22

LIMITED DEPENDENT VARIABLE AND DURATION MODELS

22.1 INTRODUCTION

This chapter is concerned with truncation and censoring.[1] The effect of truncation occurs when sample data are drawn from a subset of a larger population of interest. 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 problem in 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.

This chapter will discuss four broad topics: truncation, censoring, a form of truncation called the **sample selection** problem, and a class of models called **duration models.** Although most empirical work on the first three 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 fourth topic, on models of duration—When will a spell of unemployment or a strike end?—could reasonably stand alone. It does in countless articles and a library of books.[2] We include our introduction to this subject in this chapter because in most applications, duration modeling involves censored data and it is thus convenient to treat duration here (and because we are nearing the end of our survey and yet another chapter seems unwarranted).

22.2 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.

[1] Five of the many surveys of these topics are Dhrymes (1984), Maddala (1977b, 1983, 1984), and Amemiya (1984). The last is part of a symposium on censored and truncated regression models. A survey that is oriented toward applications and techniques is Long (1997). Some recent results on non- and semiparametric estimation appear in Lee (1996).

[2] For example, Lancaster (1990) and Kiefer (1985).

22.2.1 TRUNCATED DISTRIBUTIONS

A **truncated distribution** is the part of an untruncated distribution that is above or below some specified value. For instance, in Example 22.2, 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 22.1 Density of a Truncated Random Variable
If a continuous random variable x has pdf $f(x)$ and a is a constant, then

$$f(x \mid x > a) = \frac{f(x)}{\text{Prob}(x > a)}.^{3}$$

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 22.1. Another truncated distribution which 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.[4]

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

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

[4] See Shaw (1988).

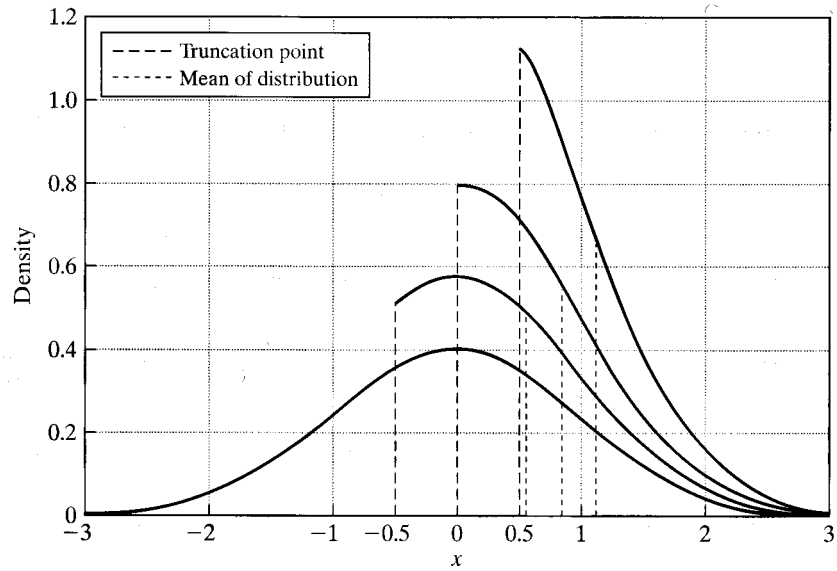


FIGURE 22.1 Truncated Normal Distributions.

22.2.2 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 22.1 Truncated Uniform Distribution

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

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

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

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

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,

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

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

Example 22.1 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. This is clearly visible in Figure 22.1.
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:[5]

THEOREM 22.2 Moments of the Truncated Normal Distribution
If $x \sim N[\mu, \sigma^2]$ *and* a *is a constant, then*

$$E[x \mid \text{truncation}] = \mu + \sigma\lambda(\alpha), \tag{22-1}$$

$$\text{Var}[x \mid \text{truncation}] = \sigma^2[1 - \delta(\alpha)], \tag{22-2}$$

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

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \quad \text{if truncation is } x > a, \tag{22-3a}$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \quad \text{if truncation is } x < a, \tag{22-3b}$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]. \tag{22-4}$$

An important result is

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

which implies point 2 after Example 22.1. 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 (22-3a) is also called the **hazard function** for the standard normal distribution.

Example 22.2 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."[6] 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." Since 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 in the population were lognormally distributed—see Section B.4.4. Then the log of income had a normal distribution with, say, mean μ and standard deviation σ. We'll deduce μ and σ then determine the population mean income. Let x = income

[5]Details may be found in Johnson, Kotz, and Balakrishnan (1994, pp. 156–158).

[6]*New York Post* (1987).

and let $y = \ln x$. Two useful numbers for this example are $\ln 100 = 4.605$ and $\ln 142 = 4.956$. Suppose that the survey was large enough for us to treat the sample average as the true mean. Then, the article stated that $E[y \mid y > 4.605] = 4.956$. It also told us that $\text{Prob}[y > 4.605] = 0.02$. From Theorem 22.2,

$$E[y \mid y > 4.605] = \mu + \frac{\sigma\phi(\alpha)}{1 - \Phi(\alpha)},$$

where $\alpha = (4.605 - \mu)/\sigma$. We also know that $\Phi(\alpha) = 0.98$, so $\alpha = \Phi^{-1}(0.98) = 2.054$. We infer, then, that (a) $2.054 = (4.605 - \mu)/\sigma$. In addition, given $\alpha = 2.054$, $\phi(\alpha) = \phi(2.054) = 0.0484$. From (22-1), then, $4.956 = \mu + \sigma(0.0484/0.02)$ or (b) $4.956 = \mu + 2.420\sigma$. The solutions to (a) and (b) are $\mu = 2.635$ and $\sigma = 0.959$.

To obtain the mean income, we now use the result that if $y \sim N[\mu, \sigma^2]$ and $x = e^y$, then $E[x] = E[e^y] = e^{\mu + \sigma^2/2}$. Inserting our values for μ and σ gives $E[x] = \$22{,}087$. The 1987 *Statistical Abstract of the United States* listed average household income 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.

22.2.3 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

$$y_i \mid \mathbf{x}_i \sim N[\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2]. \tag{22-5}$$

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 22.2. It follows that

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\frac{\phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}. \tag{22-6}$$

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

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

$$E[y_i \mid y_i > a] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda(\alpha_i), \tag{22-7}$$

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

$$\begin{aligned}\frac{\partial E[y_i \mid y_i > a]}{\partial \mathbf{x}_i} &= \boldsymbol{\beta} + \sigma(d\lambda_i/d\alpha_i)\frac{\partial \alpha_i}{\partial \mathbf{x}_i} \\ &= \boldsymbol{\beta} + \sigma\left(\lambda_i^2 - \alpha_i\lambda_i\right)(-\boldsymbol{\beta}/\sigma) \\ &= \boldsymbol{\beta}\left(1 - \lambda_i^2 + \alpha_i\lambda_i\right) \\ &= \boldsymbol{\beta}(1 - \delta_i).\end{aligned} \tag{22-8}$$

Note the appearance of the truncated variance. Since the truncated variance 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

$$\text{Var}[y_i \mid y_i > a] = \sigma^2(1 - \delta_i). \tag{22-9}$$

Whether the marginal effect in (22-7) 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 (22-7) 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 (22-6) in the form

$$y_i \mid y_i > a = E[y_i \mid y_i > a] + u_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i + u_i, \tag{22-10}$$

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 (22-10) 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.[7]

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 Cheung 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 consistent maximum likelihood estimates.[8] 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 22.4.)

22.3 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)],
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)],

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

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

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

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.

22.3.1 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 22.2 illustrates the effects.

