17

Limited Dependent Variable Models and Sample Selection Corrections

In Chapter 7, we studied the linear probability model, which is simply an application of the multiple regression model to a binary dependent variable. A binary dependent variable is an example of a **limited dependent variable** (**LDV**). An LDV is broadly defined as a dependent variable whose range of values is substantively restricted. A binary variable takes on only two values, zero and one. We have seen several other examples of limited dependent variables: participation percentage in a pension plan must be between zero and 100, the number of times an individual is arrested in a given year is a nonnegative integer, and college grade point average is between zero and 4.0 at most colleges.

Most economic variables we would like to explain are limited in some way, often because they must be positive. For example, hourly wage, housing price, and nominal interest rates must be greater than zero. But not all such variables need special treatment. If a strictly positive variable takes on many different values, a special econometric model is rarely necessary. When y is discrete and takes on a small number of values, it makes no sense to treat it as an approximately continuous variable. Discreteness of y does not in itself mean that linear models are inappropriate. However, as we saw in Chapter 7 for binary response, the linear probability model has certain drawbacks. In Section 17.1, we discuss logit and probit models, which overcome the shortcomings of the LPM; the disadvantage is that they are more difficult to interpret.

Other kinds of limited dependent variables arise in econometric analysis, especially when the behavior of individuals, families, or firms is being modeled. Optimizing behavior often leads to a **corner solution response** for some nontrivial fraction of the population. That is, it is optimal to choose a zero quantity or dollar value, for example. During any given year, a significant number of families will make zero charitable contributions. Therefore, annual family charitable contributions has a population distribution that is spread out over a large range of positive values, but with a pileup at the value zero. Although a linear model could be appropriate for capturing the expected value of charitable contributions, a linear model will likely lead to negative predictions for some families. Taking the natural log is not possible because many observations are zero. The Tobit model, which we cover in Section 17.2, is explicitly designed to model corner solution dependent variables.

Another important kind of LDV is a count variable, which takes on nonnegative integer values. Section 17.3 illustrates how Poisson regression models are well suited for modeling count variables.

In some cases, we observe limited dependent variables due to data censoring, a topic we introduce in Section 17.4. The general problem of sample selection, where we observe a nonrandom sample from the underlying population, is treated in Section 17.5.

Limited dependent variable models can be used for time series and panel data, but they are most often applied to cross-sectional data. Sample selection problems are usually confined to cross-sectional or panel data. We focus on cross-sectional applications in this chapter. Wooldridge (2002) presents these problems in the context of panel data models and provides many more details for cross-sectional and panel data applications.

17.1 Logit and Probit Models for Binary Response

The linear probability model is simple to estimate and use, but it has some drawbacks that we discussed in Section 7.5. The two most important disadvantages are that the fitted probabilities can be less than zero or greater than one and the partial effect of any explanatory variable (appearing in level form) is constant. These limitations of the LPM can be overcome by using more sophisticated **binary response models**.

In a binary response model, interest lies primarily in the **response probability**

$$P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, \ldots, x_k), \tag{17.1}$$

where we use $\mathbf{x}$ to denote the full set of explanatory variables. For example, when y is an employment indicator, $\mathbf{x}$ might contain various individual characteristics such as education, age, marital status, and other factors that affect employment status, including a binary indicator variable for participation in a recent job training program.

Specifying Logit and Probit Models

In the LPM, we assume that the response probability is linear in a set of parameters, β_j; see equation (7.27). To avoid the LPM limitations, consider a class of binary response models of the form

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \tag{17.2}$$

where G is a function taking on values strictly between zero and one: $0 < G(z) < 1$, for all real numbers z. This ensures that the estimated response probabilities are strictly between zero and one. As in earlier chapters, we write $\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$.

Various nonlinear functions have been suggested for the function G in order to make sure that the probabilities are between zero and one. The two we will cover here are used in the vast majority of applications (along with the LPM). In the **logit model**, G is the logistic function:

$$G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z), \tag{17.3}$$

which is between zero and one for all real numbers z. This is the cumulative distribution function for a standard logistic random variable. In the **probit model**, G is the standard normal cumulative distribution function (cdf), which is expressed as an integral:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)dv, \tag{17.4}$$

where $\phi(z)$ is the standard normal density

$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2). \tag{17.5}$$

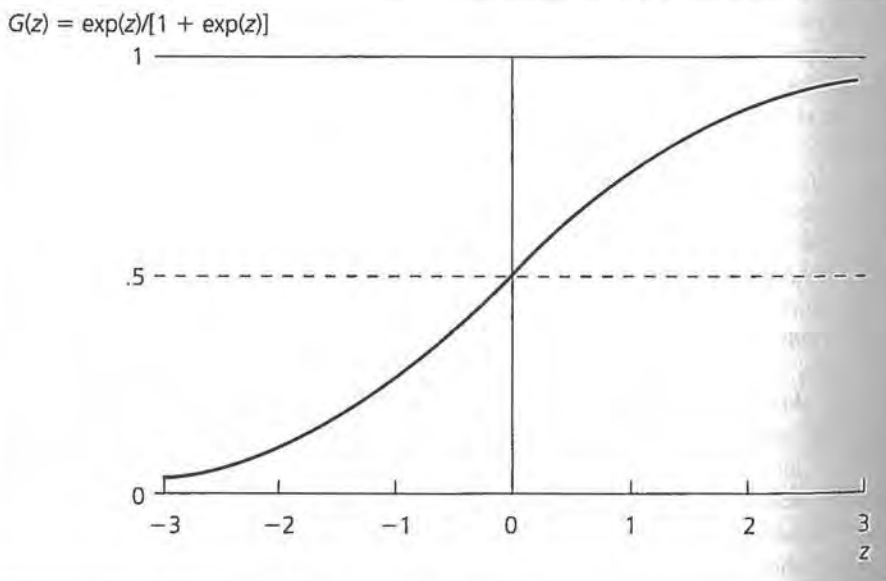
This choice of G again ensures that (17.2) is strictly between zero and one for all values of the parameters and the x_j.

The G functions in (17.3) and (17.4) are both increasing functions. Each increases most quickly at $z = 0$, $G(z) \to 0$ as $z \to -\infty$, and $G(z) \to 1$ as $z \to \infty$. The logistic function is plotted in Figure 17.1. The standard normal cdf has a shape very similar to that of the logistic cdf.

Logit and probit models can be derived from an underlying **latent variable model**. Let y^* be an unobserved, or *latent*, variable, determined by

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e, y = 1[y^* > 0], \tag{17.6}$$

FIGURE 17.1
Graph of the logistic function $G(z) = \exp(z)/[1 + \exp(z)]$.



where we introduce the notation $1[\cdot]$ to define a binary outcome. The function $1[\cdot]$ is called the *indicator function*, which takes on the value one if the event in brackets is true, and zero otherwise. Therefore, y is one if $y^* > 0$, and y is zero if $y^* \leq 0$. We assume that e is independent of $\mathbf{x}$ and that e either has the standard logistic distribution or the standard normal distribution. In either case, e is symmetrically distributed about zero, which means that $1 - G(-z) = G(z)$ for all real numbers z. Economists tend to favor the normality assumption for e, which is why the probit model is more popular than logit in econometrics. In addition, several specification problems, which we touch on later, are most easily analyzed using probit because of properties of the normal distribution.

From (17.6) and the assumptions given, we can derive the response probability for y:

$$\begin{aligned} \mathrm{P}(y = 1|\mathbf{x}) &= \mathrm{P}(y^* > 0|\mathbf{x}) = \mathrm{P}[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \end{aligned}$$

which is exactly the same as (17.2).

In most applications of binary response models, the primary goal is to explain the effects of the x_j on the response probability $\mathrm{P}(y = 1|\mathbf{x})$. The latent variable formulation tends to give the impression that we are primarily interested in the effects of each x_j on y^*. As we will see, for logit and probit, the *direction* of the effect of x_j on $\mathrm{E}(y^*|\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ and on $\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ is always the same. But the latent variable y^* rarely has a well-defined unit of measurement. (For example, y^* might be the difference in utility levels from two different actions.) Thus, the magnitudes of each β_j are not, by themselves, especially useful (in contrast to the linear probability model). For most purposes, we want to estimate the effect of x_j on the probability of success $\mathrm{P}(y = 1|\mathbf{x})$, but this is complicated by the nonlinear nature of $G(\cdot)$.

To find the partial effect of roughly continuous variables on the response probability, we must rely on calculus. If x_j is a roughly continuous variable, its partial effect on $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$ is obtained from the partial derivative:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j, \text{ where } g(z) \equiv \frac{dG}{dz}(z). \tag{17.7}$$

Because G is the cdf of a continuous random variable, g is a probability density function. In the logit and probit cases, $G(\cdot)$ is a strictly increasing cdf, and so $g(z) > 0$ for all z. Therefore, the partial effect of x_j on $p(\mathbf{x})$ depends on $\mathbf{x}$ through the positive quantity $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})$, which means that the partial effect always has the same sign as β_j.

Equation (17.7) shows that the *relative* effects of any two continuous explanatory variables do not depend on $\mathbf{x}$: the ratio of the partial effects for x_j and x_h is β_j/β_h. In the typical case that g is a symmetric density about zero, with a unique mode at zero, the largest effect occurs when $\beta_0 + \mathbf{x}\boldsymbol{\beta} = 0$. For example, in the probit case with $g(z) = \phi(z)$, $g(0) = \phi(0) = 1/\sqrt{2\pi} \approx .40$. In the logit case, $g(z) = \exp(z)/[1 + \exp(z)]^2$, and so $g(0) = .25$.

If, say, x_1 is a binary explanatory variable, then the partial effect from changing x_1 from zero to one, holding all other variables fixed, is simply

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_k x_k). \tag{17.8}$$

Again, this depends on all the values of the other x_j. For example, if y is an employment indicator and x_1 is a dummy variable indicating participation in a job training program, then (17.8) is the change in the probability of employment due to the job training program; this depends on other characteristics that affect employability, such as education and experience. Note that knowing the sign of β_1 is sufficient for determining whether the program had a positive or negative effect. But to find the *magnitude* of the effect, we have to estimate the quantity in (17.8).

We can also use the difference in (17.8) for other kinds of discrete variables (such as number of children). If x_k denotes this variable, then the effect on the probability of x_k going from c_k to $c_k + 1$ is simply

$$G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k(c_k + 1)] - G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k c_k). \tag{17.9}$$

It is straightforward to include standard functional forms among the explanatory variables. For example, in the model

$$\mathrm{P}(y = 1|\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3),$$

the partial effect of z_1 on $\mathrm{P}(y = 1|\mathbf{z})$ is $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_1 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$, and the partial effect of z_2 on the response probability is $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_2 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/z_2)$, where $\mathbf{x}\boldsymbol{\beta} = \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3$. Therefore, $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/100)$ is the approximate change in the response probability when z_2 increases by 1 percent. Models with interactions among explanatory variables, including those between discrete and continuous variables, are handled similarly. When measuring effects of discrete variables, we should use (17.9).

Maximum Likelihood Estimation of Logit and Probit Models

How should we estimate nonlinear binary response models? To estimate the LPM, we can use ordinary least squares (see Section 7.5) or, in some cases, weighted least squares (see Section 8.5). Because of the nonlinear nature of $\mathrm{E}(y|\mathbf{x})$, OLS and WLS are not applicable. We could use nonlinear versions of these methods, but it is no more difficult to use **maximum likelihood estimation** (**MLE**) (see Appendix B for a brief discussion). Up until now, we have had little need for MLE, although we did note that, under the classical linear model assumptions, the OLS estimator is the maximum likelihood estimator (conditional on the explanatory variables). For estimating limited dependent variable models, maximum likelihood methods are indispensable. Because maximum likelihood estimation is based on the distribution of y given $\mathbf{x}$, the heteroskedasticity in $\mathrm{Var}(y|\mathbf{x})$ is automatically accounted for.

Assume that we have a random sample of size n. To obtain the maximum likelihood estimator, conditional on the explanatory variables, we need the density of y_i given $\mathbf{x}_i$. We can write this as

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y}, y = 0,1, \tag{17.10}$$

where, for simplicity, we absorb the intercept into the vector $\mathbf{x}_i$. We can easily see that when $y = 1$, we get $G(\mathbf{x}_i\boldsymbol{\beta})$ and when $y = 0$, we get $1 - G(\mathbf{x}_i\boldsymbol{\beta})$. The **log-likelihood function** for observation i is a function of the parameters and the data $(\mathbf{x}_i, y_i)$ and is obtained by taking the log of (17.10):

$$\ell_i(\boldsymbol{\beta}) = y_i \log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]. \tag{17.11}$$

Because $G(\cdot)$ is strictly between zero and one for logit and probit, $\ell_i(\boldsymbol{\beta})$ is well defined for all values of $\boldsymbol{\beta}$.

