CHAPTER 1

Introduction

The purpose of this book is to present methods for the analysis of some econometric models in which the dependent variables are either qualitative or limited in their range. These models are commonly encountered in empirical work that analyzes survey data, although we shall also give examples of some time-series models. In a certain sense every variable we consider in practice, at least in econometric work, is limited in its range. However, it is not necessary to apply the complicated analysis described in this book to all these problems. For instance, if we believe that prices are necessarily positive, we might postulate that they have a log-normal distribution rather than the normal. On the other hand, in the limited-dependent-variable models discussed in this book, the variables are limited to their range because of some underlying stochastic choice mechanism. It is models of this kind that we shall be concerned with in this book. Similarly, there are many qualitative variables that are often used in econometric work. These are all usually known as dummy variables. What we shall be concerned with in this book are models in which the dummy variables are endogenous rather than exogenous. The following simple examples will illustrate the types of models that we shall be talking about. These examples can be conveniently classified into three categories: (a) truncated regression models, (b) censored regression models, and (c) dummy endogenous models.

1.1 Truncated regression models

Example 1: Negative-income-tax experiment

The negative-income-tax experiment provides detailed information on a sample of households with incomes below some threshold. Suppose we

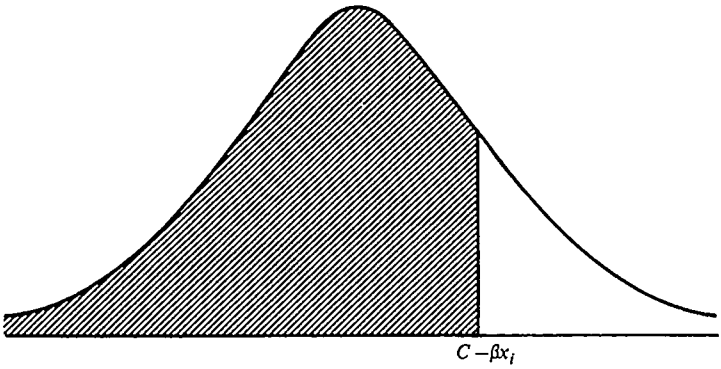

Figure 1.1. Truncated normal distribution

wish to use these data to estimate an earnings equation:

$$y = f(\text{education, age, experience, etc.})$$

Then we need to take into consideration the fact that the dependent variable is truncated at a certain point. Observations with values of y above a threshold value are not included in the sample. This problem has been analyzed by Hausman and Wise (1976, 1977) and will be discussed in detail in Chapter 6.

To see why the ordinary least-squares (OLS) method gives biased estimates in this case, suppose we want to estimate the effect on earnings (y) of years of schooling (x). The regression equation is

$$y_i = \beta x_i + u_i \tag{1.1}$$

where $u_i \sim IN(0, \sigma^2)$. We observe y_i only if $y_i \leqslant c$, where c is a given constant. This condition implies that

$$\beta x_i + u_i \leqslant c \quad \text{or} \quad u_i \leqslant c - \beta x_i \tag{1.2}$$

Clearly, $E(u_i \mid u_i \leqslant c - \beta x_i)$ is not equal to zero. In fact, it will be a function of x_i. Thus the residual is correlated with the explanatory variable x_i, and we get inconsistent estimates of the parameter β if we use the OLS method. In this case, because β is expected to be positive, and because $E(u_i \mid u_i \leqslant c - \beta x_i)$ decreases with increasing values of x_i, we get the result that the OLS estimator of β will be downward-biased;[1] that is,

[1] This follows from the well-known result about the least-squares bias from omitted variables, which says E(OLS estimate) = (true coefficient of the included variable) + (true coefficient of the omitted variable) × (the regression coefficient from a regression of the excluded variable on the included variable).

estimating the effect of years of schooling on income from the data generated by the negative-income-tax experiment gives us an underestimate of the true effect.

Example 2: Schooling and earnings of low achievers

On this topic, see Hansen et al. (1970). Suppose we have a sample of military rejects – those who scored below the thirtieth percentile on the Armed Forces Qualification Test (AFQT). We wish to estimate an "intelligence" equation:

$$\text{AFQT} = f(\text{education, age, socioeconomic characteristics, etc.})$$

Again, we need to take into account the fact that the dependent variable is truncated. The OLS method will give biased estimates.

The least-squares bias arising from the truncation of the dependent variable also arises when we classify observations on the basis of the values of the dependent variable or when we aggregate observations on the basis of the dependent variable. The aggregation bias problem has been discussed by Feige and Watts (1972).