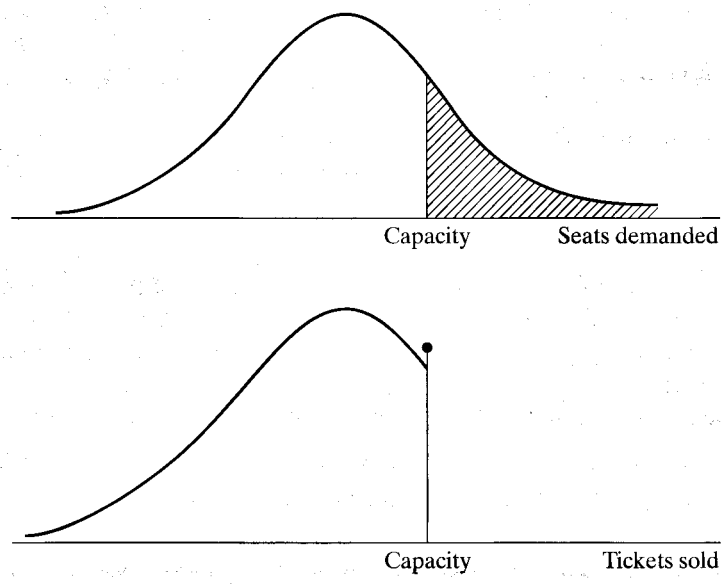
To analyze this distribution, we define a new random variable y transformed from the original one, y^*, by

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

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.

FIGURE 22.2 Partially Censored Distribution.



THEOREM 22.3 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 22.2. For the variance, we use a counterpart to the decomposition in (B-70), that is, Var[y] = E[conditional variance] + Var[conditional mean], *and Theorem 22.2.*

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 22.2.

Example 22.3 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 22.3, the 18,000 is an estimate of

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

Since 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 $\mu = 18{,}772$.

22.3.2 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,

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \\ y_i &= 0 \quad \text{if } y_i^* \le 0, \\ y_i &= y_i^* \quad \text{if } y_i^* > 0. \end{aligned} \tag{22-11}$$

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*, $E[y_i^* | \mathbf{x}_i]$ is $\mathbf{x}_i'\boldsymbol{\beta}$. If the data are always censored, however, then this result will usually not be useful. Consistent with Theorem 22.3, 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

$$\lambda_i = \frac{\phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(0 - \mathbf{x}_i'\boldsymbol{\beta})/\sigma]} = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}. \tag{22-12}$$

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 22.3, 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 marginal 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, since 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 22.4 Marginal 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*

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

Since $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma[(y^* - \boldsymbol{\beta}'\mathbf{x})/\sigma]$, *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,*

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

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,*

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

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

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

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

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

where $\alpha_i = \mathbf{x}_i'\boldsymbol{\beta}$, $\Phi_i = \Phi(\alpha_i)$ 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.

Example 22.4 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 22.1. 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 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.

22.3.3 ESTIMATION

Estimation of this model is very similar to that of truncated regression. 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

TABLE 22.1 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			

ordinary linear regression.[11] The log-likelihood for the censored regression model is

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right] + \sum_{y_i=0} \ln\left[1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \quad \textbf{(22-13)}$$

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, since 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 $\log L$ 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

$$\ln L = \sum_{y_i>0} -\frac{1}{2}[\ln(2\pi) - \ln\theta^2 + (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2] + \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma})]. \quad \textbf{(22-14)}$$

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 $\boldsymbol{\beta} = \boldsymbol{\gamma}/\theta$. The asymptotic covariance matrix for these estimates can be obtained from that for the estimates of $[\boldsymbol{\gamma}, \theta]$ using Est.Asy. Var$[\hat{\boldsymbol{\beta}}, \hat{\sigma}] =$ $\hat{\mathbf{J}}$ Asy. Var$[\hat{\boldsymbol{\gamma}}, \hat{\theta}]\hat{\mathbf{J}}'$, where

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

[11] See Hall (1984).

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.[12] The effect is illustrated in the last two columns of Table 22.1. Another strategy is to discard the limit observations, but we now see that just trades the censoring problem for the truncation problem.

22.3.4 SOME ISSUES IN SPECIFICATION

Two issues that commonly arise in microeconomic data, heteroscedasticity and nonnormality, have been analyzed at length in the tobit setting.[13]

22.3.4.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 22.5 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 OPTIONs. They report the results listed in Table 22.2 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.

---

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

[13] 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).

TABLE 22.2 Estimates of a Tobit Model (Standard errors in parentheses)

	Homoscedastic	*Heteroscedastic*	
	β	β	α
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)
Log L	−547.30	−466.27	
Sample size	200	200	

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

$$\sigma_i^2 = \sigma^2 e^{\boldsymbol{\alpha}'\mathbf{w}_i}. \tag{22-15}$$

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 19.4.1.b and in the linear regression model in Section 17.7.1.) 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,

$$\begin{aligned}
a_i &= z_i\left(\frac{\varepsilon_i}{\sigma^2}\right) + (1 - z_i)\left(\frac{(-1)\lambda_i}{\sigma}\right),\\
b_i &= z_i\left(\frac{(\varepsilon_i^2/\sigma^2 - 1)}{2\sigma^2}\right) + (1 - z_i)\left(\frac{(\mathbf{x}_i'\boldsymbol{\beta})\lambda_i}{2\sigma^3}\right),\\
\lambda_i &= \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1 - \boldsymbol{\Phi}_i(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}.
\end{aligned} \tag{22-16}$$

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 ($\varepsilon, \phi, \mathbf{x}_i'\boldsymbol{\beta}, \lambda_i$, etc.) are evaluated at the restricted (homoscedastic) maximum likelihood estimates. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}[\mathbf{G}'\mathbf{G}]^{-1}\mathbf{G}'\mathbf{i} = nR^2 \tag{22-17}$$

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 11.4.3.) 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 $\mathbf{w}_i$.

22.3.4.b Misspecification of Prob[$y^* < 0$]

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 on whether or not to purchase a car as being different from the decision on how much to spend on the car, having decided to buy one. A related problem raised by Fin and Schmidt (1984) is that in the tobit model, a variable that increases the probability of an observation being a nonlimit observation also increases the mean of the variable. They cite as an example loss due to fire in buildings. Older buildings might be more likely to have fires, so that $\partial \text{ Prob}[y_i > 0]/\partial \text{ age}_i > 0$, but, because of the greater value of newer buildings, older ones incur smaller losses when they do have fires, so that $\partial E[y_i \mid y_i > 0]/\partial \text{ age}_i < 0$. 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.

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

1. Decision equation:

$$\begin{aligned} \text{Prob}[y_i^* > 0] &= \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), \qquad z_i = 1 \text{ if } y_i^* > 0, \\ \text{Prob}[y_i^* \leq 0] &= 1 - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}), \quad z_i = 0 \text{ if } y_i^* \leq 0. \end{aligned} \tag{22-18}$$

2. Regression equation for nonlimit observations:

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

according to Theorem 22.2.

This model is a combination of the truncated regression model of Section 22.2 and the univariate probit model of Section 21.3, which suggests a method of analyzing it. The tobit model of this section arises if $\boldsymbol{\gamma}$ equals $\boldsymbol{\beta}/\sigma$. The parameters of the regression equation can be estimated independently using the truncated regression model of Section 22.2. A recent application is Melenberg and van Soest (1996).

Fin 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 $\sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$ in (22-13). This produces the log-likelihood for the truncated regression model plus (21-20) for the probit model.[14]] Therefore, a likelihood ratio statistic can be computed using

$$\lambda = -2[\ln L_T - (\ln L_P + \ln L_{TR})],$$

where

L_T = likelihood for the tobit model in (22-13), with the same coefficients,

L_P = likelihood for the probit model in (19-20), fit separately,

L_{TR} = likelihood for the truncated regression model, fit separately.

[14]The likelihood function for the truncated regression model is considered in the exercises.