The log-likelihood for a sample size of n is obtained by summing (17.11) across all observations: $\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$. The MLE of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, maximizes this log-likelihood. If $G(\cdot)$ is the standard logit cdf, then $\hat{\boldsymbol{\beta}}$ is the *logit estimator*; if $G(\cdot)$ is the standard normal cdf, then $\hat{\boldsymbol{\beta}}$ is the *probit estimator*.

Because of the nonlinear nature of the maximization problem, we cannot write formulas for the logit or probit maximum likelihood estimates. In addition to raising computational issues, this makes the statistical theory for logit and probit much more difficult than OLS or even 2SLS. Nevertheless, the general theory of MLE for random samples implies that, under very general conditions, the MLE is consistent, asymptotically normal, and asymptotically efficient. (See Wooldridge [2002, Chapter 13] for a general discussion.) We will just use the results here; applying logit and probit models is fairly easy, provided we understand what the statistics mean.

Each $\hat{\beta}_j$ comes with an (asymptotic) standard error, the formula for which is complicated and presented in the chapter appendix. Once we have the standard errors—and these are reported along with the coefficient estimates by any package that supports logit and probit—we can construct (asymptotic) t tests and confidence intervals, just as with OLS, 2SLS, and the other estimators we have encountered. In particular, to test H_0: $\beta_j = 0$, we form the t statistic $\hat{\beta}_j/\text{se}(\hat{\beta}_j)$ and carry out the test in the usual way, once we have decided on a one- or two-sided alternative.

Testing Multiple Hypotheses

We can also test multiple restrictions in logit and probit models. In most cases, these are tests of multiple exclusion restrictions, as in Section 4.5. We will focus on exclusion restrictions here.

There are three ways to test exclusion restrictions for logit and probit models. The Lagrange multiplier or score test only requires estimating the model under the null hypothesis, just as in the linear case in Section 5.2; we will not cover the score test here, since it is rarely needed to test exclusion restrictions. (See Wooldridge [2002, Chapter 15] for other uses of the score test in binary response models.)

The Wald test requires estimation of only the unrestricted model. In the linear model case, the **Wald statistic**, after a simple transformation, is essentially the F statistic, so there is no need to cover the Wald statistic separately. The formula for the Wald statistic is given in Wooldridge (2002, Chapter 15). This statistic is computed by econometrics packages that allow exclusion restrictions to be tested after the unrestricted model has been estimated.

It has an asymptotic chi-square distribution, with *df* equal to the number of restrictions being tested.

If both the restricted and unrestricted models are easy to estimate—as is usually the case with exclusion restrictions—then the *likelihood ratio* (*LR*) *test* becomes very attractive. The *LR* test is based on the same concept as the *F* test in a linear model. The *F* test measures the increase in the sum of squared residuals when variables are dropped from the model. The *LR* test is based on the difference in the log-likelihood functions for the unrestricted and restricted models. The idea is this. Because the MLE maximizes the log-likelihood function, dropping variables generally leads to a *smaller*—or at least no larger—log-likelihood. (This is similar to the fact that the *R*-squared never increases when variables are dropped from a regression.) The question is whether the fall in the log-likelihood is large enough to conclude that the dropped variables are important. We can make this decision once we have a test statistic and a set of critical values.

The **likelihood ratio statistic** is *twice* the difference in the log-likelihoods:

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r), \tag{17.12}$$

where $\mathcal{L}_{ur}$ is the log-likelihood value for the unrestricted model and $\mathcal{L}_r$ is the log-likelihood value for the restricted model. Because $\mathcal{L}_{ur} \geq \mathcal{L}_r$, *LR* is nonnegative and usually strictly positive. In computing the *LR* statistic for binary response models, it is important to know that the log-likelihood function is always a negative number. This fact follows from equation (17.11), because y_i is either zero or one and both variables inside the log function are strictly between zero and one, which means their natural logs are negative. That the log-likelihood functions are both negative does not change the way we compute the *LR* statistic; we simply preserve the negative signs in equation (17.12).

The multiplication by two in (17.12) is needed so that *LR* has an approximate chi-square distribution under H_0. If we are testing q exclusion restrictions, $LR \overset{a}{\sim} \chi^2_q$. This means that, to test H_0 at the 5% level, we use as our critical value the 95th percentile in the χ^2_q distribution. Computing p-values is easy with most software packages.

QUESTION 17.1

A probit model to explain whether a firm is taken over by another firm during a given year is

$$P(takeover = 1|\mathbf{x} = \Phi(\beta_0 + \beta_1 avgprof + \beta_2 mktval + \beta_3 debtearn + \beta_4 ceoten + \beta_5 ceosal + \beta_6 ceoage),$$

where *takeover* is a binary response variable, *avgprof* is the firm's average profit margin over several prior years, *mktval* is market value of the firm, *debtearn* is the debt-to-earnings ratio, and *ceoten*, *ceosal*, and *ceoage* are the tenure, annual salary, and age of the chief executive officer, respectively. State the null hypothesis that, other factors being equal, variables related to the CEO have no effect on the probability of takeover. How many *df* are in the chi-square distribution for the *LR* or Wald test?

Interpreting the Logit and Probit Estimates

Given modern computers, from a practical perspective the most difficult aspect of logit or probit models is presenting and interpreting the results. The coefficient estimates, their standard errors, and the value of the log-likelihood function are reported by all software packages that do logit and probit, and these should be reported in any application. The coefficients give the signs of the partial effects of each x_j on the response probability, and

the statistical significance of x_j is determined by whether we can reject H_0: $\beta_j = 0$ at a sufficiently small significance level.

As we briefly discussed in Section 7.5 for the linear probability model, we can compute a goodness-of-fit measure called the **percent correctly predicted**. As before, we define a binary predictor of y_i to be one if the predicted probability is at least .5, and zero otherwise. Mathematically, $\tilde{y}_i = 1$ if $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .5$ and $\tilde{y}_i = 0$ if $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) < .5$. Given $\{\tilde{y}_i: i = 1,2,...,n\}$, we can see how well $\tilde{y}_i$ predicts y_i across all observations. There are four possible outcomes on each pair, $(y_i,\tilde{y}_i)$; when both are zero or both are one, we make the correct prediction. In the two cases where one of the pair is zero and the other is one, we make the incorrect prediction. The percent correctly predicted is the percentage of times that $\tilde{y}_i = y_i$.

Although the percent correctly predicted is useful as a goodness-of-fit measure, it can be misleading. In particular, it is possible to get rather high percentages correctly predicted even when the least likely outcome is very poorly predicted. For example, suppose that $n = 200$, 160 observations have $y_i = 0$, and, out of these 160 observations, 140 of the $\tilde{y}_i$ are also zero (so we correctly predict 87.5% of the zero outcomes). Even if *none* of the predictions is correct when $y_i = 1$, we still correctly predict 70% of all outcomes (140/200 = .70). Often, we hope to have some ability to predict the least likely outcome (such as whether someone is arrested for committing a crime), and so we should be up front about how well we do in predicting each outcome. Therefore, it makes sense to also compute the percent correctly predicted for each of the outcomes. Problem 17.1 asks you to show that the overall percent correctly predicted is a weighted average of $\hat{q}_0$ (the percent correctly predicted for $y_i = 0$) and $\hat{q}_1$ (the percent correctly predicted for $y_i = 1$), where the weights are the fractions of zeros and ones in the sample, respectively.

Some have criticized the prediction rule just described for using a threshold value of .5, especially when one of the outcomes is unlikely. For example, if $\bar{y} = .08$ (only 8% "successes" in the sample), it could be that we *never* predict $y_i = 1$ because the estimated probability of success is never greater than .5. One alternative is to use the fraction of successes in the sample as the threshold—.08 in the previous example. In other words, define $\tilde{y}_i = 1$ when $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .08$ and zero otherwise. Using this rule will certainly increase the number of predicted successes, but not without cost: we will necessarily make more mistakes—perhaps many more—in predicting zeros ("failures"). In terms of the overall percent correctly predicted, we may do worse than using the .5 threshold.

A third possibility is to choose the threshold such that the fraction of $\tilde{y}_i = 1$ in the sample is the same as (or very close to) $\bar{y}$. In other words, search over threshold values τ, $0 < \tau < 1$, such that if we define $\tilde{y}_i = 1$ when $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq \tau$, then $\sum_{i=1}^{n} \tilde{y}_i \approx \sum_{i=1}^{n} y_i$. (The trial-and-error required to find the desired value of τ can be tedious but it is feasible. In some cases, it will not be possible to make the number of predicted successes exactly the same as the number of successes in the sample.) Now, given this set of $\tilde{y}_i$, we can compute the percent correctly predicted for each of the two outcomes as well as the overall percent correctly predicted.

There are also various **pseudo *R*-squared** measures for binary response. McFadden (1974) suggests the measure $1 - \mathcal{L}_{ur}/\mathcal{L}_o$, where $\mathcal{L}_{ur}$ is the log-likelihood function for the estimated model, and $\mathcal{L}_o$ is the log-likelihood function in the model with only an intercept. Why does this measure make sense? Recall that the log-likelihoods are negative,

and so $\mathcal{L}_{ur}/\mathcal{L}_o = |\mathcal{L}_{ur}|/|\mathcal{L}_o|$. Further, $|\mathcal{L}_{ur}| \leq |\mathcal{L}_o|$. If the covariates have no explanatory power, then $\mathcal{L}_{ur}/\mathcal{L}_o = 1$, and the pseudo R-squared is zero, just as the usual R-squared is zero in a linear regression when the covariates have no explanatory power. Usually, $|\mathcal{L}_{ur}| < |\mathcal{L}_o|$, in which case $1 - \mathcal{L}_{ur}/\mathcal{L}_o > 0$. If $\mathcal{L}_{ur}$ were zero, the pseudo R-squared would equal unity. In fact, $\mathcal{L}_{ur}$ cannot reach zero in a probit or logit model, as that would require the estimated probabilities when $y_i = 1$ all to be unity and the estimated probabilities when $y_i = 0$ all to be zero.

Alternative pseudo R-squareds for probit and logit are more directly related to the usual R-squared from OLS estimation of a linear probability model. For either probit or logit, let $\hat{y}_i = G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ be the fitted probabilities. Since these probabilities are also estimates of $\mathrm{E}(y_i|\mathbf{x}_i)$, we can base an R-squared on how close the $\hat{y}_i$ are to the y_i. One possibility that suggests itself from standard regression analysis is to compute the squared correlation between y_i and $\hat{y}_i$. Remember, in a linear regression framework, this is an algebraically equivalent way to obtain the usual R-squared; see equation (3.29). Therefore, we can compute a pseudo R-squared for probit and logit that is directly comparable to the usual R-squared from estimation of a linear probability model. In any case, goodness-of-fit is usually less important than trying to obtain convincing estimates of the ceteris paribus effects of the explanatory variables.

Often, we want to estimate the effects of the x_j on the response probabilities, $\mathrm{P}(y = 1|\mathbf{x})$. If x_j is (roughly) continuous, then

$$\widehat{\Delta \mathrm{P}(y = 1|\mathbf{x})} \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j, \tag{17.13}$$

for "small" changes in x_j. So, for $\Delta x_j = 1$, the change in the estimated success probability is roughly $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j$. Compared with the linear probability model, the cost of using probit and logit models is that the partial effects in equation (17.13) are harder to summarize because the scale factor, $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$, depends on $\mathbf{x}$ (that is, on all of the explanatory variables). One possibility is to plug in interesting values for the x_j—such as means, medians, minimums, maximums, and lower and upper quartiles—and then see how $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ changes. Although attractive, this can be tedious and result in too much information even if the number of explanatory variables is moderate.

As a quick summary for getting at the magnitudes of the partial effects, it is handy to have a single scale factor that can be used to multiply each $\hat{\beta}_j$(or at least those coefficients on roughly continuous variables). One method, commonly used in econometrics packages that routinely estimate probit and logit models, is to replace each explanatory variable with its sample average. In other words, the adjustment factor is

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k), \tag{17.14}$$

where $g(\cdot)$ is the standard normal density in the probit case and $g(z) = \exp(z)/[1 + \exp(z)]^2$ in the logit case. The idea behind (17.14) is that, when it is multiplied by $\hat{\beta}_j$, we obtain the partial effect of x_j for the "average" person in the sample. There are two potential problems with this motivation. First, if some of the explanatory variables are discrete, the averages of them represent no one in the sample (or population, for that matter). For example, if x_1 = *female* and 47.5% of the sample is female, what sense does it make to

plug in $\bar{x}_1 = .475$ to represent the "average" person? Second, if a continuous explanatory variable appears as a nonlinear function—say, as a natural log or in a quadratic—it is not clear whether we want to average the nonlinear function or plug the average into the nonlinear function. For example, should we use $\log(\overline{sales})$ or $\log(sales)$ to represent average firm size? Econometrics packages that compute the scale factor in (17.14) default to the former: the software is written to compute the averages of the regressors included in the probit or logit estimation.