1.2 Censored regression models

Example 3: Demand for durable goods

If we have survey data on consumer expenditures, we find that most households report zero expenditures on automobiles or major household goods during any year. However, among those households that make any such expenditures, there will be wide variation in the amounts. Thus there will be a lot of observations concentrated around zero. Tobin (1958) analyzed this problem by formulating the regression model

$$\begin{aligned} y &= x\beta + u \quad \text{if } y > 0 \\ y &= 0 \quad \text{otherwise} \end{aligned} \tag{1.3}$$

where y is expenditures and x is a set of explanatory variables. An alternative and perhaps a better way of formulating the model is the following:

Let y be the expenditure the individual can afford and y^* be "threshold" expenditure (the price of the cheapest automobile acceptable to the individual):

$$\begin{aligned} y &= x\beta + u \\ y^* &= z\gamma + v \end{aligned} \tag{1.4}$$

The observed expenditures are y if $y > y^*$ and zero otherwise. In this alternative formulation the threshold y^* is not necessarily zero and can vary from individual to individual. We shall describe the methods for analyzing both models (1.3) and (1.4) in Chapter 6.

Example 4: Changes in holdings of liquid assets

Consider a change in a household's holding of liquid assets during a year. This variable can be positive or negative, but it cannot be smaller than the negative of the household's assets at the beginning of the year, because one cannot liquidate more assets than one owns. Here again the threshold is nonzero and is different for different individuals.

Example 5: Married women in the labor force

Let y^* be the reservation wage of the housewife based on her valuation of time in the household. Let y be the market wage based on an employer's valuation of her effort. The woman participates in the labor force if $y > y^*$. Otherwise she does not. In any given sample we have observations on y for those women who participate in the labor force, and we have no observations on y for those who do not. For these women we know only that $y^* \geqslant y$. Given these data, we have to estimate the coefficients in the equations explaining y^* and y, as in model (1.4). The analysis is similar to that in the case of automobile demand. This problem has been analyzed by Nelson (1975, 1977). There are more complicated versions of the labor-supply model that have been analyzed by Heckman (1974), Hanoch (1980*a*, *b*), and others. One interesting example of the problem of women in the labor force is that involving nurses. There are some nurses (particularly married nurses) who choose not to work because the value of their time at home is greater than their market wage. There is an interesting policy problem here concerning whether to spend money in training more nurses or to raise the wages of nurses so that some of those already trained who have chosen to stay at home can be brought into the labor force.

Examples 1 and 2 are illustrations of truncated regression models, and examples 3, 4, and 5 are illustrations of censored regression models. The distinction between the two is as follows:[2]

[2] See Aitchison and Brown (1957, Chapter 9). See also Kendall and Stuart (1967, p. 522) for a discussion of the difference between truncating and censoring.

Suppose that y is $N(\mu, \sigma^2)$ and that all our observations are for $y \geq T$. We do not have any observations for $y < T$. Then T is called the point of truncation, and the density of y, which is called a truncated normal distribution, is

$$g(y) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \Big/ \int_T^\infty \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] dy$$

$$= \frac{1}{\sigma}\phi\frac{y-\mu}{\sigma} \Big/ \left[1-\Phi\left(\frac{T-\mu}{\sigma}\right)\right] \tag{1.5}$$

where ϕ refers to the standard normal density function and Φ refers to the cumulative normal.

On the other hand, suppose we have a sample of size n, of which n_1 observations are less than T and $n_2 = n - n_1$ observations are equal to or greater than T, and only for these n_2 observations are the exact values known. This is the case of a censored distribution.

The joint density of the observation is

$$\binom{n}{n_1}\left[\Phi\left(\frac{T-\mu}{\sigma}\right)\right]^{n_1} \prod_{i=1}^{n_2} \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \tag{1.6}$$

In this case we do not have the exact values for n_1 observations, but we do have the information that they are all less than T. The maximum-likelihood estimates of the parameters μ and σ^2 in this case are obtained by maximizing (1.6) with respect to these parameters.

In the econometric literature it is customary to use the term *truncated normal distribution* to describe both of these cases; the term *censored distribution* is rarely used.[3] This is perhaps justified, because in analyses of both these models we shall be making use of the properties of the truncated normal distribution. In any case, it is important to note the definitions commonly used in the statistical literature and to bear in mind the distinction between the two problems. In the regression context, we do not have any observations on either the explained variable y or the explanatory variables x in the case of the truncated regression model if the value of y is above (or below) a threshold. This is the case with the data generated by the negative-income-tax experiment. In the case of the censored regression model, we have data on the explanatory

[3] Amemiya (1973) discussed a regression model in which the dependent variable is truncated normal. However, the model he considered is what is known in the statistical literature as the censored model. Nelson (1977) discussed a censored regression model (although the model is not a simple regression model).

variables x for all the observations. As for the explained variable y, we have actual observations for some, but for others we know only whether or not they are above (or below) a certain threshold. This is the model considered by Tobin.