22.3.4.c 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 (22-13) is inconsistent. Research is ongoing both on alternative estimators and on methods for testing for this type of misspecification.[15]

One approach to the estimation is to use an alternative distribution. Kalbfleisch and Prentice (1980) 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.[16] 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)]. The conditional moment tests are described in the next section.

To employ a Hausman test, we require an estimator that is consistent and efficient under the null hypothesis but inconsistent under the alternative—the tobit estimator with normality—and an estimator that is consistent under both hypotheses but inefficient under the null hypothesis. Thus, we will require a robust estimator of $\boldsymbol{\beta}$, which restores the difficulties of the previous paragraph. Recent applications [e.g., Melenberg and van Soest (1996)] have used the Hausman test to compare the tobit/normal estimator with Powell's consistent, but inefficient (robust), LAD estimator. Another approach to testing is to embed the normal distribution in some other distribution and then use an LM test for the normal specification. Chesher and Irish (1987) have devised an LM test of normality in the tobit model based on **generalized residuals.** In many models, including the tobit model, the generalized residuals can be computed as the derivatives of the log-densities with respect to the constant term, so

$$e_i = \frac{1}{\sigma^2}[z_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) - (1 - z_i)\sigma\lambda_i],$$

where z_i is defined in (22-18) and λ_i is defined in (22-16). This residual is an estimate of ε_i that accounts for the censoring in the distribution. By construction, $E[e_i \mid \mathbf{x}_i] = 0$, and if the model actually does contain a constant term, then $\sum_{i=1}^{n} e_i = 0$; this is the first of the necessary conditions for the MLE. The test is then carried out by regressing a column of 1s on $\mathbf{d}_i = [e_i\mathbf{x}_i', b_i, e_i^3, e_i^4 - 3e_i^4]'$, where b_i is defined in (22-16). Note that the first $K + 1$ variables in $\mathbf{d}_i$ are the derivatives of the tobit log-likelihood. Let $\mathbf{D}$ be the $n \times (K+3)$ matrix with ith row equal to $\mathbf{d}_i'$. Then $\mathbf{D} = [\mathbf{G}, \mathbf{M}]$, where the $K+1$ columns

[15] See Duncan (1983, 1986b), Goldberger (1983), Pagan and Vella (1989), Lee (1996), and Fernandez (1986). We will examine one of the tests more closely in the following section.

[16] 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).

of $\mathbf{G}$ are the derivatives of the tobit log-likelihood and the two columns in $\mathbf{M}$ are the last two variables in $\mathbf{a}_i$. Then the chi-squared statistic is nR^2; that is,

$$\text{LM} = \mathbf{i}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{i}.$$

The necessary conditions that define the MLE are $\mathbf{i}'\mathbf{G} = \mathbf{0}$, so the first $K + 1$ elements of $\mathbf{i}'\mathbf{D}$ are zero. Using (B-66), then, the LM statistic becomes

$$\text{LM} = \mathbf{i}'\mathbf{M}[\mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{M}]^{-1}\mathbf{M}'\mathbf{i},$$

which is a chi-squared statistic with two degrees of freedom. Note the similarity to (22-17), where a test for homoscedasticity is carried out by the same method. As emerges so often in this framework, the test of the distribution actually focuses on the skewness and kurtosis of the residuals.

22.3.4.d Conditional Moment Tests

Pagan and Vella (1989) [see, as well, Ruud (1984)] describe a set of conditional moment tests of the specification of the tobit model.[17] We will consider three:

1. The variables $\mathbf{z}$ have not been erroneously omitted from the model.
2. The disturbances in the model are homoscedastic.
3. The underlying disturbances in the model are normally distributed.

For the third of these, we will take the standard approach of examining the third and fourth moments, which for the normal distribution are 0 and $3\sigma^4$, respectively. The underlying motivation for the tests can be made with reference to the regression part of the tobit model in (22-11),

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

Neglecting for the moment that we only observe y_i^* subject to the censoring, the three hypotheses imply the following expectations:

1. $E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$,
2. $E\{\mathbf{z}_i[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2]\} = \mathbf{0}$,
3. $E[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^3] = 0$ and $E[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^4 - 3\sigma^4] = 0$.

In (1), the variables in $\mathbf{z}_i$ would be one or more variables not already in the model. We are interested in assessing whether or not they should be. In (2), presumably, although not necessarily, $\mathbf{z}_i$ would be the regressors in the model. For the present, we will assume that y_i^* is observed directly, without censoring. That is, we will construct the CM tests for the classical linear regression model. Then we will go back to the necessary step and make the modification needed to account for the censoring of the dependent variable.

[17] Their survey is quite general and includes other models, specifications, and estimation methods. We will consider only the simplest cases here. The reader is referred to their paper for formal presentation of these results.

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 and Vuong (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).

Conditional moment tests are described in Section 17.6.4. To review, for a model estimated by maximum likelihood, the statistic is

$$C = \mathbf{i}'\mathbf{M}[\mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{M}]^{-1}\mathbf{M}'\mathbf{i},$$

where the rows of $\mathbf{G}$ are the terms in the gradient of the log-likelihood function, $(\mathbf{G}'\mathbf{G})^{-1}$ is the BHHH estimator of the asymptotic covariance matrix of the MLE of the model parameters, and the rows of $\mathbf{M}$ are the individual terms in the sample moment conditions. Note that this construction is the same as the LM statistic just discussed. The difference is in how the rows of $\mathbf{M}$ are constructed.

For a regression model without censoring, the sample counterparts to the moment restrictions in (1) to (3) would be

$$\mathbf{r}_1 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_i, \qquad \text{where } e_i = y_i - \mathbf{x}_i'\mathbf{b} \text{ and } \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\mathbf{r}_2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\left(e_i^2 - s^2\right), \quad \text{where } s^2 = \frac{\mathbf{e}'\mathbf{e}}{n},$$

$$\mathbf{r}_3 = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} e_i^3 \\ e_i^4 - 3s^4 \end{bmatrix}.$$

For the positive observations, we observe y^*, so the observations in $\mathbf{M}$ are the same as for the classical regression model; that is,

1. $\mathbf{m}_i = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$,
2. $\mathbf{m}_i = \mathbf{z}_i[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2]$,
3. $\mathbf{m}_i = [(y_i - \mathbf{x}_i'\boldsymbol{\beta})^3, (y_i - \mathbf{x}_i'\boldsymbol{\beta})^4 - 3\sigma^4]'$.

For the limit observations, these observations are replaced with their expected values, conditioned on $y = 0$, which means that $y^* \leq 0$ or $e_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$. Let $q_i = (\mathbf{x}_i'\boldsymbol{\beta})/\sigma$ and $\lambda_i = \phi_i/(1 - \Phi_i)$. Then from (22-2), (22-3b), and (22-4),

1. $\mathbf{m}_i = \mathbf{z}_i E[(y_i^* - \mathbf{x}_i'\boldsymbol{\beta}) \mid y = 0] = \mathbf{z}_i[(\mathbf{x}_i'\boldsymbol{\beta} - \sigma\lambda_i) - \mathbf{x}_i'\boldsymbol{\beta}] = \mathbf{z}_i(2\sigma\lambda_i)$.
2. $\mathbf{m}_i = \mathbf{z}_i E[(y_i^* - \mathbf{x}_i'\boldsymbol{\beta})^2 - \sigma^2 \mid y = 0] = \mathbf{z}_i[\sigma^2(1 + q_i\lambda_i) - \sigma^2] = \mathbf{z}_i(\sigma^2 q_i\lambda_i)$.

$E[\varepsilon_i^2 \mid y = 0, \mathbf{x}_i]$ is not the variance, since the mean is not zero.) For the third and fourth moments, we simply reproduce Pagan and Vella's results [see also Greene (1995a, pp. 618–619)]:

3. $\mathbf{m}_i = \sigma^3\lambda_i\left[-\left(2 + q_i^2\right), \sigma q_i\left(3 + q_i^2\right)\right]'$.