A different approach to computing a scale factor circumvents the issue of which values to plug in for the explanatory variables. Instead, the second scale factor results from averaging the individual partial effects across the sample (leading to what is sometimes called the **average partial effect**). For a continuous explanatory variable x_j, the average partial effect is $n^{-1}\sum_{i=1}^{n}[g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\hat{\beta}_j] = \left[n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\right]\hat{\beta}_j$. The term multiplying $\hat{\beta}_j$ acts as a scale factor:

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}). \tag{17.15}$$

Equation (17.15) is easily computed after probit or logit estimation, where $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \phi(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ in the probit case and $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})]^2$ in the logit case. The two scale factors differ—and are possibly quite different—because in (17.15) we are using the average of the nonlinear function rather than the nonlinear function of the average [as in (17.14)].

Because both of the scale factors just described depend on the calculus approximation in (17.13), neither makes much sense for discrete explanatory variables. Instead, it is better to use equation (17.9) to directly estimate the change in the probability. For a change in x_k from c_k to $c_k + 1$, the discrete analog of the partial effect based on (17.14) is

$$\begin{aligned} &G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k(c_k + 1)] \\ &- G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k c_k), \end{aligned} \tag{17.16}$$

where G is the standard normal cdf in the probit case and $G(z) = \exp(z)/[1 + \exp(z)]$ in the logit case. [For binary x_k, (17.16) is computed routinely by certain econometrics packages, such as Stata®.] The average partial effect, which usually is more comparable to LPM estimates, is

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} \{&G[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1}x_{ik-1} + \hat{\beta}_k(c_k + 1)] \\ &- G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1}x_{ik-1} + \hat{\beta}_k c_k)\}. \end{aligned} \tag{17.17}$$

Obtaining equation (17.17) for either probit or logit is actually rather simple. First, for each observation, we estimate the probability of success for the two chosen values of x_k, plugging in the actual outcomes for the other explanatory variables. (So, we would have n estimated differences.) Then, we average the differences in estimated probabilities across all observations. If x_k is binary, we plug in one and zero as the only two possible values.

In applications where one applies probit, logit, and the LPM, it makes sense to compute the scale factors described above for probit and logit in making comparisons of partial effects. Still, sometimes one wants a quicker way to compare magnitudes of the different estimates. As mentioned earlier, for probit $g(0) \approx .4$ and for logit, $g(0) = .25$. Thus, to make the magnitudes of probit and logit roughly comparable, we can multiply the probit coefficients by $.4/.25 = 1.6$, or we can multiply the logit estimates by .625. In the LPM, $g(0)$ is effectively one, so the logit slope estimates can be divided by four to make them comparable to the LPM estimates; the probit slope estimates can be divided by 2.5 to make them comparable to the LPM estimates. Still, in most cases, we want the more accurate comparisons obtained by using the scale factors in (17.15) for logit and probit.

EXAMPLE 17.1

(Married Women's Labor Force Participation)

We now use the MROZ.RAW data to estimate the labor force participation model from Example 8.8—see also Section 7.5—by logit and probit. We also report the linear probability model estimates from Example 8.8, using the heteroskedasticity-robust standard errors. The results, with standard errors in parentheses, are given in Table 17.1.

The estimates from the three models tell a consistent story. The signs of the coefficients are the same across models, and the same variables are statistically significant in each model. The pseudo *R*-squared for the LPM is just the usual *R*-squared reported for OLS; for logit and probit, the pseudo *R*-squared is the measure based on the log-likelihoods described earlier.

As we have already emphasized, the *magnitudes* of the coefficient estimates across models are not directly comparable. Instead, we compute the scale factors in equations (17.14) and (17.15). If we evaluate the standard normal probability density function $\phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k)$ at the sample averages of the explanatory variables (including the average of *exper*^{2}, *kidslt6*, and *kidsge6*), the result is approximately .391. When we compute (17.14) for the logit case, we obtain about .243. The ratio of these, $.391/.243 = 1.61$, is very close to the simple rule of thumb for scaling up the probit estimates to make them comparable to the logit estimates: multiply the probit estimates by 1.6. Nevertheless, for comparing probit and logit to the LPM estimates, it is better to use (17.15). These scale factors are about .301 (probit) and .179 (logit). For example, the scaled logit coefficient on *educ* is about $.179(.221) \approx .040$, and the scaled probit coefficient on *educ* is about $.301(.131) \approx .039$; both are remarkably close to the LPM estimate of .038. Even on the discrete variable *kidslt6*, the scaled logit and probit coefficients are similar to the LPM coefficient of $-.262$. These are $.179(-1.443) \approx -.258$(logit) and $.301(-.868) = -.261$ (probit).

The biggest difference between the LPM model and the logit and probit models is that the LPM assumes *constant* marginal effects for *educ*, *kidslt6*, and so on, while the logit and probit models imply diminishing magnitudes of the partial effects. In the LPM, one more small child is estimated to reduce the probability of labor force participation by about .262, regardless of how many young children the woman already has (and regardless

QUESTION 17.2

Using the probit estimates and the calculus approximation, what is the approximate change in the response probability when *exper* increases from 10 to 11?

TABLE 17.1

LPM, Logit, and Probit Estimates of Labor Force Participation

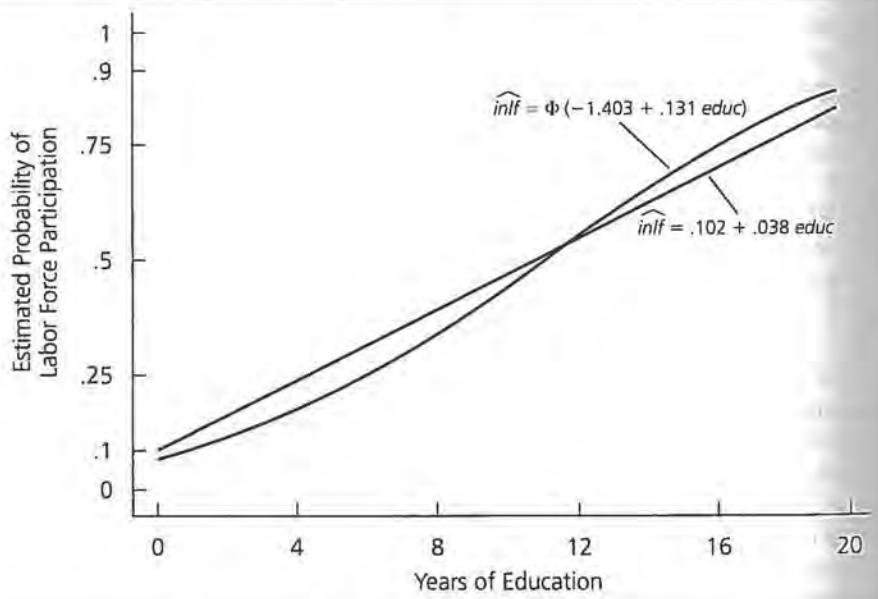
Dependent Variable: *inlf*			
Independent Variables	**LPM (OLS)**	**Logit (MLE)**	**Probit (MLE)**
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
$exper^2$	−.00060 (.00018)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (.204)	−.868 (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constant	.586 (.151)	.425 (.860)	.270 (.509)
Percent Correctly Predicted Log-Likelihood Value Pseudo R-Squared	73.4 — .264	73.6 −401.77 .220	73.4 −401.30 .221

of the levels of the other explanatory variables). We can contrast this with the estimated marginal effect from probit. For concreteness, take a woman with *nwifeinc* = 20.13, *educ* = 12.3, *exper* = 10.6, and *age* = 42.5—which are roughly the sample averages—and *kidsge6* = 1. What is the estimated decrease in the probability of working in going from zero to one small child? We evaluate the standard normal cdf, $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)$, with *kidslt6* = 1 and *kidslt6* = 0, and the other independent variables set at the preceding values. We get roughly .373 − .707 = −.334, which means that the labor force participation probability is about .334 lower when a woman has one young child. If the woman goes from one to two young children,

the probability falls even more, but the marginal effect is not as large: .117 − .373 = −.256. Interestingly, the estimate from the linear probability model, which is supposed to estimate the effect near the average, is in fact between these two estimates.

Figure 17.2 illustrates how the estimated response probabilities from nonlinear binary response models can differ from the linear probability model. The estimated probability of labor force participation is graphed against years of education for the linear probability model and the probit model. (The graph for the logit model is very similar to that for the probit model.) In both cases, the explanatory variables, other than *educ*, are set at their sample averages. In particular, the two equations graphed are $\widehat{inlf} = .102 + .038\ educ$ for the linear model and $\widehat{inlf} = \Phi(-1.403 + .131\ educ)$. At lower levels of education, the linear probability model estimates higher labor force participation probabilities than the probit model. For example, at eight years of education, the linear probability model estimates a .406 labor force participation probability while the probit model estimates about .361. The estimates are the same at around 11 1/3 years of education. At higher levels of education, the probit model gives higher labor force participation probabilities. In this sample, the smallest years of education is 5 and the largest is 17, so we really should not make comparisons outside of this range.

FIGURE 17.2
Estimated Response Probabilities with Respect to Education for the Linear Probability and Probit Models



The same issues concerning endogenous explanatory variables in linear models also arise in logit and probit models. We do not have the space to cover them, but it is possible to test and correct for endogenous explanatory variables using methods related to two stage least squares. Evans and Schwab (1995) estimated a probit model for whether a student attends college, where the key explanatory variable is a dummy variable for whether the student attends a Catholic school. Evans and Schwab estimated a model by maximum likelihood that allows attending a Catholic school to be considered endogenous. (See Wooldridge [2002, Chapter 15] for an explanation of these methods.)

Two other issues have received attention in the context of probit models. The first is nonnormality of e in the latent variable model (17.6). Naturally, if e does not have a standard normal distribution, the response probability will not have the probit form. Some authors tend to emphasize the inconsistency in estimating the β_j, but this is the wrong focus unless we are only interested in the direction of the effects. Because the response probability is unknown, we could not estimate the magnitude of partial effects even if we had consistent estimates of the β_j.

A second specification problem, also defined in terms of the latent variable model, is heteroskedasticity in e. If $\text{Var}(e|\mathbf{x})$ depends on $\mathbf{x}$, the response probability no longer has the form $G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$; instead, it depends on the form of the variance and requires more general estimation. Such models are not often used in practice, since logit and probit with flexible functional forms in the independent variables tend to work well.

Binary response models apply with little modification to independently pooled cross sections or to other data sets where the observations are independent but not necessarily identically distributed. Often, year or other time period dummy variables are included to account for aggregate time effects. Just as with linear models, logit and probit can be used to evaluate the impact of certain policies in the context of a natural experiment.

The linear probability model can be applied with panel data; typically, it would be estimated by fixed effects (see Chapter 14). Logit and probit models with unobserved effects have recently become popular. These models are complicated by the nonlinear nature of the response probabilities, and they are difficult to estimate and interpret. (See Wooldridge [2002, Chapter 15].)

17.2 The Tobit Model for Corner Solution Responses

As mentioned in the chapter introduction, another important kind of limited dependent variable is a corner solution response. Such a variable is zero for a nontrivial fraction of the population but is roughly continuously distributed over positive values. An example is the amount an individual spends on alcohol in a given month. In the population of people over age 21 in the United States, this variable takes on a wide range of values. For some significant fraction, the amount spent on alcohol is zero. The following treatment omits verification of some details concerning the Tobit model. (These are given in Wooldridge [2002, Chapter 16].)

Let y be a variable that is essentially continuous over strictly positive values but that takes on zero with positive probability. Nothing prevents us from using a linear model for y.

In fact, a linear model might be a good approximation to $E(y|x_1,x_2,\ldots,x_k)$, especially for x_j near the mean values. But we would possibly obtain negative fitted values, which leads to negative predictions for y; this is analogous to the problems with the LPM for binary outcomes. Also, the assumption that an explanatory variable appearing in level form has a constant partial effect on $E(y|\mathbf{x})$ can be misleading. Probably, $Var(y|\mathbf{x})$ would be heteroskedastic, although we can easily deal with general heteroskedasticity by computing robust standard errors and test statistics. Because the distribution of y piles up at zero, y clearly cannot have a conditional normal distribution. So all inference would have only asymptotic justification, as with the linear probability model.

