19

LIMITED DEPENDENT VARIABLES—TRUNCATION, CENSORING, AND SAMPLE SELECTION

19.1 INTRODUCTION

This chapter is concerned with **truncation** and **censoring**. As we saw in Section 18.4.6, these features complicate the analysis of data that might otherwise be amenable to conventional estimation methods such as regression. "Truncation" effects arise when one attempts to make inferences about a larger population from a sample that is drawn from a distinct subpopulation. For example, studies of income based on incomes above or below some poverty line may be of limited usefulness for inference about the whole population. Truncation is essentially a characteristic of the distribution from which the sample data are drawn. Censoring is a more common feature of recent studies. To continue the example, suppose that instead of being unobserved, all incomes below the poverty line are reported as if they were *at* the poverty line. The censoring of a range of values of the variable of interest introduces a distortion into conventional statistical results that is similar to that of truncation. Unlike truncation, however, censoring is essentially a defect in the sample data. Presumably, if they were not censored, the data would be a representative sample from the population of interest. We will also examine a form of truncation called the **sample selection** problem. Although most empirical work in this area involves censoring rather than truncation, we will study the simpler model of truncation first. It provides most of the theoretical tools we need to analyze models of censoring and sample selection.

The discussion will examine the general characteristics of truncation, censoring, and sample selection, and then, in each case, develop a major area of application of the principles. The stochastic frontier model [Aigner, Lovell, and Schmidt (1977), Fried, Lovell, and Schmidt (2008)] is a leading application of results for truncated distributions in empirical models. Censoring appears prominently in the analysis of labor supply and in modeling of duration data. Finally, the sample selection model has appeared in all areas of the social sciences and plays a significant role in the evaluation of treatment effects and program evaluation.

19.2 TRUNCATION

In this section, we are concerned with inferring the characteristics of a full population from a sample drawn from a restricted part of that population.

19.2.1 TRUNCATED DISTRIBUTIONS

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value. For instance, in Example 19.2, we are given a characteristic of the distribution of incomes above $100,000. This subset is a part of the full distribution of incomes which range from zero to (essentially) infinity.

THEOREM 19.1 Density of a Truncated Random Variable
If a continuous random variable x has pdf $f(x)$ and a is a constant, then[1]

$$f(x \mid x > a) = \frac{f(x)}{\text{Prob}(x > a)}.$$

The proof follows from the definition of conditional probability and amounts merely to scaling the density so that it integrates to one over the range above a. Note that the truncated distribution is a conditional distribution.

Most recent applications based on continuous random variables use the **truncated normal distribution**. If x has a normal distribution with mean μ and standard deviation σ, then

$$\text{Prob}(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha),$$

where $\alpha = (a - \mu)/\sigma$ and $\Phi(.)$ is the standard normal cdf. The density of the truncated normal distribution is then

$$f(x \mid x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right)}{1 - \Phi(\alpha)},$$

where $\phi(.)$ is the standard normal pdf. The **truncated standard normal distribution**, with $\mu = 0$ and $\sigma = 1$, is illustrated for $a = -0.5, 0$, and 0.5 in Figure 19.1. Another truncated distribution that has appeared in the recent literature, this one for a discrete random variable, is the truncated at zero Poisson distribution,

$$\text{Prob}[Y = y \mid y > 0] = \frac{(e^{-\lambda}\lambda^y)/y!}{\text{Prob}[Y > 0]} = \frac{(e^{-\lambda}\lambda^y)/y!}{1 - \text{Prob}[Y = 0]}$$
$$= \frac{(e^{-\lambda}\lambda^y)/y!}{1 - e^{-\lambda}}, \quad \lambda > 0, y = 1, \ldots$$

This distribution is used in models of uses of recreation and other kinds of facilities where observations of zero uses are discarded.[2]

For convenience in what follows, we shall call a random variable whose distribution is truncated a **truncated random variable**.

[1] The case of truncation from above instead of below is handled in an analogous fashion and does not require any new results.

[2] See Shaw (1988). An application of this model appears in Section 18.4.6 and Example 18.8.

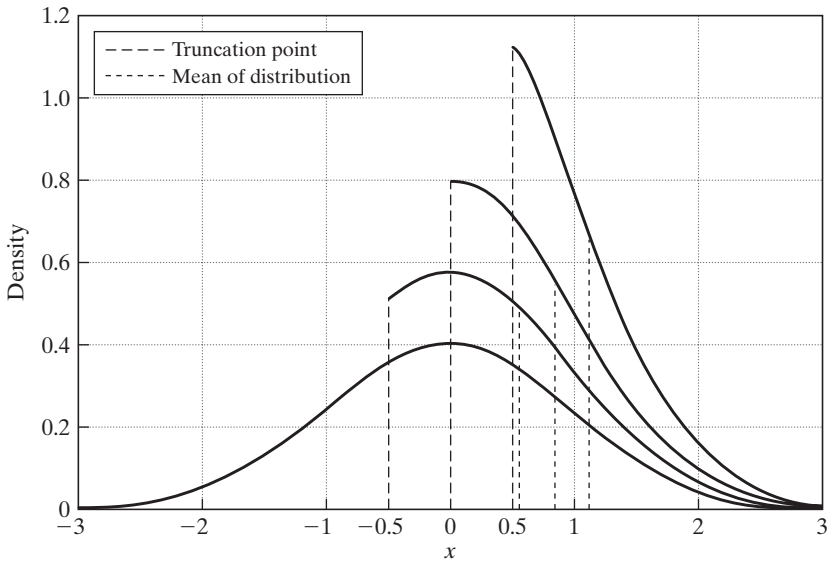


FIGURE 19.1 Truncated Normal Distributions.

19.2.2 MOMENTS OF TRUNCATED DISTRIBUTIONS

We are usually interested in the mean and variance of the truncated random variable. They would be obtained by the general formula:

$$E[x \mid x > a] = \int_a^{\infty} x f(x \mid x > a)\, dx$$

for the mean and likewise for the variance.

Example 19.1 Truncated Uniform Distribution

If x has a *standard* uniform distribution, denoted $U(0, 1)$, then

$$f(x) = 1, \quad 0 \le x \le 1.$$

The truncated at $x = \frac{1}{3}$ distribution is also uniform:

$$f\left(x \mid x > \frac{1}{3}\right) = \frac{f(x)}{\text{Prob}\left(x > \frac{1}{3}\right)} = \frac{1}{\left(\frac{2}{3}\right)} = \frac{3}{2}, \quad \frac{1}{3} \le x \le 1.$$

The expected value is

$$E\left[x \mid x > \frac{1}{3}\right] = \int_{1/3}^{1} x\left(\frac{3}{2}\right) dx = \frac{2}{3}.$$

For a variable distributed uniformly between L and U, the variance is $(U - L)^2/12$. Thus,

$$\text{Var}\left[x \mid x > \frac{1}{3}\right] = \frac{1}{27}.$$

The mean and variance of the untruncated distribution are $\frac{1}{2}$ and $\frac{1}{12}$, respectively.

Example 19.1 illustrates two results.

1. If the truncation is from below, then the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, then the mean of the truncated variable is smaller than the mean of the original one.
2. Truncation reduces the variance compared with the variance in the untruncated distribution.

Henceforth, we shall use the terms **truncated mean** and **truncated variance** to refer to the mean and variance of the random variable with a truncated distribution.

For the truncated normal distribution, we have the following theorem:[3]

THEOREM 19.2 Moments of the Truncated Normal Distribution
If $x \sim N[\mu, \sigma^2]$ *and* a *is a constant, then*

$$E[x \mid \text{truncation}] = \mu + \sigma\lambda(\alpha), \tag{19-1}$$

$$\text{Var}[x \mid \text{truncation}] = \sigma^2[1 - \delta(\alpha)], \tag{19-2}$$

where $\alpha = (a - \mu)/\sigma$, $\phi(\alpha)$ *is the standard normal density and*

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \quad \text{if truncation is } x > a, \tag{19-3a}$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \quad \text{if truncation is } x < a, \tag{19-3b}$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]. \tag{19-4}$$

An important result is

$$0 < \delta(\alpha) < 1 \quad \text{for all values of } \alpha,$$

which implies point 2 after Example 19.1. A result that we will use at several points below is $d\phi(\alpha)/d\alpha = -\alpha\phi(\alpha)$. The function $\lambda(\alpha)$ is called the **inverse Mills ratio**. The function in (19-3a) is also called the **hazard function** for the standard normal distribution.

Example 19.2 A Truncated Lognormal Income Distribution

"The typical 'upper affluent American' . . . makes $142,000 per year . . . . The people surveyed had household income of at least $100,000."[4] Would this statistic tell us anything about the "typical American"? As it stands, it probably does not (popular impressions notwithstanding). The 1987 article where this appeared went on to state, "If you're in that category, pat yourself on the back—only 2 percent of American households make the grade, according to the survey." Because the **degree of truncation** in the sample is 98 percent, the $142,000 was probably quite far from the mean in the full population.

Suppose that incomes, x, in the population were lognormally distributed—see Section B.4.4. Then the log of income, y, had a normal distribution with, say, mean μ and

[3] Details may be found in Johnson, Kotz, and Balakrishnan (1994, pp. 156–158). Proofs appear in Cameron and Trivedi (2005).

[4] See *New York Post* (1987).

standard deviation, σ. Suppose that the survey was large enough for us to treat the sample average as the true mean. Assuming so, we'll deduce μ and σ and then determine the population mean income.

Two useful numbers for this example are ln 100 = 4.605 and ln 142 = 4.956. The article states that

$$\text{Prob}[x \geq 100] = \text{Prob}[\exp(y) \geq 100] = 0.02,$$

or

$$\text{Prob}(y < 4.605) = 0.98.$$

This implies that

$$\text{Prob}[(y - \mu)/\sigma < (4.605 - \mu)/\sigma] = 0.98.$$

Because $\Phi[(4.605 - \mu)/\sigma] = 0.98$, we know that

$$\Phi^{-1}(0.98) = 2.054 = (4.605 - \mu)/\sigma,$$

or

$$4.605 = \mu + 2.054\sigma.$$

The article also states that

$$E[x \mid x > 100] = E[\exp(y) \mid \exp(y) > 100] = 142,$$

or

$$E[\exp(y) \mid y > 4.645] = 142.$$

To proceed, we need another result for the lognormal distribution:

$$\text{If } y \sim \text{N}[\mu, \sigma^2], \text{ then } E[\exp(y) \mid y > a] = \exp(\mu + \sigma^2/2) \times \frac{\Phi(\sigma - (a - \mu)/\sigma)}{1 - \Phi((a - \mu)/\sigma)}.$$

[See Johnson, Kotz and Balakrishnan (1995, p. 241).] For our application, we would equate this expression to 142, and a to ln 100 = 4.605. This provides a second equation. To estimate the two parameters, we used the method of moments. We solved the minimization problem

$$\text{Minimize}_{\mu,\sigma}\ [4.605 - (\mu + 2.054\sigma)]^2$$
$$+ [142\Phi((\mu - 4.605)/\sigma) - \exp(\mu + \sigma^2/2)\,\Phi(\sigma - (4.605 - \mu)/\sigma)]^2.$$

The two solutions are 2.89372 and 0.83314 for μ and σ, respectively. To obtain the mean income, we now use the result that if $y \sim \text{N}[\mu, \sigma^2]$ and $x = \exp(y)$, then $E[x] = \exp(\mu + \sigma^2/2)$. Inserting our values for μ and σ gives $E[x]$ = \$25,554. The 1987 Statistical Abstract of the United States gives the mean of household incomes across all groups for the United States as about \$25,000. So, the estimate based on surprisingly little information would have been relatively good. These meager data did, indeed, tell us something about the average American.

19.2.3 THE TRUNCATED REGRESSION MODEL

In the model of the earlier examples, we now assume that

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$$

is the deterministic part of the classical regression model. Then

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where

$$\varepsilon_i \mid \mathbf{x}_i \sim N[0, \sigma^2],$$

so that

$$y_i \mid \mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2]. \tag{19-5}$$

We are interested in the distribution of y_i given that y_i is greater than the truncation point a. This is the result described in Theorem 19.2. It follows that

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma \frac{\phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}. \tag{19-6}$$

The conditional mean is therefore a nonlinear function of a, σ, $\mathbf{x}$, and $\boldsymbol{\beta}$.

The partial effects in this model *in the subpopulation* can be obtained by writing

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda(\alpha_i), \tag{19-7}$$

where now $\alpha_i = (a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma$. For convenience, let $\lambda_i = \lambda(\alpha_i)$ and $\delta_i = \delta(\alpha_i)$. Then

$$\begin{aligned} \frac{\partial E[y_i \mid y_i > a]}{\partial \mathbf{x}_i} &= \boldsymbol{\beta} + \sigma(d\lambda_i/d\alpha_i)\frac{\partial \boldsymbol{\alpha}_i}{\partial \mathbf{x}_i} \\ &= \boldsymbol{\beta} + \sigma\left(\lambda_i^2 - \alpha_i\lambda_i\right)(-\boldsymbol{\beta}/\sigma) \\ &= \boldsymbol{\beta}\left(1 - \lambda_i^2 + \alpha_i\lambda_i\right) \\ &= \boldsymbol{\beta}(1 - \delta_i). \end{aligned} \tag{19-8}$$

Note the appearance of the scale factor $1 - \delta_i$ from the truncated variance. Because $(1 - \delta_i)$ is between zero and one, we conclude that for every element of $\mathbf{x}_i$, the marginal effect is less than the corresponding coefficient. There is a similar **attenuation** of the variance. In the subpopulation $y_i > a$, the regression variance is not σ^2 but

$$\text{Var}[y_i \mid y_i > a] = \sigma^2(1 - \delta_i). \tag{19-9}$$

Whether the partial effect in (19-7) or the coefficient $\boldsymbol{\beta}$ itself is of interest depends on the intended inferences of the study. If the analysis is to be confined to the subpopulation, then (19-7) is of interest. If the study is intended to extend to the entire population, however, then it is the coefficients $\boldsymbol{\beta}$ that are actually of interest.

One's first inclination might be to use ordinary least squares to estimate the parameters of this regression model. For the subpopulation from which the data are drawn, we could write (19-6) in the form

$$y_i \mid y_i > a = E[y_i \mid y_i > a] + u_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i + u_i, \tag{19-10}$$

where u_i is y_i minus its conditional expectation. By construction, u_i has a zero mean, but it is heteroscedastic:

$$\text{Var}[u_i] = \sigma^2\left(1 - \lambda_i^2 + \lambda_i\alpha_i\right) = \sigma^2(1 - \delta_i),$$

which is a function of $\mathbf{x}_i$. If we estimate (19-10) by ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$, then we have omitted a variable, the nonlinear term λ_i. All the biases that arise because of an omitted variable can be expected.[5]

Without some knowledge of the distribution of $\mathbf{x}$, it is not possible to determine how serious the bias is likely to be. A result obtained by Chung and Goldberger (1984) is broadly suggestive. If $E[\mathbf{x} \mid y]$ in the full population is a linear function of y, then plim $\mathbf{b} = \boldsymbol{\beta}\tau$ for some proportionality constant τ. This result is consistent with the widely observed (albeit rather rough) proportionality relationship between least squares estimates of this model and maximum likelihood estimates.[6] The proportionality result appears to be quite general. In applications, it is usually found that, compared with consistent maximum likelihood estimates, the OLS estimates are biased toward zero. (See Example 19.5.)

19.2.4 THE STOCHASTIC FRONTIER MODEL

A lengthy literature commencing with theoretical work by Knight (1933), Debreu (1951), and Farrell (1957) and the pioneering empirical study by Aigner, Lovell, and Schmidt (ALS, 1977) has been directed at models of production that specifically account for the textbook proposition that a production function is a theoretical ideal.[7] If $y = f(\mathbf{x})$ defines a production relationship between inputs, $\mathbf{x}$, and an output, y, then for any given $\mathbf{x}$, the observed value of y must be less than or equal to $f(\mathbf{x})$. The implication for an empirical regression model is that in a formulation such as $y = h(\mathbf{x}, \boldsymbol{\beta}) + u$, u must be negative. Because the theoretical production function is an ideal—the frontier of efficient production—any nonzero disturbance must be interpreted as the result of inefficiency. A strictly orthodox interpretation embedded in a Cobb–Douglas production model might produce an empirical frontier production model such as

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u, \quad u \geq 0.$$

The gamma model described in Example 4.7 was an application. One-sided disturbances such as this one present a particularly difficult estimation problem. The primary theoretical problem is that any measurement error in $\ln y$ must be embedded in the disturbance. The practical problem is that the entire estimated function becomes a slave to any single errantly measured data point.

Aigner, Lovell, and Schmidt proposed instead a formulation within which observed deviations from the production function could arise from two sources: (1) productive inefficiency, as we have defined it earlier and that would necessarily be negative, and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign. The end result was what they labeled the **stochastic frontier**:

$$\begin{aligned}\ln y &= \beta_1 + \sum_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N\left[0, \sigma_v^2\right].\\ &= \beta_1 + \sum_k \beta_k \ln x_k + \varepsilon.\end{aligned}$$

[5] See Heckman (1979) who formulates this as a "specification error."

[6] See the appendix in Hausman and Wise (1977) and Greene (1983) as well.

[7] A survey by Greene (2008a) appears in Fried, Lovell, and Schmidt (2008). Kumbhakar and Lovell (2000) is a comprehensive reference on the subject.

The frontier for any particular firm is $h(\mathbf{x}, \boldsymbol{\beta}) + v$, hence the name *stochastic frontier*. The inefficiency term is u, a random variable of particular interest in this setting. Because the data are in log terms, u is a measure of the percentage by which the particular observation fails to achieve the frontier, ideal production rate.

To complete the specification, they suggested two possible distributions for the inefficiency term: the absolute value of a normally distributed variable, which has the truncated at zero distribution shown in Figure 19.1, and an exponentially distributed variable. The density functions for these two compound variables are given by Aigner, Lovell, and Schmidt; let $\varepsilon = v - u$, $\lambda = \sigma_u/\sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(z) =$ the probability to the left of z in the standard normal distribution (see Section B.4.1). For the "half-normal" model,

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \lambda, \sigma) = \left[-\ln \sigma + \left(\frac{1}{2}\right) \ln \frac{2}{\pi} - \frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \ln \Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)\right],$$

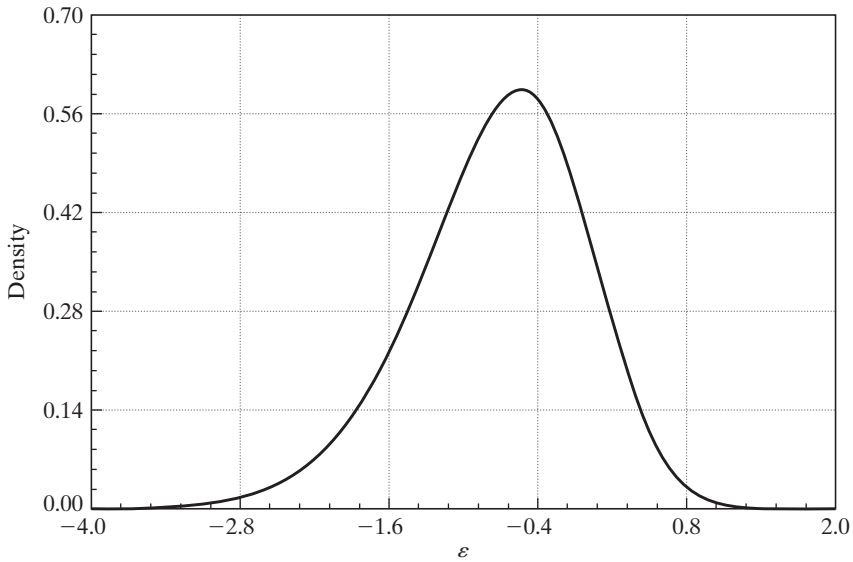
whereas for the exponential model

$$\ln h(\varepsilon_i \mid \boldsymbol{\beta}, \theta, \sigma_v) = \left[\ln \theta + \frac{1}{2}\theta^2 \sigma_v^2 + \theta \varepsilon_i + \ln \Phi\left(-\frac{\varepsilon_i}{\sigma_v} - \theta \sigma_v\right)\right].$$

Both these distributions are asymmetric. We thus have a regression model with a nonnormal distribution specified for the disturbance. The disturbance, ε, has a nonzero mean as well; $E[\varepsilon] = -\sigma_u(2/\pi)^{1/2}$ for the half-normal model and $-1/\theta$ for the exponential model. Figure 19.2 illustrates the density for the half-normal model with $\sigma = 1$ and $\lambda = 2$. By writing $\beta_0 = \beta_1 + E[\varepsilon]$ and $\varepsilon^* = \varepsilon - E[\varepsilon]$, we obtain a more conventional formulation

$$\ln y = \beta_0 + \sum_k \beta_k \ln x_k + \varepsilon^*,$$

FIGURE 19.2 Density for the Disturbance in the Stochastic Frontier Model.



which does have a disturbance with a zero mean but an asymmetric, nonnormal distribution. The asymmetry of the distribution of ε^* does not negate our basic results for least squares in this classical regression model. This model satisfies the assumptions of the Gauss–Markov theorem, so least squares is unbiased and consistent (save for the constant term) and efficient among linear unbiased estimators. In this model, however, the maximum likelihood estimator is not linear, and it is more efficient than least squares.

The log-likelihood function for the half normal model is given in ALS (1977):

$$\ln L = -n \ln \sigma + \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^{n} \ln \Phi \left(\frac{-\varepsilon_i \lambda}{\sigma} \right). \qquad \textbf{(19-11)}$$

Maximization programs for this model are built into modern software packages such as *Stata*, *NLOGIT*, and *TSP*. The log-likelihood is simple enough that it can also be readily adapted to the generic optimization routines in, for example, *MatLab* or *Gauss*. Some treatments in the literature use the parameterization employed by Battese and Coelli (1992) and Coelli (1996), $\gamma = \sigma_u^2/\sigma^2$. This is a one-to-one transformation of λ; $\lambda = (\gamma/(1-\gamma))^{1/2}$, so which parameterization is employed is a matter of convenience; the empirical results will be the same. The log-likelihood function for the exponential model can be built up from the density given earlier. For the half-normal model, we would also rely on the invariance of maximum likelihood estimators to recover estimates of the structural variance parameters, $\sigma_v^2 = \sigma^2/(1+\lambda^2)$ and $\sigma_u^2 = \sigma^2\lambda^2/(1+\lambda^2)$.[8] (Note, the variance of the truncated variable, u_i, is not σ_u^2; using (19-2), it reduces to $(1-2/\pi)\sigma_u^2$.) In addition, a structural parameter of interest is the proportion of the total variance of ε that is due to the inefficiency term. For the half-normal model, $\text{Var}[\varepsilon] = \text{Var}[u] + \text{Var}[v] = (1-2/\pi)\sigma_u^2 + \sigma_v^2$ whereas for the exponential model, the counterpart is $1/\theta^2 + \sigma_v^2$.

Modeling in the stochastic frontier setting is rather unlike what we are accustomed to up to this point, in that the disturbance, specifically u_i, not the model parameters, is the central focus of the analysis. The reason is that in this context, the disturbance, u_i, rather than being the catchall for the unknown and unknowable factors omitted from the equation, has a particular interpretation—it is the firm-specific inefficiency. Ideally, we would like to estimate u_i for each firm in the sample to compare them on the basis of their productive efficiency. Unfortunately, the data do not permit a direct estimate, because with estimates of $\boldsymbol{\beta}$ in hand, we are only able to compute a direct estimate of $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. Jondrow et al. (1982), however, have derived a useful approximation that is now the standard measure in these settings,

$$E[u_i|\varepsilon_i] = \frac{\sigma\lambda}{1+\lambda^2} \left[\frac{\phi(z_i)}{1-\Phi(z_i)} - z_i \right], \; z_i = \frac{\varepsilon_i \lambda}{\sigma}$$

for the half-normal model, and

$$E[u_i|\varepsilon_i] = z_i + \sigma_v \frac{\phi(z_i/\sigma_v)}{\Phi(z_i/\sigma_v)}, \; z_i = -\left(\varepsilon_i + \theta\sigma_v^2\right)$$

for the exponential model.These values can be computed using the maximum likelihood estimates of the structural parameters in the model. In some cases in which researchers

[8] A vexing problem for estimation of the model is that if the ordinary least squares residuals are skewed in the positive (wrong) direction (See Figure 19.2), OLS with $\hat{\lambda} = 0$ will be the MLE. OLS residuals with a positive skew are apparently inconsistent with a model in which, in theory, they should have a negative skew. [See Waldman (1982) for theoretical development of this result.]

are interested in discovering best practice [e.g., WHO (2000), Tandon et al. (2000)], the estimated values are sorted and the ranks of the individuals in the sample become of interest.