These three items are the remaining terms needed to compute $\mathbf{M}$.

22.3.5 CENSORING AND TRUNCATION IN MODELS FOR COUNTS

Truncation and censoring are relatively common in applications of models for counts (see Section 21.9). Truncation often arises as a consequence of discarding what appear to be unusable data, such as the zero values in survey data on the number of uses of recreation facilities [Shaw (1988) and Bockstael et al. (1990)]. The zero values in this setting might represent a discrete decision not to visit the site, which is a qualitatively different decision from the positive number for someone who had decided to make at

least one visit. In such a case, it might make sense to confine attention to the nonzero observations, thereby truncating the distribution. Censoring, in contrast, is often employed to make survey data more convenient to gather and analyze. For example, survey data on access to medical facilities might ask, "How many trips to the doctor did you make in the last year?" The responses might be 0, 1, 2, 3 or more.

Models with these characteristics can be handled within the Poisson and negative binomial regression frameworks by using the laws of probability to modify the likelihood. For example, in the *censored* data case,

$$P_i(j) = \text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{j!}, \quad j = 0, 1, 2$$

$$P_i(3) = \text{Prob}[y_i \geq 3] = 1 - [\text{Prob}(y_i = 0) + \text{Prob}(y_i = 1) + \text{Prob}(y_i = 2)].$$

The probabilities in the model with *truncation* above zero would be

$$P_i(j) = \text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{[1 - P_i(0)]j!} = \frac{e^{-\lambda_i}\lambda_i^j}{[1 - e^{-\lambda_i}]j!}, \quad j = 1, 2, \ldots.$$

These models are not appreciably more complicated to analyze than the basic Poisson or negative binomial models. [See Terza (1985b), Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1998), Lambert (1992), and Winkelmann (1997).] They do, however, bring substantive changes to the familiar characteristics of the models. For example, the conditional means are no longer λ_i; in the censoring case,

$$E[y_i \mid \mathbf{x}_i] = \lambda_i - \sum_{j=3}^{\infty}(j-3)P_i(j) < \lambda_i.$$

Marginal effects are changed as well. Recall that our earlier result for the **count data** models was $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i = \lambda_i \boldsymbol{\beta}$. With censoring or truncation, it is straightforward in general to show that $\partial E[y_i \mid \mathbf{x}_i]/\partial \mathbf{x}_i = \delta_i \boldsymbol{\beta}$, but the new scale factor need not be smaller than λ_i.

22.3.6 APPLICATION: CENSORING IN THE TOBIT AND POISSON REGRESSION MODELS

In 1969, the popular magazine *Psychology Today* published a 101-question survey on sex and asked its readers to mail in their answers. The results of the survey were discussed in the July 1970 issue. From the approximately 2,000 replies that were collected in electronic form (of about 20,000 received), Professor Ray Fair (1978) extracted a sample of 601 observations on men and women then currently married for the first time and analyzed their responses to a question about extramarital affairs. He used the tobit model as a platform. Fair's analysis in this frequently cited study suggests several interesting econometric questions. [In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model contributed to the development of the EM algorithm, which was published by and is usually associated with Dempster, Laird, and Rubin (1977).]

As noted, Fair used the tobit model as his estimation framework for this study. The nonexperimental nature of the data (which can be downloaded from the Internet at http://fairmodel.econ.yale.edu/rayfair/work.ss.htm) provides a fine laboratory case that

we can use to examine the relationships among the tobit, truncated regression, and probit models. In addition, as we will explore below, although the tobit model seems to be a natural choice for the model for these data, a closer look suggests that the models for counts we have examined at several points earlier might be yet a better choice. Finally, the preponderance of zeros in the data that initially motivated the tobit model suggests that even the standard Poisson model, although an improvement, might still be inadequate. In this example, we will reestimate Fair's original model and then apply some of the specification tests and modified models for count data as alternatives.

The study was based on 601 observations on the following variables (full details on data coding are given in the data file and Appendix Table F22.2):

y = number of affairs in the past year, 0, 1, 2, 3, 4–10 coded as 7, "monthly, weekly, or daily," coded as 12. Sample mean = 1.46. Frequencies = (451, 34, 17, 19, 42, 38).

z_1 = sex = 0 for female, 1 for male. Sample mean = 0.476.

z_2 = age. Sample mean = 32.5.

z_3 = number of years married. Sample mean = 8.18.

z_4 = children, 0 = no, 1 = yes. Sample mean = 0.715.

z_5 = religiousness, 1 = anti, ..., 5 = very. Sample mean = 3.12.

z_6 = education, years, 9 = grade school, 12 = high school, ..., 20 = Ph.D or other. Sample mean = 16.2.

z_7 = occupation, "Hollingshead scale," 1–7. Sample mean = 4.19.

z_8 = self-rating of marriage, 1 = very unhappy, ..., 5 = very happy. Sample mean = 3.93.

The tobit model was fit to y using a constant term and all eight variables. A restricted model was fit by excluding z_1, z_4, and z_6, none of which was individually statistically significant in the model. We are able to match exactly Fair's results for both equations. The log-likelihood functions for the full and restricted models are 2704.7311 and 2705.5762. The chi-squared statistic for testing the hypothesis that the three coefficients are zero is twice the difference, 1.6902. The critical value from the chi-squared distribution with three degrees of freedom is 7.81, so the hypothesis that the coefficients on these three variables are all zero is not rejected. The Wald and Lagrange multiplier statistics are likewise small, 6.59 and 1.681. Based on these results, we will continue the analysis using the restricted set of variables, $\mathbf{Z} = (\mathbf{1}, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_5, \mathbf{z}_7, \mathbf{z}_8)$. Our interest is solely in the numerical results of different modeling approaches. Readers may draw their own conclusions and interpretations from the estimates.

Table 22.3 presents parameter estimates based on Fair's specification of the normal distribution. The inconsistent least squares estimates appear at the left as a basis for comparison. The maximum likelihood tobit estimates appear next. The sample is heavily dominated by observations with $y = 0$ (451 of 601, or 75 percent), so the marginal effects are very different from the coefficients, by a multiple of roughly 0.766. The scale factor is computed using the results of Theorem 22.4 for left censoring at zero and the upper limit of $+\infty$, with all variables evaluated at the sample means and the parameters equal

TABLE 22.3 Model Estimates Based on the Normal Distribution (Standard Errors in Parentheses)

		Tobit			*Probit*	*Truncated Regression*	
Variable	*Least Squares (1)*	*Estimate (2)*	*Marginal Effect (3)*	*Scaled by $1/\sigma$ (4)*	*Estimate (5)*	*Estimate (6)*	*Marginal Effect (7)*
Constant	5.61	8.18	—	0.991	0.997	8.32	—
	(0.797)	(2.74)	—	(0.336)	(0.361)	(3.96)	—
z_2	−0.0504	−0.179	−0.042	−0.022	−0.022	−0.0841	−0.0407
	(0.0221)	(0.079)	(0.184)	(0.010)	(0.102)	(0.119)	(0.0578)
z_3	0.162	0.554	0.130	0.0672	0.0599	0.560	0.271
	(0.0369)	(0.135)	(0.0312)	(0.0161)	(0.0171)	(0.219)	(0.106)
z_5	−0.476	−1.69	−0.394	−0.2004	−0.184	−1.502	−0.728
	(0.111)	(0.404)	(0.093)	(0.484)	(0.0515)	(0.617)	(0.299)
z_7	0.106	0.326	0.0762	0.0395	0.0375	0.189	0.0916
	(0.0711)	(0.254)	(0.0595)	(0.0308)	(0.0328)	(0.377)	(0.182)
z_8	−0.712	−2.29	−0.534	−0.277	−0.273	−1.35	−0.653
	(0.118)	(0.408)	(0.0949)	(0.0483)	(0.0525)	(0.565)	(0.273)
σ	3.09	8.25				5.53	
log L			−705.5762		−307.2955	−329.7103	

to the maximum likelihood estimates:

$$\text{scale} = \Phi\left[\frac{+\infty - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\text{ML}}}{\hat{\sigma}_{\text{ML}}}\right] - \Phi\left[\frac{0 - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\text{ML}}}{\hat{\sigma}_{\text{ML}}}\right] = 1 - \Phi\left[\frac{0 - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\text{ML}}}{\hat{\sigma}_{\text{ML}}}\right] = \Phi\left[\frac{\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\text{ML}}}{\hat{\sigma}_{\text{ML}}}\right] = 0.234.$$