1.3 Dummy endogenous variables

Example 6: Effects of unions on wages

Suppose we are given data on wages and personal characteristics of workers and we are told whether or not they are unionized. A naive way of estimating the effects of unionization on wages is to estimate a regression of wages on the personal characteristics of the workers (age, race, sex, education, experience, etc.) and a dummy variable that is defined as

$D = 1$ for unionized workers
$D = 0$ otherwise

The coefficient of D then measures the effect of unions on wages. Here the dummy variable D is exogenous. However, this is not a satisfactory method for analyzing the problem, because the dummy variable D is not exogenous but endogenous. The decision to join or not to join the union is determined by the expected gain. This is the formulation used by Lee (1978). Alternative formulations of endogenizing the dummy variable D in this example are somewhat mechanical approaches to this problem. All these studies will be discussed in Chapter 11.

Both examples 5 and 6 can also be called self-selectivity models. The data are generated by self-selection of individuals – the choices of individuals whether or not to participate in the labor force, whether or not to join the union,[4] and so forth. The earliest discussion of this problem in the context of labor supply was that of Lewis (1974). Some further examples of this follow.[5]

Example 7: Returns to college education

If we are given data on incomes of a sample of individuals, some with college educations and others without, our analysis must take into

[4] Of course, in the unions example, there are other problems, such as restriction of entry and union selectivity, in addition to the self-selectivity of the workers. There may also be some employer selectivity that makes employers more selective in their choices of employees if the employees are unionized.

[5] Models with self-selection are discussed in Chapter 9.

account the fact that those who have college educations are those who chose to go to college, and those who do not have college educations are those who chose (for their own reasons) not to go to college. A naive and commonly used way of analyzing these differences is to define a dummy variable

$D = 1$ if the person went to college
$D = 0$ otherwise

and estimate an earnings function with D as an extra explanatory variable. However, this is not a satisfactory solution to the problem, because the dummy variable D is itself determined by the choices of individuals on the basis of expected income and other factors. There have been many efforts made to include ability as an extra variable in the equations for earnings. However, in addition to the ability bias there is also a selectivity bias. The two are not mutually exclusive. Perhaps one should model the two factors taking into account the fact that persons with more ability are also capable of making the correct choices, so that the difference between expected incomes from the different choices is almost the same as the actual difference.

There has been a lot of discussion on returns to college education and whether or not college matters. However, in all these studies the usual regressions have been dummy-variable regressions (with perhaps some explanatory variables and IQ scores added). This is not a correct measure, because of the self-selectivity involved in the data. Also, a question is often asked: What would have been the average income for those who did not go to college if they had gone to college? Again, the answers to such questions are not obtained from simple dummy-variable regressions. See the work of Griliches et al. (1978), Kenny et al. (1979), and Willis and Rosen (1979).

Example 8: Demand for housing

In studies of demand for housing it is customary to analyze the demand for owner-occupied housing and the demand for rental housing separately, or else to regress housing expenditures (imputed rent for owner-occupied housing) on other explanatory variables and a dummy variable defined as

$D = 1$ for owner-occupied housing
$D = 0$ otherwise

Again, there is a self-selectivity problem here: Some individuals choose to own houses, and others choose to rent. In the estimated demand func-

tions this choice must be taken into account. This problem has been analyzed by Trost (1977), Lee and Trost (1978), and Rosen (1979).

Example 9: Effects of fair-employment laws

Landes (1968) studied the effects of fair-employment legislation on the status of blacks. Landes used the regression model $y_i = \alpha X_i + \beta D_i + u_i$, where y_i is the wage of blacks relative to that for whites in state i, X_i is the vector of exogenous variables for state i, $D_i = 1$ if the state i has a fair-employment law ($D_i = 0$ otherwise), and u_i is a residual. The regression coefficient β is hypothesized to be positive. Landes found a marginally significant positive coefficient for β.