Research in this area since the methodological developments beginning in the 1930s and the building of the empirical foundations in 1977 and 1982 has proceeded in several directions. Most theoretical treatments of "inefficiency" as envisioned here attribute it to aspects of management of the firm. It remains to establish a firm theoretical connection between the theory of firm behavior and the stochastic frontier model as a device for measurement of inefficiency.

In the context of the model, many studies have developed alternative, more flexible functional forms that (it is hoped) can provide a more realistic model for inefficiency. Two that are relevant in this chapter are Stevenson's (1980) truncated normal model and the normal-gamma frontier. One intuitively appealing form of the truncated normal model is

$$U_i \sim N\left[\mu + \mathbf{z}_i'\boldsymbol{\alpha}, \sigma_u^2\right],$$
$$u_i = |U_i|.$$

The original normal–half-normal model results if μ equals zero and $\boldsymbol{\alpha}$ equals zero. This is a device by which the environmental variables noted in the next paragraph can enter the model of inefficiency. A truncated normal model is presented in Example 19.3. The half-normal, truncated normal, and exponential models all take the form of distribution shown in Figure 19.1. The gamma model,

$$f(u) = [\theta^P/\Gamma(P)]\exp(-\theta u)u^{P-1},$$

is a flexible model that presents the advantage that the distribution of inefficiency can move away from zero. If P is greater than one, then the density at $u = 0$ equals zero and the entire distribution moves away from the origin. The implication is that the distribution of inefficiency among firms can move away from zero. The gamma model is estimated by simulation methods—either Bayesian MCMC [Huang (2003) and Tsionas (2002)] or maximum simulated likelihood [Greene (2003)]. Many other functional forms have been proposed. [See Greene (2008) for a survey.]

There are usually elements in the environment in which the firm operates that impact the firm's output and/or costs but are not, themselves, outputs, inputs, or input prices. In example 19.3, the costs of the Swiss railroads are affected by three variables; track width, long tunnels, and curvature. It is not yet specified how such factors should be incorporated into the model; four candidates are in the mean and variance of u_i, directly in the function, or in the variance of v_i. [See Hadri, Guermat, and Whittaker (2003) and Kumbhakar (1997c).] All of these can be found in the received studies. This aspect of the model was prominent in the discussion of the famous World Health Organization efficiency study of world health systems [WHO (2000), Tandon, Murray, Lauer, and Evans (2000), and Greene (2004)]. In Example 19.3, we have placed the environmental factors in the mean of the inefficiency distribution. This produces a rather extreme set of results for the JLMS estimates of inefficiency—many railroads are estimated to be extremely inefficient. An alternative formulation would be a "heteroscedastic" model in which $\sigma_{u,i} = \sigma_u \exp(z_i'\boldsymbol{\delta})$ or $\sigma_{v,i} = \sigma_v \exp(\mathbf{z}_i'\boldsymbol{\eta})$, or both. We can see from the JLMS formula that the term heteroscedastic is actually a bit misleading, since both

standard deviations enter (now) λ_i, which is, in turn, a crucial parameter in the mean of inefficiency.

How should inefficiency be modeled in panel data, such as in our example? It might be tempting to treat it as a time-invariant "effect" [as in Schmidt and Sickles (1984) and Pitt and Lee (1984) in two pioneering papers]. Greene (2004) argued that a preferable approach would be to allow inefficiency to vary freely over time in a panel, and to the extent that there is a common time-invariant effect in the model, that should be treated as unobserved heterogeneity, not inefficiency. A string of studies, including Battese and Coelli (1992, 1995), Cuesta (2000), Kumbhakar (1997a) Kumbhakar and Orea (2004), and many others have proposed hybrid forms that treat the core random part of inefficiency as a time-invariant firm-specific effect that is modified over time by a deterministic, possibly firm-specific, function. The Battese-Coelli form,

$$u_{it} = \exp[-\eta(t - T)]|U_i| \text{ where } U_i\, N\left[0, \sigma_u^2\right],$$

has been used in a number of applications. Cuesta (2000) suggests allowing η to vary across firms, producing a model that bears some relationship to a fixed-effects specification. This thread of the literature is one of the most active ongoing pursuits.

Is it reasonable to use a possibly restrictive parametric approach to modeling inefficiency? Sickles (2005) and Kumbhakar, Simar, Park, and Tsionas (2007) are among numerous studies that have explored less parametric approaches to efficiency analysis. Proponents of **data envelopment analysis** [see, e.g., Simar and Wilson (2000, 2007)] have developed methods that impose absolutely no parametric structure on the production function. Among the costs of this high degree of flexibility is a difficulty to include environmental effects anywhere in the analysis, and the uncomfortable implication that any unmeasured heterogeneity of any sort is necessarily included in the measure of inefficiency. That is, data envelopment analysis returns to the deterministic frontier approach where this section began.

Example 19.3 Stochastic Cost Frontier for Swiss Railroads

Farsi, Filippini, and Greene (2005) analyzed the cost efficiency of Swiss railroads. In order to use the stochastic frontier approach to analyze costs of production, rather than production, we rely on the fundamental duality of production and cost [see Samuelson (1938), Shephard (1953), and Kumbhakar and Lovell (2000)]. An appropriate cost frontier model for a firm that produces more than one output—the Swiss railroads carry both freight and passengers—will appear as the following:

$$\ln(C/P_K) = \alpha + \sum_{k=1}^{K-1} \beta_k \ln(P_k/P_K) + \sum_{m=1}^{M} \gamma_m \ln Q_m + v + u.$$

The requirement that the cost function be homogeneous of degree one in the input prices has been imposed by normalizing total cost, C, and the first $K - 1$ prices by the Kth input price. In this application, the three factors are labor, capital, and electricity—the third is used as the numeraire in the cost function. Notice that the inefficiency term, u, enters the cost function positively; actual cost is above the frontier cost. [The MLE is modified simply by replacing ε_i with $-\varepsilon_i$ in (19-11).] In analyzing costs of production, we recognize that there is an additional source of inefficiency that is absent when we analyze production. On the production side, inefficiency measures the difference between output and frontier output, which arises because of technical inefficiency. By construction, if output fails to reach the efficient level for the given input usage, then costs must be higher than frontier costs. However, costs can be excessive even if the firm is technically efficient if it is "allocatively inefficient." That is, the firm can be technically efficient while not using inputs in the cost minimizing mix (equating

the ratio of marginal products to the input price ratios). It follows that on the cost side, "u" can contain both elements of inefficiency while on the production side, we would expect to measure only technical inefficiency. [See Kumbhakar (1997b).]

The data for this study are an unbalanced panel of 50 railroads with T_i ranging from 1 to 13. (Thirty-seven of the firms are observed 13 times, 8 are observed 12 times, and the remaining 5 are observed 10, 7, 7, 3, and 1 times.) The variables we will use here are

CT: Total costs adjusted for inflation (1,000 Swiss franc)
QP: Total passenger-output in passenger-kilometers
QF: Total goods-output in ton-kilometers
PL: Labor price adjusted for inflation (in Swiss Francs per person per year)
PK: Capital price with capital stock proxied by total number of seats
PE: Price of electricity (Swiss franc per kWh)

Logs of costs and prices (ln *CT*, ln *PK*, ln *PL*) are normalized by *PE*. We will also use these environmental variables:

NARROW_T: Dummy for the networks with narrow track (1 m wide) The usual width is 1.435m.
TUNNEL: Dummy for networks that have tunnels with an average length of more than 300 meters.
VIRAGE: Dummy for the networks whose minimum radius of curvature is 100 meters or less.

The full data set is given in Appendix Table F19.1. Several other variables not used here are presented in the appendix table. In what follows, we will ignore the panel data aspect of the data set. This would be a focal point of a more extensive study.

There have been dozens of models proposed for the inefficiency component of the stochastic frontier model. Table 19.1 presents several different forms. The basic half-normal model is given in the first column. The estimated cost function parameters across the different

TABLE 19.1 Estimated Stochastic Frontier Cost Functions[a]

	Model					
Variable	*Half Normal*	*Truncated Normal*	*Exponential*	*Gamma*	*Heterosced*	*Heterogen*
Constant	−10.0799	−9.80624	−10.1838	−10.1944	−9.82189	−10.2891
ln QP	0.64220	0.62573	0.64403	0.64401	0.61976	0.63576
ln QF	0.06904	0.07708	0.06803	0.06810	0.07970	0.07526
ln PK	0.26005	0.26625	0.25883	0.25886	0.25464	0.25893
ln PL	0.53845	0.50474	0.56138	0.56047	0.53953	0.56036
Constant		0.44116			−2.48218[b]	
Narrow		0.29881			2.16264[b]	0.14355
Virage		−0.20738			−1.52964[b]	−0.10483
Tunnel		0.01118			0.35748[b]	−0.01914
σ	0.44240	0.38547	(0.34325)	(0.34288)	0.45392[c]	0.40597
λ	1.27944	2.35055				0.91763
P			1.0000	1.22920		
θ			13.2922	12.6915		
σ_u	(0.34857)	(0.35471)	(0.07523)	(0.09685)	0.37480[c]	0.27448
σ_v	(0.27244)	(0.15090)	0.33490	0.33197	0.25606	0.29912
Mean E[u|ε]	0.27908	0.52858	0.075232	0.096616	0.29499	0.21926
ln L	−210.495	−200.67	−211.42	−211.091	−201.731	−208.349

[a]Estimates in parentheses are derived from other MLEs.
[b]Estimates used in computation of σ_u.
[c]Obtained by averaging $\lambda = \sigma_{u,i}/\sigma_v$ over observations.

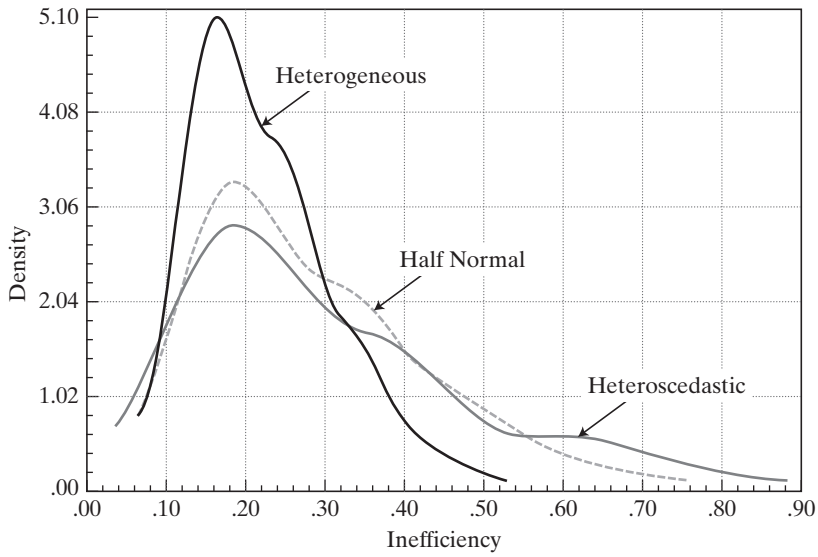


FIGURE 19.3 Kernel Density Estimator for JLMS Estimates.

forms are broadly similar, as might be expected as (α, $\boldsymbol{\beta}$) are consistently estimated in all cases. There are fairly pronounced differences in the implications for the components of ε, however.

There is an ambiguity in the model as to whether modifications to the distribution of u_i will affect the mean of the distribution, the variance, or both. The following results suggest that it is both for these data. The gamma and exponential models appear to remove most of the inefficiency from the data. Note that the estimates of σ_u are considerably smaller under these specifications, and σ_v is correspondingly larger. The second to last row shows the sample averages of the Jondrow estimators—this estimates $E_\varepsilon E[u|\varepsilon] = E[u]$. There is substantial difference across the specifications.

The estimates in the rightmost two columns illustrate two different placements of the measured heterogeneity: in the variance of u_i and directly in the cost function. The log-likelihood function appears to favor the first of these. However, the models are not nested and involve the same number of parameters. We used the Vuong test (see Section 14.6.6), instead and obtained a value of -2.65 in favor of the heteroscedasticity model. Figure 19.3 describes the values of $E[u_i|\varepsilon_i]$ estimated for the sample observations for the half-normal, heteroscedastic and heterogeneous models. The smaller estimate of σ_u for the third of these is evident in the figure, which suggests a somewhat tighter concentration of values than the other two.

19.3 CENSORED DATA

A very common problem in microeconomic data is **censoring** of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Some examples that have appeared in the empirical literature are as follows:[9]

1. Household purchases of durable goods [Tobin (1958)]
2. The number of extramarital affairs [Fair (1977, 1978)]

[9]More extensive listings may be found in Amemiya (1984) and Maddala (1983).

3. The number of hours worked by a woman in the labor force [Quester and Greene (1982)]
4. The number of arrests after release from prison [Witte (1980)]
5. Household expenditure on various commodity groups [Jarque (1987)]
6. Vacation expenditures [Melenberg and van Soest (1996)]

Each of these studies analyzes a dependent variable that is zero for a significant fraction of the observations. Conventional regression methods fail to account for the qualitative difference between *limit* (zero) observations and *nonlimit* (continuous) observations.

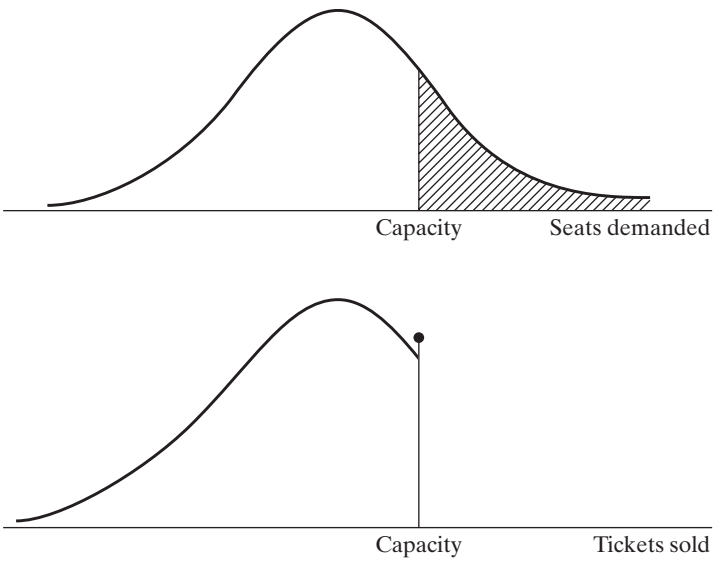
19.3.1 THE CENSORED NORMAL DISTRIBUTION

The relevant distribution theory for a **censored variable** is similar to that for a truncated one. Once again, we begin with the normal distribution, as much of the received work has been based on an assumption of normality. We also assume that the censoring point is zero, although this is only a convenient normalization. In a truncated distribution, only the part of distribution above $y = 0$ is relevant to our computations. To make the distribution integrate to one, we scale it up by the probability that an observation in the untruncated population falls in the range that interests us. When data are censored, the distribution *that applies to the sample data* is a mixture of discrete and continuous distributions. Figure 19.4 illustrates the effects.

To analyze this distribution, we define a new random variable y transformed from the original one, y^*, by

$$y = 0 \quad \text{if } y^* \leq 0,$$
$$y = y^* \quad \text{if } y^* > 0.$$

FIGURE 19.4 Partially Censored Distribution.



The distribution that applies if $y^* \sim N[\mu, \sigma^2]$ is $\text{Prob}(y = 0) = \text{Prob}(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - \Phi(\mu/\sigma)$, and if $y^* > 0$, then y has the density of y^*.

This distribution is a mixture of discrete and continuous parts. The total probability is one, as required, but instead of scaling the second part, we simply assign the full probability in the censored region to the censoring point, in this case, zero.

THEOREM 19.3 Moments of the Censored Normal Variable

If $y^* \sim N[\mu, \sigma^2]$ *and* $y = a$ *if* $y^* \leq a$ *or else* $y = y^*$, *then*

$$E[y] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda),$$

and

$$\text{Var}[y] = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi],$$

where

$$\Phi[(a - \mu)/\sigma] = \Phi(\alpha) = \text{Prob}(y^* \leq a) = \Phi, \quad \lambda = \phi/(1 - \Phi),$$

and

$$\delta = \lambda^2 - \lambda\alpha.$$

Proof: *For the mean,*

$$\begin{aligned} E[y] &= \text{Prob}(y = a) \times E[y \mid y = a] + \text{Prob}(y > a) \times E[y \mid y > a] \\ &= \text{Prob}(y^* \leq a) \times a + \text{Prob}(y^* > a) \times E[y^* \mid y^* > a] \\ &= \Phi a + (1 - \Phi)(\mu + \sigma\lambda) \end{aligned}$$

using Theorem 19.2. For the variance, we use a counterpart to the decomposition in (B-69), that is, $\text{Var}[y] = E[\text{conditional variance}] + \text{Var}[\text{conditional mean}]$, *and Theorem 19.2.*

For the special case of $a = 0$, the mean simplifies to

$$E[y \mid a = 0] = \Phi(\mu/\sigma)(\mu + \sigma\lambda), \quad \text{where } \lambda = \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}.$$

For censoring of the upper part of the distribution instead of the lower, it is only necessary to reverse the role of Φ and $1 - \Phi$ and redefine λ as in Theorem 19.2.

Example 19.4 Censored Random Variable

We are interested in the number of tickets *demanded* for events at a certain arena. Our only measure is the number actually *sold*. Whenever an event sells out, however, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored when it is transformed to obtain the number sold. Suppose that the arena in question has 20,000 seats and, in a recent season, sold out 25 percent of the time. If the average attendance, including sellouts, was 18,000, then what are the mean and standard deviation of the demand for seats? According to Theorem 19.3, the 18,000 is an estimate of

$$E[\text{sales}] = 20{,}000(1 - \Phi) + [\mu + \sigma\lambda]\Phi.$$

Because this is censoring from above, rather than below, $\lambda = -\phi(\alpha)/\Phi(\alpha)$. The argument of Φ, ϕ, and λ is $\alpha = (20{,}000 - \mu)/\sigma$. If 25 percent of the events are sellouts, then $\Phi = 0.75$. Inverting the standard normal at 0.75 gives $\alpha = 0.675$. In addition, if $\alpha = 0.675$,

then $-\phi(0.675)/0.75 = \lambda = -0.424$. This result provides two equations in μ and σ, (a) $18{,}000 = 0.25(20{,}000) + 0.75(\mu - 0.424\sigma)$ and (b) $0.675\sigma = 20{,}000 - \mu$. The solutions are $\sigma = 2426$ and $\mu = 18{,}362$.

For comparison, suppose that we were told that the mean of 18,000 applies only to the events that were *not* sold out and that, on average, the arena sells out 25 percent of the time. Now our estimates would be obtained from the equations (a) $18{,}000 = \mu - 0.424\sigma$ and (b) $0.675\sigma = 20{,}000 - \mu$. The solutions are $\sigma = 1820$ and $\mu = 18{,}772$.

19.3.2 THE CENSORED REGRESSION (TOBIT) MODEL

The regression model based on the preceding discussion is referred to as the **censored regression model** or the **tobit model** [in reference to Tobin (1958), where the model was first proposed]. The regression is obtained by making the mean in the preceding correspond to a classical regression model. The general formulation is usually given in terms of an index function,

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$
$$y_i = 0 \quad \text{if } y_i^* \le 0,$$
$$y_i = y_i^* \quad \text{if } y_i^* > 0.$$

There are potentially three conditional mean functions to consider, depending on the purpose of the study. For the index variable, sometimes called the *latent variable,* $E[y_i^* \mid \mathbf{x}_i]$ is $\mathbf{x}_i'\boldsymbol{\beta}$. If the data are always censored, however, then this result will usually not be useful. Consistent with Theorem 19.3, for an observation randomly drawn from the population, which may or may not be censored,

$$E[y_i \mid \mathbf{x}_i] = \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i),$$

where

$$\lambda_i = \frac{\phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]} = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}. \tag{19-12}$$

Finally, if we intend to confine our attention to uncensored observations, then the results for the truncated regression model apply. The limit observations should not be discarded, however, because the truncated regression model is no more amenable to least squares than the censored data model. It is an unresolved question which of these functions should be used for computing predicted values from this model. Intuition suggests that $E[y_i \mid \mathbf{x}_i]$ is correct, but authors differ on this point. For the setting in Example 19.4, for predicting the number of tickets sold, say, to plan for an upcoming event, the censored mean is obviously the relevant quantity. On the other hand, if the objective is to study the need for a new facility, then the mean of the latent variable y_i^* would be more interesting.

There are differences in the partial effects as well. For the index variable,

$$\frac{\partial E[y_i^* \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}.$$

But this result is not what will usually be of interest, because y_i^* is unobserved. For the observed data, y_i, the following general result will be useful:[10]

[10]See Greene (1999) for the general result and Rosett and Nelson (1975) and Nakamura and Nakamura (1983) for applications based on the normal distribution.

THEOREM 19.4 Partial Effects in the Censored Regression Model

In the censored regression model with latent regression $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ and observed dependent variable, $y = a$ if $y^* \leq a$, $y = b$ if $y^* \geq b$, and $y = y^*$ otherwise, where a and b are constants, let $f(\varepsilon)$ and $F(\varepsilon)$ denote the density and cdf of ε. Assume that ε is a continuous random variable with mean 0 and variance σ^2, and $f(\varepsilon \mid \mathbf{x}) = f(\varepsilon)$. Then

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta} \times \text{Prob}[a < y^* < b].$$

Proof: *By definition,*

$$E[y \mid \mathbf{x}] = a\,\text{Prob}[y^* \leq a \mid \mathbf{x}] + b\,\text{Prob}[y^* \geq b \mid \mathbf{x}] + \text{Prob}[a < y^* < b \mid \mathbf{x}]E[y^* \mid a < y^* < b \mid \mathbf{x}].$$

Let $\alpha_j = (j - \mathbf{x}'\boldsymbol{\beta})/\sigma$, $F_j = F(\alpha_j)$, $f_j = f(\alpha_j)$, and $j = a, b$. Then

$$E[y \mid \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)E[y^* \mid a < y^* < b, \mathbf{x}].$$

Because $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma[(y^* - \boldsymbol{\beta}'\mathbf{x})/\sigma]$, the conditional mean may be written

$$E[y^* \mid a < y^* < b, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma E\left[\frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \,\middle|\, \frac{a - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{b - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right]$$

$$= \mathbf{x}'\boldsymbol{\beta} + \sigma \int_{\alpha_a}^{\alpha_b} \frac{(\varepsilon/\sigma) f(\varepsilon/\sigma)}{F_b - F_a} d\left(\frac{\varepsilon}{\sigma}\right).$$

Collecting terms, we have

$$E[y \mid \mathbf{x}] = aF_a + b(1 - F_b) + (F_b - F_a)\boldsymbol{\beta}'\mathbf{x} + \sigma \int_{\alpha_a}^{\alpha_b} \left(\frac{\varepsilon}{\sigma}\right) f\left(\frac{\varepsilon}{\sigma}\right) d\left(\frac{\varepsilon}{\sigma}\right).$$

Now, differentiate with respect to $\mathbf{x}$. The only complication is the last term, for which the differentiation is with respect to the limits of integration. We use Leibnitz's theorem and use the assumption that $f(\varepsilon)$ does not involve $\mathbf{x}$. Thus,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) a f_a - \left(\frac{-\boldsymbol{\beta}}{\sigma}\right) b f_b + (F_b - F_a)\boldsymbol{\beta} + (\mathbf{x}'\boldsymbol{\beta})(f_b - f_a)\left(\frac{-\boldsymbol{\beta}}{\sigma}\right) + \sigma[\alpha_b f_b - \alpha_a f_a]\left(\frac{-\boldsymbol{\beta}}{\sigma}\right).$$

After inserting the definitions of α_a and α_b, and collecting terms, we find all terms sum to zero save for the desired result,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = (F_b - F_a)\boldsymbol{\beta} = \boldsymbol{\beta} \times \text{Prob}[a < y_i^* < b].$$

Note that this general result includes censoring in either or both tails of the distribution, and it does not assume that ε is normally distributed. For the standard case with censoring at zero and normally distributed disturbances, the result specializes to

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right).$$

Although not a formal result, this does suggest a reason why, in general, least squares estimates of the coefficients in a tobit model usually resemble the MLEs times the proportion of nonlimit observations in the sample.