These estimates are shown in the third column. As expected, they resemble the least squares estimates, although not enough that one would be content to use OLS for estimation. The fifth column in Table 22.3 gives estimates of the probit model estimated for the dependent variable $q_i = 0$ if $y_i = 0$, $q_i = 1$ if $y_i > 0$. If the specification of the tobit model is correct, then the probit estimators should be consistent for $(1/\sigma)\boldsymbol{\beta}$ from the tobit model. These estimates, with standard errors computed using the **delta method,** are shown in column 4. The results are surprisingly close, especially given the results of the specification test considered later. Finally, columns 6 and 7 give the estimates for the truncated regression model that applies to the 150 nonlimit observations if the specification of the model is correct. Here the results seem a bit less consistent.

Several specification tests were suggested for this model. The Cragg/Greene test for appropriate specification of Prob$[y_i = 0]$ is given in Section 22.3.4.b. This test is easily carried out using the log-likelihood values in the table. The chi-squared statistic, which has seven degrees of freedom is $-2\{-705.5762 - [-307.2955 + (-392.7103)]\} = 11.141$, which is smaller than the critical value of 14.067. We conclude that the tobit model is correctly specified (the decision of whether or not is not different from the decision of how many, given "whether"). We now turn to the normality tests. We emphasize that these tests are nonconstructive tests of the skewness and kurtosis of the distribution of ε. A fortiori, if we do reject the hypothesis that these values are 0.0 and 3.0, respectively, then we can reject normality. But that does not suggest what to do next. We turn to that issue later. The Chesher–Irish and Pagan–Vella chi-squared statistics are 562.218 and 22.314, respectively. The critical value is 5.99, so on the basis of both of these

values, the hypothesis of normality is rejected. Thus, both the probability model and the distributional framework are rejected by these tests.

Before leaving the tobit model, we consider one additional aspect of the original specification. The values above 4 in the observed data are not true observations on the response; 7 is an estimate of the mean of observations that fall in the range 4 to 10, whereas 12 was chosen more or less arbitrarily for observations that were greater than 10. These observations represent 80 of the 601 observations, or about 13 percent of the sample. To some extent, this coding scheme might be driving the results. [This point was not overlooked in the original study; "[a] linear specification was used for the estimated equation, and it did not seem reasonable in this case, given the range of explanatory variables, to have a dependent variable that ranged from, say, 0 to 365" [Fair (1978), p. 55]. The tobit model allows for censoring in both tails of the distribution. Ignoring the results of the specification tests for the moment, we will examine a doubly censored regression by recoding all observations that take the values 7, or 12 as 4. The model is thus

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \\ y &= 0 \quad \text{if } y^* \le 0, \\ y &= y^* \quad \text{if } 0 < y^* < 4, \\ y &= 4 \quad \text{if } y^* \ge 4. \end{aligned}$$

The log-likelihood is built up from three sets of terms:

$$\ln L = \sum_{y=0} \ln \Phi\left[\frac{0 - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right] + \sum_{0<y<4} \ln \frac{1}{\sigma}\phi\left[\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right] + \sum_{y=4} \ln\left[1 - \Phi\left(\frac{4 - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right].$$

Maximum likelihood estimates of the parameters of this model based on the doubly censored data appear in Table 22.4. The effect on the coefficient estimates is relatively minor, but the effect on the estimates of the marginal effects is very large; they are reduced by about 50 percent, which makes sense. With the original data, increases in the index were associated with increases in y that could be from 3 to 7 or from 3 to 12. But with the data treated as censored, y cannot increase past 4. Thus, the range of variation is greatly reduced. The numerical results are also suggestive. Recall that the scale factor for the singly censored data was 0.2338. For the doubly censored variable, this factor is $\Phi[(4 - \boldsymbol{\beta}'\mathbf{x})/\sigma] - \Phi[(0 - \boldsymbol{\beta}'\mathbf{x})/\sigma] = 0.8930 - 0.7701 = 0.1229$. The regression model

TABLE 22.4 Estimates of a Doubly Censored Tobit Model

	Left Censored at 0 Only			*Censored at Both 0 and 4*		
Variable	*Estimate*	*Standard Error*	*Marginal Effect*	*Estimate*	*Standard Error*	*Marginal Effect*
Constant	8.18	0.797	—	7.90	2.804	—
z_2	−0.179	0.079	−0.0420	−0.178	0.080	−0.0218
z_3	0.554	0.135	0.130	0.532	0.141	0.0654
z_5	−1.69	0.404	−0.394	−1.62	0.424	−0.199
z_7	0.326	0.254	0.0762	0.324	0.254	0.0399
z_8	−2.29	0.408	−0.534	−2.21	0.459	−0.271
σ	8.25 Prob(nonlimit) = 0.2338			7.94 Prob(nonlimit) = 0.1229		
$E[y \mid \mathbf{x} = E[\mathbf{x}]]$	1.126			0.226		

for y^* has not changed much, but the effect now is to assign the upper tail area to the censored region, whereas before it was in the uncensored region. The effect, then, is to reduce the scale roughly by this 0.107, from 0.234 to about 0.123.

By construction, the tobit model should only be viewed as an approximation for these data. The dependent variable is a count, not a continuous measurement. (Thus, the testing results obtained earlier are not surprising.) The Poisson regression model, or perhaps one of the many variants of it, should be a preferable modeling framework. Table 22.5 presents estimates of the Poisson and negative binomial regression model. There is ample evidence of overdispersion in these data; the t-ratio on the estimated overdispersion parameter is $7.014/0.945 = 7.42$, which is strongly suggestive. The large absolute value of the coefficient is likewise suggestive.

Before proceeding to a model that specifically accounts for overdispersion, we can find a candidate for its source, at least to some degree. As discussed earlier, responses of 7 and 12 do not represent the actual counts. It is unclear what the effect of the first recoding would be, since it might well be the mean of the observations in this group. But the second is clearly a censored observation. To remove both of these effects, we have recoded both the values 7 and 12 as 4 and treated this observation (appropriately) as a censored observation, with 4 denoting "4 or more." As shown in the third and fourth sets of results in Table 22.5, the effect of this treatment of the data is greatly to reduce the measured effects, which is the same effect we observed for the tobit model. Although this step does remove a deficiency in the data, it does not remove the overdispersion; at this point, the negative binomial model is still the preferred specification.