In some cases, it is important to have a model that implies nonnegative predicted values for y, and which has sensible partial effects over a wide range of the explanatory variables. Plus, we sometimes want to estimate features of the distribution of y given $x_1, \ldots, x_k$ other than the conditional expectation. The **Tobit model** is quite convenient for these purposes. Typically, the Tobit model expresses the observed response, y, in terms of an underlying latent variable:

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u,\ u|\mathbf{x} \sim \text{Normal}(0,\sigma^2) \tag{17.18}$$

$$y = \max(0,y^*). \tag{17.19}$$

The latent variable y^* satisfies the classical linear model assumptions; in particular, it has a normal, homoskedastic distribution with a linear conditional mean. Equation (17.19) implies that the observed variable, y, equals y^* when $y^* \geq 0$, but $y = 0$ when $y^* < 0$. Because y^* is normally distributed, y has a continuous distribution over strictly positive values. In particular, the density of y given $\mathbf{x}$ is the same as the density of y^* given $\mathbf{x}$ for positive values. Further,

$$\begin{aligned} P(y = 0|\mathbf{x}) &= P(y^* < 0|\mathbf{x}) = P(u < -\mathbf{x}\boldsymbol{\beta}|\mathbf{x}) \\ &= P(u/\sigma < -\mathbf{x}\boldsymbol{\beta}/\sigma|\mathbf{x}) = \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \end{aligned}$$

because u/σ has a standard normal distribution and is independent of $\mathbf{x}$; we have absorbed the intercept into $\mathbf{x}$ for notational simplicity. Therefore, if $(\mathbf{x}_i,y_i)$ is a random draw from the population, the density of y_i given $\mathbf{x}_i$ is

$$(2\pi\sigma^2)^{-1/2}\exp[-(y - \mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)] = (1/\sigma)\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma],\ y > 0 \tag{17.20}$$

$$P(y_i = 0|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma), \tag{17.21}$$

where ϕ is the standard normal density function.

From (17.20) and (17.21), we can obtain the log-likelihood function for each observation i:

$$\begin{aligned} \ell_i(\boldsymbol{\beta},\sigma) &= 1(y_i = 0)\log[1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)] \\ &+ 1(y_i > 0)\log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}; \end{aligned} \tag{17.22}$$

notice how this depends on σ, the standard deviation of u, as well as on the β_j. The log-likelihood for a random sample of size n is obtained by summing (17.22) across all i. The

QUESTION 17.3

Let y be the number of extramarital affairs for a married woman from the U.S. population; we would like to explain this variable in terms of other characteristics of the woman—in particular, whether she works outside of the home—her husband, and her family. Is this a good candidate for a Tobit model?

maximum likelihood estimates of $\boldsymbol{\beta}$ and σ are obtained by maximizing the log-likelihood; this requires numerical methods, although in most cases this is easily done using a packaged routine.

As in the case of logit and probit, each Tobit estimate comes with a standard error, and these can be used to construct t statistics for each $\hat{\beta}_j$; the matrix formula used to find the standard errors is complicated and will not be presented here. (See, for example, Wooldridge [2002, Chapter 16].)

Testing multiple exclusion restrictions is easily done using the Wald test or the likelihood ratio test. The Wald test has a similar form to the logit or probit case; the *LR* test is always given by (17.12), where, of course, we use the Tobit log-likelihood functions for the restricted and unrestricted models.

Interpreting the Tobit Estimates

Using modern computers, the maximum likelihood estimates for Tobit models are usually not much more difficult to obtain than the OLS estimates of a linear model. Further, the outputs from Tobit and OLS are often similar. This makes it tempting to interpret the $\hat{\beta}_j$ from Tobit as if these were estimates from a linear regression. Unfortunately, things are not so easy.

From equation (17.18), we see that the β_j measure the partial effects of the x_j on $E(y^*|\mathbf{x})$, where y^* is the latent variable. Sometimes, y^* has an interesting economic meaning, but more often it does not. The variable we want to explain is y, as this is the observed outcome (such as hours worked or amount of charitable contributions). For example, as a policy matter, we are interested in the sensitivity of hours worked to changes in marginal tax rates.

We can estimate $P(y = 0|\mathbf{x})$ from (17.21), which, of course, allows us to estimate $P(y > 0|\mathbf{x})$. What happens if we want to estimate the expected value of y as a function of $\mathbf{x}$? In Tobit models, two expectations are of particular interest: $E(y|y > 0,\mathbf{x})$, which is sometimes called the "conditional expectation" because it is conditional on $y > 0$, and $E(y|\mathbf{x})$, which is, unfortunately, called the "unconditional expectation." (Both expectations are conditional on the explanatory variables.) The expectation $E(y|y > 0,\mathbf{x})$ tells us, for given values of $\mathbf{x}$, the expected value of y for the subpopulation where y is positive. Given $E(y|y > 0,\mathbf{x})$, we can easily find $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = P(y > 0|\mathbf{x}) \cdot E(y|y > 0,\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma) \cdot E(y|y > 0,\mathbf{x}). \qquad (17.23)$$

To obtain $E(y|y > 0,\mathbf{x})$, we use a result for normally distributed random variables: if $z \sim \text{Normal}(0,1)$, then $E(z|z > c) = \phi(c)/[1 - \Phi(c)]$ for any constant c. But $E(y|y > 0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + E(u|u > -\mathbf{x}\boldsymbol{\beta}) = \mathbf{x}\boldsymbol{\beta} + \sigma E[(u/\sigma)|(u/\sigma) > -\mathbf{x}\boldsymbol{\beta}/\sigma] = \mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$, because $\phi(-c) = \phi(c)$, $1 - \Phi(-c) = \Phi(c)$, and u/σ has a standard normal distribution independent of $\mathbf{x}$.

We can summarize this as

$$E(y|y > 0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma), \tag{17.24}$$

where $\lambda(c) = \phi(c)/\Phi(c)$ is called the **inverse Mills ratio**; it is the ratio between the standard normal pdf and standard normal cdf, each evaluated at c.

Equation (17.24) is important. It shows that the expected value of y conditional on $y > 0$ is equal to $\mathbf{x}\boldsymbol{\beta}$ plus a strictly positive term, which is σ times the inverse Mills ratio evaluated at $\mathbf{x}\boldsymbol{\beta}/\sigma$. This equation also shows why using OLS only for observations where $y_i > 0$ will not always consistently estimate $\boldsymbol{\beta}$; essentially, the inverse Mills ratio is an omitted variable, and it is generally correlated with the elements of $\mathbf{x}$.

Combining (17.23) and (17.24) gives

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)] = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \tag{17.25}$$

where the second equality follows because $\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma) = \phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$. This equation shows that when y follows a Tobit model, $E(y|\mathbf{x})$ is a nonlinear function of $\mathbf{x}$ *and* $\boldsymbol{\beta}$. Although it is not obvious, the right-hand side of equation (17.25) can be shown to be positive for any values of $\mathbf{x}$ and $\boldsymbol{\beta}$. Therefore, once we have estimates of $\boldsymbol{\beta}$, we can be sure that predicted values for y—that is, estimates of $E(y|\mathbf{x})$—are positive. The cost of ensuring positive predictions for y is that equation (17.25) is more complicated than a linear model for $E(y|\mathbf{x})$. Even more importantly, the partial effects from (17.25) are more complicated than for a linear model. As we will see, the partial effects of x_j on $E(y|y > 0,\mathbf{x})$ and $E(y|\mathbf{x})$ have the same sign as the coefficient, β_j, but the magnitude of the effects depends on the values of *all* explanatory variables and parameters. Because σ appears in (17.25), it is not surprising that the partial effects depend on σ, too.

If x_j is a continuous variable, we can find the partial effects using calculus. First,

$$\partial E(y|y > 0,\mathbf{x})/\partial x_j = \beta_j + \beta_j \cdot \frac{d\lambda}{dc}(\mathbf{x}\boldsymbol{\beta}/\sigma),$$

assuming that x_j is not functionally related to other regressors. By differentiating $\lambda(c) = \phi(c)/\Phi(c)$ and using $d\Phi/dc = \phi(c)$ and $d\phi/dc = -c\phi(c)$, it can be shown that $d\lambda/dc = -\lambda(c)[c + \lambda(c)]$. Therefore,

$$\partial E(y|y > 0,\mathbf{x})/\partial x_j = \beta_j\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}. \tag{17.26}$$

This shows that the partial effect of x_j on $E(y|y > 0,\mathbf{x})$ is not determined just by β_j. The adjustment factor is given by the term in brackets, $\{\cdot\}$, and depends on a linear function of $\mathbf{x}$, $\mathbf{x}\boldsymbol{\beta}/\sigma = (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)/\sigma$. It can be shown that the adjustment factor is strictly between zero and one. In practice, we can estimate (17.26) by plugging in the MLEs of the β_j and σ. As with logit and probit models, we must plug in values for the x_j, usually the mean values or other interesting values. Equation (17.26) reveals a subtle point that is sometimes lost in applying the Tobit model to corner solution responses: the

parameter σ appears directly in the partial effects, so having an estimate of σ is crucial for estimating the partial effects. Sometimes, σ is called an "ancillary" parameter (which means it is auxiliary, or unimportant). Although it is true that the value of σ does not affect the sign of the partial effects, it does affect the magnitudes, and we are often interested in the economic importance of the explanatory variables. Therefore, characterizing σ as ancillary is misleading and comes from a confusion between the Tobit model for corner solution applications and applications to true data censoring. (See Section 17.4.)

All of the usual economic quantities, such as elasticities, can be computed. For example, the elasticity of y with respect to x_1, conditional on $y > 0$, is

$$\frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial x_1} \cdot \frac{x_1}{\mathrm{E}(y|y > 0, \mathbf{x})}. \tag{17.27}$$

This can be computed when x_1 appears in various functional forms, including level, logarithmic, and quadratic forms.

If x_1 is a binary variable, the effect of interest is obtained as the difference between $\mathrm{E}(y|y > 0, \mathbf{x})$, with $x_1 = 1$ and $x_1 = 0$. Partial effects involving other discrete variables (such as number of children) can be handled similarly.

We can use (17.25) to find the partial derivative of $\mathrm{E}(y|\mathbf{x})$ with respect to continuous x_j. This derivative accounts for the fact that people starting at $y = 0$ might choose $y > 0$ when x_j changes:

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \frac{\partial \mathrm{P}(y > 0|\mathbf{x})}{\partial x_j} \cdot \mathrm{E}(y|y > 0, \mathbf{x}) + \mathrm{P}(y > 0|\mathbf{x}) \cdot \frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial x_j}. \tag{17.28}$$

Because $\mathrm{P}(y > 0|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$,

$$\frac{\partial \mathrm{P}(y > 0|\mathbf{x})}{\partial x_j} = (\beta_j/\sigma)\phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \tag{17.29}$$

so we can estimate each term in (17.28), once we plug in the MLEs of the β_j and σ and particular values of the x_j.

Remarkably, when we plug (17.26) and (17.29) into (17.28) and use the fact that $\Phi(c)\lambda(c) = \phi(c)$ for any c, we obtain

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \beta_j \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \tag{17.30}$$

Equation (17.30) allows us to roughly compare OLS and Tobit estimates. [Equation (17.30) also can be derived directly from equation (17.25) using the fact that $\mathrm{d}\phi(z)/\mathrm{d}z = -z\phi(z)$.] The OLS slope coefficients, say, $\hat{\gamma}_j$, from the regression of y_i on $x_{i1}, x_{i2}, \ldots, x_{ik}$,

$i = 1,...,n$—that is, using all of the data—are direct estimates of $\partial E(y|\mathbf{x})/\partial x_j$. To make the Tobit coefficient, $\hat{\beta}_j$, comparable to $\hat{\gamma}_j$, we must multiply $\hat{\beta}_j$ by an adjustment factor.

As in the probit and logit cases, there are two common approaches for computing an adjustment factor for the coefficients on the continuous explanatory variables. First, we can evaluate $\Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma})$ at the sample averages to obtain $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Or, second, we can average the individual adjustment factors: $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$. For comparing scaled Tobit coefficients to OLS coefficients, the latter scale factor generally is more appropriate. Because $P(y > 0|\mathbf{x}) = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ both scale factors will tend to be closer to one when there are relatively few observations with $y_i = 0$. In the extreme case that all $y_i > 0$, the Tobit and OLS estimates are identical.

Unfortunately, for discrete explanatory variables, comparing OLS and Tobit estimates is not so easy (although using the scale factor for continuous explanatory variables often is a useful approximation). For Tobit, the partial effect of a discrete explanatory variable, for example, a binary variable, should really be obtained by estimating $E(y|\mathbf{x})$ from equation (17.25). For example, if x_1 is a binary, we should first plug in $x_1 = 1$ and then $x_1 = 0$. If we set the other explanatory variables at their sample averages, we obtain a measure analogous to (17.16) for the logit and probit cases. If we compute the difference in expected values for each individual, and then average the difference, we get a measure analogous to (17.17).

EXAMPLE 17.2

(Married Women's Annual Labor Supply)

The file MROZ.RAW includes data on hours worked for 753 married women, 428 of whom worked for a wage outside the home during the year; 325 of the women worked zero hours. For the women who worked positive hours, the range is fairly broad, extending from 12 to 4,950. Thus, annual hours worked is a good candidate for a Tobit model. We also estimate a linear model (using all 753 observations) by OLS. The results are given in Table 17.2.