McDonald and Moffitt (1980) suggested a useful decomposition of $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i$,

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \times \left\{\Phi_i[1 - \lambda_i(\alpha_i + \lambda_i)] + \phi_i(\alpha_i + \lambda_i)\right\},$$

where $\alpha_i = \mathbf{x}_i'\boldsymbol{\beta}/\sigma$, $\Phi_i = \Phi(\alpha_i)$ and $\lambda_i = \phi_i/\Phi_i$. Taking the two parts separately, this result decomposes the slope vector into

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \text{Prob}[y_i > 0]\frac{\partial E[y_i \mid \mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} + E[y_i \mid \mathbf{x}_i, y_i > 0]\frac{\partial \text{Prob}[y_i > 0]}{\partial \mathbf{x}_i}.$$

Thus, a change in $\mathbf{x}_i$ has two effects: It affects the conditional mean of y_i^* in the positive part of the distribution, and it affects the probability that the observation will fall in that part of the distribution.

19.3.3 ESTIMATION

The tobit model has become so routine and been incorporated in so many computer packages that despite formidable obstacles in years past, estimation is now essentially on the level of ordinary linear regression. The log-likelihood for the censored regression model is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right] + \sum_{y_i=0} \ln\left[1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \quad \textbf{(19-13)}$$

The two parts correspond to the classical regression for the nonlimit observations and the relevant probabilities for the limit observations, respectively. This likelihood is a nonstandard type, because it is a mixture of discrete and continuous distributions. In a seminal paper, Amemiya (1973) showed that despite the complications, proceeding in the usual fashion to maximize $\ln L$ would produce an estimator with all the familiar desirable properties attained by MLEs.

The log-likelihood function is fairly involved, but **Olsen's** (1978) **reparameterization** simplifies things considerably. With $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\theta = 1/\sigma$, the log-likelihood is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}[\ln(2\pi) - \ln\,\theta^2 + (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2] + \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma})]. \quad \textbf{(19-14)}$$

The results in this setting are now very similar to those for the truncated regression. The Hessian is always negative definite, so Newton's method is simple to use and usually converges quickly. After convergence, the original parameters can be recovered using $\sigma = 1/\theta$ and $\beta = \boldsymbol{\gamma}/\theta$. The asymptotic covariance matrix for these estimates can be obtained from that for the estimates of $[\boldsymbol{\gamma}, \theta]$ using the **delta method**:

TABLE 19.2 Tobit Estimates of an Hours Worked Equation

	White Wives		*Black Wives*		*Least*	*Scaled*
	Coefficient	***Slope***	***Coefficient***	***Slope***	***Squares***	***OLS***
Constant	−1803.13		−2753.87			
	(−8.64)		(−9.68)			
Small kids	−1324.84	−385.89	−824.19	−376.53	−352.63	−766.56
	(−19.78)		(−10.14)			
Education	−48.08	−14.00	22.59	10.32	11.47	24.93
difference	(−4.77)		(1.96)			
Relative wage	312.07	90.90	286.39	130.93	123.95	269.46
	(5.71)		(3.32)			
Second marriage	175.85	51.51	25.33	11.57	13.14	28.57
	(3.47)		(0.41)			
Mean divorce	417.39	121.58	481.02	219.75	219.22	476.57
probability	(6.52)		(5.28)			
High divorce	670.22	195.22	578.66	264.36	244.17	530.80
probability	(8.40)		(5.33)			
σ	1559	618	1511	826		
Sample size	7459		2798			
Proportion working	0.29		0.46			

Est. Asy. Var$[\hat{\boldsymbol{\beta}}, \hat{\sigma}] = \hat{\mathbf{J}}$ Asy. Var$[\hat{\boldsymbol{\gamma}}, \hat{\theta}]\hat{\mathbf{J}}'$, where

$$\mathbf{J} = \begin{bmatrix} \partial\boldsymbol{\beta}/\partial\boldsymbol{\gamma}' & \partial\beta/\partial\theta \\ \partial\sigma/\partial\boldsymbol{\gamma}' & \partial\sigma/\partial\theta \end{bmatrix} = \begin{bmatrix} (1/\theta)\mathbf{I} & (-1/\theta^2)\gamma \\ \mathbf{0}' & (-1/\theta^2) \end{bmatrix}.$$

Researchers often compute ordinary least squares estimates despite their inconsistency. Almost without exception, it is found that the OLS estimates are smaller in absolute value than the MLEs. A striking empirical regularity is that the maximum likelihood estimates can often be approximated by dividing the OLS estimates by the proportion of nonlimit observations in the sample.[11] The effect is illustrated in the last two columns of Table 19.2. Another strategy is to discard the limit observations, but we now see that just trades the censoring problem for the truncation problem.

Example 19.5 Estimated Tobit Equations for Hours Worked

In their study of the number of hours worked in a survey year by a large sample of wives, Quester and Greene (1982) were interested in whether wives whose marriages were statistically more likely to dissolve hedged against that possibility by spending, on average, more time working. They reported the tobit estimates given in Table 19.2. The last figure in the table implies that a very large proportion of the women reported zero hours, so least squares regression would be inappropriate.

The figures in parentheses are the ratio of the coefficient estimate to the estimated asymptotic standard error. The dependent variable is hours worked in the survey year. "Small kids" is a dummy variable indicating whether there were children in the household. The "education difference" and "relative wage" variables compare husband and wife on these two dimensions. The wage rate used for wives was predicted using a previously estimated regression model and is thus available for all individuals, whether working or not. "Second marriage" is a

[11] This concept is explored further in Greene (1980b), Goldberger (1981), and Chung and Goldberger (1984).

dummy variable. Divorce probabilities were produced by a large microsimulation model presented in another study [Orcutt, Caldwell, and Wertheimer (1976)]. The variables used here were dummy variables indicating "mean" if the predicted probability was between 0.01 and 0.03 and "high" if it was greater than 0.03. The "slopes" are the marginal effects described earlier.

Note the marginal effects compared with the tobit coefficients. Likewise, the estimate of σ is quite misleading as an estimate of the standard deviation of hours worked.

The effects of the divorce probability variables were as expected and were quite large. One of the questions raised in connection with this study was whether the divorce probabilities could reasonably be treated as independent variables. It might be that for these individuals, the number of hours worked was a significant determinant of the probability.

19.3.4 TWO-PART MODELS AND CORNER SOLUTIONS

The tobit model contains a restriction that might be unreasonable in an economic setting. Consider a behavioral outcome, $y =$ charitable donation. Two implications of the tobit model are that

$$\text{Prob}(y > 0 \mid \mathbf{x}) = \text{Prob}(\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$$

and [from (19-7)]

$$E[y \mid y > 0, \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma).$$

Differentiating both of these, we find from (17-11) and (19-8),

$$\partial \text{Prob}(y > 0 \mid \mathbf{x})/\partial \mathbf{x} = [\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)/\sigma]\boldsymbol{\beta} = \text{ a positive multiple of } \boldsymbol{\beta},$$

$$\partial E[y \mid y > 0, \mathbf{x}]/\partial \mathbf{x} = \{[1 - \delta(\mathbf{x}'\boldsymbol{\beta}/\sigma)]/\sigma\}\beta = \text{ a positive multiple of } \boldsymbol{\beta}.$$

Thus, any variable that appears in the model affects the participation probability and the intensity equation with the same sign. In the case suggested, for example, it is conceivable that age might affect participation and intensity in different directions. Fin and Schmidt (1984) suggest another application, loss due to fire in buildings; older buildings might be more likely to have fires but, because of the greater value of newer buildings, the actual damage might be greater in newer buildings. This fact would require the coefficient on age to have different signs in the two functions, which is impossible in the tobit model because they are the same coefficient.

In an early study in this literature, Cragg (1971) proposed a somewhat more general model in which the probability of a limit observation is independent of the regression model for the nonlimit data. One can imagine, for instance, the decision of whether or not to purchase a car as being different from the decision of how much to spend on the car, having decided to buy one.

A more general model that accommodates these objections is as follows:

1. Participation equation

$$\begin{aligned} &\text{Prob}[y_i^* > 0] = \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), && d_i = 1 \text{ if } y_i^* > 0, \\ &\text{Prob}[y_i^* \le 0] = 1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), && d_i = 0 \text{ if } y_i^* \le 0. \end{aligned} \qquad \textbf{(19-15)}$$

2. Intensity equation for nonlimit observations

$$E[y_i \mid d_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i,$$

according to Theorem 19.2. This two-part model is a combination of the truncated regression model of Section 19.2 and the univariate probit model of Section 17.3, which suggests a method of analyzing it. Note that it is precisely the same approach we considered in Section 18.4.8 and Example 18.12 where we used a hurdle model to model doctor visits. The tobit model returns if $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$. The parameters of the regression (intensity) equation can be estimated independently using the truncated regression model of Section 19.2. An application is Melenberg and van Soest (1996).

Lin and Schmidt (1984) considered testing the restriction of the tobit model. Based only on the tobit model, they devised a Lagrange multiplier statistic that, although a bit cumbersome algebraically, can be computed without great difficulty. If one is able to estimate the truncated regression model, the tobit model, and the probit model separately, then there is a simpler way to test the hypothesis. The tobit log-likelihood is the sum of the log-likelihoods for the truncated regression and probit models. To show this result, add and subtract $\sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ in (19-13). This produces the log-likelihood for the truncated regression model (considered in the exercises) plus (17-20) for the probit model. Therefore, a likelihood ratio statistic can be computed using

$$\lambda = -2[\ln LT - (\ln LP + \ln LTR)],$$

where

LT = likelihood for the tobit model in (19-13), with the same coefficients

LP = likelihood for the probit model in (17-17), fit separately

LTR = likelihood for the truncated regression model, fit separately

The two-part model just considered extends the tobit model, but it stops a bit short of the generality we might achieve. In the preceding hurdle model, we have assumed that the same regressors appear in both equations. Although this produces a convenient way to retreat to the tobit model as a parametric restriction, it couples the two decisions perhaps unreasonably. In our example to follow, where we model extramarital affairs, the decision whether or not to spend any time in an affair may well be an entirely different decision from how much time to spend having once made that commitment. The obvious way to proceed is to reformulate the hurdle model as

1. Participation equation

$$\begin{aligned} &\text{Prob}[d_i^* > 0] = \Phi(\mathbf{z}_i'\boldsymbol{\gamma}), && d_i = 1 \text{ if } d_i^* > 0, \\ &\text{Prob}[d_i^* \le 0] = 1 - \Phi(\mathbf{z}_i'\boldsymbol{\gamma}), && d_i = 0 \text{ if } d_i^* \le 0. \end{aligned} \qquad \textbf{(19-16)}$$

2. Intensity equation for nonlimit observations

$$E[y_i \mid d_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i.$$

This extension, however, omits an important element; it seems unlikely that the two decisions would be uncorrelated; that is, the implicit disturbances in the equations should be correlated. The combination of these produces what has been labeled a **type-II tobit model**. [Amemiya (1985) identified five possible permutations of the model specification and observation mechanism. The familiar tobit model is type I; this is type-II.] The full model is

1. Participation equation

$$d_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + u_i, \qquad u_i \sim N[0,1]$$
$$d_i = 1 \text{ if } d_i^* > 0, \qquad 0 \text{ otherwise.}$$

2. Intensity equation

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N[0, \sigma^2].$$

3. Observation mechanism

(a) $y_i^* = 0$ if $d_i = 0$ and $y_i = y_i^*$ if $d_i = 1$.

(b) $y_i = y_i^*$ if $d_i = 1$ and y_i is unobserved if $d_i = 0$.

4. Endogeneity

$$(u_i, \varepsilon_i) \sim \text{ bivariate normal with correlation } \rho.$$

Mechanism (a) produces Amemiya's type II model. (Amemiya blends these two interpretations. In the statement of the model, he presents (a), but in the subsequent discussion, assumes (b). The difference is substantive if $\mathbf{x}_i$ is observed in case (b). Otherwise, they are the same, and "$y_i = 0$" is not actually meaningful. Amemiya notes, "$y_i^* = 0$ merely signifies the event $d_i^* \leq 0$." If $\mathbf{x}_i$ is observed when $d_i = 0$, then these observations will contribute to the likelihood for the full sample. If not, then they will not. We will develop this idea later when we consider Heckman's selection model [which is case (b) without observed $\mathbf{x}_i$ when $d_i = 0$].

There are two estimation strategies that can be used to fit the type II model. A two-step method can proceed as follows: The probit model for d_i can be estimated using maximum likelihood as shown in Section 17.3. For the second step, we make use of our theorems on truncation (and Theorem 19.5 that will appear later) to write

$$\begin{aligned} E[y_i \mid d_i = 1,\ \mathbf{x}_i,\ \mathbf{z}_i] &= \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i \mid d_i = 1,\ \mathbf{x}_i,\ \mathbf{z}_i] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})} \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma\lambda_i. \end{aligned} \tag{19-17}$$

Since we have estimated γ at step 1, we can compute $\hat{\lambda}_i = \phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})$ using the first-step estimates, and we can estimate β and $\theta = (\rho\sigma)$ by least squares regression of y_i on $\mathbf{x}_i$ and $\hat{\lambda}_i$. It will be necessary to correct the asymptotic covariance matrix that is computed for $(\hat{\boldsymbol{\beta}}, \hat{\theta})$. This is a template application of the Murphy and Topel (2002) results that appear in Section 14.7. The second approach is full information maximum likelihood, estimating all the parameters in both equations simultaneously. We will return to the details of estimation of the type II tobit model in Section 19.5 where we examine Heckman's model of "sample selection" model (which is the type II tobit model).

Many of the applications of the tobit model in the received literature are constructed not to accommodate censoring of the underlying data, but, rather, to model the appearance of a large cluster of zeros. Cragg's application is clearly related to this phenomenon. Consider, for example, survey data on purchases of consumer durables, firm expenditure on research and development, or consumer savings. In each case, the

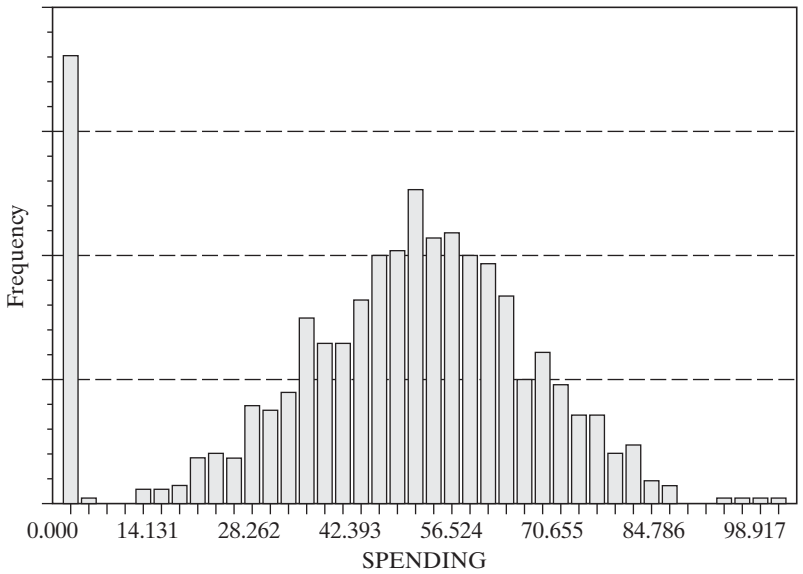


FIGURE 19.5 Hypothetical Spending Data.

observed data will consist of zero or some positive amount. Arguably, there are two decisions at work in these scenarios: First, whether to engage in the activity or not, and second, given that the answer to the first question is yes, how intensively to engage in it—how much to spend, for example. This is precisely the motivation behind the hurdle model. This specification has been labeled a "**corner solution model**"; see Wooldridge (2002a, pp. 518–519).

In practical terms, the difference between the **hurdle model** and the tobit model should be evident in the data. Often overlooked in tobit analyses is that the model predicts not only a cluster of zeros (or limit observations), but also a grouping of observations *near zero* (or the limit point). For example, the tobit model is surely misspecified for the sort of (hypothetical) spending data shown in Figure 19.5 for a sample of 1,000 observations. Neglecting for the moment the earlier point about the underlying decision process, Figure 19.6 shows the characteristic appearance of a (substantively) censored variable. The implication for the model builder is that an appropriate specification would consist of two equations, one for the "participation decision," and one for the distribution of the positive dependent variable. Formally, we might, continuing the development of Cragg's specification, model the first decision with a binary choice (e.g., probit or logit model). The second equation is a model for $y \mid y > 0$, for which the truncated regression model of Section 19.2.3 is a natural candidate. As we will see, this is essentially the model behind the sample selection treatment developed in Section 19.5.

Two practical issues frequently intervene at this point. First, one might well have a model in mind for the intensity (regression) equation, but none for the participation equation. This is the usual backdrop for the uses of the tobit model, which produces the considerations in the previous section. The second issue concerns the appropriateness

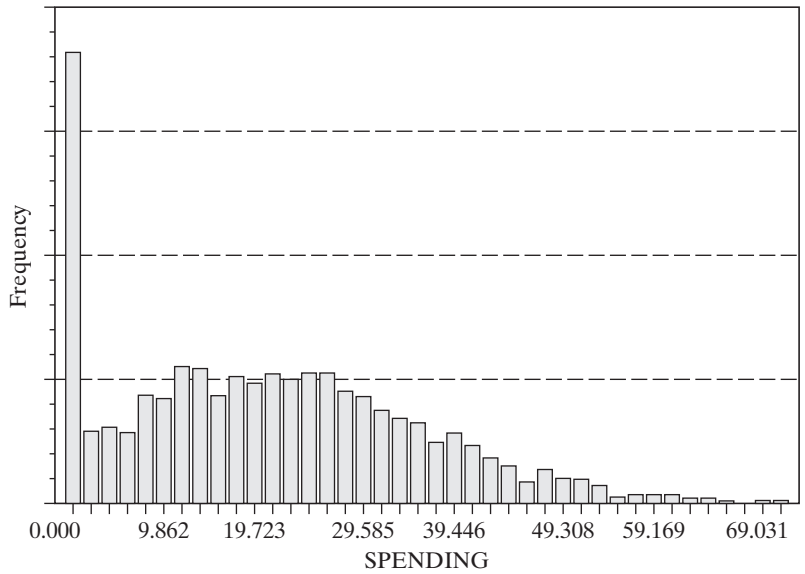


FIGURE 19.6 Hypothetical Censored Data.

of the truncation or censoring model to data such as those in Figure 19.6. If we consider only the nonlimit observations in Figure 19.5, the underlying distribution does not appear to be truncated at all. The truncated regression model in Section 19.2.3 fit to these data will not depart significantly from ordinary least squares [because the underlying probability in the denominator of (19-6) will equal one and the numerator will equal zero]. But, this is not the case of a tobit model forced on these same data. Forcing the model in (19-13) on data such as these will significantly distort the estimator—all else equal, it will significantly attenuate the coefficients, the more so the larger is the proportion of limit observations in the sample. Once again, this stands as a caveat for the model builder. The tobit model is manifestly misspecified for data such as those in Figure 19.5.

Example 19.6 Two-Part Model for Extramarital Affairs

In Example 18.9, we examined Fair's (1977) *Psychology Today* survey data on extramarital affairs. The 601 observations in the data set are mostly zero—451 of the 601. This feature of the data motivated the author to use a tobit model to analyze these data. In our example, we reconsidered the model, since the nonzero observations were a count, not a continuous variable. Another data set in Fair's study was the *Redbook Magazine* survey of 6,366 married women. Once again, the outcome variable of interest was extramarital affairs. However, in this instance, the outcome data were transformed to a measure of time spent, which, being continuous, lends itself more naturally to the tobit model we are studying here. The variables in the data set are as follows (excluding three unidentified and not used):

id = Identification number
C = Constant, value = 1
yrb = Constructed measure of time spent in extramarital affairs
v_1 = Rating of the marriage, coded 1 to 4
v_2 = Age, in years, aggregated
v_3 = Number of years married

TABLE 19.3 Estimated Censored Regression Models (*t*-ratios in parentheses)

	Model						
	Linear OLS	***Tobit***	***Truncated Regression***	***Probit***	***Tobit/σ***	***Hurdle Participation***	***Hurdle Intensity***
Constant	3.62346	7.83653	8.89449	2.21010	1.74189	1.56419	4.84602
	(13.63)	(10.98)	(2.90)	(12.60)		(17.75)	(5.87)
RateMarr	−0.42053	−1.53071	−0.44303	−0.42874	−0.34024	−0.42582	−0.24603
	(−14.79)	(−20.85)	(−1.45)	(−23.40)		(−23.61)	(−.46)
Age	−0.01457	−0.10514	−0.22394	−0.03542	−0.02337		−0.01903
	(−1.59)	(−4.24)	(−1.83)	(−5.87)			(−.77)
YrsMarr	−0.01599	0.12829	−0.94437	0.06563	0.02852		−0.16822
	(−1.62)	(4.86)	(−7.27)	(10.18)			(−6.52)
NumKids	−0.01705	−0.02777	−0.02280	−0.00394	−0.00617	0.14024	−0.28365
	(−.57)	(−0.36)	(−0.06)	(−0.21)		(11.55)	(−1.49)
Religious	−0.24374	−0.94350	−0.50490	−0.22281	−0.20972	−0.21466	−0.05452
	(−7.83)	(−11.11)	(−1.29)	(−10.88)		(−10.64)	(−0.19)
Education	−0.01743	−0.08598	−0.06406	−0.02373	−0.01911		0.00338
	(−1.24)	(−2.28)	(−0.38)	(−2.60)			(0.09)
Wife Occ.	0.06577	0.31284	0.00805	0.09539	0.06954		0.01505
	(2.10)	(3.82)	(0.02)	(4.75)			(0.19)
Hus. Occ.	0.00405	0.01421	−0.09946	0.00659	0.00316		−0.02911
	(0.19)	(0.26)	(−0.41)	(0.49)			(−0.53)
σ	2.14351	4.49887	5.46846				3.43748
ln L	$R^2 = 0.05479$	−7804.38	−3463.71	−3469.58			

v_4 = Number of children, top coded at 5
v_5 = Religiosity, 1 to 4, 1 = not, 4 = very
v_6 = Education, coded 9, 12, 14, 16, 17, 20
v_7 = Wife's Occupation—Hollingshead scale
v_8 = Husband's occupation—Hollingshead scale

This is a cross section of 6,366 observations with 4,313 zeros and 2,053 positive values.

Table 19.3 presents estimates of various models for *yrb*. The leftmost column presents the OLS estimates. The least squares estimator is inconsistent in this model. The empirical regularity that the OLS estimator appears to be biased toward zero, the more so is the smaller the proportion of limit observations. Here, the ratio, based on the tobit estimates in the second column, appears to be about 4 or 5 to 1. Likewise, the OLS estimator of σ appears to be greatly underestimated. This would be expected, as the OLS estimator is treating the limit observations, which have no variation in the dependent variable, as if they were nonlimit observations. The third set of results is the truncated regression estimator. In principle, the truncated regression estimator is also consistent. However, it will be less efficient as it is based on less information. In our example, this estimator seems to be quite erratic, again compared to the tobit estimator. Note, for example, the coefficient on years married, which, although it is "significant" in both cases, changes sign. The *t* ratio on Religiousness falls from −11.11 to −1.29 in the truncation model. The probit estimator based on $yrb > 0$ appears next. As a rough check on the corner solution aspect of our model, we would expect the normalized tobit coefficients $(\boldsymbol{\beta}/\sigma)$ to approximate the probit coefficients, which they appear to. However, the likelihood ratio statistic for testing the internal consistency based on the three estimated models is $2[7804.38 - 3463.71 - 3469.58] = 1742.18$ with nine degrees of freedom. The hypothesis of parameter constancy implied by the tobit model is rejected. The last two sets of results are for a hurdle model in which the intensity equation is fit by the two-step method.

19.3.5 SOME ISSUES IN SPECIFICATION

Two issues that commonly arise in microeconomic data, heteroscedasticity and nonnormality, have been analyzed at length in the tobit setting.[12]

19.3.5.a Heteroscedasticity

Maddala and Nelson (1975), Hurd (1979), Arabmazar and Schmidt (1982a,b), and Brown and Moffitt (1982) all have varying degrees of pessimism regarding how inconsistent the maximum likelihood estimator will be when **heteroscedasticity** occurs. Not surprisingly, the degree of censoring is the primary determinant. Unfortunately, all the analyses have been carried out in the setting of very specific models—for example, involving only a single dummy variable or one with groupwise heteroscedasticity—so the primary lesson is the very general conclusion that heteroscedasticity emerges as an obviously serious problem.