The tobit model remains the standard approach to modeling a dependent variable that displays a large cluster of limit values, usually zeros. But in these data, it is clear that

TABLE 22.5 Model Estimates Based on the Poisson Distribution

	Poisson Regression			*Negative Binomial Regression*		
Variable	*Estimate*	*Standard Error*	*Marginal Effect*	*Estimate*	*Standard Error*	*Marginal Effect*
			Based on Uncensored Poisson Distribution			
Constant	2.53	0.197	—	2.19	0.859	—
z_2	−0.0322	0.00585	−0.0470	−0.0262	0.0180	−0.00393
z_3	0.116	0.00991	0.168	0.0848	0.0400	0.127
z_5	−0.354	0.0309	−0.515	−0.422	0.171	−0.632
z_7	0.0798	0.0194	0.116	0.0604	0.0908	0.0906
z_8	−0.409	0.0274	−0.0596	−0.431	0.167	−0.646
α				7.01	0.945	
log L	−1427.037			−728.2441		
			Based on Poisson Distribution Right Censored at y = 4			
Constant	1.90	0.283	—	4.79	1.16	—
z_2	−0.0328	0.00838	−0.0235	−0.0166	0.0250	−0.00428
z_3	0.105	0.0140	0.0754	0.174	0.0568	0.045
z_5	−0.323	0.0437	−0.232	−0.723	0.198	−0.186
z_7	0.0798	0.0275	0.0521	0.0900	0.116	0.0232
z_8	−0.390	0.0391	−0.279	−0.854	0.216	−0.220
α				9.39	1.36	
log L	−747.7541			−482.0505		

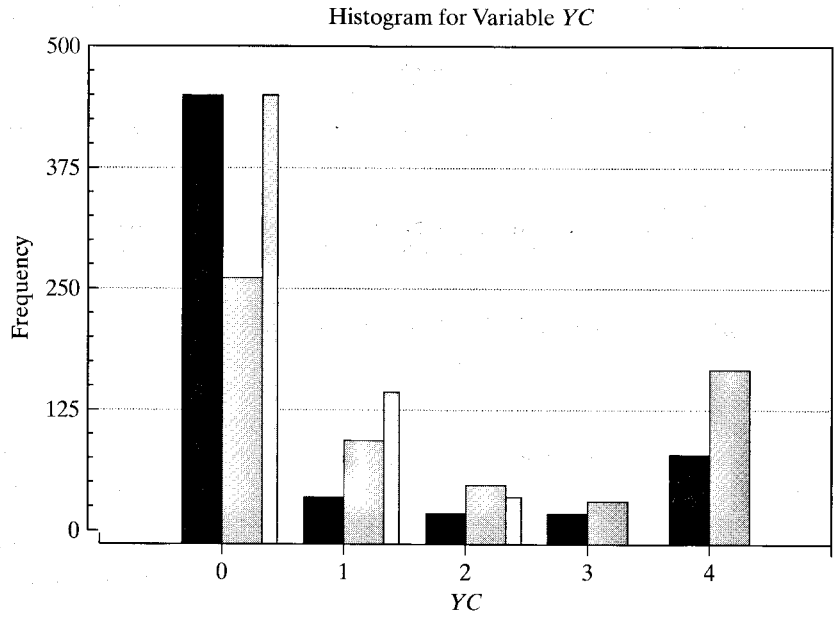


FIGURE 22.3 Histogram for Model Predictions.

the zero value represents something other than a censoring; it is the outcome of a discrete decision. Thus, for this reason and based on the preceding results, it seems appropriate to turn to a different model for this dependent variable. The Poisson and negative binomial models look like an improvement, but there remains a noteworthy problem. Figure 22.3 shows a histogram of the actual values (solid dark bars) and predicted values from the negative binomial model estimated with the censored data (lighter bars). Predictions from the latter model are the integer values of $E[y \mid \mathbf{x}] = \exp(\boldsymbol{\beta}'\mathbf{x})$. As in the actual data, values larger than 4 are censored to 4. Evidently, the negative binomial model predicts the data fairly poorly. In fact, it is not hard to see why. The source of the overdispersion in the data is not the extreme values on the right of the distribution; it is the very large number of zeros on the left.

There are a large variety of models and permutations that one might turn to at this point. We will conclude with just one of these, Lambert's (1992) zero-inflated Poisson (ZIP) model with a logit "splitting" model discussed in Section 21.9.6 and Example 21.12. The doubly censored count is the dependent variable in this model. (Mullahy's (1986) hurdle model is an alternative that might be considered. The difference between these two is in the interpretation of the zero observations. In the ZIP formulation, the zero observations would be a mixture of "never" and "not in the last year," whereas the hurdle model assumes two distinct decisions, "whether or not" and "how many, given yes.") The estimates of the parameters of the ZIP model are shown in Table 22.6. The Vuong statistic of 21.64 strongly supports the ZIP model over the Poisson model. (An attempt to combine the ZIP model with the negative binomial was unsuccessful. Since, as expected, the ancillary model for the zeros accounted for the overdispersion in the data, the negative binomial model degenerated to the Poisson form.) Finally,

TABLE 22.6 Estimates of a Zero-Inflated Poisson Model

	Poisson Regression		*Logit Splitting Model*		*Marginal Effects*		
Variable	*1.1Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*ZIP*	*Tobit (0)*	*Tobit (0, 4)*
Constant	1.27	0.439	−1.85	0.664	—	—	
Age	−0.00422	0.0122	0.0397	0.0190	−0.0252	−0.0420	−0.0218
Years	0.0331	0.0231	−0.0981	0.0318	0.0987	0.130	0.0654
Religion	−0.0909	0.0721	0.306	0.0951	−0.288	−0.394	−0.199
Occupation	0.0205	0.0441	0.0677	0.0607	0.0644	0.0762	0.0399
Happiness	−0.817	0.0666	0.458	0.0949	−0.344	−0.534	−0.271

the marginal effects, $\delta = \partial E[y \mid \mathbf{x}]/\partial \mathbf{x}$, are shown in Table 22.6 for three models: the ZIP model, Fair's original tobit model, and the tobit model estimated with the doubly censored count. The estimates for the ZIP model are considerably lower than those for Fair's tobit model. When the tobit model is reestimated with the censoring on the right, however, the resulting marginal effects are reasonably close to those from the ZIP model, though uniformly smaller. (This result may be from not building the censoring into the ZIP model, a refinement that would be relatively straightforward.)

We conclude that the original tobit model provided only a fair approximation to the marginal effects produced by (we contend) the more appropriate specification of the Poisson model. But the approximation became much better when the data were recorded and treated as censored. Figure 22.3 also shows the predictions from the ZIP model (narrow bars). As might be expected, it provides a much better prediction of the dependent variable. (The integer values of the conditional mean function for the tobit model were roughly evenly split between zeros and ones, whereas the doubly censored model always predicted $y = 0$.) Surprisingly, the treatment of the highest observations does greatly affect the outcome. If the ZIP model is fit to the original uncensored data, then the vector of marginal effects is $\delta = [-0.0586, 0.2446, -0.692, 0.115, -0.787]$, which is extremely large. Thus, perhaps more analysis is called for—the ZIP model can be further improved, and one might reconsider the hurdle model—but we have tortured Fair's data enough. Further exploration is left for the reader.

22.4 THE SAMPLE SELECTION MODEL

The topic of sample selection, or **incidental truncation,** has been the subject of an enormous recent literature, both theoretical and applied.[18] This analysis combines both of the previous topics.

Example 22.6 Incidental Truncation

In the high-income survey discussed in Example 22.2, respondents were also included in the survey if their net worth, not including their homes, was at least $500,000. Suppose that

[18] A large proportion of the analysis in this framework has been in the area of labor economics. 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"). The four surveys noted in the introduction to this chapter provide fairly extensive, although far from exhaustive, lists of the studies. Some studies that comment on methodological issues are Heckman (1990), Manski (1989, 1990, 1992), and Newey, Powell, and Walker (1990).

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

22.4.1 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. Since it has dominated the empirical literature, we will focus first on the bivariate normal distribution.[19]

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 22.5.[20]

THEOREM 22.5 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

$$E[y \mid z > a] = \mu_y + \rho\sigma_y\lambda(\alpha_z),$$
$$\text{Var}[y \mid z > a] = \sigma_y^2[1 - \rho^2\delta(\alpha_z)], \tag{22-19}$$

where

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

[19] We will reconsider the issue of the normality assumption in Section 22.4.5.

