date after the experimental scheme has begun. Thus non-response can occur at two stages. A household might refuse to cooperate with the survey at all, or it might initially cooperate but then drop out of the programme before being re-interviewed. The latter phenomenon, the gradual decay of participation in a panel survey, is known as sample attrition.

A simple picture of this process is as follows: \( x_n \) and \( x_{na} \) are the dependent variable and the unobserved 'willingness to participate' variable at the time of the first interview; \( x_n \) and \( x_{na} \) are their analogues observed after the experimental treatment has begun. Assume that these four variables are generated by a conventional linear regression structure:

\[
x_n, x_{na}, x_{na1}, x_{na2}, \xi_n, \xi_{na}, \xi_{na1}, \xi_{na2} \sim N(\mu, \Sigma)
\]

(2.122)
of bivariate normal distributions (see appendix 3), and it is necessary to use information on individuals who refuse to co-operate even at the initial stage of the experiment.

For the last of these reasons, this sequential two-period model has not been estimated in practice. The nearest approach to it is the study by Hausman and Wise (1979) which concentrates on the problem of attrition and neglects the difficulties raised by initial non-response. The reason for this is that considerable information is available on people who drop out of the experiment after the first interview, whereas little is known about those who refuse any cooperation. The Hausman and Wise model can be seen as a special case of (2.122) and (2.123) with \( \sigma_2 = \sigma_3 = \sigma_6 = 0 \). In this case, \( x_{na} \) is independent of \( x_{1a}, x_{2a} \) and \( x_{3a} \), and thus the joint density (2.124) which relates to full participants can be written

\[
f(x_{na}, x_{2a} | \xi_1, \xi_2, \xi_3) = \frac{1}{\sqrt{2\pi} \sigma_2} \exp \left( - \frac{(x_{2a} - \mu_2)^2}{2\sigma_2^2} \right) \Phi(\xi_1)
\]

(2.130)

where

\[
\mu_2 = [\gamma_1, \gamma_2, \gamma_3]'
\]

(2.131)

and

\[
\Sigma^* = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}
\]

(2.132)

Expression (2.125), which is the density of \( x_{na} \) for those who drop out of the panel after the first interview, simplifies to the form

\[
f(x_{na} | \xi_1, \xi_2, \xi_3) = \left[ 1 - \Phi(\xi_1) \right] \frac{x_{na} - \mu_2}{\sigma_2} \Phi(\xi_1)
\]

(2.133)

In deriving (2.137), we have used the results of section A.3.3 of appendix 3. Note that the integral term in (2.130) can also be expressed in terms of the univariate \( \Phi(\cdot) \) and \( \phi(\cdot) \) functions by decomposing the multivariate function into marginal and conditional components (see appendix 3 or Hausman and Wise, 1979). This yields a log-likelihood composed of two elements:

\[
\log L(\gamma, \beta, \gamma_1, \beta_1, \Sigma^*) = \log L(\beta) + \log L(\gamma, \beta_1, \Sigma^*)
\]

(2.134)
where \( L^2 (\theta) \) is the log-likelihood for a simple probit model of the initial participation decision, and

\[
L(\gamma, \delta, \Sigma) = K + \sum \left[ \log f(x_{it}, \xi_{it}) - \log f(x_{it}, \xi_{it}) \right] + \sum \log f(x_{it}, \xi_{it}, \xi_{it}). \tag{2.135}
\]

Since \( L^2 \) and \( L \) have no arguments in common, \( L(\gamma, \delta, \Sigma) \) can be maximized independently of \( \delta \) and without knowledge of \( \xi_{it} \).

Hausman and Wise use this ML technique on an earnings model applied to a sample containing both individuals from a control group (subjected to no experimental treatment) and individuals experiencing an experimental change in their insured income. They assume \( \omega_i = \omega_j \) and \( \gamma_i = \gamma_j \), but allow for shifts in behavior between control group and experimental group individuals and/or individuals at interview one and interview two by including two appropriate dummy variables. Apart from the enforced exclusion of the temporal dummy variable, \( \xi_{it} \) is identical with \( \xi_{it} \). They find that inclusion in the experimental group has a small but significant direct negative effect on earnings quite apart from its indirect effect through the change in unearned income; this could be interpreted as a consequence of delayed adjustment to the experimental change. Experimental treatment also has a strong negative effect on the probability of drop-out: as one would expect, people selected to receive additional payments are much less likely to leave the experiment than members of the control group. The parameter \( \omega_i \) is small but statistically significant, and there is evidence that the estimation of \( \gamma_i \) using data on \( x_{it} \) and \( x_{it} \) without correction for attrition leads to appreciable biases.

Considerably more complex structures than this can arise from experimental surveys. People may withdraw from the experiment for a variety of reasons (ill-health, change of locality etc.), each governed by a different stochastic model. The allocation of people to experimental and control groups may be endogenous, either through an element of self-selection or because the programme administrator makes the allocation on the basis of endogenous characteristics. There may also be an element of truncated endogenous sampling, of the type discussed in the previous section. The generalization of our methods to these more complicated settings is straightforward in principle, although the practical problem of estimating the parameters of all the stochastic relations present in the data is likely to be formidable, particularly if no information is available on those who are excluded from the sample.
\[ I = \varphi(y_1; \mu_1, \Sigma_1) \ldots \varphi(y_n; \mu_n, \Sigma_n) \, dx_1 \ldots dx_n \]  
(A3.32)

The first part of (A3.32) is merely a \((q - p)\)-dimensional normal p.d.f., and the second is a \(p\)-dimensional c.d.f. which can be evaluated by the methods discussed above.

In the important special case of the bivariate normal density integrated with respect to one of its arguments, this result is

\[ I = \Phi \left[ \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 - \rho \sigma_x \sigma_y}} \right] \varphi \left( \frac{\sigma_x}{\sqrt{\sigma_y^2 + 
\rho \sigma_x \sigma_y}} \right) \]  
(A3.33)

\[ = \Phi \left( \frac{\sigma_x \rho - \rho \sigma_y}{\sqrt{\sigma_x^2 + \rho \sigma_x \sigma_y}} \right) \varphi \left( \frac{\sigma_x \rho - \rho \sigma_y}{\sqrt{\sigma_x^2 + \rho \sigma_x \sigma_y}} \right) \]  
(A3.34)

where \( \rho = \text{corr}(x, y) = \sigma_x \sigma_y + \gamma \) and \( \Phi(\cdot) \) here represent the c.d.f. and p.d.f. of the univariate \( N(0, 1) \) distribution.

References


Aigner, D. J. (1985), 'The residential electricity time-of-use pricing experiments: what have we learned?'. In J. A. Hausman and D. A. Wise (eds), Social Experimentation, Chicago, IL: Chicago University Press.


Amemiya, T. (1973), 'Regression analysis when the dependent variable is truncated normal', Econometrica, 41, 919-38.

Amemiya, T. (1974), 'Multivariate regression and simultaneous equation models when the dependent variables are truncated normal', Econometrica, 42, 996-1012.


References 339


Mills, J. F. (1926), 'Table of the ratio: area to rounding ordinate for any portion of normal curve', Biometrika, 18, 399-400.


Quandt, R. E. (1968), 'Estimation of modal splits', Transportation Research, 2, 41-50.


Rosen, H. S. (1979), 'Housing decisions and the US income tax: an econometric
Rothbard, E. (1941), 'The measurement of changes in real income under conditions of rationing', Review of Economic Studies, 8, 100-7.