One can approach the heteroscedasticity problem directly. Petersen and Waldman (1981) present the computations needed to estimate a tobit model with heteroscedasticity of several types. Replacing σ with σ_i in the log-likelihood function and including σ_i^2 in the summations produces the needed generality. Specification of a particular model for σ_i provides the empirical model for estimation.

Example 19.7 Multiplicative Heteroscedasticity in the Tobit Model

Petersen and Waldman (1981) analyzed the volume of short interest in a cross section of common stocks. The regressors included a measure of the market component of heterogeneous expectations as measured by the firm's *BETA* coefficient; a company-specific measure of heterogeneous expectations, *NONMARKET*; the *NUMBER* of analysts making earnings forecasts for the company; the number of common shares to be issued for the acquisition of another firm, *MERGER*; and a dummy variable for the existence of *OPTION*s. They report the results listed in Table 19.4 for a model in which the variance is assumed to be of the form $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$. The values in parentheses are the ratio of the coefficient to the estimated asymptotic standard error.

The effect of heteroscedasticity on the estimates is extremely large. We do note, however, a common misconception in the literature. The change in the coefficients is often misleading. The marginal effects in the heteroscedasticity model will generally be very similar to those computed from the model which assumes homoscedasticity. (The calculation is pursued in the exercises.)

A test of the hypothesis that $\boldsymbol{\alpha} = \mathbf{0}$ (except for the constant term) can be based on the likelihood ratio statistic. For these results, the statistic is $-2[-547.3 - (-466.27)] = 162.06$. This statistic has a limiting chi-squared distribution with five degrees of freedom. The sample value exceeds the critical value in the table of 11.07, so the hypothesis can be rejected.

In the preceding example, we carried out a likelihood ratio test against the hypothesis of homoscedasticity. It would be desirable to be able to carry out the test without having to estimate the unrestricted model. A **Lagrange multiplier test** can be used for

[12]Two symposia that contain numerous results on these subjects are Blundell (1987) and Duncan (1986b). An application that explores these two issues in detail is Melenberg and van Soest (1996). Developing specification tests for the tobit model has been a popular enterprise. A sampling of the received literature includes Nelson (1981); Bera, Jarque, and Lee (1982); Chesher and Irish (1987); Chesher, Lancaster, and Irish (1985); Gourieroux et al. (1984, 1987); Newey (1986); Rivers andVuong (1988); Horowitz and Neumann (1989); and Pagan and Vella (1989). Newey (1985a,b) are useful references on the general subject of conditional moment testing. More general treatments of specification testing are Godfrey (1988) and Ruud (1984).

TABLE 19.4 Estimates of a Tobit Model (standard errors in parentheses)

	Homoscedastic	*Heteroscedastic*	
	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
Constant	−18.28 (5.10)	−4.11 (3.28)	−0.47 (0.60)
Beta	10.97 (3.61)	2.22 (2.00)	1.20 (1.81)
Nonmarket	0.65 (7.41)	0.12 (1.90)	0.08 (7.55)
Number	0.75 (5.74)	0.33 (4.50)	0.15 (4.58)
Merger	0.50 (5.90)	0.24 (3.00)	0.06 (4.17)
Option	2.56 (1.51)	2.96 (2.99)	0.83 (1.70)
ln L	−547.30	−466.27	
Sample size	200	200	

that purpose. Consider the heteroscedastic tobit model in which we specify that

$$\sigma_i^2 = \sigma^2[\exp(\mathbf{w}_i'\boldsymbol{\alpha})]^2. \tag{19-18}$$

This model is a fairly general specification that includes many familiar ones as special cases. The null hypothesis of homoscedasticity is $\boldsymbol{\alpha} = \mathbf{0}$. (We used this specification in the probit model in Section 17.3.7 and in the linear regression model in Section 9.7.1) Using the BHHH estimator of the Hessian as usual, we can produce a Lagrange multiplier statistic as follows: Let $z_i = 1$ if y_i is positive and 0 otherwise,

$$\begin{aligned}
a_i &= z_i\left(\frac{\varepsilon_i}{\sigma^2}\right) + (1-z_i)\left(\frac{(-1)\lambda_i}{\sigma}\right),\\
b_i &= z_i\left(\frac{(\varepsilon_i^2/\sigma^2 - 1)}{2\sigma^2}\right) + (1-z_i)\left(\frac{(\mathbf{x}_i'\boldsymbol{\beta})\lambda_i}{2\sigma^3}\right),\\
\lambda_i &= \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1-\boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}.
\end{aligned} \tag{19-19}$$

The data vector is $\mathbf{g}_i = [a_i\mathbf{x}_i', b_i, b_i\mathbf{w}_i']'$. The sums are taken over all observations, and all functions involving unknown parameters $(\varepsilon_i, \phi_i, \Phi_i, \mathbf{x}_i'\boldsymbol{\beta}, \sigma, \lambda_i)$ are evaluated at the restricted (homoscedastic) maximum likelihood estimates. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}[\mathbf{G}'\mathbf{G}]^{-1}\mathbf{G}'\mathbf{i} = nR^2 \tag{19-20}$$

in the regression of a column of ones on the $K + 1 + P$ derivatives of the log-likelihood function for the model with multiplicative heteroscedasticity, evaluated at the estimates from the restricted model. (If there were no limit observations, then it would reduce to the Breusch–Pagan statistic discussed in Section 9.5.2.) Given the maximum likelihood estimates of the tobit model coefficients, it is quite simple to compute. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\mathbf{w}_i$.

19.3.5.b Nonnormality

Nonnormality is an especially difficult problem in this setting. It has been shown that if the underlying disturbances are not normally distributed, then the estimator based

on (19-13) is inconsistent. Research is ongoing both on alternative estimators and on methods for testing for this type of misspecification.[13]

One approach to the estimation is to use an alternative distribution. Kalbfleisch and Prentice (2002) present a unifying treatment that includes several distributions such as the exponential, lognormal, and Weibull. (Their primary focus is on survival analysis in a medical statistics setting, which is an interesting convergence of the techniques in very different disciplines.) Of course, assuming some other specific distribution does not necessarily solve the problem and may make it worse. A preferable alternative would be to devise an estimator that is robust to changes in the distribution. Powell's (1981, 1984) least absolute deviations (LAD) estimator appears to offer some promise.[14] The main drawback to its use is its computational complexity. An extensive application of the LAD estimator is Melenberg and van Soest (1996). Although estimation in the nonnormal case is relatively difficult, testing for this failure of the model is worthwhile to assess the estimates obtained by the conventional methods. Among the tests that have been developed are Hausman tests, Lagrange multiplier tests [Bera and Jarque (1981, 1982), Bera, Jarque, and Lee (1982)], and **conditional moment tests** [Nelson (1981)].

19.3.6 PANEL DATA APPLICATIONS

Extension of the familiar panel data results to the tobit model parallel the probit model, with the attendant problems. The random effects or random parameters models discussed in Chapter 17 can be adapted to the censored regression model using simulation or quadrature. The same reservations with respect to the orthogonality of the effects and the regressors will apply here, as will the applicability of the Mundlak (1978) correction to accommodate it.

Most of the attention in the theoretical literature on panel data methods for the tobit model has been focused on fixed effects. The departure point would be the maximum likelihood estimator for the static fixed effects model,

$$y_{it}^{*} = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \varepsilon_{it} \sim \mathrm{N}[0, \sigma^2],$$
$$y_{it} = \mathrm{Max}(0, y_{it}).$$

However, there are no firm theoretical results on the behavior of the MLE in this model. Intuition might suggest, based on the findings for the binary probit model, that the MLE would be biased in the same fashion, away from zero. Perhaps surprisingly, the results in Greene (2004) persistently found that not to be the case in a variety of model specifications. Rather, the incidental parameters, such as it is, manifests in a downward bias in the estimator of σ, not an upward (or downward) bias in the MLE of $\boldsymbol{\beta}$. However, this is less surprising when the tobit estimator is juxtaposed with the MLE in the linear regression model with fixed effects. In that model, the MLE is the within-groups (LSDV) estimator which is unbiased and consistent. But, the ML estimator of the disturbance variance in the linear regression model is $\mathbf{e}_{\mathrm{LSDV}}'\mathbf{e}_{\mathrm{LSDV}}/(nT)$, which is biased downward

[13] See Duncan (1983, 1986b), Goldberger (1983), Pagan and Vella (1989), Lee (1996), and Fernandez (1986).

[14] See Duncan (1986a,b) for a symposium on the subject and Amemiya (1984). Additional references are Newey, Powell, and Walker (1990); Lee (1996); and Robinson (1988).

by a factor of $(T-1)/T$. [This is the result found in the original source on the incidental parameters problem, Neyman and Scott (1948).] So, what evidence there is suggests that unconditional estimation of the tobit model behaves essentially like that for the linear regression model. That does not settle the problem, however; if the evidence is correct, then it implies that although consistent estimation of $\boldsymbol{\beta}$ is possible, appropriate statistical inference is not. The bias in the estimation of σ shows up in any estimator of the asymptotic covariance of the MLE of $\boldsymbol{\beta}$.

Unfortunately, there is no conditional estimator of $\boldsymbol{\beta}$ for the tobit (or truncated regression) model. First differencing or taking group mean deviations does not preserve the model. Because the latent variable is censored before observation, these transformations are not meaningful. Some progress has been made on theoretical, **semiparametric estimators** for this model. See, for example, Honorè and Kyriazidou (2000) for a survey. Much of the theoretical development has also been directed at dynamic models where the benign result of the previous paragraph (such as it is) is lost once again. Arellano (2001) contains some general results. Hahn and Kuersteiner (2004) have characterized the bias of the MLE, and suggested methods of reducing the bias of the estimators in dynamic binary choice and censored regression models.

19.4 MODELS FOR DURATION

The leading application of the censoring models we examined in Section 19.3 is models for durations and events. We consider the time until some kind of transition as the duration, and the transition, itself, as the event. The length of a spell of unemployment (until rehire or exit from the market), the duration of a strike, the amount of time until a patient ends a health-related spell in connection with a disease or operation, and the length of time between origination and termination (via prepayment, default, or some other mechanism) of a mortgage are all examples of durations and transitions. The role that censoring plays in these scenarios is that in almost all cases in which we as analysts study duration data, some or even many of the spells we observe do not end in transitions. For example, in studying the lengths of unemployment spells, many of the individuals in the sample may still be unemployed at the time the study ends—the analyst observes (or believes) that the spell will end some time after the observation window closes. These data on spell lengths are, by construction, censored. Models of duration will generally account explicitly for censoring of the duration data.

This section is concerned with models of duration. In some aspects, the regression-like models we have studied, such as the discrete choice models, are the appropriate tools. As in the previous two chapters, however, the models are nonlinear, and the familiar regression methods are not appropriate. Most of this analysis focuses on maximum likelihood estimators. In modeling duration, although an underlying regression model is, in fact, at work, it is generally not the conditional mean function that is of interest. More likely, as we will explore next, the objects of estimation are certain probabilities of events, for example in the conditional probability of a transition in a given interval given that the spell has lasted up to the point of interest. These are known as "hazard models"—the probability is labeled the hazard function—and are a central focus of this type of analysis.

19.4.1 MODELS FOR DURATION DATA[15]

Intuition might suggest that the longer a strike persists, the more likely it is that it will end within, say, the next week. Or is it? It seems equally plausible to suggest that the longer a strike has lasted, the more difficult must be the problems that led to it in the first place, and hence the *less* likely it is that it will end in the next short time interval. A similar kind of reasoning could be applied to spells of unemployment or the interval between conceptions. In each of these cases, it is not only the duration of the event, per se, that is interesting, but also the likelihood that the event will end in "the next period" given that it has lasted as long as it has.

Analysis of the length of *time until failure* has interested engineers for decades. For example, the models discussed in this section were applied to the durability of electric and electronic components long before economists discovered their usefulness. Likewise, the analysis of *survival times*—for example, the length of survival after the onset of a disease or after an operation such as a heart transplant—has long been a staple of biomedical research. Social scientists have recently applied the same body of techniques to strike duration, length of unemployment spells, intervals between conception, time until business failure, length of time between arrests, length of time from purchase until a warranty claim is made, intervals between purchases, and so on.

This section will give a brief introduction to the econometric analysis of duration data. As usual, we will restrict our attention to a few straightforward, relatively uncomplicated techniques and applications, primarily to introduce terms and concepts. The reader can then wade into the literature to find the extensions and variations. We will concentrate primarily on what are known as **parametric models**. These apply familiar inference techniques and provide a convenient departure point. Alternative approaches are considered at the end of the discussion.

19.4.2 DURATION DATA

The variable of interest in the analysis of duration is the length of time that elapses from the beginning of some event either until its end or until the measurement is taken, which may precede termination. Observations will typically consist of a cross section of durations, $t_1, t_2, \ldots, t_n$. The process being observed may have begun at different points in calendar time for the different individuals in the sample. For example, the strike duration data examined in Example 19.8 are drawn from nine different years.

Censoring is a pervasive and usually unavoidable problem in the analysis of duration data. The common cause is that the measurement is made while the process is ongoing. An obvious example can be drawn from medical research. Consider analyzing the survival times of heart transplant patients. Although the beginning times may be known with precision, at the time of the measurement, observations on any individuals who are still alive are necessarily censored. Likewise, samples of spells of unemployment drawn from surveys will probably include some individuals who are still unemployed at the time the survey is taken. For these individuals, duration, or survival, is at least the

[15]There are a large number of highly technical articles on this topic, but relatively few accessible sources for the uninitiated. A particularly useful introductory survey is Kiefer (1988), upon which we have drawn heavily for this section. Other useful sources are Kalbfleisch and Prentice (2002), Heckman and Singer (1984a), Lancaster (1990), Florens, Fougere, and Mouchart (1996) and Cameron and Trivedi (2005, Chapters 17–19).

observed t_i, but not equal to it. Estimation must account for the censored nature of the data for the same reasons as considered in Section 19.3. The consequences of ignoring censoring in duration data are similar to those that arise in regression analysis.

In a conventional regression model that characterizes the conditional mean and variance of a distribution, the regressors can be taken as fixed characteristics at the point in time or for the individual for which the measurement is taken. When measuring duration, the observation is implicitly on a process that has been under way for an interval of time from zero to t. If the analysis is conditioned on a set of covariates (the counterparts to regressors) $\mathbf{x}_t$, then the duration is implicitly a function of the entire time path of the variable $\mathbf{x}(t), t = (0, t)$, which may have changed during the interval. For example, the observed duration of employment in a job may be a function of the individual's rank in the firm. But their rank may have changed several times between the time they were hired and when the observation was made. As such, observed rank at the end of the job tenure is not necessarily a complete description of the individual's rank *while they were employed*. Likewise, marital status, family size, and amount of education are all variables that can change during the duration of unemployment and that one would like to account for in the duration model. The treatment of **time-varying covariates** is a considerable complication.[16]

19.4.3 A REGRESSION-LIKE APPROACH: PARAMETRIC MODELS OF DURATION

We will use the term *spell* as a catchall for the different duration variables we might measure. Spell length is represented by the random variable T. A simple approach to duration analysis would be to apply regression analysis to the sample of observed spells. By this device, we could characterize the expected duration, perhaps conditioned on a set of covariates whose values were measured at the end of the period. We could also assume that conditioned on an $\mathbf{x}$ that has remained fixed from $T = 0$ to $T = t, t$ has a normal distribution, as we commonly do in regression. We could then characterize the probability distribution of observed duration times. But, normality turns out not to be particularly attractive in this setting for a number of reasons, not least of which is that duration is positive by construction, while a normally distributed variable can take negative values. (*log*normality turns out to be a palatable alternative, but it is only one among a long list of candidates.)

19.4.3.a Theoretical Background

Suppose that the random variable T has a continuous probability distribution $f(t)$, where t is a realization of T. The cumulative probability is

$$F(t) = \int_0^t f(s)\, ds = \text{Prob}(T \leq t).$$

We will usually be more interested in the probability that the spell is of length *at least* t, which is given by the **survival function**,

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t).$$

[16] See Petersen (1986) for one approach to this problem.

Consider the question raised in the introduction: Given that the spell has lasted until time t, what is the probability that it will end in the next short interval of time, say, Δt? It is

$$l(t, \Delta t) = \text{Prob}(t \leq T \leq t + \Delta t \mid T \geq t).$$

A useful function for characterizing this aspect of the distribution is the **hazard rate**,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t\ S(t)} = \frac{f(t)}{S(t)}.$$

Roughly, the hazard rate is the rate at which spells are completed after duration t, given that they last at least until t. As such, the hazard function gives an answer to our original question.

The hazard function, the density, the CDF, and the survival function are all related. The hazard function is

$$\lambda(t) = \frac{-d \ln S(t)}{dt},$$

so

$$f(t) = S(t)\lambda(t).$$

Another useful function is the **integrated hazard function**

$$\Lambda(t) = \int_0^t \lambda(s)\, ds,$$

for which

$$S(t) = e^{-\Lambda(t)},$$

so

$$\Lambda(t) = -\ln S(t).$$

The integrated hazard function is **generalized residual** in this setting. [See Chesher and Irish (1987) and Example 19.8.]

19.4.3.b Models of the Hazard Function

For present purposes, the hazard function is more interesting than the survival rate or the density. Based on the previous results, one might consider modeling the hazard function itself, rather than, say, modeling the survival function and then obtaining the density and the hazard. For example, the base case for many analyses is a hazard rate that does not vary over time. That is, $\lambda(t)$ is a constant λ. This is characteristic of a process that has no memory; the *conditional* probability of "failure" in a given short interval is the same regardless of when the observation is made. Thus,

$$\lambda(t) = \lambda.$$

From the earlier definition, we obtain the simple differential equation,

$$\frac{-d \ln S(t)}{dt} = \lambda.$$

The solution is

$$\ln S(t) = k - \lambda t,$$

or

$$S(t) = Ke^{-\lambda t},$$

where K is the constant of integration. The terminal condition that $S(0) = 1$ implies that $K = 1$, and the solution is

$$S(t) = e^{-\lambda t}.$$

This solution is the **exponential** distribution, which has been used to model the time until failure of electronic components. Estimation of λ is simple, because with an exponential distribution, $E[t] = 1/\lambda$. The maximum likelihood estimator of λ would be the reciprocal of the sample mean.

A natural extension might be to model the hazard rate as a linear function, $\lambda(t) = \alpha + \beta t$. Then $\Lambda(t) = \alpha t + \frac{1}{2}\beta t^2$ and $f(t) = \lambda(t)S(t) = \lambda(t)\exp[-\Lambda(t)]$. To avoid a negative hazard function, one might depart from $\lambda(t) = \exp[g(t, \boldsymbol{\theta})]$, where $\boldsymbol{\theta}$ is a vector of parameters to be estimated. With an observed sample of durations, estimation of α and β is, at least in principle, a straightforward problem in maximum likelihood. [Kennan (1985) used a similar approach.]

A distribution whose hazard function slopes upward is said to have **positive duration dependence**. For such distributions, the likelihood of failure at time t, conditional upon duration up to time t, is increasing in t. The opposite case is that of decreasing hazard or **negative duration dependence**. Our question in the introduction about whether the strike is more or less likely to end at time t given that it has lasted until time t can be framed in terms of positive or negative duration dependence. The assumed distribution has a considerable bearing on the answer. If one is unsure at the outset of the analysis whether the data can be characterized by positive or negative duration dependence, then it is counterproductive to assume a distribution that displays one characteristic or the other over the entire range of t. Thus, the exponential distribution and our suggested extension could be problematic. The literature contains a cornucopia of choices for duration models: normal, inverse normal [inverse Gaussian; see Lancaster (1990)], lognormal, F, gamma, Weibull (which is a popular choice), and many others.[17] To illustrate the differences, we will examine a few of the simpler ones. Table 19.5 lists the hazard functions and survival functions for four commonly used distributions. Each involves two parameters, a location parameter λ, and a scale parameter, p. [Note that in the benchmark case of the exponential distribution, λ is the hazard function. In all other cases, the hazard function is a function of λ, p, and, where there is duration dependence, t as well. Different authors, for example, Kiefer (1988), use different parameterizations of these models. We follow the convention of Kalbfleisch and Prentice (2002).]

All these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors, as can be seen in Figure 19.7. The hazard function for the exponential distribution is constant, that for the Weibull is monotonically increasing or decreasing depending on p, and the hazards for lognormal and loglogistic distributions first increase and then decrease. Which among these or the many alternatives is likely to be best in any application is uncertain.

[17] Three sources that contain numerous specifications are Kalbfleisch and Prentice (2002), Cox and Oakes (1985), and Lancaster (1990).

TABLE 19.5 Survival Distributions

Distribution	*Hazard Function, $\lambda(t)$*	*Survival Function, $S(t)$*
Exponential	λ,	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1}$,	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p\ln(\lambda t)]$	$S(t) = \Phi[-p\ln(\lambda t)]$
	[$\ln t$ is normally distributed with mean $-\ln\lambda$ and standard deviation $1/p$.]	
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p]$,	$S(t) = 1/[1 + (\lambda t)^p]$
	[$\ln t$ has a logistic distribution with mean $-\ln\lambda$ and variance $\pi^2/(3p^2)$.]	

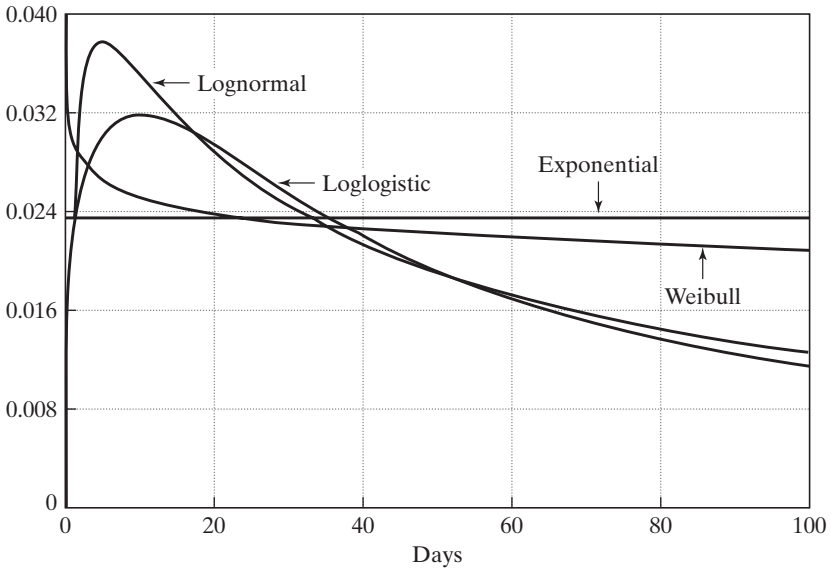


FIGURE 19.7 Parametric Hazard Functions.

19.4.3.c Maximum Likelihood Estimation

The parameters λ and p of these models can be estimated by maximum likelihood. For observed duration data, $t_1, t_2, \ldots, t_n$, the log-likelihood function can be formulated and maximized in the ways we have become familiar with in earlier chapters. Censored observations can be incorporated as in Section 19.3 for the tobit model. [See (19-13).] As such,

$$\ln L(\boldsymbol{\theta}) = \sum_{\substack{\text{uncensored}\\ \text{observations}}} \ln f(t \mid \boldsymbol{\theta}) + \sum_{\substack{\text{censored}\\ \text{observations}}} \ln S(t \mid \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\lambda, p)$. For some distributions, it is convenient to formulate the log-likelihood function in terms of $f(t) = \lambda(t)S(t)$ so that

$$\ln L = \sum_{\substack{\text{uncensored}\\ \text{observations}}} \ln \lambda(t \mid \boldsymbol{\theta}) + \sum_{\substack{\text{all}\\ \text{observations}}} \ln S(t \mid \boldsymbol{\theta}).$$

Inference about the parameters can be done in the usual way. Either the BHHH estimator or actual second derivatives can be used to estimate asymptotic standard errors for the estimates. The transformation $w = p(\ln t + \ln \lambda)$ for these distributions greatly facilitates maximum likelihood estimation. For example, for the Weibull model, by defining $w = p(\ln t + \ln \lambda)$, we obtain the very simple density $f(w) = \exp[w - \exp(w)]$ and survival function $S(w) = \exp(-\exp(w))$.[18] Therefore, by using $\ln t$ instead of t, we greatly simplify the log-likelihood function. Details for these and several other distributions may be found in Kalbfleisch and Prentice (2002, pp. 68–70). The Weibull distribution is examined in detail in the next section.