[20] 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).

Note that the expressions involving z are analogous to the moments of the truncated distribution of x given in Theorem 22.2. 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.

22.4.2 REGRESSION IN A MODEL OF SELECTION

To motivate a regression model that corresponds to the results in Theorem 22.5, we consider two examples.

Example 22.7 A Model of Labor Supply

A simple model of female labor supply that has been examined in many studies consists of two equations:[21]

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

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

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 22.5 to obtain the model *that applies to the observations in our sample:*

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

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

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 22.5 implies that even if λ_i were observed, then least squares would be inefficient. The disturbance v_i is heteroscedastic.

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

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

where

$$\delta_i = \lambda_i^2 - \alpha_i \lambda_i.^{22}$$

Suppose that ρ is positive and $E[y_i]$ is greater when z_i^* is positive than when it is negative. Since $0 < \delta_i < 1$, 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 22.9, 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. Since 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 21 in connection with the probit model.)

[22] We have reversed the sign of α_μ in (22-19) since $a = 0$, and $\alpha = \boldsymbol{\gamma}'\mathbf{w}/\sigma_M$ is somewhat more convenient. Also, as such, $\partial\lambda/\partial\alpha = -\delta$.

Thus, we reformulate the model as follows:

$$\begin{aligned}
&\text{selection mechanism: } z_i^* = \mathbf{w}_i'\boldsymbol{\gamma} + u_i,\ z_i = 1 \text{ if } z_i^* > 0 \text{ and 0 otherwise;}\\
&\qquad\qquad \text{Prob}(z_i = 1 \mid \mathbf{w}_i) = \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\gamma}) \text{ and}\\
&\qquad\qquad \text{Prob}(z_i = 0 \mid \mathbf{w}_i) = 1 - \boldsymbol{\Phi}(\mathbf{w}_i'\boldsymbol{\gamma}). \qquad \textbf{(22-20)}\\
&\text{regression model: } y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ observed only if } z_i = 1,\\
&\qquad\qquad (u_i, \varepsilon_i) \sim \text{bivariate normal } [0, 0, 1, \sigma_\varepsilon, \rho].
\end{aligned}$$

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 examined earlier, with

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

22.4.3 ESTIMATION

The parameters of the sample selection model can be estimated by maximum likelihood.[23] However, Heckman's (1979) **two-step estimation** procedure is usually used instead. Heckman's method is as follows:[24]

1. Estimate the probit equation by maximum likelihood to obtain estimates of $\boldsymbol{\gamma}$. For each observation in the selected sample, compute $\hat{\lambda}_i = \phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{w}_i'\hat{\boldsymbol{\gamma}})$ and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i - \mathbf{w}_i'\hat{\boldsymbol{\gamma}})$.
2. Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_e$ 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

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

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

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

whereas based on the probit results we have

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

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

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

[23] See Greene (1995a).

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

Finally, an estimator of ρ^2 is

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

which provides a complete set of estimators of the model's parameters.[25]

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 22.5 that the disturbance term in

$$(y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i + v_i \qquad \textbf{(22-21)}$$

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 11, 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 (22-21) as

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

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 (21-22) to (21-24) may be used to compute $\hat{\mathbf{V}}$. The complete expression is

$$\text{Est.Asy. Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\boldsymbol{\Delta}})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}.$$[26]

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

[26] This matrix formulation is derived in Greene (1981). Note that the Murphy and Topel (1985) results for two-step estimators given in Theorem 10.3 would apply here as well. Asymptotically, this method would give the same answer. The Heckman formulation has become standard in the literature.

TABLE 22.7 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)	−0.632	(1.063)	−2.56	(0.929)
β_2	0.021	(0.0625)	0.00897	(0.000678)	0.0325	(0.0616)
β_3	0.000137	(0.00188)	−0.334d − 4	(0.782d − 7)	−0.000260	(0.00184)
β_4	0.417	(0.100)	0.147	(0.0142)	0.481	(0.0669)
β_5	0.444	(0.316)	0.144	(0.0614)	0.449	(0.449)
$(\rho\sigma)$	−1.100	(0.127)				
ρ	−0.340		−0.131	(0.218)	0.000	
σ	3.200		0.321	(0.00866)	3.111	

Example 22.8 Female Labor Supply

Examples 21.1 and 21.4 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:

$$wage = \beta_1 + \beta_2\, Exper + \beta_3\, Exper^2 + \beta_4\, Education + \beta_5\, City + \varepsilon$$

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 22.7. 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 below. Note as well that the two-step estimator estimates the single coefficient on λ_i and the structural parameters σ 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 22.7 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 0.149 in total for the maximum likelihood estimates. For the kids variable, the marginal effect is −.293 for the two-step estimates and only −0.0113 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 22.9 A Mover Stayer Model for Migration

The model of migration analyzed by Nakosteen and Zimmer (1980) fits into the framework described above. 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

TABLE 22.8 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 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 2 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.

The earnings equations included ΔSIC and SE. The authors reported the results given in Table 22.8. The figures in parentheses are asymptotic t ratios.

22.4.4 TREATMENT EFFECTS

The basic model of selectivity outlined earlier has been extended in an impressive variety of directions.[27] An interesting application that has found wide use is the measurement of **treatment effects** and program effectiveness.[28]

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

[27] For a survey, see Maddala (1983).

[28] This is one of the fundamental applications of this body of techniques, and is also the setting for the most longstanding and contentious debate on the subject. A *Journal of Business and Economic Statistics* symposium [Angrist et al. (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

(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

$$\begin{aligned} E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] &= \mathbf{x}_i'\boldsymbol{\beta} + \delta + E[\varepsilon_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \delta + \rho\sigma_\varepsilon \lambda(-\mathbf{w}_i'\boldsymbol{\gamma}) \end{aligned} \tag{22-22}$$

once again. [See (22-19).] 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 that will account for the self-selected nature of program participation. For nonparticipants, the counterpart to (22-22) is

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

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

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] - E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{z}_i] = \delta + \rho\sigma_\varepsilon \left[\frac{\phi_i}{\Phi_i(1 - \Phi_i)}\right].$$

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 since (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 22.9 by least squares, which, as we have seen, would lead to inconsistent estimates of both sets of parameters.

There are many variations of this model in the empirical literature. They have been applied to the analysis of education,[29] the Head Start program,[30] and a host of other settings.[31] This strand of literature is particularly important because the use of dummy variable models to analyze treatment effects and program participation has a long

[29] Willis and Rosen (1979).

[30] Goldberger (1972).

[31] A useful summary of the issues is Barnow, Cain, and Goldberger (1981). See also Maddala (1983) for a long list of applications. A related application is the switching regression model. See, for example, Quandt (1982, 1988).

history in empirical economics. This analysis has called into question the interpretation of a number of received studies.

22.4.5 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 specification.

Recent research includes specific attempts to move away from the normality assumption.[32] An example is Martins (2001), building on Newey (1991), which takes the core specification as given in (22-20) as the platform, but constructs an alternative to the assumption of bivariate normality. Martins' 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 21.5.4. 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' results, based on two-step 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, e.g., Honore and Kyriazidou (1999, 2000) represent more substantial departures from the normal model, but are much less operational.[33] 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.

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

[33] 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.

22.4.6 SELECTION IN QUALITATIVE RESPONSE MODELS

The problem of sample selection has been modeled in other settings besides the linear regression model. In Section 21.6.4, we saw, for example, an application of what amounts to a model of sample selection in a bivariate probit model; a binary response variable $y_i = 1$ if an individual defaults on a loan is observed only if a related variable z_i equals one (the individual is granted a loan). Greene's (1992) application to credit card applications and defaults is similar.