This table has several noteworthy features. First, the Tobit coefficient estimates have the same sign as the corresponding OLS estimates, and the statistical significance of the estimates is similar. (Possible exceptions are the coefficients on *nwifeinc* and *kidsge6*, but the *t* statistics have similar magnitudes.) Second, though it is tempting to compare the magnitudes of the OLS and Tobit estimates, this is not very informative. We must be careful not to think that, because the Tobit coefficient on *kidslt6* is roughly twice that of the OLS coefficient, the Tobit model implies a much greater response of hours worked to young children.

We can multiply the Tobit estimates by appropriate adjustment factors to make them roughly comparable to the OLS estimates. The factor $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ turns out to be about .589, which we can use to obtain the average partial effects for the Tobit estimation. If, for example, we multiply the *educ* coefficient by .589 we get .589(80.65) = 47.50 (that is, 47.5 hours more), which is quite a bit larger than the OLS partial effect, about 28.8 hours. So, even for estimating an average effect, the Tobit estimates are notably larger in magnitude than the corresponding OLS estimate. If, instead, we want the estimated effect of another year of education starting at the average values of all explanatory variables, then we compute the scale

TABLE 17.2

OLS and Tobit Estimation of Annual Hours Worked

Dependent Variable: *hours*		
Independent Variables	**Linear (OLS)**	**Tobit (MLE)**
nwifeinc	−3.45 (2.54)	−8.81 (4.46)
educ	28.76 (12.95)	80.65 (21.58)
exper	65.67 (9.96)	131.56 (17.28)
$exper^2$	−.700 (.325)	−1.86 (0.54)
age	−30.51 (4.36)	−54.41 (7.42)
kidslt6	−442.09 (58.85)	−894.02 (111.88)
kidsge6	−32.78 (23.18)	−16.22 (38.64)
constant	1,330.48 (270.78)	965.31 (446.44)
Log-Likelihood Value *R*-Squared $\hat{\sigma}$	— .266 750.18	−3,819.09 .274 1,122.02

factor $\Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. This turns out to be about .645 [when we use the squared average of experience, $(\overline{exper})^2$, rather than the average of $exper^2$] This partial effect, which is about 52 hours, is almost twice as large as the OLS estimate. With the exception of *kidsge6*, the scaled Tobit slope coefficients are all greater in magnitude than the corresponding OLS coefficient.

We have reported an *R*-squared for both the linear regression and the Tobit models. The *R*-squared for OLS is the usual one. For Tobit, the *R*-squared is the square of the correlation

coefficient between y_i and $\hat{y}_i$, where $\hat{y}_i = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})\mathbf{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}\phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ is the estimate of $E(y|\mathbf{x} = \mathbf{x}_i)$. This is motivated by the fact that the usual *R*-squared for OLS is equal to the squared correlation between the y_i and the fitted values [see equation (3.29)]. In nonlinear models such as the Tobit model, the squared correlation coefficient is not identical to an *R*-squared based on a sum of squared residuals as in (3.28). This is because the fitted values, as defined earlier, and the residuals, $y_i - \hat{y}_i$, are not uncorrelated in the sample. An *R*-squared defined as the squared correlation coefficient between y_i and $\hat{y}_i$ has the advantage of always being between zero and one; an *R*-squared based on a sum of squared residuals need not have this feature.

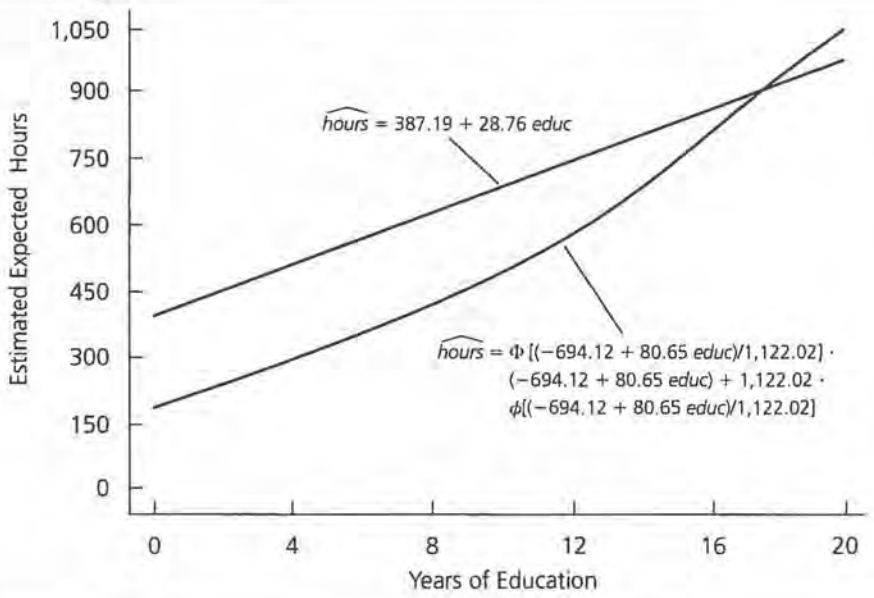
We can see that, based on the *R*-squared measures, the Tobit conditional mean function fits the hours data somewhat, but not substantially, better. However, we should remember that the Tobit estimates are not chosen to maximize an *R*-squared—they maximize the log-likelihood function—whereas the OLS estimates are the values that do produce the highest *R*-squared given the linear functional form.

By construction, all of the Tobit fitted values for *hours* are positive. By contrast, 39 of the OLS fitted values are negative. Although negative predictions are of some concern, 39 out of 753 is just over 5% of the observations. It is not entirely clear how negative fitted values for OLS translate into differences in estimated partial effects. Figure 17.3 plots estimates of $E(hours|\mathbf{x})$ as a function of education; for the Tobit model, the other explanatory variables are set at their average values. For the linear model, the equation graphed is $\widehat{hours} = 387.19 + 28.76\,educ$. For the Tobit model, the equation graphed is $\widehat{hours} = \Phi[(-694.12 + 80.65\,educ)/1{,}122.02] \cdot (-694.12 + 80.65\,educ) + 1{,}122.02 \cdot \phi[(-694.12 + 80.65\,educ)/1{,}122.02]$. As can be seen from the figure, the linear model gives notably higher estimates of the expected hours worked at even fairly high levels of education. For example, at eight years of education, the OLS predicted value of hours is about 617.5, while the Tobit estimate is about 423.9. At 12 years of education, the predicted *hours* are about 732.7 and 598.3, respectively. The two prediction lines cross after 17 years of education, but no woman in the sample has more than 17 years of education. The increasing slope of the Tobit line clearly indicates the increasing marginal effect of education on expected hours worked.

Specification Issues in Tobit Models

The Tobit model, and in particular the formulas for the expectations in (17.24) and (17.25), rely crucially on normality and homoskedasticity in the underlying latent variable model. When $E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, we know from Chapter 5 that conditional normality of y does not play a role in unbiasedness, consistency, or large sample inference. Heteroskedasticity does not affect unbiasedness or consistency of OLS, although we must compute robust standard errors and test statistics to perform approximate inference. In a Tobit model, if any of the assumptions in (17.18) fail, then it is hard to know what the Tobit MLE is estimating. Nevertheless, for moderate departures from the assumptions, the Tobit model is likely to provide good estimates of the partial effects on the conditional means. It is possible to allow for more general assumptions in (17.18), but such models are much more complicated to estimate and interpret.

FIGURE 17.3
Estimated Expected Values of Hours with Respect to Education for the Linear and Tobit Models.



One potentially important limitation of the Tobit model, at least in certain applications, is that the expected value conditional on $y > 0$ is closely linked to the probability that $y > 0$. This is clear from equations (17.26) and (17.29). In particular, the effect of x_j on $P(y > 0|\mathbf{x})$ is proportional to β_j, as is the effect on $E(y|y > 0, \mathbf{x})$, where both functions multiplying β_j are positive and depend on $\mathbf{x}$ only through $\mathbf{x}\boldsymbol{\beta}/\sigma$. This rules out some interesting possibilities. For example, consider the relationship between amount of life insurance coverage and a person's age. Young people may be less likely to have life insurance at all, so the probability that $y > 0$ increases with age (at least up to a point). Conditional on having life insurance, the value of policies might decrease with age, since life insurance becomes less important as people near the end of their lives. This possibility is not allowed for in the Tobit model.

One way to informally evaluate whether the Tobit model is appropriate is to estimate a probit model where the binary outcome, say, w, equals one if $y > 0$, and $w = 0$ if $y = 0$. Then, from (17.21), w follows a probit model, where the coefficient on x_j is $\gamma_j = \beta_j/\sigma$. This means we can estimate the ratio of β_j to σ by probit, for each j. If the Tobit model holds, the probit estimate, $\hat{\gamma}_j$, should be "close" to $\hat{\beta}_j/\hat{\sigma}$, where $\hat{\beta}_j$ and $\hat{\sigma}$ are the Tobit estimates. These will never be identical because of sampling error. But we can look for certain problematic signs. For example, if $\hat{\gamma}_j$ is significant and negative, but $\hat{\beta}_j$ is positive, the Tobit model might not be appropriate. Or, if $\hat{\gamma}_j$ and $\hat{\beta}_j$ are the same sign, but

$|\hat{\beta}_j/\hat{\sigma}|$ is much larger or smaller than $|\hat{\gamma}_j|$, this could also indicate problems. We should not worry too much about sign changes or magnitude differences on explanatory variables that are insignificant in both models.

In the annual hours worked example, $\hat{\sigma} = 1{,}122.02$. When we divide the Tobit co-efficient on *nwifeinc* by $\hat{\sigma}$, we obtain $-8.81/1{,}122.02 \approx -.0079$; the probit coefficient on *nwifeinc* is about $-.012$, which is different, but not dramatically so. On *kidslt6*, the coefficient estimate over $\hat{\sigma}$ is about $-.797$, compared with the probit estimate of $-.868$. Again, this is not a huge difference, but it indicates that having small children has a larger effect on the initial labor force participation decision than on how many hours a woman chooses to work once she is in the labor force. (Tobit effectively averages these two effects together.) We do not know whether the effects are statistically different, but they are of the same order of magnitude.

What happens if we conclude that the Tobit model is inappropriate? There are models, usually called *hurdle* or *two-part* models, that can be used when Tobit seems unsuitable. These all have the property that $P(y > 0|\mathbf{x})$ and $E(y|y > 0,\mathbf{x})$ depend on different parameters, so x_j can have dissimilar effects on these two functions. (See Wooldridge [2002, Chapter 16] for a description of these models.)

17.3 The Poisson Regression Model

Another kind of nonnegative dependent variable is a **count variable**, which can take on nonnegative integer values: {0,1,2,...}. We are especially interested in cases where y takes on relatively few values, including zero. Examples include the number of children ever born to a woman, the number of times someone is arrested in a year, or the number of patents applied for by a firm in a year. For the same reasons discussed for binary and Tobit responses, a linear model for $E(y|x_1,\ldots,x_k)$ might not provide the best fit over all values of the explanatory variables. (Nevertheless, it is always informative to start with a linear model, as we did in Example 3.5.)

As with a Tobit outcome, we cannot take the logarithm of a count variable because it takes on the value zero. A profitable approach is to model the expected value as an exponential function:

$$E(y|x_1,x_2,\ldots,x_k) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k). \tag{17.31}$$

Because $\exp(\cdot)$ is always positive, (17.31) ensures that predicted values for y will also be positive. The exponential function is graphed in Figure A.5 of Appendix A.

Athough (17.31) is more complicated than a linear model, we basically already know how to interpret the coefficients. Taking the log of equation (17.31) shows that

$$\log[E(y|x_1,x_2,\ldots,x_k)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{17.32}$$

so that the log of the expected value is linear. Therefore, using the approximation properties of the log function that we have used often in previous chapters,

$$\%\Delta E(y|\mathbf{x}) \approx (100\beta_j)\Delta x_j.$$

In other words, $100\beta_j$ is roughly the percentage change in $E(y|\mathbf{x})$, given a one-unit increase in x_j. Sometimes, a more accurate estimate is needed, and we can easily find one by looking at discrete changes in the expected value. Keep all explanatory variables except x_k fixed and let x_k^0 be the initial value and x_k^1 the subsequent value. Then, the proportionate change in the expected value is

$$[\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^1)/\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^0)] - 1 = \exp(\beta_k \Delta x_k) - 1,$$

where $\mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1}$ is shorthand for $\beta_1 x_1 + \ldots + \beta_{k-1}x_{k-1}$, and $\Delta x_k = x_k^1 - x_k^0$. When $\Delta x_k = 1$—for example, if x_k is a dummy variable that we change from zero to one—then the change is $\exp(\beta_k) - 1$. Given $\hat{\beta}_k$, we obtain $\exp(\hat{\beta}_k) - 1$ and multiply this by 100 to turn the proportionate change into a percentage change.