19.4.3.d Exogenous Variables

One limitation of the models given earlier is that external factors are not given a role in the survival distribution. The addition of "covariates" to duration models is fairly straightforward, although the interpretation of the coefficients in the model is less so. Consider, for example, the Weibull model. (The extension to other distributions will be similar.) Let

$$\lambda_i = e^{-\mathbf{x}_i'\boldsymbol{\beta}},$$

where $\mathbf{x}_i$ is a constant term and a set of variables that are assumed not to change from time $T = 0$ until the "failure time," $T = t_i$. Making λ_i a function of a set of regressors is equivalent to changing the units of measurement on the time axis. For this reason, these models are sometimes called **accelerated failure time models**. Note as well that in all the models listed (and generally), the regressors do not bear on the question of duration dependence, which is a function of p.

Let $\sigma = 1/p$ and let $\delta_i = 1$ if the spell is completed and $\delta_i = 0$ if it is censored. As before, let

$$w_i = p \ln(\lambda_i t_i) = \frac{(\ln t_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sigma},$$

and denote the density and survival functions $f(w_i)$ and $S(w_i)$. The observed random variable is

$$\ln t_i = \sigma w_i + \mathbf{x}_i'\boldsymbol{\beta}.$$

The Jacobian of the transformation from w_i to $\ln t_i$ is $\mathrm{d}w_i/\mathrm{d}\ln t_i = 1/\sigma$, so the density and survival functions for $\ln t_i$ are

$$f(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} f\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right), \quad \text{and} \quad S(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) = S\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right).$$

The log-likelihood for the observed data is

$$\ln L(\boldsymbol{\beta}, \sigma \mid \text{data}) = \sum_{i=1}^{n} [\delta_i \ln f(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma) + (1 - \delta_i) \ln S(\ln t_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma)].$$

[18] The transformation is $\exp(w) = (\lambda t)^p$ so $t = (1/\lambda)[\exp(w)]^{1/p}$. The Jacobian of the transformation is $\mathrm{d}t/\mathrm{d}w = [\exp(w)]^{1/p}/(\lambda p)$. The density in Table 19.5 is $\lambda p[\exp(w)]^{-(1/p)-1}[\exp(-\exp(w))]$. Multiplying by the Jacobian produces the result, $f(w) = \exp[w - \exp(w)]$. The survival function is the antiderivative, $[\exp(-\exp(w))]$.

For the **Weibull model**, for example (see footnote 18),

$$f(w_i) = \exp(w_i - e^{w_i}),$$

and

$$S(w_i) = \exp(-e^{w_i}).$$

Making the transformation to $\ln t_i$ and collecting terms reduces the log-likelihood to

$$\ln L(\boldsymbol{\beta}, \sigma \mid \text{data}) = \sum_i \left[\delta_i \left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} - \ln \sigma \right) - \exp\left(\frac{\ln t_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right].$$

(Many other distributions, including the others in Table 19.5, simplify in the same way. The exponential model is obtained by setting σ to one.) The derivatives can be equated to zero using the methods described in Section E.3. The individual terms can also be used to form the BHHH estimator of the asymptotic covariance matrix for the estimator.[19] The Hessian is also simple to derive, so Newton's method could be used instead.[20]

Note that the hazard function generally depends on t, p, and $\mathbf{x}$. The sign of an estimated coefficient suggests the direction of the effect of the variable on the hazard function when the hazard is monotonic. But in those cases, such as the loglogistic, in which the hazard is nonmonotonic, even this may be ambiguous. The magnitudes of the effects may also be difficult to interpret in terms of the hazard function. In a few cases, we do get a regression-like interpretation. In the Weibull and exponential models, $E[t \mid \mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\boldsymbol{\Gamma}[(1/p) + 1]$, whereas for the lognormal and loglogistic models, $E[\ln t \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$. In these cases, β_k is the derivative (or a multiple of the derivative) of this conditional mean. For some other distributions, the conditional median of t is easily obtained. Numerous cases are discussed by Kiefer (1988), Kalbfleisch and Prentice (2002), and Lancaster (1990).

19.4.3.e Heterogeneity

The problem of **heterogeneity** in duration models can be viewed essentially as the result of an incomplete specification. Individual specific covariates are intended to incorporate observation specific effects. But if the model specification is incomplete and if systematic individual differences in the distribution remain after the observed effects are accounted for, then inference based on the improperly specified model is likely to be problematic. We have already encountered several settings in which the possibility of heterogeneity mandated a change in the model specification; the fixed and random effects regression, logit, and probit models all incorporate observation-specific effects. Indeed, all the failures of the linear regression model discussed in the preceding chapters can be interpreted as a consequence of heterogeneity arising from an incomplete specification.

There are a number of ways of extending duration models to account for heterogeneity. The strictly nonparametric approach of the Kaplan–Meier estimator (see Section 19.4.4) is largely immune to the problem, but it is also rather limited in how

[19] Note that the log-likelihood function has the same form as that for the tobit model in Section 19.3.2. By just reinterpreting the nonlimit observations in a tobit setting, we can, therefore, use this framework to apply a wide range of distributions to the tobit model. [See Greene (1995a) and references given therein.]

[20] See Kalbfleisch and Prentice (2002) for numerous other examples.

much information can be culled from it. One direct approach is to model heterogeneity in the parametric model. Suppose that we posit a survival function conditioned on the individual specific effect v_i. We treat the survival function as $S(t_i|v_i)$. Then add to that a model for the unobserved heterogeneity $f(v_i)$. (Note that this is a counterpart to the incorporation of a disturbance in a regression model and follows the same procedures that we used in the Poisson model with random effects.) Then

$$S(t) = E_v[S(t \mid v)] = \int_v S(t \mid v)\, f(v)\, dv.$$

The gamma distribution is frequently used for this purpose.[21] Consider, for example, using this device to incorporate heterogeneity into the Weibull model we used earlier. As is typical, we assume that v has a gamma distribution with mean 1 and variance $\theta = 1/k$. Then

$$f(v) = \frac{k^k}{\Gamma(k)} e^{-kv} v^{k-1},$$

and

$$S(t \mid v) = e^{-(v\lambda t)^p}.$$

After a bit of manipulation, we obtain the unconditional distribution,

$$S(t) = \int_0^\infty S(t \mid v)\, f(v)\, dv = [1 + \theta(\lambda t)^p]^{-1/\theta}.$$

The limiting value, with $\theta = 0$, is the **Weibull survival model**, so $\theta = 0$ corresponds to $\text{Var}[v] = 0$, or no heterogeneity.[22] The hazard function for this model is

$$\lambda(t) = \lambda p(\lambda t)^{p-1}[S(t)]^\theta,$$

which shows the relationship to the Weibull model.

This approach is common in parametric modeling of heterogeneity. In an important paper on this subject, Heckman and Singer (1984b) argued that this approach tends to overparameterize the survival distribution and can lead to rather serious errors in inference. They gave some dramatic examples to make the point. They also expressed some concern that researchers tend to choose the distribution of heterogeneity more on the basis of mathematical convenience than on any sensible economic basis.

19.4.4 NONPARAMETRIC AND SEMIPARAMETRIC APPROACHES

The parametric models are attractive for their simplicity. But by imposing as much structure on the data as they do, the models may distort the estimated hazard rates. It may be that a more accurate representation can be obtained by imposing fewer restrictions.

[21] See, for example, Hausman, Hall, and Griliches (1984), who use it to incorporate heterogeneity in the Poisson regression model. The application is developed in Section 18.4.4.

[22] For the strike data analyzed in Figure 19.7, the maximum likelihood estimate of θ is 0.0004, which suggests that at least in the context of the Weibull model, latent heterogeneity does not appear to be a feature of these data.

The Kaplan–Meier (1958) **product limit estimator** is a strictly empirical, nonparametric approach to survival and hazard function estimation. Assume that the observations on duration are sorted in ascending order so that $t_1 \leq t_2$ and so on and, for now, that no observations are censored. Suppose as well that there are K distinct survival times in the data, denoted T_k; K will equal n unless there are ties. Let n_k denote the number of individuals whose observed duration is at least T_k. The set of individuals whose duration is at least T_k is called the **risk set** at this duration. (We borrow, once again, from biostatistics, where the risk set is those individuals still "at risk" at time T_k). Thus, n_k is the size of the risk set at time T_k. Let h_k denote the number of observed spells completed at time T_k. A strictly empirical estimate of the survivor function would be

$$\hat{S}(T_k) = \prod_{i=1}^{k} \frac{n_i - h_i}{n_i} = \frac{n_i - h_i}{n_1}.$$

The estimator of the hazard rate is

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}. \tag{19-21}$$

Corrections are necessary for observations that are censored. Lawless (1982), Kalbfleisch and Prentice (2002), Kiefer (1988), and Greene (1995a) give details. Susin (2001) points out a fundamental ambiguity in this calculation (one which he argues appears in the 1958 source). The estimator in (19-21) is not a "rate" as such, as the width of the time window is undefined, and could be very different at different points in the chain of calculations. Because many intervals, particularly those late in the observation period, might have zeros, the failure to acknowledge these intervals should impart an upward bias to the estimator. His proposed alternative computes the counterpart to (19-21) over a mesh of defined intervals as follows:

$$\hat{\lambda}\left(I_a^b\right) = \frac{\sum_{j=a}^{b} h_j}{\sum_{j=a}^{b} n_j b_j},$$

where the interval is from $t = a$ to $t = b$, h_j is the number of failures in each period in this interval, n_j is the number of individuals at risk in that period and b_j is the width of the period. Thus, an interval (a, b) is likely to include several "periods."

Cox's (1972) approach to the **proportional hazard** model is another popular, **semiparametric** method of analyzing the effect of covariates on the hazard rate. The model specifies that

$$\lambda(t_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})\lambda_0(t_i)$$

The function λ_0 is the "baseline" hazard, which is the individual heterogeneity. In principle, this hazard is a parameter for each observation that must be estimated. Cox's **partial likelihood** estimator provides a method of estimating $\boldsymbol{\beta}$ without requiring estimation of λ_0. The estimator is somewhat similar to Chamberlain's estimator for the logit model with panel data in that a conditioning operation is used to remove the heterogeneity. (See Section 17.4.4.) Suppose that the sample contains K distinct exit times, $T_1, \ldots, T_K$. For any time T_k, the risk set, denoted R_k, is all individuals whose exit time is at least T_k. The risk set is defined with respect to any moment in time T as the set of individuals who

have not yet exited just prior to that time. For every individual i in risk set R_k, $t_i \geq T_k$. The probability that an individual exits at time T_k *given that exactly one individual exits at this time* (which is the counterpart to the conditioning in the binary logit model in Chapter 17) is

$$\text{Prob}[t_i = T_k \mid \text{risk set}_k] = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\sum_{j \in R_k} e^{\mathbf{x}_j'\boldsymbol{\beta}}}.$$

Thus, the conditioning sweeps out the baseline hazard functions. For the simplest case in which exactly one individual exits at each distinct exit time and there are no censored observations, the partial log-likelihood is

$$\ln L = \sum_{k=1}^{K} \left[\mathbf{x}_k'\boldsymbol{\beta} - \ln \sum_{j \in R_k} e^{\mathbf{x}_j'\boldsymbol{\beta}} \right].$$

If m_k individuals exit at time T_k, then the contribution to the log-likelihood is the sum of the terms for each of these individuals.

The proportional hazard model is a common choice for modeling durations because it is a reasonable compromise between the Kaplan–Meier estimator and the possibly excessively structured parametric models. Hausman and Han (1990) and Meyer (1988), among others, have devised other, "semiparametric" specifications for hazard models.

Example 19.8 Survival Models for Strike Duration

The strike duration data given in Kennan (1985, pp. 14–16) have become a familiar standard for the demonstration of hazard models. Appendix Table F19.2 lists the durations, in days, of 62 strikes that commenced in June of the years 1968 to 1976. Each involved at least 1,000 workers and began at the expiration or reopening of a contract. Kennan reported the actual duration. In his survey, Kiefer (1985), using the same observations, censored the data at 80 days to demonstrate the effects of censoring. We have kept the data in their original form; the interested reader is referred to Kiefer for further analysis of the censoring problem.[23]

Parameter estimates for the four duration models are given in Table 19.6. The estimate of the median of the survival distribution is obtained by solving the equation $S(t) = 0.5$. For example, for the Weibull model,

$$S(M) = 0.5 = \exp[-(\lambda M)^P],$$

or

$$M = [(\ln 2)^{1/p}]/\lambda.$$

For the exponential model, $p = 1$. For the lognormal and loglogistic models, $M = 1/\lambda$. The delta method is then used to estimate the standard error of this function of the parameter estimates. (See Section 4.4.4.) All these distributions are skewed to the right. As such, $E[t]$ is greater than the median. For the exponential and Weibull models, $E[t] = [1/\lambda]\Gamma[(1/p) + 1]$; for the normal, $E[t] = (1/\lambda)[\exp(1/p^2)]^{1/2}$. The implied hazard functions are shown in Figure 19.7.

The variable x reported with the strike duration data is a measure of unanticipated aggregate industrial production net of seasonal and trend components. It is computed as the residual in a regression of the log of industrial production in manufacturing on time, time squared, and monthly dummy variables. With the industrial production variable included as

[23] Our statistical results are nearly the same as Kiefer's despite the censoring.

TABLE 19.6 Estimated Duration Models (estimated standard errors in parentheses)

	λ	p	*Median Duration*
Exponential	0.02344 (0.00298)	1.00000 (0.00000)	29.571 (3.522)
Weibull	0.02439 (0.00354)	0.92083 (0.11086)	27.543 (3.997)
Loglogistic	0.04153 (0.00707)	1.33148 (0.17201)	24.079 (4.102)
Lognormal	0.04514 (0.00806)	0.77206 (0.08865)	22.152 (3.954)

a covariate, the estimated Weibull model is

$$-\ln\lambda = 3.7772 - 9.3515\,x, \qquad p = 1.00288,$$
$$(0.1394)\ \ (2.973) \qquad\qquad (0.1217),$$
$$\text{median strike length} = 27.35(3.667)\ \text{days}, E[t] = 39.83\ \text{days}.$$

Note that the Weibull model is now almost identical to the exponential model ($p = 1$). Because the hazard conditioned on x is approximately equal to λ_i, it follows that the hazard function is increasing in "unexpected" industrial production. A 1 percent increase in x leads to a 9.35 percent increase in λ, which because $p \approx 1$ translates into a 9.35 percent decrease in the median strike length or about 2.6 days. (Note that $M = \ln 2/\lambda$.)

The proportional hazard model does not have a constant term. (The baseline hazard is an individual specific constant.) The estimate of β is -9.0726, with an estimated standard error of 3.225. This is very similar to the estimate obtained for the Weibull model.

19.5 INCIDENTAL TRUNCATION AND SAMPLE SELECTION

The topic of sample selection, or **incidental truncation**, has been the subject of an enormous recent literature, both theoretical and applied.[24] This analysis combines both of the previous topics.

Example 19.9 Incidental Truncation

In the high-income survey discussed in Example 19.2, respondents were also included in the survey if their net worth, not including their homes, was at least \$500,000. Suppose that the survey of incomes was based *only* on people whose net worth was at least \$500,000. This selection is a form of truncation, but not quite the same as in Section 19.2. This selection criterion does not necessarily exclude individuals whose incomes at the time might be quite low. Still, one would expect that, on average, individuals with a high net worth would have a high income as well. Thus, the average income in this subpopulation would in all likelihood also be misleading as an indication of the income of the typical American. The data in such a survey would be nonrandomly selected or incidentally truncated.

Econometric studies of nonrandom sampling have analyzed the deleterious effects of sample selection on the properties of conventional estimators such as least squares; have produced a variety of alternative estimation techniques; and, in the process, have

[24] A large proportion of the analysis in this framework has been in the area of labor economics. See, for example, Vella (1998), which is an extensive survey for practitioners. The results, however, have been applied in many other fields, including, for example, long series of stock market returns by financial economists ("survivorship bias") and medical treatment and response in long-term studies by clinical researchers ("attrition bias"). Some studies that comment on methodological issues are Heckman (1990), Manski (1989, 1990, 1992), and Newey, Powell, and Walker (1990).

yielded a rich crop of empirical models. In some cases, the analysis has led to a reinterpretation of earlier results.

19.5.1 INCIDENTAL TRUNCATION IN A BIVARIATE DISTRIBUTION

Suppose that y and z have a bivariate distribution with correlation ρ. We are interested in the distribution of y given that z exceeds a particular value. Intuition suggests that if y and z are positively correlated, then the truncation of z should push the distribution of y to the right. As before, we are interested in (1) the form of the incidentally truncated distribution and (2) the mean and variance of the incidentally truncated random variable. Because it has dominated the empirical literature, we will focus first on the bivariate normal distribution.

The truncated *joint* density of y and z is

$$f(y, z \mid z > a) = \frac{f(y, z)}{\text{Prob}(z > a)}.$$

To obtain the incidentally truncated marginal density for y, we would then integrate z out of this expression. The moments of the incidentally truncated normal distribution are given in Theorem 19.5.[25]

THEOREM 19.5 Moments of the Incidentally Truncated Bivariate Normal Distribution

If y and z have a bivariate normal distribution with means μ_y and μ_z, standard deviations σ_y and σ_z, and correlation ρ, then

$$E[y \mid z > a] = \mu_y + \rho\sigma_y\lambda(\alpha_z),$$
$$\text{Var}[y \mid z > a] = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

where

$$\alpha_z = (a - \mu_z)/\sigma_z, \lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)], \text{ and } \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z].$$

Note that the expressions involving z are analogous to the moments of the truncated distribution of x given in Theorem 19.2. If the truncation is $z < a$, then we make the replacement $\lambda(\alpha_z) = -\phi(\alpha_z)/\Phi(\alpha_z)$.

As expected, the truncated mean is pushed in the direction of the correlation if the truncation is from below and in the opposite direction if it is from above. In addition, the incidental truncation reduces the variance, because both $\delta(\alpha)$ and ρ^2 are between zero and one.

19.5.2 REGRESSION IN A MODEL OF SELECTION

To motivate a regression model that corresponds to the results in Theorem 19.5, we consider the following example.

[25] Much more general forms of the result that apply to multivariate distributions are given in Johnson and Kotz (1974). See also Maddala (1983, pp. 266–267).

Example 19.10 A Model of Labor Supply

A simple model of female labor supply that has been examined in many studies consists of two equations:[26]

1. *Wage equation*. The difference between a person's *market wage,* what she could command in the labor market, and her *reservation wage,* the wage rate necessary to make her choose to participate in the labor market, is a function of characteristics such as age and education as well as, for example, number of children and where a person lives.
2. *Hours equation*. The desired number of labor hours supplied depends on the wage, home characteristics such as whether there are small children present, marital status, and so on.

The problem of truncation surfaces when we consider that the second equation describes desired hours, but an actual figure is observed only if the individual is working. (In most such studies, only a *participation equation,* that is, whether hours are positive or zero, is observable.) We infer from this that the market wage exceeds the reservation wage. Thus, the hours variable in the second equation is incidentally truncated.

To put the preceding examples in a general framework, let the equation that determines the sample selection be

$$z_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$

and let the equation of primary interest be

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

The sample rule is that y_i is observed only when z_i^* is greater than zero. Suppose as well that ε_i and u_i have a bivariate normal distribution with zero means and correlation ρ. Then we may insert these in Theorem 19.5 to obtain the model *that applies to the observations in our sample:*

$$\begin{aligned} E[y_i \mid y_i \text{ is observed}] &= E[y_i \mid z_i^* > 0] \\ &= E[y_i \mid u_i > -\mathbf{w}_i'\boldsymbol{\gamma}] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i \mid u_i > -\mathbf{w}_i'\boldsymbol{\gamma}] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i(\alpha_u) \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u), \end{aligned}$$

where $\alpha_u = -\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u$ and $\lambda(\alpha_u) = \phi(\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u)/\Phi(\mathbf{w}_i'\boldsymbol{\gamma}/\sigma_u)$. So,

$$\begin{aligned} y_i \mid z_i^* > 0 &= E[y_i \mid z_i^* > 0] + v_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u) + v_i. \end{aligned}$$

Least squares regression using the observed data—for instance, OLS regression of hours on its determinants, using only data for women who are working—produces inconsistent estimates of $\boldsymbol{\beta}$. Once again, we can view the problem as an omitted variable. Least squares regression of y on $\mathbf{x}$ *and* λ would be a consistent estimator, but if λ is omitted, then the **specification error** of an omitted variable is committed. Finally, note that the second part of Theorem 19.5 implies that even if λ_i were observed, then least squares would be inefficient. The disturbance v_i is heteroscedastic.

[26] See, for example, Heckman (1976). This strand of literature begins with an exchange by Gronau (1974) and Lewis (1974).

The marginal effect of the regressors on y_i *in the observed sample* consists of two components. There is the direct effect on the mean of y_i, which is $\boldsymbol{\beta}$. In addition, for a particular independent variable, if it appears in the probability that z_i^* is positive, then it will influence y_i through its presence in λ_i. The full effect of changes in a regressor that appears in both $\mathbf{x}_i$ and $\mathbf{w}_i$ on y is

$$\frac{\partial E[y_i \mid z_i^* > 0]}{\partial x_{ik}} = \beta_k - \gamma_k \left(\frac{\rho\sigma_\varepsilon}{\sigma_u}\right) \delta_i(\alpha_u),$$

where[27]

$$\delta_i = \lambda_i^2 - \alpha_i \lambda_i.$$

Suppose that ρ is positive and $E[y_i]$ is greater when z_i^* is positive than when it is negative. Because $0 < \delta_i < 1$, the additional term serves to reduce the marginal effect. The change in the probability affects the mean of y_i in that the mean in the group $z_i^* > 0$ is higher. The second term in the derivative compensates for this effect, leaving only the marginal effect of a change *given that* $z_i^* > 0$ *to begin with*. Consider Example 19.12, and suppose that education affects both the probability of migration and the income in either state. If we suppose that the income of migrants is higher than that of otherwise identical people who do not migrate, then the marginal effect of education has two parts, one due to its influence in increasing the probability of the individual's entering a higher-income group and one due to its influence on income within the group. As such, the coefficient on education in the regression overstates the marginal effect of the education of migrants and understates it for nonmigrants. The sizes of the various parts depend on the setting. It is quite possible that the magnitude, sign, and statistical significance of the effect might all be different from those of the estimate of $\boldsymbol{\beta}$, a point that appears frequently to be overlooked in empirical studies.

In most cases, the selection variable z^* is not observed. Rather, we observe only its sign. To consider our two examples, we typically observe only whether a woman is working or not working or whether an individual migrated or not. We can infer the sign of z^*, but not its magnitude, from such information. Because there is no information on the scale of z^*, the disturbance variance in the selection equation cannot be estimated. (We encountered this problem in Chapter 17 in connection with the probit model.) Thus, we reformulate the model as follows:

$$\begin{aligned}
&\text{selection mechanism: } z_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,\ z_i = 1 \quad \text{if } z_i^* > 0 \text{ and 0 otherwise;}\\
&\qquad\qquad \text{Prob}(z_i = 1 \mid \mathbf{w}_i) = \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\gamma});\ \text{and}\\
&\qquad\qquad \text{Prob}(z_i = 0 \mid \mathbf{w}_i) = 1 - \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\gamma}). \qquad \textbf{(19-22)}\\
&\text{regression model:} \quad y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ observed only if } z_i = 1,\\
&\qquad\qquad (u_i, \varepsilon_i) \sim \text{bivariate normal } [0, 0, 1, \sigma_\varepsilon, \rho].
\end{aligned}$$

Suppose that, as in many of these studies, z_i and $\mathbf{w}_i$ are observed for a random sample of individuals but y_i is observed only when $z_i = 1$. This model is precisely the one we

[27] We have reversed the sign of α_u in (Theorem 19.5) because $a = 0$, and $\alpha = \mathbf{w}'\boldsymbol{\gamma}/\sigma_M$ is somewhat more convenient. Also, as such, $\partial\lambda/\partial\alpha = -\delta$.

examined earlier, with

$$E[y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda(\mathbf{w}_i'\boldsymbol{\gamma}).$$

19.5.3 TWO-STEP AND MAXIMUM LIKELIHOOD ESTIMATION

The parameters of the sample selection model can be estimated by maximum likelihood.[28] However, Heckman's (1979) **two-step estimation** procedure is usually used instead. Heckman's method is as follows:[29]

1. Estimate the probit equation by maximum likelihood to obtain estimates of γ. For each observation in the selected sample, compute $\hat{\lambda}_i = \phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})/\boldsymbol{\Phi}(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})$ and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \mathbf{w}_i'\hat{\boldsymbol{\gamma}})$.
2. Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_\varepsilon$ by least squares regression of y on $\mathbf{x}$ and $\hat{\lambda}$.