In this formulation, D_i is exongenous. However, the presence of a law is not an exogenous event. States in which blacks would fare well without a fair-employment law may be more likely to pass such a law if legislation depends on the concensus. On the other hand, one can argue for the reverse case that in states with much market discrimination, the demand for antidiscrimination legislation on the part of blacks is high, and this leads to a greater probability of fair-employment legislation in such states. Heckman (1976*a*) reanalyzed the Landes data allowing for the endogeneity of the dummy variable D_i.

Example 10: Compulsory school attendance laws

This example is similar to example 9 involving fair-employment laws. The passage of the legislation is itself an endogenous variable. This example has been discussed by Edwards (1978).

The foregoing examples illustrate the wide class of models in which limited-dependent-variable methods are applicable. We still have not discussed the multivariate log-linear models considered by Nerlove and Press (1973, 1976) or the several applications of the conditional logit model pioneered by McFadden (1974). These will be discussed in Chapters 5 and 3, respectively. We also have not discussed disequilibrium models, which will be considered in Chapter 10. Elaborating the illustrative examples for all these models would make this list far too lengthy.

Much of what will be discussed in subsequent chapters depends on the assumption that the underlying random variables have a (univariate or multivariate) normal distribution. Not much is known about the robustness of the results to departures from normality. We shall discuss the available results at the appropriate places.

In the case in which there is only one underlying random variable u,

some alternative functional forms have been suggested. These are the following:

1. *Logistic distribution.* Actually, this is the cumulative distribution of the hyperbolic-secant-square (sech^2) distribution whose density function is given by

$$f(u) = \frac{e^u}{(1+e^u)^2}\, du \quad -\infty < u < \infty \tag{1.7}$$

The cumulative distribution is

$$F(Z) = \frac{e^Z}{1+e^Z} \tag{1.8}$$

which is the logistic function. The advantage of this distribution is that the distribution function has, unlike the normal, a closed-form expression given by (1.8). However, given the fast computer programs to evaluate the cumulative normal, this is not a major advantage today. The logistic and cumulative normal differ very little, and only at the tails. Hence, unless the sample size is very large, the empirical results obtained from the two will be very close.

The logistic function has been used frequently only in cases in which the dependent variable is binary (taking only two values, say 0 and 1), and then we refer to it as logit analysis (as opposed to probit analysis, in which the underlying variable u has a normal distribution). In actual practice, one need not confine the use of this function to only binary variables. One can as well use it to estimate any of the models given by equations (1.3)–(1.6).

2. *Cauchy distribution.* The main thing that characterizes the sech^2 distribution in equation (1.7) is that it has thicker tails than the normal. This is all the more true of the Cauchy distribution, whose density function is (in the standard form):

$$f(u) = \frac{1}{\pi} \cdot \frac{1}{1+u^2} \tag{1.9}$$

The cumulative distribution function of this is

$$F(z) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} z$$

The general form of (1.9) is

$$f(u) = \frac{1}{\pi} \frac{1}{[1+(x-\theta)/\lambda]^2} \tag{1.10}$$

Table 1.1. *Comparisons of Cauchy and normal distributions*

	$\Pr(X \leq x)$	
x	Cauchy	Normal
0	0.5000	0.5000
0.4	0.6211	0.6063
0.6	0.6720	0.6571
1.0	0.7500	0.7500
2.0	0.8524	0.9113
3.0	0.8976	0.9785
4.0	0.9220	0.9965

A comparison of the Cauchy and the normal distributions is given in Table 1.1.[6] This shows how much thicker the tails of the Cauchy distribution are as compared with those of the normal. The cumulative normal and the logistic, on the other hand, agree much more closely (see Cox, 1970, Table 2.1., p. 28).

3. *Burr distribution.* This has been discussed by Burr (1942). The density function is given by

$$f(u) = \frac{cku^{c-1}}{(1+u^c)^{k+1}} \quad (c, k > 0,\ u > 0) \tag{1.11}$$

and the cumulative distribution function is given by

$$F(z) = 1 - \frac{1}{(1+z^c)^k} \quad (c, k > 0) \tag{1.12}$$

All these distributions have the property that the cumulative distributions have closed-form expressions. However, as mentioned earlier, this is not important, because of present-day computer technology. The Burr distribution has the advantage that it can handle random variables that take on only positive values. With the normal and the logistic distributions, one can handle such variables by considering the appropriate log-transforms, that is, by assuming that $\log u$ instead of u follows a normal (or sech^2) distribution. Other feasible alternatives are the gamma distribution and the beta distribution.

As mentioned earlier, unless samples are very large and the observations at the tails exert a large influence, one obtains similar results using

[6] This is based on Table 1 from Johnson and Kotz (1970, p. 155).