By reasoning similar to the linear model, if β_j multiplies $\log(x_j)$, then β_j is an elasticity. The bottom line is that, for practical purposes, we can interpret the coefficients in equation (17.31) as if we have a linear model, with $\log(y)$ as the dependent variable. There are some subtle differences that we need not study here.

Because (17.31) is nonlinear in its parameters—remember, $\exp(\cdot)$ is a nonlinear function—we cannot use linear regression methods. We could use *nonlinear least squares*, which, just as with OLS, minimizes the sum of squared residuals. It turns out, however, that all standard count data distributions exhibit heteroskedasticity, and nonlinear least squares does not exploit this (see Wooldridge [2002, Chapter 12]). Instead, we will rely on maximum likelihood and the important related method of *quasi-maximum likelihood estimation*.

In Chapter 4, we introduced normality as the standard distributional assumption for linear regression. The normality assumption is reasonable for (roughly) continuous dependent variables that can take on a large range of values. A count variable cannot have a normal distribution (because the normal distribution is for continuous variables that can take on all values), and if it takes on very few values, the distribution can be very different from normal. Instead, the nominal distribution for count data is the **Poisson distribution**.

Because we are interested in the effect of explanatory variables on y, we must look at the Poisson distribution conditional on $\mathbf{x}$. The Poisson distribution is entirely determined by its mean, so we only need to specify $E(y|\mathbf{x})$. We assume this has the same form as (17.31), which we write in shorthand as $\exp(\mathbf{x}\boldsymbol{\beta})$. Then, the probability that y equals the value h, conditional on $\mathbf{x}$, is

$$P(y = h|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})][\exp(\mathbf{x}\boldsymbol{\beta})]^h/h!, \; h = 0,1,\ldots,$$

where $h!$ denotes factorial (see Appendix B). This distribution, which is the basis for the **Poisson regression model**, allows us to find conditional probabilities for any values of the explanatory variables. For example, $P(y = 0|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$. Once we have estimates of the β_j, we can plug them into the probabilities for various values of $\mathbf{x}$.

Given a random sample $\{(\mathbf{x}_i, y_i): i = 1,2,\ldots,n\}$, we can construct the log-likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i \boldsymbol{\beta} - \exp(\mathbf{x}_i\boldsymbol{\beta})\}, \tag{17.33}$$

where we drop the term $-\log(y_i!)$ because it does not depend on $\boldsymbol{\beta}$. This log-likelihood function is simple to maximize, although the Poisson MLEs are not obtained in closed form.

The standard errors of the Poisson estimates $\hat{\beta}_j$ are easy to obtain after the log-likelihood function has been maximized; the formula is in the chapter appendix. These are reported along with the $\hat{\beta}_j$ by any software package.

As with the probit, logit, and Tobit models, we cannot directly compare the magnitudes of the Poisson estimates of an exponential function with the OLS estimates of a linear function. Nevertheless, a rough comparison is possible, at least for continuous explanatory variables. If (17.31) holds, then the partial effect of x_j with respect to $E(y|x_1, x_2,...,x_k)$ is $\partial E(y|x_1,x_2,...,x_k)/x_j = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \cdot \beta_j$. This expression follows from the chain rule in calculus because the derivative of the exponential function is just the exponential function. If we let $\hat{\gamma}_j$ denote an OLS slope coefficient from the regression y on $x_1,x_2,...,x_k$, then we can roughly compare the magnitude of the $\hat{\gamma}_j$ and the average partial effect for an exponential regression function, namely, $[n^{-1} \sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik})] \hat{\beta}_j$.

Although Poisson MLE analysis is a natural first step for count data, it is often much too restrictive. All of the probabilities and higher moments of the Poisson distribution are determined entirely by the mean. In particular, the variance is equal to the mean:

$$\text{Var}(y|\mathbf{x}) = \text{E}(y|\mathbf{x}). \tag{17.34}$$

This is restrictive and has been shown to be violated in many applications. Fortunately, the Poisson distribution has a very nice robustness property: whether or not the Poisson distribution holds, we still get consistent, asymptotically normal estimators of the β_j. (See Wooldridge [2002, Chapter 19] for details.) This is analogous to the OLS estimator, which is consistent and asymptotically normal whether or not the normality assumption holds; yet OLS is the MLE under normality.

When we use Poisson MLE, but we do not assume that the Poisson distribution is entirely correct, we call the analysis **quasi-maximum likelihood estimation** (**QMLE**). The Poisson QMLE is very handy because it is programmed in many econometrics packages. However, unless the Poisson variance assumption (17.34) holds, the standard errors need to be adjusted.

A simple adjustment to the standard errors is available when we assume that the variance is proportional to the mean:

$$\text{Var}(y|\mathbf{x}) = \sigma^2 \text{E}(y|\mathbf{x}), \tag{17.35}$$

where $\sigma^2 > 0$ is an unknown parameter. When $\sigma^2 = 1$, we obtain the Poisson variance assumption. When $\sigma^2 > 1$, the variance is greater than the mean for all $\mathbf{x}$; this is called **overdispersion** because the variance is larger than in the Poisson case, and it is observed in many applications of count regressions. The case $\sigma^2 < 1$, called *underdispersion*, is less common but is allowed in (17.35).

Under (17.35), it is easy to adjust the usual Poisson MLE standard errors. Let $\hat{\beta}_j$ denote the Poisson QMLE and define the residuals as $\hat{u}_i = y_i - \hat{y}_i$, where $\hat{y}_i = \exp(\hat{\beta}_0 +$

$\hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$) is the fitted value. As usual, the residual for observation i is the difference between y_i and its fitted value. A consistent estimator of σ^2 is $(n-k-1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2/\hat{y}_i$, where the division by $\hat{y}_i$ is the proper heteroskedasticity adjustment, and $n-k-1$ is the *df* given n observations and $k+1$ estimates $\hat{\beta}_0$, $\hat{\beta}_1, \ldots, \hat{\beta}_k$. Letting $\hat{\sigma}$ be the positive square root of $\hat{\sigma}^2$, we multiply the usual Poisson standard errors by $\hat{\sigma}$. If $\hat{\sigma}$ is notably greater than one, the corrected standard errors can be much bigger than the nominal, generally incorrect, Poisson MLE standard errors.

Even (17.35) is not entirely general. Just as in the linear model, we can obtain standard errors for the Poisson QMLE that do not restrict the variance at all. (See Wooldridge [2002, Chapter 19] for further explanation.)

Under the Poisson distributional assumption, we can use the likelihood ratio statistic to test exclusion restrictions, which, as always, has the form in (17.12). If we have q exclusion restrictions, the statistic is distributed approximately as χ_q^2 under the null. Under the less restrictive assumption (17.35), a simple adjustment is available (and then we call the statistic the **quasi-likelihood ratio statistic**): we divide (17.12) by $\hat{\sigma}^2$, where $\hat{\sigma}^2$ is obtained from the unrestricted model.

QUESTION 17.4

Suppose that we obtain $\hat{\sigma}^2 = 2$. How will the adjusted standard errors compare with the usual Poisson MLE standard errors? How will the quasi-*LR* statistic compare with the usual *LR* statistic?

EXAMPLE 17.3

(Poisson Regression for Number of Arrests)

We now apply the Poisson regression model to the arrest data in CRIME1.RAW, used, among other places, in Example 9.1. The dependent variable, *narr86*, is the number of times a man is arrested during 1986. This variable is zero for 1,970 of the 2,725 men in the sample, and only eight values of *narr86* are greater than five. Thus, a Poisson regression model is more appropriate than a linear regression model. Table 17.3 also presents the results of OLS estimation of a linear regression model.

The standard errors for OLS are the usual ones; we could certainly have made these robust to heteroskedasticity. The standard errors for Poisson regression are the usual maximum likelihood standard errors. Because $\hat{\sigma} = 1.232$, the standard errors for Poisson regression should be inflated by this factor (so each corrected standard error is about 23% higher). For example, a more reliable standard error for *tottime* is 1.23(.015) ≈ .0185, which gives a t statistic of about 1.3. The adjustment to the standard errors reduces the significance of all variables, but several of them are still very statistically significant.

The OLS and Poisson coefficients are not directly comparable, and they have very different meanings. For example, the coefficient on *pcnv* implies that, if $\Delta pcnv = .10$, the expected number of arrests falls by .013 (*pcnv* is the proportion of prior arrests that led to conviction). The Poisson coefficient implies that $\Delta pcnv = .10$ reduces expected arrests by about 4% [.402(.10) = .0402, and we multiply this by 100 to get the percentage effect]. As a policy matter, this suggests we can reduce overall arrests by about 4% if we can increase the probability of conviction by .1.

TABLE 17.3
Determinants of Number of Arrests for Young Men

Dependent Variable: *narr86*		
Independent Variables	**Linear (OLS)**	**Exponential (Poisson QMLE)**
pcnv	−.132 (.040)	−.402 (.085)
avgsen	−.011 (.012)	−.024 (.020)
tottime	.012 (.009)	.024 (.015)
ptime86	−.041 (.009)	−.099 (.021)
qemp86	−.051 (.014)	−.038 (.029)
inc86	−.0015 (.0003)	−.0081 (.0010)
black	.327 (.045)	.661 (.074)
hispan	.194 (.040)	.500 (.074)
born60	−.022 (.033)	−.051 (.064)
constant	.577 (.038)	−.600 (.067)
Log-Likelihood Value R-Squared $\hat{\sigma}$	— .073 .829	−2,248.76 .077 1.232

The Poisson coefficient on *black* implies that, other factors being equal, the expected number of arrests for a black man is estimated to be about $100 \cdot [\exp(.661) - 1] \approx 93.7\%$ higher than for a white man with the same values for the other explanatory variables.

As with the Tobit application in Table 17.2, we report an *R*-squared for Poisson regression: the squared correlation coefficient between y_i and $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$. The motivation for this goodness-of-fit measure is the same as for the Tobit model. We see that the exponential regression model, estimated by Poisson QMLE, fits slightly better. Remember that the OLS estimates are chosen to maximize the *R*-squared, but the Poisson estimates are not. (They are selected to maximize the log-likelihood function.)

Other count data regression models have been proposed and used in applications, which generalize the Poisson distribution in a variety of ways. If we are interested in the effects of the x_j on the mean response, there is little reason to go beyond Poisson regression: it is simple, often gives good results, and has the robustness property discussed earlier. In fact, we could apply Poisson regression to a y that is a Tobit-like outcome, provided (17.31) holds. This might give good estimates of the mean effects. Extensions of Poisson regression are more useful when we are interested in estimating probabilities, such as $P(y > 1|\mathbf{x})$. (See, for example, Cameron and Trivedi [1998].)

17.4 Censored and Truncated Regression Models

The models in Sections 17.1, 17.2, and 17.3 apply to various kinds of limited dependent variables that arise frequently in applied econometric work. In using these methods, it is important to remember that we use a probit or logit model for a binary response, a Tobit model for a corner solution outcome, or a Poisson regression model for a count response because we want models that account for important features of the distribution of y. There is no issue of data observability. For example, in the Tobit application to women's labor supply in Example 17.2, there is no problem with observing hours worked: it is simply the case that a nontrivial fraction of married women in the population choose not to work for a wage. In the Poisson regression application to annual arrests, we observe the dependent variable for every young man in a random sample from the population, but the dependent variable can be zero as well as other small integer values.

Unfortunately, the distinction between lumpiness in an outcome variable (such as taking on the value zero for a nontrivial fraction of the population) and problems of data censoring can be confusing. This is particularly true when applying the Tobit model. In this book, the standard Tobit model described in Section 17.2 is only for corner solution outcomes. But the literature on Tobit models usually treats another situation within the same framework: the response variable has been censored above or below some threshold. Typically, the censoring is due to survey design and, in some cases, institutional constraints. Rather than treat data censoring problems along with corner solution outcomes, we solve data censoring by applying a **censored regression model**. Essentially, the problem solved by a censored regression model is one of missing data on the response variable, y, but where we have information about the variable when it is missing, namely, whether it is above or below a known threshold.

A **truncated regression model** arises when we exclude, on the basis of y, a subset of the population in our sampling scheme. In other words, we do not have a random sample from the underlying population, but we know the rule that was used to include units in the sample. This rule is determined by whether y is above or below a certain threshold. We explain more fully the difference between censored and truncated regression models later.

Censored Regression Models

While censored regression models can be defined without distributional assumptions, in this subsection we study the **censored normal regression model**. The variable we would like to explain, y, follows the classical linear model. For emphasis, we put an i subscript on a random draw from the population:

$$y_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i,\ u_i|\mathbf{x}_i,c_i \sim \text{Normal}(0,\sigma^2) \tag{17.36}$$

$$w_i = \min(y_i,c_i). \tag{17.37}$$

Rather than observing y_i, we only observe it if it is less than a censoring value, c_i. Notice that (17.36) includes the assumption that u_i is independent of c_i. (For concreteness, we explicitly consider censoring from above, or *right censoring*; the problem of censoring from below, or *left censoring*, is handled similarly.)