It is possible also to construct consistent estimators of the individual parameters ρ and σ_ε. At each observation, the true conditional variance of the disturbance would be

$$\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

The average conditional variance for the sample would converge to

$$\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 = \sigma_\varepsilon^2(1 - \rho^2\bar{\delta}),$$

which is what is estimated by the least squares residual variance $\boldsymbol{e}'\boldsymbol{e}/n$. For the square of the coefficient on λ, we have

$$\operatorname{plim} b_\lambda^2 = \rho^2\sigma_\varepsilon^2,$$

whereas based on the probit results we have

$$\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_i = \bar{\delta}.$$

We can then obtain a consistent estimator of σ_ε^2 using

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} + \bar{\hat{\delta}}b_\lambda^2.$$

Finally, an estimator of ρ^2 is

$$\hat{\rho}^2 = \frac{b_\lambda^2}{\hat{\sigma}_\varepsilon^2}, \tag{19-23}$$

which provides a complete set of estimators of the model's parameters.[30]

To test hypotheses, an estimate of the asymptotic covariance matrix of $[\mathbf{b}', b_\lambda]$ is needed. We have two problems to contend with. First, we can see in Theorem 19.5 that

[28] See Greene (1995a).

[29] Perhaps in a mimicry of the "tobit" estimator described earlier, this procedure has come to be known as the "Heckit" estimator.

[30] Note that $\hat{\rho}^2$ is not a sample correlation and, as such, is not limited to $[0, 1]$. See Greene (1981) for discussion.

the disturbance term in

$$(y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i + v_i \qquad \textbf{(19-24)}$$

is heteroscedastic;

$$\text{Var}[v_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \sigma_\varepsilon^2(1 - \rho^2\delta_i).$$

Second, there are unknown parameters in λ_i. Suppose that we assume for the moment that λ_i and δ_i are known (i.e., we do not have to estimate $\boldsymbol{\gamma}$). For convenience, let $\mathbf{x}_i^* = [\mathbf{x}_i, \lambda_i]$, and let $\mathbf{b}^*$ be the least squares coefficient vector in the regression of y on $\mathbf{x}^*$ in the selected data. Then, using the appropriate form of the variance of ordinary least squares in a heteroscedastic model from Chapter 9, we would have to estimate

$$\begin{aligned}\text{Var}[\mathbf{b}^*] &= \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}\left[\sum_{i=1}^{n}(1-\rho^2\delta_i)\mathbf{x}_i^*\mathbf{x}_i^{*\prime}\right][\mathbf{X}_*'\mathbf{X}_*]^{-1}\\ &= \sigma_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I}-\rho^2\boldsymbol{\Delta})\mathbf{X}_*][\mathbf{X}_*'\mathbf{X}_*]^{-1},\end{aligned}$$

where $\mathbf{I} - \rho^2\boldsymbol{\Delta}$ is a diagonal matrix with $(1 - \rho^2\delta_i)$ on the diagonal. Without any other complications, this result could be computed fairly easily using $\mathbf{X}$, the sample estimates of σ_ε^2 and ρ^2, and the assumed known values of λ_i and δ_i.

The parameters in $\boldsymbol{\gamma}$ do have to be estimated using the probit equation. Rewrite (19-24) as

$$(y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda\hat{\lambda}_i + v_i - \beta_\lambda(\hat{\lambda}_i - \lambda_i).$$

In this form, we see that in the preceding expression we have ignored both an additional source of variation in the compound disturbance and correlation across observations; the same estimate of $\boldsymbol{\gamma}$ is used to compute $\hat{\lambda}_i$ for every observation. Heckman has shown that the earlier covariance matrix can be appropriately corrected by adding a term inside the brackets,

$$\mathbf{Q} = \hat{\rho}^2(\mathbf{X}_*'\hat{\boldsymbol{\Delta}}\mathbf{W})\text{Est. Asy. Var}[\hat{\boldsymbol{\gamma}}](\mathbf{W}'\hat{\boldsymbol{\Delta}}\mathbf{X}_*) = \hat{\rho}^2\hat{\mathbf{F}}\hat{\mathbf{V}}\hat{\mathbf{F}}',$$

where $\hat{\mathbf{V}} = \text{Est. Asy. Var}[\hat{\boldsymbol{\gamma}}]$, the estimator of the asymptotic covariance of the probit coefficients. Any of the estimators in (17-22) to (17-24) may be used to compute $\hat{\mathbf{V}}$. The complete expression is[31]

$$\text{Est. Asy. Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\boldsymbol{\Delta}})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}.$$

The sample selection model can also be estimated by maximum likelihood. The full log-likelihood function for the data is built up from

$$\text{Prob(selection)} \times \text{density} \mid \text{selection for observations with } z_i = 1,$$

and

$$\text{Prob(nonselection) for observations with } z_i = 0.$$

[31] This matrix formulation is derived in Greene (1981). Note that the Murphy and Topel (1985) results for two-step estimators given in Theorem 14.8 would apply here as well. Asymptotically, this method would give the same answer. The Heckman formulation has become standard in the literature.

Combining the parts produces the full log-likelihood function,

$$\ln L = \sum_{z=1} \ln \left[\frac{\exp\left(-(1/2)\varepsilon_i^2/\sigma_\varepsilon^2\right)}{\sigma_\varepsilon\sqrt{2\pi}} \Phi\left(\frac{\rho\varepsilon_i/\sigma_\varepsilon + \mathbf{w}_i'\boldsymbol{\gamma}}{\sqrt{1-\rho^2}} \right) \right] + \sum_{z=0} [1 - \ln \Phi(\mathbf{w}_i'\boldsymbol{\gamma})],$$

where $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. Note, the FIML estimator with its assumption of bivariate normality is not less robust than the two-step estimator. because the latter also requires bivariate normality to form the conditional mean for the regression.

Two virtues of the FIML estimator will be the greater efficiency brought by using the likelihood function rather than the method of moments and, second, the estimation of ρ subject to the constraint $-1 < \rho < 1$. (This is typically done by reparameterizing the model in terms of the monotonic inverse hyperbolic tangent, $\tau = (1/2)\ln[(1+\rho)/(1-\rho)] = \text{atanh}(\rho)$. The transformed parameter, τ, is unrestricted. The inverse transformation is $\rho = [\exp(2\tau) - 1]/[\exp(2\tau) + 1]$ which is bounded between zero and one.) One possible drawback (it might be argued) could be the complexity of the likelihood function that would make estimation more difficult than the two-step estimator. However, the MLE for the selection model appears as a built-in procedure in modern software such as *Stata* and *NLOGIT*, and it is straightforward to implement in *Gauss* and *MatLab*, so this might be a moot point. Surprisingly, the MLE is by far less common than the two-step estimator in the received applications. The estimation of ρ is the difficult part of the estimaton process (this is often the case). It is quite common for the method of moments estimator and the FIML estimator to be very different—our application in Example 19.11 is a case. Perhaps surprisingly so, the moment-based estimator of ρ in (19-23) is not bounded by zero and one. [See Greene (1981).] This would seem to recommend the MLE.

The fully parametric bivariate normality assumption of the model has been viewed as a potential drawback. However, relatively little progress has been made on devising informative semi- and nonparametric estimators—see, for one example, Gallant and Nychka (1987). The obstacle here is that, ultimately, the model hangs on a parameterization of the correlation of the unobservables in the two equations. So, method of moment estimators or kernel-based estimators must still incorporate this feature of a bivariate distribution. Some results have been obtained using the method of copula functions. [See Smith (2003, 2005) and Trivedi and Zimmer (2007).]

Example 19.11 Female Labor Supply

Examples 17.1 and 17.8 proposed a labor force participation model for a sample of 753 married women in a sample analyzed by Mroz (1987). The data set contains wage and hours information for the 428 women who participated in the formal market (LFP=1). Following Mroz, we suppose that for these 428 individuals, the offered wage exceeded the reservation wage and, moreover, the unobserved effects in the two wage equations are correlated. As such, a wage equation based on the market data should account for the sample selection problem. We specify a simple wage model:

$$\textit{Wage} = \beta_1 + \beta_2\, \textit{Exper} + \beta_3\, \textit{Exper}^2 + \beta_4\, \textit{Education} + \beta_5\, \textit{City} + \varepsilon$$

where Exper is labor market experience and City is a dummy variable indicating that the individual lived in a large urban area. Maximum likelihood, Heckman two-step, and ordinary least squares estimates of the wage equation are shown in Table 19.7. The maximum likelihood estimates are FIML estimates—the labor force participation equation is reestimated at the same time. Only the parameters of the wage equation are shown next. Note as well that the two-step estimator estimates the single coefficient on λ_i and the structural parameters σ

TABLE 19.7 Estimated Selection Corrected Wage Equation

	Two-Step		*Maximum Likelihood*		*Least Squares*	
	Estimate	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
β_1	−0.971	(2.06)	−1.963	(1.684)	−2.56	(0.929)
β_2	0.021	(0.0625)	0.0279	(0.0756)	0.0325	(0.0616)
β_3	0.000137	(0.00188)	−0.0001	(0.00234)	−0.000260	(0.00184)
β_4	0.417	(0.100)	0.457	(0.0964)	0.481	(0.0669)
β_5	0.444	(0.316)	0.447	(0.427)	(0.449)	0.318
$(\rho\sigma)$	−1.098	(1.266)				
ρ	−0.343		−0.132	(0.224)	0.000	
σ	3.200		3.108	(0.0837)	3.111	

and ρ are deduced by the method of moments. The maximum likelihood estimator computes estimates of these parameters directly. [Details on maximum likelihood estimation may be found in Maddala (1983).]

The differences between the two-step and maximum likelihood estimates in Table 19.7 are surprisingly large. The difference is even more striking in the marginal effects. The effect for education is estimated as 0.417 + 0.0641 for the two-step estimators and 0.480 in total for the maximum likelihood estimates. For the kids variable, the marginal effect is −.293 for the two-step estimates and only −.11003 for the MLEs. Surprisingly, the direct test for a selection effect in the maximum likelihood estimates, a nonzero ρ, fails to reject the hypothesis that ρ equals zero.

In some settings, the selection process is a nonrandom sorting of individuals into two or more groups. The mover-stayer model in the next example is a familiar case.

Example 19.12 A Mover-Stayer Model for Migration

The model of migration analyzed by Nakosteen and Zimmer (1980) fits into the framework described in this section. The equations of the model are

$$\text{net benefit of moving:} \quad M_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$
$$\text{income if moves:} \quad I_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \varepsilon_{i1},$$
$$\text{income if stays:} \quad I_{i0} = \mathbf{x}_{i0}'\boldsymbol{\beta}_0 + \varepsilon_{i0}.$$

One component of the net benefit is the market wage individuals could achieve if they move, compared with what they could obtain if they stay. Therefore, among the determinants of the net benefit are factors that also affect the income received in either place. An analysis of income in a sample of migrants must account for the incidental truncation of the mover's income on a positive net benefit. Likewise, the income of the stayer is incidentally truncated on a nonpositive net benefit. The model implies an income after moving for all observations, but we observe it only for those who actually do move. Nakosteen and Zimmer (1980) applied the selectivity model to a sample of 9,223 individuals with data for two years (1971 and 1973) sampled from the Social Security Administration's Continuous Work History Sample. Over the period, 1,078 individuals migrated and the remaining 8,145 did not. The independent variables in the migration equation were as follows:

SE = self-employment dummy variable; 1 if yes
ΔEMP = rate of growth of state employment
ΔPCI = growth of state per capita income
$\mathbf{x}$ = age, race (nonwhite= 1), sex (female= 1)
ΔSIC = 1 if individual changes industry

TABLE 19.8 Estimated Earnings Equations

	Migration	*Migrant Earnings*	*Nonmigrant Earnings*
Constant	−1.509	9.041	8.593
SE	−0.708 (−5.72)	−4.104 (−9.54)	−4.161 (−57.71)
ΔEMP	−1.488 (−2.60)	—	—
ΔPCI	1.455 (3.14)	—	—
Age	−0.008 (−5.29)	—	—
Race	−0.065 (−1.17)	—	—
Sex	−0.082 (−2.14)	—	—
ΔSIC	0.948 (24.15)	−0.790 (−2.24)	−0.927 (−9.35)
λ	—	0.212 (0.50)	0.863 (2.84)

The earnings equations included ΔSIC and *SE*. The authors reported the results given in Table 19.8. The figures in parentheses are asymptotic *t* ratios.

19.5.4 SAMPLE SELECTION IN NONLINEAR MODELS

The preceding analysis has focused on an extension of the linear regression (or the estimation of simple averages of the data). The method of analysis changes in nonlinear models. To begin, it is not necessarily obvious what the impact of the sample selection is on the response variable, or how it can be accommodated in a model. Consider the model analyzed by Boyes, Hoffman, and Lowe (1989):

$$y_{i1} = 1 \text{ if individual } i \text{ defaults on a loan, 0 otherwise,}$$
$$y_{i2} = 1 \text{ if the individual is granted a loan, 0 otherwise.}$$

Wynand and van Praag (1981) also used this framework to analyze consumer insurance purchases in the first application of the selection methodology in a nonlinear model. Greene (1992) applied the same model to $y_1 =$ default on credit card loans, in which y_{i2} denotes whether an application for the card was accepted or not. [Mohanty (2002) also used this model to analyze teen employment in California.] For a given individual, y_1 is not observed unless $y_{i2} = 1$. Following the lead of the linear regression case in Section 19.5.3, a natural approach might seem to be to fit the second (selection) equation using a univariate probit model, compute the inverse Mills ratio, λ_i, and add it to the first equation as an additional "control" variable to accommodate the selection effect. [This is the approach used by Wynand and van Praag (1981) and Greene (1994).] The problems with this control function approach are, first, it is unclear what in the model is being "controlled" and, second, assuming the first model is correct, the appropriate model conditioned on the sample selection is unlikely to contain an inverse Mills ratio anywhere in it. [See Terza (2010) for discussion.] That result is specific to the linear model, where it arises as $E[\varepsilon_i \mid \text{selection}]$. What would seem to be the apparent counterpart for this probit model,

$$\text{Prob}(y_{i1} = 1 \mid \text{ selection on } y_{i2} = 1) = \Phi(\mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \theta\lambda_i),$$

is not, in fact, the appropriate conditional mean, or probability. For this particular application, the appropriate conditional probability (extending the bivariate probit model

of Section 17.5) would be

$$\text{Prob}[y_{i1} = 1 \mid y_{i2} = 1] = \frac{\Phi_2(\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\boldsymbol{\beta}_2, \rho)}{\Phi(\mathbf{x}_{i2}'\boldsymbol{\beta}_2)}.$$

We would use this result to build up the likelihood function for the three observed outcomes, as follows: The three types of observations in the sample, with their unconditional probabilities, are

$$\begin{aligned}
y_{i2} = 0\text{: Prob}(y_{i2} = 0 \qquad\quad &\mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = 1 - \Phi(\mathbf{x}_{i2}'\boldsymbol{\beta}_2), \\
y_{i1} = 0,\ y_{i2} = 1\text{: Prob}(y_{i1} = 0,\ y_{i2} = 1 &\mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = \boldsymbol{\Phi}_2(-\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\boldsymbol{\beta}_2, -\rho), \\
y_{i1} = 1,\ y_{i2} = 1\text{: Prob}(y_{i1} = 1,\ y_{i2} = 1 &\mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = \boldsymbol{\Phi}_2(\mathbf{x}_{i1}'\boldsymbol{\beta}_1, \mathbf{x}_{i2}'\boldsymbol{\beta}_2, \rho).
\end{aligned} \tag{19-25}$$

The log-likelihood function is based on these probabilities.[32] An application appears in Section 17.5.6.

Example 19.13 Doctor Visits and Insurance

Continuing our analysis of the utilization of the German health care system, we observe that the data set contains an indicator of whether the individual subscribes to the "Public" health insurance or not. Roughly 87 percent of the observations in the sample do. We might ask whether the selection on public insurance reveals any substantive difference in visits to the physician. We estimated a logit specification for this model in Example 17.4. Using (19-25) as the framework, we define $\mathbf{y}_{i2}$ to be presence of insurance and $\mathbf{y}_{i1}$ to be the binary variable defined to equal 1 if the individual makes at least one visit to the doctor in the survey year.

The estimation results are given in Table 19.9. Based on these results, there does appear to be a very strong relationship. The coefficients do change somewhat in the conditional model. A Wald test for the presence of the selection effect against the null hypothesis that ρ equals zero produces a test statistic of $(-7.188)^2 = 51.667$, which is larger than the critical value of 3.84. Thus, the hypothesis is rejected. A likelihood ratio statistic is computed as the difference between the log-likelihood for the full model and the sum of the two separate log-likelihoods for the independent probit models when ρ equals zero. The result is

$$\lambda_{\text{LR}} = 2[-23969.58 - (-15536.39 + (-8471.508)) = 77.796$$

The hypothesis is rejected once again. Partial effects were computed using the results in Section 17.5.3.

The large correlation coefficient can be misleading. The estimated -0.9299 does not state that the presence of insurance makes it much less likely to go to the doctor. This is the correlation among the unobserved factors in each equation. The factors that make it more likely to purchase insurance make it less likely to use a physician. To obtain a simple correlation between the two variables, we might use the tetrachoric correlation defined in Example 17.18. This would be computed by fitting a bivariate probit model for the two binary variables without any other variables. The estimated value is 0.120.

More general cases are typically much less straightforward. Greene (2005, 2006, 2010) and Terza (1998, 2010) present sample selection models for nonlinear specifications based on the underlying logic of the Heckman model in Section 19.5.3, that the influence of the incidental truncation acts on the unobservable variables in the model. (That is the source of the "selection bias" in conventional estimators.) The modeling extension introduces the unobservables into the model in a natural fashion that parallels the regression model. Terza (2010) presents a survey of the general results.

[32] Extensions of the bivariate probit model to other types of censoring are discussed in Poirier (1980) and Abowd and Farber (1982).

TABLE 19.9 Estimated Probit Equations for Doctor Visits

	Independent: No Selection			*Sample Selection Model*		
Variable	*Estimate*	*Standard Error*	*Partial Effect*	*Estimate*	*Standard Error*	*Partial Effect*
Constant	0.05588	0.06564		−9.4366	0.06760	
Age	0.01331	0.0008399	0.004971	0.01284	0.0008131	0.005042
Income	−0.1034	0.05089	−0.03860	−0.1030	0.04582	−0.04060
Kids	−0.1349	0.01947	−0.05059	−0.1264	0.01790	−0.04979
Education	−0.01920	0.004254	−0.007170	0.03660	0.004744	0.002703
Married	0.03586	0.02172	0.01343	0.03564	0.02016	0.01404
ln L	−15536.39					
Constant	3.3585	0.06959		3.2699	0.06916	
Age	0.0001868	0.0009744		−0.0002679	0.001036	
Education	−0.1854	0.003941		−0.1807	0.003936	
Female	0.1150	0.02186	0.0000[a]	0.2230	0.02101	0.01446[a]
ln L	−8471.508					
ρ	0.0000	0.0000		−0.9299	0.1294	
ln L	−24007.90			−23969.58		

[a]Indirect effect from second equation.

The generic model will take the form

1. Probit selection equation:

$$\begin{aligned} z_i^* &= \mathbf{w}_i'\boldsymbol{\alpha} + u_i \text{ in which } u_i \sim \mathrm{N}[0,1], \\ z_i &= 1 \text{ if } z_i^* > 0, 0 \text{ otherwise.} \end{aligned} \tag{19-26}$$

2. Nonlinear index function model with unobserved heterogeneity and sample selection:

$$\begin{aligned} &\mu_i \mid \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i, \varepsilon_i \sim N[0,1], \\ &y_i \mid \mathbf{x}_i, \varepsilon_i \sim \text{density } g(y_i \mid \mathbf{x}_i, \varepsilon_i) = f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i), \\ &y_i, \mathbf{x}_i \text{ are observed only when } z_i = 1, \\ &[u_i, \varepsilon_i] \sim \mathrm{N}[(0,1),(1,\rho,1)]. \end{aligned} \tag{19-27}$$

For example, in a Poisson regression model, the conditional mean function becomes $E(y_i \mid \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i) = \exp(\mu_i)$. (We used this specification of the model in Chapter 18 to introduce random effects in the Poisson regression model for panel data.)

The log-likelihood function for the full model is the joint density for the observed data. When z_i equals one, $(y_i, \mathbf{x}_i, z_i, \mathbf{w}_i)$ are all observed. To obtain the joint density $p(y_i, z_i = 1 \mid \mathbf{x}_i, \mathbf{w}_i)$, we proceed as follows:

$$p(y_i, z_i = 1 \mid \mathbf{x}_i, \mathbf{w}_i) = \int_{-\infty}^{\infty} p(y_i, z_i = 1 \mid \mathbf{x}_i, \mathbf{w}_i, \varepsilon_i)\, f(\varepsilon_i) d\varepsilon_i.$$

Conditioned on ε_i, z_i and y_i are independent. Therefore, the joint density is the product,

$$p(y_i, z_i = 1 \mid \mathbf{x}_i, \mathbf{w}_i, \varepsilon_i) = f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)\mathrm{Prob}(z_i = 1 \mid \mathbf{w}_i, \varepsilon_i).$$

The first part, $f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)$ is the conditional index function model in (19-27). By joint normality, $f(u_i \mid \varepsilon_i) = N[\rho\varepsilon_i, (1-\rho^2)]$, so $u_i \mid \varepsilon_i = \rho\varepsilon_i + (u_i - \rho\varepsilon_i) = \rho\varepsilon_i + v_i$ where $E[v_i] = 0$ and $\text{Var}[v_i] = (1-\rho^2)$. Therefore,

$$\text{Prob}(z_i = 1 \mid \mathbf{w}_i, \varepsilon_i) = \boldsymbol{\Phi}\left(\frac{\mathbf{w}_i'\boldsymbol{\alpha} + \rho\varepsilon_i}{\sqrt{1-\rho^2}}\right).$$

Combining terms and using the earlier approach, the unconditional joint density is

$$p(y_i, z_i = 1 \mid \mathbf{x}_i, \mathbf{w}_i) = \int_{-\infty}^{\infty} f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)\Phi\left(\frac{\mathbf{w}_i'\boldsymbol{\alpha} + \rho\varepsilon_i}{\sqrt{1-\rho^2}}\right)\frac{\exp(-\varepsilon_i^2/2)}{\sqrt{2\pi}}d\varepsilon_i. \quad \textbf{(19-28)}$$

The other part of the likelihood function for the observations with $z_i = 0$ will be

$$\begin{aligned}\text{Prob}(z_i = 0 \mid \mathbf{w}_i) &= \int_{-\infty}^{\infty} \text{Prob}(z_i = 0 \mid \mathbf{w}_i, \varepsilon_i)\, f(\varepsilon_i)d\varepsilon_i. \\ &= \int_{-\infty}^{\infty}\left[1 - \boldsymbol{\Phi}\left(\frac{\mathbf{w}_i'\boldsymbol{\alpha} + \rho\varepsilon_i}{\sqrt{1-\rho^2}}\right)\right] f(\varepsilon_i)d\varepsilon_i \qquad \textbf{(19-29)} \\ &= \int_{-\infty}^{\infty}\Phi\left(\frac{-(\mathbf{w}_i'\boldsymbol{\alpha} + \rho\varepsilon_i)}{\sqrt{1-\rho^2}}\right)\frac{\exp(-\varepsilon_i^2/2)}{\sqrt{2\pi}}d\varepsilon_i.\end{aligned}$$

For convenience, we can use the invariance principle to reparameterize the likelihood function in terms of $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sqrt{1-\rho^2}$ and $\tau = \rho/\sqrt{1-\rho^2}$. Combining all the preceding terms, the log-likelihood function to be maximized is

$$\ln L = \sum_{i=1}^{n}\ln\int_{-\infty}^{\infty}[(1-z_i) + z_i f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)]\boldsymbol{\Phi}[(2z_i - 1)(\mathbf{w}_i'\boldsymbol{\gamma} + \boldsymbol{\tau}\varepsilon_i)]\phi(\varepsilon_i)d\varepsilon_i. \quad \textbf{(19-30)}$$

This can be maximized with respect to $(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}, \boldsymbol{\tau})$ using quadrature or simulation. When done, ρ can be recovered from $\rho = \tau/(1+\tau^2)^{1/2}$ and $\boldsymbol{\alpha} = (1-\rho^2)^{1/2}\boldsymbol{\gamma}$. All that differs from one model to another is the specification of $f(y_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i)$. This is the specification used in Terza (1998) and Terza and Kenkel (2001). (In these two papers, the authors also analyzed $E[y_i \mid z_i = 1]$. This estimator was based on nonlinear least squares, but as earlier, it is necessary to integrate the unobserved heterogeneity out of the conditional mean function.) Greene (2010) applies the method to a stochastic frontier model.