A current strand of literature has developed several models of sample selection for count data models.[34] Terza (1995) models the phenomenon as a form of heterogeneity in the Poisson model. We write

$$\begin{aligned} y_i \mid \varepsilon_i &\sim \text{Poisson}(\lambda_i), \\ \ln \lambda_i \mid \varepsilon_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \end{aligned} \tag{22-23}$$

Then the sample selection is similar to that discussed in the previous sections, with

$$\begin{aligned} z_i^* &= \mathbf{w}_i'\boldsymbol{\gamma} + u_i, \\ z_i &= 1 \quad \text{if } z_i^* > 0, 0 \text{ otherwise} \end{aligned}$$

and $[\varepsilon_i, u_i]$ have a bivariate normal distribution with the same specification as in our earlier model. As before, we assume that $[y_i, \mathbf{x}_i]$ are only observed when $z_i = 1$. Thus, the effect of the selection is to affect the mean (and variance) of y_i, although the effect on the distribution is unclear. In the observed data, y_i no longer has a Poisson distribution. Terza (1998), Terza and Kenkel (2001) and Greene (1997a) suggested a maximum likelihood approach for estimation.

22.5 MODELS FOR DURATION DATA[35]

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.

[34] See, for example, Bockstael et al. (1990), Smith (1988), Brannas (1995), Greene (1994, 1995c, 1997a), Weiss (1995), and Terza (1995, 1998), and Winkelmann (1997).

[35] 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 (1980), Heckman and Singer (1984a), Lancaster (1990) and Florens, Fougere, and Mouchart (1996).

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.

22.5.1 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 22.10 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 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 22.3. 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) $\mathbf{x}_i$, 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.[36]

22.5.2 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.)

22.5.2.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,**

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

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

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

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

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 22.10.]

22.5.2.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 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, since 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 $\lambda(t) = \exp[g(t, \theta)]$, where θ is a vector of parameters to be estimated. With an observed sample of durations, estimation of α and

TABLE 22.9 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 $\pi^2/(3p^2)$.]	

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.[37] To illustrate the differences, we will examine a few of the simpler ones. Table 22.9 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, e.g., Kiefer (1988), use different parameterizations of these models; We follow the convention of Kalbfleisch and Prentice (1980).]

All these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors, as can be seen in Figure 22.4. 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.

22.5.2.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 22.3 for the tobit model. [See (22-13).]

[37] Three sources that contain numerous specifications are Kalbfleisch and Prentice (1980), Cox and Oakes (1985), and Lancaster (1990).

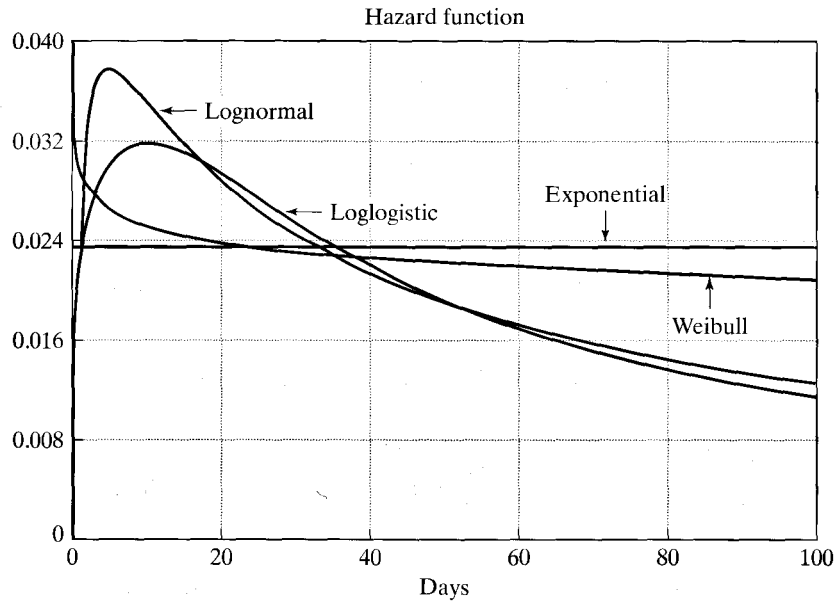


FIGURE 22.4 Parametric Hazard Functions.

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

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

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 $S(w) = \exp(-\exp(w))$.[38] 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 (1980, pp. 56–60). The Weibull distribution is examined in detail in the next section.

[38]The transformation is $\exp(w) = (\lambda t)^p$ so $t = (1/\lambda)[\exp(w)]^{1/p}$. The Jacobian of the transformation is $dt/dw = [\exp(w)]^{1/p}/(\lambda p)$. The density in Table 22.9 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, $[\exp(-\exp(w))]$.

22.5.2.d Exogenous Variables

One limitation of the models given above 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

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

For the **Weibull model,** for example (see footnote 38)

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

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 22.9, 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 Appendix E. The individual terms can also be used

to form the BHHH estimator of the asymptotic covariance matrix for the estimator.[39] The Hessian is also simple to derive, so Newton's method could be used instead.[40]

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, $E[t \mid \mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\Gamma[(1/p) + 1]$, 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 (1980), and Lancaster (1990).

22.5.2.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 22.5.3) is largely immune to the problem, but it is also rather limited in how 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.[41] 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

[39] Note that the log-likelihood function has the same form as that for the tobit model in Section 22.3. 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.]

[40] See Kalbfleisch and Prentice (1980) for numerous other examples.

[41] 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 21.9.5.

$\theta = 1/k$. 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.[42] 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.

22.5.3 OTHER 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.

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

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

[42] For the strike data analyzed earlier, the maximum likelihood estimate of θ is 0.0004, which suggests that at least in the context of the Weibull model, heterogeneity does not appear to be a problem.

The estimator of the hazard rate is

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}. \tag{22-24}$$

Corrections are necessary for observations that are censored. Lawless (1982), Kalbfleisch and Prentice (1980), 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 (22-24) 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. Since 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 (22-24) over a mesh of defined intervals as follows:

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

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 21.5.1.b.) 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 21) is

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

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

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

TABLE 22.10 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)

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 22.10 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 F22.1 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, 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.[43]

Parameter estimates for the four duration models are given in Table 22.10. 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 5.2.4.) 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 22.4.

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 a covariate, the estimated Weibull model is

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

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

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

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.

22.6 SUMMARY AND CONCLUSIONS

This chapter has examined three settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form. 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. 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. 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. In the final section, we examined an application, models of duration, which describe variables with limited (nonnegative) ranges of variation and which are often observed subject to censoring.

Key Terms and Concepts

- Accelerated failure time
- Attenuation
- Censored regression
- Censored variable
- Censoring
- Conditional moment test
- Count data
- Degree of truncation
- Delta method
- Duration dependence
- Duration model
- Generalized residual
- Hazard function
- Hazard rate
- Heterogeneity
- Heteroscedasticity
- Incidental truncation
- Integrated hazard function
- Inverse Mills ratio
- Lagrange multiplier test
- Marginal effects
- Negative duration dependence
- Olsen's reparameterization
- Parametric model
- Partial likelihood
- Positive duration dependence
- Product limit
- Proportional hazard
- Risk set
- Sample selection
- Semiparametric model
- Specification error
- Survival function
- Time varying covariate
- Tobit model
- Treatment effect
- Truncated bivariate normal distribution
- Truncated distribution
- Truncated mean
- Truncated random variable
- Truncated variance
- Two step estimation
- Weibull 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 22.2. 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 22.5.]
5. Derive the marginal effects for the tobit model with heteroscedasticity that is described in Section 22.3.4.a.
6. Prove that the Hessian for the tobit model in (22-14) is negative definite after Olsen's transformation is applied to the parameters.