QUESTION 17.5

Let mvp_i be the marginal value product for worker i; this is the price of a firm's good multiplied by the marginal product of the worker. Assume mvp_i is a linear function of exogenous variables, such as education, experience, and so on, and an unobservable error. Under perfect competition and without institutional constraints, each worker is paid his or her marginal value product. Let $minwage_i$ denote the minimum wage for worker i, which varies by state. We observe $wage_i$, which is the larger of mvp_i and $minwage_i$. Write the appropriate model for the observed wage.

One example of right data censoring is **top coding**. When a variable is top coded, we know its value only up to a certain threshold. For responses greater than the threshold, we only know that the variable is at least as large as the threshold. For example, in some surveys, family wealth is top coded. Suppose that respondents are asked their wealth, but people are allowed to respond with "more than $500,000." Then, we observe actual wealth for those respondents whose wealth is less than $500,000 but not for those whose wealth is greater than $500,000. In this case, the censoring threshold, c_i, is the same for all i. In many situations, the censoring threshold changes with individual or family characteristics.

If we observed a random sample for $(\mathbf{x},y)$, we would simply estimate $\boldsymbol{\beta}$ by OLS, and statistical inference would be standard. (We again absorb the intercept into $\mathbf{x}$ for simplicity.) The censoring causes problems. Using arguments similar to the Tobit model, an OLS regression using only the uncensored observations—that is, those with $y_i < c_i$—produces inconsistent estimators of the β_j. An OLS regression of w_i on $\mathbf{x}_i$, using all observations, does not consistently estimate the β_j, unless there is no censoring. This is similar to the Tobit case, but the problem is much different. In the Tobit model, we are modeling economic behavior, which often yields zero outcomes; the Tobit model is supposed to reflect this. With censored regression, we have a data collection problem because, for some reason, the data are censored.

Under the assumptions in (17.36) and (17.37), we can estimate $\boldsymbol{\beta}$ (and σ^2) by maximum likelihood, given a random sample on $(\mathbf{x}_i,w_i)$. For this, we need the density of w_i, given $(\mathbf{x}_i,c_i)$. For uncensored observations, $w_i = y_i$, and the density of w_i is the same as that for y_i: Normal$(\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$. For censored observations, we need the probability that w_i equals the censoring value, c_i, given $\mathbf{x}_i$:

$$P(w_i = c_i|\mathbf{x}_i) = P(y_i \geq c_i|\mathbf{x}_i) = P(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma].$$

We can combine these two parts to obtain the density of w_i, given $\mathbf{x}_i$ and c_i:

$$f(w|\mathbf{x}_i,c_i) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \quad w = c_i, \tag{17.38}$$

$$= (1/\sigma)\phi\,[(w - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \quad w < c_i. \tag{17.39}$$

The log-likelihood for observation i is obtained by taking the natural log of the density for each i. We can maximize the sum of these across i, with respect to the β_j and σ, to obtain the MLEs.

It is important to know that we can interpret the β_j just as in a linear regression model under random sampling. This is much different than the Tobit applications, where the expectations of interest are nonlinear functions of the β_j.

An important application of censored regression models is **duration analysis**. A *duration* is a variable that measures the time before a certain event occurs. For example, we might wish to explain the number of days before a felon released from prison is arrested. For some felons, this may never happen, or it may happen after such a long time that we must censor the duration in order to analyze the data.

In duration applications of censored normal regression, as well as in top coding, we often use the natural log as the dependent variable, which means we also take the log of the censoring threshold in (17.37). As we have seen throughout this text, using the log transformation for the dependent variable causes the parameters to be interpreted as percentage changes. Further, as with many positive variables, the log of a duration typically has a distribution closer to normal than the duration itself.

EXAMPLE 17.4

(Duration of Recidivism)

The file RECID.RAW contains data on the time in months until an inmate in a North Carolina prison is arrested after being released from prison; call this *durat*. Some inmates participated in a work program while in prison. We also control for a variety of demographic variables, as well as for measures of prison and criminal history.

Of 1,445 inmates, 893 had not been arrested during the period they were followed; therefore, these observations are censored. The censoring times differed among inmates, ranging from 70 to 81 months.

Table 17.4 gives the results of censored normal regression for log(*durat*). Each of the coefficients, when multiplied by 100, gives the estimated percentage change in expected duration given a ceteris paribus increase of one unit in the corresponding explanatory variable.

TABLE 17.4

Censored Regression Estimation of Criminal Recidivism

Dependent Variable: log(*durat*)	
Independent Variables	**Coefficient (Standard Error)**
workprg	−.063 (.120)
priors	−.137 (.021)
tserved	−.019 (.003)
felon	.444 (.145)
alcohol	−.635 (.144)
drugs	−.298 (.133)
black	−.543 (.117)
married	.341 (.140)
educ	.023 (.025)
age	.0039 (.0006)
constant	4.099 (.348)
Log-Likelihood Value $\hat{\sigma}$	−1,597.06 1.810

Several of the coefficients in Table 17.4 are interesting. The variables *priors* (number of prior convictions) and *tserved* (total months spent in prison) have negative effects on the time until the next arrest occurs. This suggests that these variables measure proclivity for criminal activity rather than representing a deterrent effect. For example, an inmate with one more prior conviction has a duration until next arrest that is almost 14% less. A year of time served reduces duration by about $100 \cdot 12(.019) = 22.8\%$. A somewhat surprising finding is that a man serving time for a felony has an estimated expected duration that is almost 56% ($\exp(.444) - 1 \approx .56$) *longer* than a man serving time for a nonfelony.

Those with a history of drug or alcohol abuse have substantially shorter expected durations until the next arrest. (The variables *alcohol* and *drugs* are binary variables.) Older men, and men who were married at the time of incarceration, are expected to have significantly longer durations until their next arrest. Black men have substantially shorter durations, on the order of 42% [$\exp(-.543) - 1 \approx -.42$].

The key policy variable, *workprg*, does not have the desired effect. The point estimate is that, other things being equal, men who participated in the work program have estimated recidivism durations that are about 6.3% shorter than men who did not participate. The coefficient has a small t statistic, so we would probably conclude that the work program has no effect. This could be due to a self-selection problem, or it could be a product of the way men were assigned to the program. Of course, it may simply be that the program was ineffective.

In this example, it is crucial to account for the censoring, especially because almost 62% of the durations are censored. If we apply straight OLS to the entire sample and treat the censored durations as if they were uncensored, the coefficient estimates are markedly different. In fact, they are all shrunk toward zero. For example, the coefficient on *priors* becomes $-.059$ (se = .009), and that on *alcohol* becomes $-.262$ (se = .060). Although the directions of the effects are the same, the importance of these variables is greatly diminished. The censored regression estimates are much more reliable.

There are other ways of measuring the effects of each of the explanatory variables in Table 17.4 on the duration, rather than focusing only on the expected duration. A treatment of modern duration analysis is beyond the scope of this text. (For an introduction, see Wooldridge [2002, Chapter 20].)

If any of the assumptions of the censored normal regression model are violated—in particular, if there is heteroskedasticity or nonnormality—the MLEs are generally inconsistent. This shows that the censoring is potentially very costly, as OLS using an uncensored sample requires neither normality nor homoskedasticity for consistency. There are methods that do not require us to assume a distribution, but they are more advanced. (See Wooldridge [2002, Chapter 16].)

Truncated Regression Models

A truncated regression model is similar to a censored regression model, but it differs in one major respect: in a truncated regression model, we do not observe any information about a certain segment of the population. This typically happens when a survey targets a particular subset of the population and, perhaps due to cost considerations, entirely ignores the other part of the population.

For example, Hausman and Wise (1977) used data from a negative income tax experiment to study various determinants of earnings. To be included in the study, a family had to have income less than 1.5 times the 1967 poverty line, where the poverty line depended on family size.

The **truncated normal regression model** begins with an underlying population model that satisfies the classical linear model assumptions:

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u,\ u|\mathbf{x} \sim \text{Normal}(0,\sigma^2). \tag{17.40}$$

Recall that this is a strong set of assumptions, because u must not only be independent of $\mathbf{x}$, but also normally distributed. We focus on this model because relaxing the assumptions is difficult.

Under (17.40) we know that, given a random sample from the population, OLS is the most efficient estimation procedure. The problem arises because we do not observe a random sample from the population: Assumption MLR.2 is violated. In particular, a random draw $(\mathbf{x}_i,y_i)$ is observed only if $y_i \leq c_i$, where c_i is the truncation threshold that can depend on exogenous variables—in particular, the $\mathbf{x}_i$. (In the Hausman and Wise example, c_i depends on family size.) This means that, if $\{(\mathbf{x}_i,y_i): i = 1,\ldots,n\}$ is our *observed* sample, then y_i is necessarily less than or equal to c_i. This differs from the censored regression model: in a censored regression model, we observe $\mathbf{x}_i$ for any randomly drawn observation from the population; in the truncated model, we only observe $\mathbf{x}_i$ if $y_i \leq c_i$.

To estimate the β_j (along with σ), we need the distribution of y_i, given that $y_i \leq c_i$ and $\mathbf{x}_i$. This is written as

$$g(y|\mathbf{x}_i,c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)},\ y \leq c_i, \tag{17.41}$$

where $f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ denotes the normal density with mean $\beta_0 + \mathbf{x}_i\boldsymbol{\beta}$ and variance σ^2, and $F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ is the normal cdf with the same mean and variance, evaluated at c_i. This expression for the density, conditional on $y_i \leq c_i$, makes intuitive sense: it is the population density for y, given $\mathbf{x}$, divided by the probability that y_i is less than or equal to c_i (given $\mathbf{x}_i$), $P(y_i \leq c_i|\mathbf{x}_i)$. In effect, we renormalize the density by dividing by the area under $f(\cdot|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ that is to the left of c_i.

If we take the log of (17.41), sum across all i, and maximize the result with respect to the β_j and σ^2, we obtain the maximum likelihood estimators. This leads to consistent, approximately normal estimators. The inference, including standard errors and log-likelihood statistics, is standard.

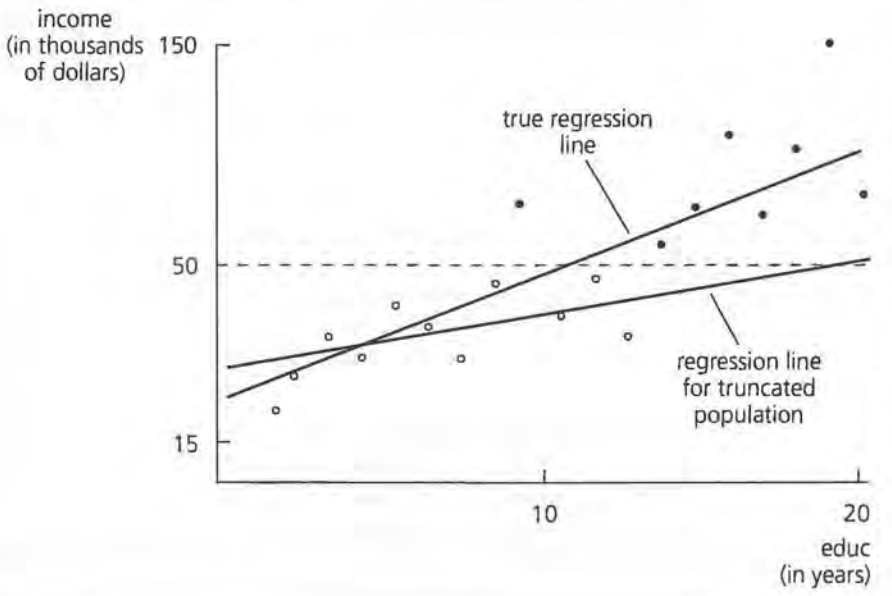
We could analyze the data from Example 17.4 as a truncated sample if we drop all data on an observation whenever it is censored. This would give us 552 observations from a truncated normal distribution, where the truncation point differs across i. However, we would never analyze duration data (or top-coded data) in this way, as it eliminates useful information. The fact that we know a lower bound for 893 durations, along with the explanatory variables, is useful information; censored regression uses this information, while truncated regression does not.

A better example of truncated regression is given in Hausman and Wise (1977), where they emphasize that OLS applied to a sample truncated from above generally produces estimators biased toward zero. Intuitively, this makes sense. Suppose that the relationship of interest is between income and education levels. If we only observe people whose income is below a certain threshold, we are lopping off the upper end. This tends to flatten the estimated line relative to the true regression line in the whole population. Figure 17.4 illustrates the problem when income is truncated from above at $50,000. Although we observe the data points represented by the open circles, we do not observe the data sets represented by the darkened circles. A regression analysis using the truncated sample does not lead to consistent estimators. Incidentally, if the sample in Figure 17.4 was censored rather than truncated—that is, we had top-coded data—we would observe education levels for all points in Figure 17.4, but for individuals with incomes above $50,000 we would not know the exact income amount. We would only know that income was at least $50,000. In effect, all observations represented by the darkened circles would be brought down to the horizontal line at *income* = 50.