19.5.5 PANEL DATA APPLICATIONS OF SAMPLE SELECTION MODELS

The development of methods for extending sample selection models to panel data settings parallels the literature on cross-section methods. It begins with Hausman and Wise (1979) who devised a maximum likelihood estimator for a two-period model with attrition—the "selection equation" was a formal model for attrition from the sample. Subsequent research has drawn the analogy between attrition and sample selection in a variety of applications, such as Keane et al. (1988) and Verbeek and Nijman (1992), and produced theoretical developments including Wooldridge (2002a, b).

The direct extension of panel data methods to sample selection brings several new issues for the modeler. An immediate question arises concerning the nature of the

selection itself. Although much of the theoretical literature [e.g., Kyriazidou (1997, 2001)] treats the panel as if the selection mechanism is run anew in every period, in practice, the selection process often comes in two very different forms. First, selection may take the form of selection of the entire group of observations into the panel data set. Thus, the selection mechanism operates once, perhaps even before the observation window opens. Consider the entry (or not) of eligible candidates for a job training program. In this case, it is not appropriate to build the model to allow entry, exit, and then reentry. Second, for most applications, selection comes in the form of attrition or retention. Once an observation is "deselected," it does not return. Leading examples would include "survivorship" in time-series–cross-section models of firm performance and attrition in medical trials and in panel data applications involving large national survey data bases, such as Contoyannis et al. (2004). Each of these cases suggests the utility of a more structured approach to the selection mechanism.

19.5.5.a Common Effects in Sample Selection Models

A formal "effects" treatment for sample selection was first suggested in complete form by Verbeek (1990), who formulated a random effects model for the probit equation and a fixed effects approach for the main regression. Zabel (1992) criticized the specification for its asymmetry in the treatment of the effects in the two equations. He also argued that the likelihood function that neglected correlation between the effects and regressors in the probit model would render the FIML estimator inconsistent. His proposal involved fixed effects in both equations. Recognizing the difficulty of fitting such a model, he then proposed using the Mundlak correction. The full model is

$$\begin{aligned} y_{it}^* &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad \eta_i = \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \tau w_i, w_i \sim N[0,1], \\ d_{it}^* &= \theta_i + \mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it}, \quad \theta_i = \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i, v_i \sim N[0,1], \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0,0), (\sigma^2, \mathbf{1}, \boldsymbol{\rho\sigma})]. \end{aligned} \tag{19-31}$$

The "selectivity" in the model is carried through the correlation between ε_{it} and u_{it}. The resulting log-likelihood is built up from the contribution of individual i,

$$\begin{aligned} L_i = &\int_{-\infty}^{\infty} \prod_{d_{it}=0} \boldsymbol{\Phi}[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i'\boldsymbol{\delta} - \omega v_i]\phi(v_i)dv_i \\ &\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \prod_{d_{it}=1} \boldsymbol{\Phi}\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1-\rho^2}}\right] \\ &\times \frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it}}{\sigma}\right)\phi_2(v_i, w_i)dv_i dw_i, \\ \varepsilon_{it} = &\, y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \tau w_i. \end{aligned} \tag{19-32}$$

The log-likelihood is then $\ln L = \sum_i \ln L_i$.

The log-likelihood requires integration in two dimensions for any selected observations. Vella (1998) suggested two-step procedures to avoid the integration. However, the bivariate normal integration is actually the product of two univariate normals, because in the preceding specification, v_i and w_i are assumed to be uncorrelated. As such, the likelihood function in (19-32) can be readily evaluated using familiar simulation or quadrature techniques. [See Sections 14.9.6.c and 15.6. Vella and Verbeek (1999)

suggest this in a footnote, but do not pursue it.] To show this, note that the first line in the log-likelihood is of the form $E_v[\prod_{d=0} \Phi(\ldots)]$ and the second line is of the form $E_w[E_v[\Phi(\ldots)\phi(\ldots)/\sigma]]$. Either of these expectations can be satisfactorily approximated with the average of a sufficient number of draws from the standard normal populations that generate w_i and v_i. The term in the simulated likelihood that follows this prescription is

$$L_i^S = \frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=0}\Phi[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i'\boldsymbol{\delta} - \omega v_{i,r}]$$
$$\times \frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=1}\Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_{i,r} + (\rho/\sigma)\varepsilon_{it,r}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it,r}}{\sigma}\right), \qquad \textbf{(19-33)}$$
$$\varepsilon_{it,r} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \boldsymbol{\tau} w_{i,r}.$$

Maximization of this log-likelihood with respect to $(\boldsymbol{\beta},\sigma,\rho,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\pi},\tau,\boldsymbol{\omega})$ by conventional gradient methods is quite feasible. Indeed, this formulation provides a means by which the likely correlation between v_i and w_i can be accommodated in the model. Suppose that w_i and v_i are bivariate standard normal with correlation ρ_{vw}. We can project w_i on v_i and write

$$w_i = \rho_{vw} v_i + \left(1 - \rho_{vw}^2\right)^{1/2} h_i,$$

where h_i has a standard normal distribution. To allow the correlation, we now simply substitute this expression for w_i in the simulated (or original) log-likelihood and add ρ_{vw} to the list of parameters to be estimated. The simulation is still over independent normal variates, v_i and h_i.

Notwithstanding the preceding derivation, much of the recent attention has focused on simpler two-step estimators. Building on Ridder and Wansbeek (1990) and Verbeek and Nijman (1992) [see Vella (1998) for numerous additional references], Vella and Verbeek (1999) purpose a two-step methodology that involves a random effects framework similar to the one in (19-31). As they note, there is some loss in efficiency by not using the FIML estimator. But, with the sample sizes typical in contemporary panel data sets, that efficiency loss may not be large. As they note, their two-step template encompasses a variety of models including the tobit model examined in the preceding sections and the mover-stayer model noted earlier.

The Vella and Verbeek model requires some fairly intricate maximum likelihood procedures. Wooldridge (1995) proposes an estimator that, with a few probably—but not necessarily—innocent assumptions, can be based on straightforward applications of conventional, everyday methods. We depart from a fixed effects specification,

$$y_{it}^* = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$
$$d_{it}^* = \theta_i + \mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it},$$
$$(\varepsilon_{it}, u_{it}) \sim N_2[(0,0), (\sigma^2, 1, \rho\sigma)].$$

Under the **mean independence assumption** $E[\varepsilon_{it} \mid \eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{it}, v_{i1}, \ldots, v_{it}, d_{i1}, \ldots, d_{it}] = \rho u_{it}$, it will follow that

$$E[y_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{it}, v_{i1}, \ldots, v_{it}, d_{i1}, \ldots, d_{it}] = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho u_{it}.$$

This suggests an approach to estimating the model parameters; however, it requires computation of u_{it}. That would require estimation of θ_i, which cannot be done, at least not consistently—and that precludes simple estimation of u_{it}. To escape the dilemma, Wooldridge (2002c) suggests Chamberlain's approach to the fixed effects model,

$$\theta_i = f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{it}'\mathbf{f}_T + h_i.$$

With this substitution,

$$\begin{aligned} d_{it}^* &= \mathbf{z}_{it}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{it}'\mathbf{f}_T + h_i + u_{it} \\ &= \mathbf{z}_{it}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{it}'\mathbf{f}_T + w_{it}, \end{aligned}$$

where w_{it} is independent of $\mathbf{z}_{it}$, $t = 1, \ldots, T$. This now implies that

$$\begin{aligned} E[y_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{it}, v_{i1}, \ldots, v_{it}, d_{i1}, \ldots, d_{it}] &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho(w_{it} - h_i) \\ &= (\eta_i - \rho h_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho w_{it}. \end{aligned}$$

To complete the estimation procedure, we now compute T cross-sectional probit models (reestimating $f_0, \mathbf{f}_1, \ldots$ each time) and compute $\hat{\lambda}_{it}$ from each one. The resulting equation,

$$y_{it} = a_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho\hat{\lambda}_{it} + v_{it},$$

now forms the basis for estimation of $\boldsymbol{\beta}$ and ρ by using a conventional fixed effects linear regression with the observed data.

19.5.5.b Attrition

The recent literature or sample selection contains numerous analyses of two-period models, such as Kyriazidou (1997, 2001). They generally focus on non- and semiparametric analyses. An early parametric contribution of Hausman and Wise (1979) is also a two-period model of attrition, which would seem to characterize many of the studies suggested in the current literature. The model formulation is a two-period random effects specification:

$$\begin{aligned} y_{i1} &= \mathbf{x}_{i1}'\boldsymbol{\beta} + \varepsilon_{i1} + u_i \quad \text{(first period regression)}, \\ y_{i2} &= \mathbf{x}_{i2}'\boldsymbol{\beta} + \varepsilon_{i2} + u_i \quad \text{(second period regression)}. \end{aligned}$$

Attrition is likely in the second period (to begin the study, the individual must have been observed in the first period). The authors suggest that the probability that an observation is made in the second period varies with the value of y_{i2} as well as some other variables,

$$z_{i2}^* = \delta y_{i2} + \mathbf{x}_{i2}'\boldsymbol{\theta} + \mathbf{w}_{i2}'\boldsymbol{\alpha} + v_{i2}.$$

Attrition occurs if $z_{i2}^* \leq 0$, which produces a probit model,

$$z_{i2} = 1\left(z_{i2}^* > 0\right) \quad \text{(attrition indicator observed in period 2)}.$$

An observation is made in the second period if $z_{i2} = 1$, which makes this an early version of the familiar sample selection model. The reduced form of the observation

equation is

$$\begin{aligned} z_{i2}^{*} &= \mathbf{x}_{i2}'(\boldsymbol{\delta}\boldsymbol{\beta} + \boldsymbol{\theta}) + \mathbf{w}_{i2}'\boldsymbol{\alpha} + \delta\varepsilon_{i2} + v_{i2} \\ &= \mathbf{x}_{i2}'\boldsymbol{\pi} + \mathbf{w}_{i2}'\boldsymbol{\alpha} + h_{i2} \\ &= \mathbf{r}_{i2}'\boldsymbol{\gamma} + h_{i2}. \end{aligned}$$

The variables in the probit equation are all those in the second period regression plus any additional ones dictated by the application. The estimable parameters in this model are $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 = \text{Var}[\varepsilon_{it} + u_i]$, and two correlation coefficients,

$$\rho_{12} = \text{Corr}[\varepsilon_{i1} + u_i, \varepsilon_{i2} + u_i] = \text{Var}[u_i]/\sigma^2,$$

and

$$\rho_{23} = \text{Corr}[h_{i2}, \varepsilon_{i2} + u_i].$$

All disturbances are assumed to be normally distributed. (Readers are referred to the paper for motivation and details on this specification.)

The authors propose a full information maximum likelihood estimator. Estimation can be simplified somewhat by using two steps. The parameters of the probit model can be estimated first by maximum likelihood. Then the remaining parameters are estimated by maximum likelihood, conditionally on these first-step estimates. The Murphy and Topel adjustment is made after the second step. [See Greene (2007a).]

The Hausman and Wise model covers the case of two periods in which there is a formal mechanism in the model for retention in the second period. It is unclear how the procedure could be extended to a multiple-period application such as that in Contoyannis et al. (2004), which involved a panel data set with eight waves. In addition, in that study, the variables in the main equations were counts of hospital visits and physican visits, which complicates the use of linear regression. A workable solution to the problem of attrition in a multiperiod panel is the **inverse probability weighted estimator** [Wooldridge (2002a, 2006b) and Rotnitzky and Robins (2005).] In the Contoyannis application, there are eight waves in the panel. Attrition is taken to be "ignorable" so that the unobservables in the attrition equation and in the main equation(s) of interest are uncorrelated. (Note that Hausman and Wise do not make this assumption.) This enables Contoyannis et al. to fit a "retention" probit equation for each observation present at wave 1, for waves 2–8, using characteristics observed at the entry to the panel. (This defines, then, "selection (retention) on observables.") Defining d_{it} to be the indicator for presence ($d_{it} = 1$) or absence ($d_{it} = 0$) of observation i in wave t, it will follow that the sequence of observations will begin at 1 and either stay at 1 or change to 0 for the remaining waves. Let $\hat{p}_{it}$ denote the predicted probability from the probit estimator at wave t. Then, their full log-likelihood is constructed as

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{d_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Wooldridge (2002b) presents the underlying theory for the properties of this weighted maximum likelihood estimator. [Further details on the use of the inverse probability weighted estimator in the Contoyannis et al. (2004) study appear in Jones, Koolman, and Rice (2006) and in Section 17.4.9.]

19.6 EVALUATING TREATMENT EFFECTS

The leading recent application of models of selection and endogeneity is the evaluation of "**treatment effects**." The central focus is on analysis of the effect of participation in a treatment, T, on an outcome variable, y—examples include job training programs [LaLonde (1986), Business Week (2009; Example 19.14)] and education [e.g., test scores, Angrist and Lavy (1999), Van der Klaauw (2002)]. Wooldridge and Imbens (2009, pp. 22–23) cite a number of labor market applications. Recent more narrow examples include Munkin and Trivedi's (2007) analysis of the effect of dental insurance and Jones and Rice's (2010) survey that notes a variety of techniques and applications in health economics.

Example 19.14 German Labor Market Interventions

"Germany long had the highest ratio of unfilled jobs to unemployed people in Europe. Then, in 2003, Berlin launched the so-called Hartz reforms, ending generous unemployment benefits that went on indefinitely. Now payouts for most recipients drop sharply after a year, spurring people to look for work. From 12.7% in 2005, unemployment fell to 7.1% last November. Even now, after a year of recession, Germany's jobless rate has risen to just 8.6%.

At the same time, lawmakers introduced various programs intended to make it easier for people to learn new skills. One initiative instructed the Federal Labor Agency, which had traditionally pushed the long-term unemployed into government-funded make-work positions, to cooperate more closely with private employers to create jobs. That program last year paid Dutch staffing agency Randstad to teach 15,000 Germans information technology, business English, and other skills. And at a Daimler truck factory in Wörth, 55 miles west of Stuttgart, several dozen short-term employees at risk of being laid off got government help to continue working for the company as mechanic trainees.

Under a second initiative, Berlin pays part of the wages of workers hired from the ranks of the jobless. Such payments make employers more willing to take on the costs of training new workers. That extra training, in turn, helps those workers keep their jobs after the aid expires, a study by the government-funded Institute for Employment Research found. Café Nenninger in the city of Kassel, for instance, used the program to train an unemployed single mother. Co-owner Verena Nenninger says she was willing to take a chance on her in part because the government picked up about a third of her salary the first year. 'It was very helpful, because you never know what's going to happen,' Nenninger says" [*Business Week* (2009)].

Empirical measurement of treatment effects, such as the impact of going to college or participating in a job training program, presents a large variety of econometric complications. The natural, ultimate objective of an analysis of a "treatment" or intervention would be the "effect of treatment on the treated." For example, what is the effect of a college education on the lifetime income of someone who goes to college? Measuring this effect econometrically encounters at least two compelling computations:

Endogeneity of the treatment: The analyst risks attributing to the treatment causal effects that should be attributed to factors that motivate both the treatment and the outcome. In our example, the individual who goes to college might well have succeeded (more) in life than their counterpart who did not go to college even if they (themselves) did not attend college.

Missing counterfactual: The preceding thought experiment is not actually the effect we wish to measure. In order to measure the impact of college attendance on lifetime earnings in a pure sense, we would have to run an individual's lifetime twice, once with

college attendance and once without. Any individual is observed in only one of the two states, so the pure measurement is impossible.

Accommodating these two problems forms the focal point of this enormous and still growing literature. **Rubin's causal model** (1974, 1978) provides a useful framework for the analysis. Every individual in a population has a potential outcome, y and can be exposed to the treatment, C. We will denote by C_i the indicator whether or not the individual receives the treatment. Thus, the potential outcomes are $y_i \mid (C_i = 1) = y_{i1}$ and $y_i \mid (C_i = 0) = y_{i0}$. The **average treatment effect**, averaged across the entire population is

$$\text{ATE} = E[y_{i1} - y_{i0}].$$

The compelling complication is that the individual will exist in only one of the two states, so it is not possible to estimate ATE without further assumptions. More specifically, what the researcher would prefer see is the **average treatment effect on the treated**,

$$\text{ATET} = E[y_{i1} - y_{i0} \mid C_i = 1]$$

and note that the second term is the missing counterfactual.

One of the major themes of the recent research is to devise robust methods of estimation that do not rely heavily on fragile assumptions such as identification by functional form (e.g., relying on bivariate normality) and identification by exclusion restrictions (e.g., relying on basic instrumental variable estimators). This is a challenging exercise—we have relied heavily on these assumptions in most of the work in this book up to this point. For purposes of the general specification, we will denote by $\mathbf{x}$ the exogenous information that will be brought to bear on this estimation problem. The vector $\mathbf{x}$ may (usually will) be a set of variables that will appear in a regression model, but it is useful to think more generally than that and consider $\mathbf{x}$ rather to be an information set. Certain minimal assumptions are necessary to make any headway at all. The following appear at different points in the analysis.

Conditional independence: Receiving the treatment, C_i, does not depend on the outcome variable once the effect of $\mathbf{x}$ on the outcome is accounted for. If assignment to the treatment group is completely random, then we would omit the effect of $\mathbf{x}$ in this assumption. This assumption is extended for regression approaches with the **conditional mean assumption**: $E[y_{i0} \mid \mathbf{x}_i, C_i = 1] = E[y_{i0} \mid \mathbf{x}_i, C_i = 0] = E[y_{i0} \mid \mathbf{x}]$. This states that the outcome in the untreated state does not affect the participation.

Distribution of potential outcomes: The model that is used for the outcomes is the same for treated and nontreated, $f(y \mid \mathbf{x}, T = 1) = f(y \mid \mathbf{x}, T = 0)$. In a regression context, this would mean that the same regression applies in both states and that the disturbance is uncorrelated with T, or that T is exogenous. This is a very strong assumption that we will relax later. For the present, it removes one of the complications noted previously, so a step in the model-building exercise will be to relax this assumption.

Overlap assumption: For any value of $\mathbf{x}$, $0 < \text{Prob}(C_i = 1 \mid \mathbf{x}) < 1$. The strict inequality in this assumption means that for any $\mathbf{x}$, the population will contain a mix of treated and nontreated individuals. The usefulness of the overlap assumption is that with it, we can expect to find, for any treated individual, an individual who looks like them but is not treated. This assumption will be useful for regression approaches.

The following sections will describe three major parts of the research agenda on treatment effects: regression analysis with control functions in Section 19.6.1, propensity score matching in Section 19.6.2, and regression discontinuity design in Section 19.6.3. A fourth area, instrumental variable estimation, was developed in Chapter 8. As noted, this is a huge and rapidly growing literature. For example, Imbens and Wooldridge's (2009) survey paper runs 85 pages and includes nearly 300 references, most of them since 2000. Our purpose here is to provide some of the vocabulary and a superficial introduction to methods. The survey papers by Imbens and Wooldridge (2009) and Jones and Rice (2010) provide greater detail. The conference volume by Millment, Smith, and Vytlacil (2008) contains many theoretical contributions and empirical applications.[33] A *Journal of Business and Economic Statistics* symposium [Angrist (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

19.6.1 REGRESSION ANALYSIS OF TREATMENT EFFECTS

The basic model of selectivity outlined earlier has been extended in an impressive variety of directions. An interesting application that has found wide use is the measurement of **treatment effects** and program effectiveness.

An earnings equation that accounts for the value of a college education is

$$\text{earnings}_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta C_i + \varepsilon_i,$$

where C_i is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does δ measure the value of a college education (assuming that the rest of the regression model is correctly specified)? The answer is no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of δ will actually overestimate the treatment effect. The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

To put this in a more familiar context, suppose that we model program participation (e.g., whether or not the individual goes to college) as

$$C_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,$$
$$C_i = 1 \quad \text{if } C_i^* > 0, 0 \text{ otherwise.}$$

We also suppose that, consistent with our previous conjecture, u_i and ε_i are correlated. Coupled with our earnings equation, we find that

$$\begin{aligned} E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{w}_i] &= \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\delta} + E[\varepsilon_i \mid C_i = 1, \mathbf{x}_i, \mathbf{w}_i] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\delta} + \rho\sigma_\varepsilon\lambda(-\mathbf{w}_i'\boldsymbol{\gamma}) \end{aligned} \qquad \textbf{(19-34)}$$

once again. [See (19-24).] Evidently, a viable strategy for estimating this model is to use the two-step estimator discussed earlier. The net result will be a different estimate of δ

[33] In the initial essay in the volume, Goldberger (2008) reproduces Goldberger (1972) in which the author explores the endogeneity issue in detail with specific reference to the Head Start program of the 1960s.

that will account for the self-selected nature of program participation. For nonparticipants, the counterpart to (19-34) is

$$E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon \left[\frac{-\phi(\mathbf{w}_i'\boldsymbol{\gamma})}{1 - \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\gamma})}\right]. \tag{19-35}$$

The difference in expected earnings between participants and nonparticipants is, then,

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{w}_i] - E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{w}_i] = \delta + \rho\sigma_\varepsilon \left[\frac{\phi_i}{\boldsymbol{\Phi}_i(1 - \boldsymbol{\Phi}_i)}\right]. \tag{19-36}$$

If the selectivity correction λ_i is omitted from the least squares regression, then this difference is what is estimated by the least squares coefficient on the treatment dummy variable. But because (by assumption) all terms are positive, we see that least squares overestimates the treatment effect. Note, finally, that simply estimating separate equations for participants and nonparticipants does not solve the problem. In fact, doing so would be equivalent to estimating the two regressions of Example 19.12 by least squares, which, as we have seen, would lead to inconsistent estimates of both sets of parameters.

To describe the problem created by **selection on the unobservables**, we will drop the independence assumptions. The model with endogenous participation and different outcome equations would be

$$\begin{aligned}
&C_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,\ C_i = \mathbf{1} \text{ if } C_i^* > 0 \text{ and 0 otherwise,}\\
&y_{i0} = \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_{i0},\\
&y_{i1} = \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{i1}.
\end{aligned}$$

It is useful to combine the second and third equations in

$$y_{ij} = C_i(\mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{i1}) + (1 - C_i)(\mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_{i0}),\ j = 0, 1.$$

We assume joint normality for the three disturbances;

$$\begin{pmatrix} u_i \\ \varepsilon_{i0} \\ \varepsilon_{i1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0\theta_0 & \rho_1\theta_1 \\ \rho_0\theta_0 & \theta_0^2 & \theta_{01} \\ \rho_1\theta_1 & \theta_{01} & \theta_1^2 \end{pmatrix}\right].$$

The variance in the participation equation is normalized to one for a binary outcome, as described earlier (Section 17.2). Endogeneity of the participation is implied by the nonzero values of the correlations ρ_0 and ρ_1. The familiar problem of the missing counterfactual appears here in our inability to estimate θ_{01}. The data will never contain information on both states simultaneously, so it will be impossible to estimate a covariance of y_{i0} and y_{i1} (conditioned on $\mathbf{x}_i$ or otherwise). Thus, the parameter θ_{01} is not identified (estimable)—we normalize it to zero. The parameters of this model after the two normalizations can be estimated by two-step least squares as suggested in Section 19.XX, or by full information maximum likelihood. The average treatment effect on the treated would be

$$\begin{aligned}
\text{ATET} &= E[y_{i1} \mid C_i = 1,\ \mathbf{x}_i,\ \mathbf{w}_i] - E[y_{i0} \mid C_i = 1,\ \mathbf{x}_i,\ \mathbf{w}_i]\\
&= \mathbf{x}_i'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\rho_1\theta_1 - \rho_0\theta_0)\frac{\phi\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{w}_i'\boldsymbol{\gamma}\right)}.
\end{aligned}$$

[See (19-34).] If the treatment assignment is completely random, then $\rho_1 = \rho_0 = 0$, and we are left with the first term. But, of course, it is the nonrandomness of the treatment assignment that brought us to this point. Finally, if the two coefficient vectors differ only in their constant terms, $\beta_{0,0}$ and $\beta_{1,0}$, then we are left with the same δ that appears in (19-36)—the ATET would be $\beta_{0,1} + C_i(\beta_{1,0} - \beta_{0,0})$.

There are many variations of this model in the empirical literature. They have been applied to the analysis of education,[34] the Head Start program,[35] and a host of other settings.[36] This strand of literature is particularly important because the use of dummy variable models to analyze treatment effects and program participation has a long history in empirical economics. This analysis has called into question the interpretation of a number of received studies.

19.6.1.a The Normality Assumption

Some research has cast some skepticism on the selection model based on the normal distribution. [See Goldberger (1983) for an early salvo in this literature.] Among the findings are that the parameter estimates are surprisingly sensitive to the distributional assumption that underlies the model. Of course, this fact in itself does not invalidate the normality assumption, but it does call its generality into question. On the other hand, the received evidence is convincing that sample selection, in the abstract, raises serious problems, distributional questions aside. The literature—for example, Duncan (1986b), Manski (1989, 1990), and Heckman (1990)—has suggested some promising approaches based on robust and nonparametric estimators. These approaches obviously have the virtue of greater generality. Unfortunately, the cost is that they generally are quite limited in the breadth of the models they can accommodate. That is, one might gain the robustness of a nonparametric estimator at the cost of being unable to make use of the rich set of accompanying variables usually present in the panels to which selectivity models are often applied. For example, the nonparametric bounds approach of Manski (1990) is defined for two regressors. Other methods [e.g., Duncan (1986b)] allow more elaborate specifications.

Recent research includes specific attempts to move away from the normality assumption.[37] An example is Martins (2001), building on Newey (1991), which takes the core specification as given in (19-22) as the platform but constructs an alternative to the assumption of bivariate normality. Martins's specification modifies the Heckman model by employing an equation of the form

$$E[y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \mu(\mathbf{w}_i'\boldsymbol{\gamma})$$

where the latter "selectivity correction" is not the inverse Mills ratio, but some other result from a different model. The correction term is estimated using the Klein and Spady model discussed in Section 23.6.1. This is labeled a "semiparametric" approach. Whether the conditional mean in the selected sample should even remain a linear index function remains to be settled. Not surprisingly, Martins's results, based on two-step

[34] Willis and Rosen (1979).

[35] Goldberger (1972, 2008).

[36] A useful summary of the issues is Barnow, Cain, and Goldberger (1981). See, also, Imbens and Wooldridge (2009).

[37] Again, Angrist (2001) is an important contribution to this literature.

least squares differ only slightly from the conventional results based on normality. This approach is arguably only a fairly small step away from the tight parameterization of the Heckman model. Other non- and semiparametric specifications, for example, Honorè and Kyriazidou (1997, 2000) represent more substantial departures from the normal model, but are much less operational.[38] The upshot is that the issue remains unsettled. For better or worse, the empirical literature on the subject continues to be dominated by Heckman's original model built around the joint normal distribution.

19.6.1.b Estimating the Effect of Treatment on the Treated

Consider a regression approach to analyzing treatment effects in a two-period setting,

$$y_{it} = \theta_t + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma C_i + u_i + \varepsilon_{it}, \quad t = 0, 1,$$

where C_i is the treatment dummy variable and u_i is the unobserved individual effect. The setting is the pre- and posttreatment analysis of the sort considered in this section, where we examine the impact of a job training program on post training earnings. Because there are two periods, a natural approach to the analysis is to examine the changes,

$$\Delta y_i = (\theta_1 - \theta_0) + \gamma \Delta C_i + (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \Delta \varepsilon_{it}$$

where $\Delta C_i = 1$ for the treated and 0 for the nontreated individuals, and the first differences eliminate the unobserved individual effects. In the absence of controls (regressors, $\mathbf{x}_{it}$), or assuming that the controls are unchanged, the estimator of the effect of the treatment will be

$$\hat{\gamma} = \overline{\Delta y} \,|\, (\Delta C_i = 1) - \overline{\Delta y} \,|\, (C_i = 0),$$

which is the **difference in differences** estimator. This simplifies the problem considerably but has several shortcomings. Most important, by using the simple differences, we have lost our ability to discern what induced the change, whether it was the program or something else, presumably in $\mathbf{x}_{it}$.

Even without the normality assumption, the preceding regression approach is more tightly structured than many are comfortable with. A considerable amount of research has focused on what assumptions are needed to reach that model and whether they are likely to be appropriate in a given setting.[39] The overall objective of the analysis of the preceding two sections is to evaluate the effect of a treatment, C_i, on the individual treated. The implicit counterfactual is an observation on what the "response" (dependent variable) of the treated individual would have been had they not been treated. But, of course, an individual will be in one state or the other, not both. Denote by y_0 the random variable that is the outcome variable in the absence of the treatment and by y_1 the outcome when the treatment has taken place. The **average treatment effect**,

[38] This particular work considers selection in a "panel" (mainly two periods). But, the panel data setting for sample selection models is more involved than a cross-section analysis. In a panel data set, the "selection" is likely to be a decision at the beginning of Period 1 to be in the data set for all subsequent periods. As such, something more intricate than the model we have considered here is called for.

[39] A sampling of the more important parts of the literature on this issue includes Heckman (1992, 1997), Imbens and Angrist (1994), Manski (1996), and Wooldridge (2002a, Chapter 18).

averaged over the entire population is

$$ATE = E[y_1 - y_0].$$

This is the impact of the treatment on an individual drawn at random from the entire population. However, the desired quantity is not necessarily the *ATE*, but the **average treatment effect on the treated**, which would be

$$ATE \mid T = E[y_1 - y_0 \mid C = 1].$$

The difficulty of measuring this is, once again, the counterfactual, $E[y_0 \mid C = 1]$. Whether these two measures will be the same is at the center of the much of the discussion on this subject. If treatment is completely randomly assigned, then $E[y_j \mid C = 1] = E[y_j \mid C = 0] = E[y_j \mid C = j]$, $j = 0, 1$. This means that with completely random treatment assignment

$$ATE = E[y_1 \mid C = 1] - E[y_0 \mid C = 0].$$

To put this in our example, if college attendance were completely randomly distributed throughout the population, then the impact of college attendance on income (neglecting other covariates at this point) could be measured simply by averaging the incomes of college attendees and subtracting the average income of nonattendees. The preceding theory might work for the treatment "having brown eyes," but it is unlikely to work for college attendance. Not only is the college attendance treatment not randomly distributed, but the treatment "assignment" is surely related to expectations about y_1 versus y_0, and, at a minimum, y_0 itself. (College is expensive.) More generally, the researcher faces the difficulty in calculating treatment effects that assignment to the treatment might not be exogenous.

The **control function** approach that we used in (19-34)–(19-36) is used to account for the endogeneity of the treatment assignment in the regression context. The very specific assumptions of the bivariate normal distribution of the unobservables somewhat simplifies the estimation, because they make explicit what control function (λ_i) is appropriate to use in the regression. As Wooldridge (2002a, p. 622) points out, however, the binary variable in the treatment effects regression represents simply an endogenous variable in a linear equation, amenable to **instrumental variable estimation** (assuming suitable instruments are available). Barnow, Cain, and Goldberger (1981) proposed a two-stage least squares estimator, with instrumental variable equal to the predicted probability from the probit treatment assignment model. This is slightly less **parametric** than (19-36) because, in principle, its validity does not rely on joint normality of the disturbances. [Wooldridge (2002a, pp. 621–633) discusses the underlying assumptions.

19.6.2 PROPENSITY SCORE MATCHING

If the treatment assignment is "completely ignorable," then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common $\mathbf{x}_i$, one with $C_i = 0$, the other with $C_i = 1$. If done with a sufficient number of pairs so as to average

over the population of $\mathbf{x}_i$'s, then a **matching estimator**, the average value of $(y_i \mid C_i = 1) - (y_i \mid C_i = 0)$, would estimate $E[y_1 - y_0]$, which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of $\mathbf{x}_i$ are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a "family income" variable. Rosenbaum and Rubin (1983) and others[40] suggested, instead, matching on the **propensity score**, $F(\mathbf{x}_i) = \text{Prob}(C_i = 1 \mid \mathbf{x}_i)$. Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid. [See, e.g., Becker and Ichino (2002) and Greene (2007c).]

Example 19.15 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." [The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for details on the NSW experiments.] The training period was 1976–1977; the outcome of interest for the sample examined here was posttraining 1978 earnings. LaLonde reports a large variety of estimates of the treatment effect, for different subgroups and using different estimation methods. Nonparametric estimates for the group in our sample are roughly $900 for the income increment in the posttraining year. (See LaLonde, p. 609.) Similar results are reported from a two-step regression-based estimator similar to (19-34) to (19-36). (See LaLonde's footnote to Table 6, p. 616.)

LaLonde's data are fairly well traveled, having been used in replications and extensions in, for example, Dehejia and Wahba (1999), Becker and Ichino (2002), and Greene (2007b, c). We have reestimated the matching estimates reported in Becker and Ichino. The data in the file used there (and here) contain 2,490 control observations and 185 treatment observations on the following variables:

t	= treatment dummy variable
age	= age in years
educ	= education in years
marr	= dummy variable for married
black	= dummy variable for black
hisp	= dummy variable for Hispanic
nodegree	= dummy for no degree (not used)
*re*74	= real earnings in 1974
*re*75	= real earnings in 1975
*re*78	= real earnings in 1978

[40] Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith, and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), and Heckman and Vytlacil (2000).

Transformed variables added to the equation are

age^2 = *age* squared
$educ^2$ = *educ* squared
$re74^2$ = *re*74 squared
*re*7[illegible] = *re*75 squared
*blacku*74 = *black* times 1(*re*74 = 0)

We also scaled all earnings variables by 10,000 before beginning the analysis. (See Appendix Table F19.3. The data are downloaded from the website http://www.nber.org/%7Erdehejia/nswdata.html. The two specific subsamples are in http://www.nber.org/%7Erdehejia//psid_controls.txt and http://www.nber.org/%7Erdehejia/nswre74_treated.txt.) (We note that Becker and Ichino report they were unable to replicate Dehejia and Wahba's results, although they could come reasonably close. We, in turn, were not able to replicate either set of results, though we, likewise, obtained quite similar results.)

The analysis proceeded as follows: A logit model in which the included variables were a *constant, age, age*^{2}*, education, education*^{2}*, marr, black, hisp, re74, re75, re742, re752*, and *black74* was computed for the treatment assignment. The fitted probabilities are used for the propensity scores. By means of an iterative search, the range of propensity scores was partitioned into eight regions within which, by a simple F test, the mean scores of the treatments and controls were not statistically different. The partitioning is shown in Table 19.10. The 1,347 observations are all the treated observations and the 1,162 control observations are those whose propensity scores fell within the range of the scores for the treated observations.

Within each interval, each treated observation is paired with a small number of the nearest control observations. We found the average difference between treated observation and control to equal $1,574.35. Becker and Ichino reported $1,537.94.

As an experiment, we refit the propensity score equation using a probit model, retaining the fitted probabilities. We then used the two-step estimator described earlier to fit (19-34) and (19-35) using the entire sample. The estimates of δ, ρ, and σ were -1.01437, 0.35519, 1.38426). Using the results from the probit model, we averaged the result in (19-36) for the entire sample, obtaining an estimated treatment effect of $1,476.30.

TABLE 19.10 Empirical Distribution of Propensity Scores

Percent	*Lower*	*Upper*
0–5	0.000591	0.000783
5–10	0.000787	0.001061
10–15	0.001065	0.001377
15–20	0.001378	0.001748
20–25	0.001760	0.002321
25–30	0.002340	0.002956
30–35	0.002974	0.004057
35–40	0.004059	0.005272
40–45	0.005278	0.007486
45–50	0.007557	0.010451
50–55	0.010563	0.014643
55–60	0.014686	0.022462
60–65	0.022621	0.035060
65–70	0.035075	0.051415
70–75	0.051415	0.076188
75–80	0.076376	0.134189
80–85	0.134238	0.320638
85–90	0.321233	0.616002
90–95	0.624407	0.949418
95–100	0.949418	0.974835

Sample size = 1,347
Average score = 0.137238
Std. Dev score = 0.274079

	Lower	Upper	# Obs
1	0.000591	0.098016	1041
2	0.098016	0.195440	63
3	0.195440	0.390289	65
4	0.390289	0.585138	36
5	0.585138	0.779986	32
6	0.779986	0.877411	17
7	0.877411	0.926123	7
8	0.926123	0.974835	86

19.6.3 REGRESSION DISCONTINUITY

There are many situations in which there is no possibility of randomized assignment of treatments. Examples include student outcomes and policy interventions in schools. Angrist and Lavy (1999), for example, studied the effect of class sizes on test scores. Van der Klaauw studied financial aid offers that were tied to SAT scores and grade point averages. In these cases, the natural experiment approach advocated by Angrist and Pischke (2009) is an appealing way to proceed, when it is feasible. The **regression discontinuity design** presents an alternative strategy. The conditions under which the approach can be effective are when (1) the outcome, y, is a continuous variable; (2) the outcome varies smoothly with an assignment variable, A, and (3) treatment is "sharply" assigned based on the value of A, specifically $C = 1(A > A^*)$ where A^* is a fixed threshold or cutoff value. [A "**fuzzy design** is based on $\text{Prob}(C = 1 \mid A) = F(A)$. The identification problems with fuzzy design are much more complicated than with sharp design. Readers are referred to Van der Klaauw (2002) for further discussion of fuzzy design.] We assume, then, that

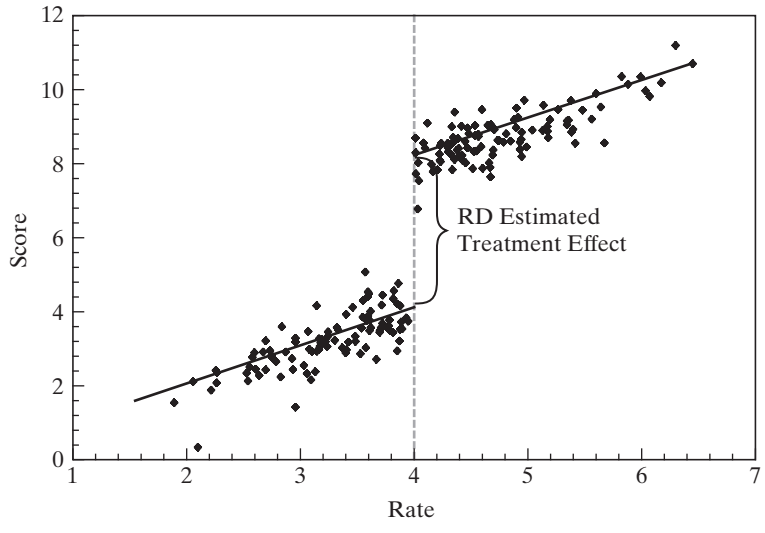
$$y = f(A, C) + \varepsilon.$$

Suppose, for example, the outcome variable is a test score, and that an administrative treatment such as a special education program is funded based on the poverty rates of certain communities. The ideal conditions for a regression discontinuity design based on these assumptions is shown in Figure 19.8. The logic of the calculation is that the points near the threshold value, which have "essentially" the same stimulus value, constitute a nearly random sample of observations which are segmented by the treatment.

The method requires that $E[\varepsilon \mid A, C] = E[\varepsilon \mid A]$—the assignment variable—be exogenous to the experiment. The result in Figure 19.8 is consistent with

$$y = f(A) + \alpha C + \varepsilon,$$

FIGURE 19.8 Regression Discontinuity.



where α will be the treatment effect to be estimated. The specification of $f(A)$ can be problematic; assuming a linear function when something more general will bias the estimate of α. For this reason, nonparametric methods, such as the LOWESS regression (see Section 12.3.5) might be attractive. This is likely to enable the analyst to make fuller use of the observations that are more distant from the cutoff point. [See Van der Klaaus (2002).] Identification of the treatment effect begins with the assumption that $f(A)$ is continuous at A^*, so that

$$\lim_{A\uparrow A^*} f(A) = \lim_{A\downarrow A^*} f(A) = f(A^*).$$

Then

$$\lim_{A\downarrow A^*} E[y \mid A] - \lim_{A\uparrow A^*} E[y \mid A] = f(A^*) + \alpha + \lim_{A\downarrow A^*} E[\varepsilon \mid A] - f(A^*) - \lim_{A\uparrow A^*} E[\varepsilon \mid A]$$
$$= \alpha.$$

With this in place, the treatment effect can be estimated by the difference of the average outcomes for those individuals "close" to the threshold value, A^*. Details on regression discontinuity design are provided by Trochim (1984, 2000) and Van der Klaauw (2002).

19.7 SUMMARY AND CONCLUSIONS

This chapter has examined settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form: truncation, censoring, and sample selection or endogenous sampling. For each case, we develop the basic theory of the effect and then use the results in a major area of research in econometrics.

In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But when data are truncated so that plausible values of the dependent variable are precluded, for example, when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. The stochastic frontier model is based on a composite disturbance in which one part follows the assumptions of the familiar regression model while the second component is built on a platform of the truncated regression.

When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Consistent estimation, and useful interpretation of estimation results are based on maximum likelihood or some other technique that explicitly accounts for the censoring mechanism. The most common case of censoring in observed data arises in the context of duration analysis, or survival functions (which borrows a term from medical statistics where this style of model building originated). It is useful to think of duration, or survival data, as the measurement of time between transitions or changes of state. We examined three modeling approaches that correspond to the description in Chapter 12; nonparametric (survival tables), semiparametric (the proportional hazard models), and parametric (various forms such as the Weibull model).

Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two-step estimator has become more common. The leading contemporary application of selection methods and endogenous sampling is in the measure of treatment effects. We considered three approaches to analysis of treatment effects; regression methods, propensity score matching, and regression discontinuity.

Key Terms and Concepts

- Accelerated failure time model
- Attenuation
- Average treatment effect
- Average treatment effect on the treated
- Censored regression model
- Censored variable
- Censoring
- Conditional mean assumption
- Conditional moment test
- Control function
- Corner solution model
- Data envelopment analysis
- Degree of truncation
- Delta method
- Difference in differences
- Duration model
- Exponential
- Exponential model
- Fuzzy design
- Generalized residual
- Hazard function
- Hazard rate
- Heterogeneity
- Heteroscedasticity
- Hurdle model
- Incidental truncation
- Instrumetal variable estimation
- Integrated hazard function
- Inverse probability weighted estimator
- Inverse Mills ratio
- Lagrange multiplier test
- Matching estimator
- Mean independence assumption
- Missing counterfactual
- Negative duration dependence
- Olsen's reparameterization
- Parametric
- Parametric model
- Partial likelihood
- Positive duration dependence
- Product limit estimator
- Propensity score
- Proportional hazard
- Regression discontinuity design
- Risk set
- Rubin causal model
- Sample selection
- Selection on observables
- Selection on unobservables
- Semiparametric estimator
- Semiparametric model
- Specification error
- Stochastic frontier model
- Survival function
- Time-varying covariate
- Tobit model
- Treatment effect
- Truncated distribution
- Truncated mean
- Truncated normal distribution
- Truncated random variable
- Truncated standard normal distribution
- Truncated variance
- Truncation
- Two-step estimation
- Type II tobit model
- Weibull model
- Weibull survival model

Exercises

1. The following 20 observations are drawn from a censored normal distribution:

3.8396	7.2040	0.00000	0.00000	4.4132	8.0230
5.7971	7.0828	0.00000	0.80260	13.0670	4.3211
0.00000	8.6801	5.4571	0.00000	8.1021	0.00000
1.2526	5.6016				

The applicable model is

$$y_i^* = \mu + \varepsilon_i,$$
$$y_i = y_i^* \quad \text{if } \mu + \varepsilon_i > 0, 0 \text{ otherwise},$$
$$\varepsilon_i \sim N[0, \sigma^2].$$

Exercises 1 through 4 in this section are based on the preceding information. The OLS estimator of μ in the context of this tobit model is simply the sample mean. Compute the mean of all 20 observations. Would you expect this estimator to over- or underestimate μ? If we consider only the nonzero observations, then the truncated regression model applies. The sample mean of the nonlimit observations is the least squares estimator in this context. Compute it and then comment on whether this sample mean should be an overestimate or an underestimate of the true mean.

2. We now consider the tobit model that applies to the full data set.
   a. Formulate the log-likelihood for this very simple tobit model.
   b. Reformulate the log-likelihood in terms of $\theta = 1/\sigma$ and $\gamma = \mu/\sigma$. Then derive the necessary conditions for maximizing the log-likelihood with respect to θ and γ.
   c. Discuss how you would obtain the values of θ and γ to solve the problem in part b.
   d. Compute the maximum likelihood estimates of μ and σ.
3. Using only the nonlimit observations, repeat Exercise 2 in the context of the truncated regression model. Estimate μ and σ by using the method of moments estimator outlined in Example 19.2. Compare your results with those in the previous exercises.
4. Continuing to use the data in Exercise 1, consider once again only the nonzero observations. Suppose that the sampling mechanism is as follows: y^* and another normally distributed random variable z have population correlation 0.7. The two variables, y^* and z, are sampled jointly. When z is greater than zero, y is reported. When z is less than zero, both z and y are discarded. Exactly 35 draws were required to obtain the preceding sample. Estimate μ and σ. (*Hint*: Use Theorem 19.5.)
5. Derive the partial effects for the tobit model with heteroscedasticity that is described in Section 19.3.5.a.
6. Prove that the Hessian for the tobit model in (19-14) is negative definite after Olsen's transformation is applied to the parameters.

Applications

1. We examined Ray Fair's famous analysis (*Journal of Political Economy*, 1978) of a *Psychology Today* survey on extramarital affairs in Example 18.9 using a Poisson regression model. Although the dependent variable used in that study was a count, Fair (1978) used the tobit model as the platform for his study. You can reproduce the tobit estimates in Fair's paper easily with any software package that contains a tobit estimator—most do. The data appear in Appendix Table F18.1. Reproduce

Fair's least squares and tobit estimates. Compute the partial effects for the model and interpret all results.

2. Fair's original study also included but did not analyze a second data set that was a similar survey conducted by *Redbook* magazine. The data are reproduced in Appendix Table F17.2. (Our thanks to Ray Fair for providing these data.) This sample contains observations on 6,366 women and the following variables:

id = an identification number
C = constant, value = 1
yrb = a constructed measure of time spent in extramarital affairs
v_1 = a rating of the marriage, coded 1 to 4
v_2 = age, in years, aggregated
v_3 = number of years married
v_4 = number of children, top coded at 5
v_5 = religiosity, 1 to 4, 1 = not, 4 = very
v_6 = education, coded 9, 12, 14, 16, 17, 20
v_7 = occupation
v_8 = husband's occupation

Three other variables were not used. Details on the variables in the model are given in Fair's (1978) *Journal of Political Economy* paper. Using these data, conduct a parallel study to the *Psychology Today* study that was done in Fair (1978). Are the results consistent? Report all results, including partial effects and relevant diagnostic statistics.

3. Continuing the analysis of the previous application, note that these data conform precisely to the description of "corner solutions" in Section 19.3.4. The dependent variable is not censored in the fashion usually assumed for a tobit model. To investigate whether the dependent variable is determined by a two-part decision process (yes/no and, if yes, how much), specify and estimate a two-equation model in which the first equation analyzes the binary decision $A = 1$ if $yrb > 0$ and 0 otherwise and the second equation analyzes $yrb \mid yrb > 0$. What is the appropriate model? What do you find? Report all results. (*Note*: If you analyze the second dependent variable using the truncated regression, you should remove some extreme observations from your sample. The truncated regression estimator refuses to converge with the full data set but works nicely for the example if you omit observations with $yrb > 5$.)

4. **StochasticFrontier Model**. Section 10.5.1 presents estimates of a Cobb–Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Appendix Table F4.3. These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model discussed in Section 19.2.4. It has been suggested that one explanation for the apparent finding of economies of

scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in Stata, TSP, and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS, MatLab, or GAUSS. (*Note*: For a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function that appears in the log-likelihood.)