As with censored regression, if the underlying homoskedastic normal assumption in (17.40) is violated, the truncated normal MLE is biased and inconsistent. Methods that do not require these assumptions are available; see Wooldridge (2002, Chapter 17) for discussion and references.

FIGURE 17.4

A true, or population, regression line and the incorrect regression line for the truncated population with incomes below $50,000.



17.5 Sample Selection Corrections

Truncated regression is a special case of a general problem known as **nonrandom sample selection**. But survey design is not the only cause of nonrandom sample selection. Often, respondents fail to provide answers to certain questions, which leads to missing data for the dependent or independent variables. Because we cannot use these observations in our estimation, we should wonder whether dropping them leads to bias in our estimators.

Another general example is usually called **incidental truncation**. Here, we do not observe y because of the outcome of another variable. The leading example is estimating the so-called *wage offer function* from labor economics. Interest lies in how various factors, such as education, affect the wage an individual could earn in the labor force. For people who are in the workforce, we observe the wage offer as the current wage. But, for those currently out of the workforce, we do not observe the wage offer. Because working may be systematically correlated with unobservables that affect the wage offer, using only working people—as we have in all wage examples so far—might produce biased estimators of the parameters in the wage offer equation.

Nonrandom sample selection can also arise when we have panel data. In the simplest case, we have two years of data, but, due to attrition, some people leave the sample. This is particularly a problem in policy analysis, where attrition may be related to the effectiveness of a program.

When Is OLS on the Selected Sample Consistent?

In Section 9.4, we provided a brief discussion of the kinds of sample selection that can be ignored. The key distinction is between *exogenous* and *endogenous* sample selection. In the truncated Tobit case, we clearly have endogenous sample selection, and OLS is biased and inconsistent. On the other hand, if our sample is determined solely by an exogenous explanatory variable, we have exogenous sample selection. Cases between these extremes are less clear, and we now provide careful definitions and assumptions for them. The population model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \; \mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0. \tag{17.42}$$

It is useful to write the population model for a *random* draw as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{17.43}$$

where we use $\mathbf{x}_i\boldsymbol{\beta}$ as shorthand for $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. Now, let n be the size of a *random sample* from the population. If we could observe y_i and each x_{ij} for all i, we would simply use OLS. Assume that, for some reason, either y_i or some of the independent variables are not observed for certain i. For at least some observations, we observe the full set of variables. Define a *selection indicator* s_i for each i by $s_i = 1$ if we observe all of $(y_i, \mathbf{x}_i)$, and $s_i = 0$ otherwise. Thus, $s_i = 1$ indicates that we will use the observation in our analysis; $s_i = 0$ means the observation will not be used. We are interested in the

statistical properties of the OLS estimators using the **selected sample**, that is, using observations for which $s_i = 1$. Therefore, we use fewer than n observations, say, n_1.

It turns out to be easy to obtain conditions under which OLS is consistent (and even unbiased). Effectively, rather than estimating (17.43), we can only estimate the equation

$$s_i y_i = s_i \mathbf{x}_i \boldsymbol{\beta} + s_i u_i. \tag{17.44}$$

When $s_i = 1$, we simply have (17.43); when $s_i = 0$, we simply have $0 = 0 + 0$, which clearly tells us nothing about $\boldsymbol{\beta}$. Regressing $s_i y_i$ on $s_i \mathbf{x}_i$ for $i = 1,2,\ldots,n$ is the same as regressing y_i on $\mathbf{x}_i$ using the observations for which $s_i = 1$. Thus, we can learn about the consistency of the $\hat{\beta}_j$ by studying (17.44) on a random sample.

From our analysis in Chapter 5, the OLS estimators from (17.44) are consistent if the error term has zero mean and is uncorrelated with each explanatory variable. In the population, the zero mean assumption is $\mathrm{E}(su) = 0$, and the zero correlation assumptions can be stated as

$$\mathrm{E}[(sx_j)(su)] = \mathrm{E}(sx_j u) = 0, \tag{17.45}$$

where s, x_j, and u are random variables representing the population; we have used the fact that $s^2 = s$ because s is a binary variable. Condition (17.45) is different from what we need if we observe all variables for a random sample: $\mathrm{E}(x_j u) = 0$. Therefore, in the population, we need u to be uncorrelated with sx_j.

The key condition for unbiasedness is $\mathrm{E}(su|sx_1, \ldots, sx_k) = 0$. As usual, this is a stronger assumption than that needed for consistency.

If s is a function only of the explanatory variables, then sx_j is just a function of x_1, $x_2, \ldots, x_k$; by the conditional mean assumption in (17.42), sx_j is also uncorrelated with u. In fact, $\mathrm{E}(su|sx_1, \ldots, sx_k) = s\mathrm{E}(u|sx_1, \ldots, sx_k) = 0$, because $\mathrm{E}(u|x_1, \ldots, x_k) = 0$. This is the case of **exogenous sample selection**, where $s_i = 1$ is determined entirely by $x_{i1}, \ldots, x_{ik}$. As an example, if we are estimating a wage equation where the explanatory variables are education, experience, tenure, gender, marital status, and so on—which are assumed to be exogenous—we can select the sample on the basis of any or all of the explanatory variables.

If sample selection is entirely random in the sense that s_i is *independent* of $(\mathbf{x}_i, u_i)$, then $\mathrm{E}(sx_j u) = \mathrm{E}(s)\mathrm{E}(x_j u) = 0$, because $\mathrm{E}(x_j u) = 0$ under (17.42). Therefore, if we begin with a random sample and randomly drop observations, OLS is still consistent. In fact, OLS is again unbiased in this case, provided there is not perfect multicollinearity in the selected sample.

If s depends on the explanatory variables and additional random terms that are independent of $\mathbf{x}$ and u, OLS is also consistent and unbiased. For example, suppose that IQ score is an explanatory variable in a wage equation, but IQ is missing for some people. Suppose we think that selection can be described by $s = 1$ if $IQ \geq v$, and $s = 0$ if $IQ < v$, where v is an unobserved random variable that is independent of IQ, u, and the other explanatory variables. This means that we are more likely to observe an IQ that is high, but there is always some chance of not observing any IQ. Conditional on the explanatory variables, s is independent of u, which means that $\mathrm{E}(u|x_1, \ldots, x_k, s) = \mathrm{E}(u|x_1, \ldots, x_k)$, and the last expectation is zero by assumption on the population model. If we add the

homoskedasticity assumption $E(u^2|\mathbf{x},s) = E(u^2) = \sigma^2$, then the usual OLS standard errors and test statistics are valid.

So far, we have shown several situations where OLS on the selected sample is unbiased, or at least consistent. When is OLS on the selected sample inconsistent? We already saw one example: regression using a truncated sample. When the truncation is from above, $s_i = 1$ if $y_i \leq c_i$, where c_i is the truncation threshold. Equivalently, $s_i = 1$ if $u_i \leq c_i - \mathbf{x}_i\boldsymbol{\beta}$. Because s_i depends directly on u_i, s_i and u_i will not be uncorrelated, even conditional on $\mathbf{x}_i$. This is why OLS on the selected sample does not consistently estimate the β_j. There are less obvious ways that s and u can be correlated; we consider this in the next subsection.

The results on consistency of OLS extend to instrumental variables estimation. If the IVs are denoted z_h in the population, the key condition for consistency of 2SLS is $E(sz_h u) = 0$, which holds if $E(u|\mathbf{z},s) = 0$. Therefore, if selection is determined entirely by the exogenous variables $\mathbf{z}$, or if s depends on other factors that are independent of u and $\mathbf{z}$, then 2SLS on the selected sample is generally consistent. We do need to assume that the explanatory and instrumental variables are appropriately correlated in the selected part of the population. Wooldridge (2002, Chapter 17) contains precise statements of these assumptions.

It can also be shown that, when selection is entirely a function of the exogenous variables, maximum likelihood estimation of a nonlinear model—such as a logit or probit model—produces consistent, asymptotically normal estimators, and the usual standard errors and test statistics are valid. (Again, see Wooldridge [2002, Chapter 17].)

Incidental Truncation

As we mentioned earlier, a common form of sample selection is called incidental truncation. We again start with the population model in (17.42). However, we assume that we will always observe the explanatory variables x_j. The problem is, we only observe y for a subset of the population. The rule determining whether we observe y does *not* depend directly on the outcome of y. A leading example is when $y = \log(wage^o)$, where $wage^o$ is the *wage offer*, or the hourly wage that an individual could receive in the labor market. If the person is actually working at the time of the survey, then we observe the wage offer because we assume it is the observed wage. But for people out of the workforce, we cannot observe $wage^o$. Therefore, the truncation of wage offer is *incidental* because it depends on another variable, namely, labor force participation. Importantly, we would generally observe all other information about an individual, such as education, prior experience, gender, marital status, and so on.

The usual approach to incidental truncation is to add an explicit selection equation to the population model of interest:

$$y = \mathbf{x}\boldsymbol{\beta} + u, \; E(u|\mathbf{x}) = 0 \tag{17.46}$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0], \tag{17.47}$$

where $s = 1$ if we observe y, and zero otherwise. We assume that elements of $\mathbf{x}$ and $\mathbf{z}$ are always observed, and we write $\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ and $\mathbf{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_m z_m$.

The equation of primary interest is (17.46), and we could estimate $\boldsymbol{\beta}$ by OLS given a random sample. The selection equation, (17.47), depends on observed variables, z_h, and an unobserved error, v. A standard assumption, which we will make, is that $\mathbf{z}$ is exogenous in (17.46):

$$\mathrm{E}(u|\mathbf{x},\mathbf{z}) = 0.$$

In fact, for the following proposed methods to work well, we will require that $\mathbf{x}$ be a strict subset of $\mathbf{z}$: any x_j is also an element of $\mathbf{z}$, and we have some elements of $\mathbf{z}$ that are not also in $\mathbf{x}$. We will see later why this is crucial.

The error term v in the sample selection equation is assumed to be independent of $\mathbf{z}$ (and therefore $\mathbf{x}$). We also assume that v has a standard normal distribution. We can easily see that correlation between u and v generally causes a sample selection problem. To see why, assume that (u,v) is independent of $\mathbf{z}$. Then, taking the expectation of (17.46), conditional on $\mathbf{z}$ and v, and using the fact that $\mathbf{x}$ is a subset of $\mathbf{z}$ gives

$$\mathrm{E}(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|v),$$

where $\mathrm{E}(u|\mathbf{z},v) = \mathrm{E}(u|v)$ because (u,v) is independent of $\mathbf{z}$. Now, if u and v are jointly normal (with zero mean), then $\mathrm{E}(u|v) = \rho v$ for some parameter ρ. Therefore,

$$\mathrm{E}(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \rho v.$$

We do not observe v, but we can use this equation to compute $\mathrm{E}(y|\mathbf{z},s)$ and then specialize this to $s = 1$. We now have:

$$\mathrm{E}(y|\mathbf{z},s) = \mathbf{x}\boldsymbol{\beta} + \rho\mathrm{E}(v|\mathbf{z},s).$$

Because s and v are related by (17.47), and v has a standard normal distribution, we can show that $\mathrm{E}(v|\mathbf{z},s)$ is simply the inverse Mills ratio, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, when $s = 1$. This leads to the important equation

$$\mathrm{E}(y|\mathbf{z},s = 1) = \mathbf{x}\boldsymbol{\beta} + \rho\lambda(\mathbf{z}\boldsymbol{\gamma}). \tag{17.48}$$

Equation (17.48) shows that the expected value of y, given $\mathbf{z}$ *and* observability of y, is equal to $\mathbf{x}\boldsymbol{\beta}$, plus an additional term that depends on the inverse Mills ratio evaluated at $\mathbf{z}\boldsymbol{\gamma}$. Remember, we hope to estimate $\boldsymbol{\beta}$. This equation shows that we can do so using only the selected sample, provided we include the term $\lambda(\mathbf{z}\boldsymbol{\gamma})$ as an additional regressor.

If $\rho = 0$, $\lambda(\mathbf{z}\boldsymbol{\gamma})$ does not appear, and OLS of y on $\mathbf{x}$ using the selected sample consistently estimates $\boldsymbol{\beta}$. Otherwise, we have effectively omitted a variable, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, which is generally correlated with $\mathbf{x}$. When does $\rho = 0$? The answer is when u and v are uncorrelated.

Because $\boldsymbol{\gamma}$ is unknown, we cannot evaluate $\lambda(\mathbf{z}_i\boldsymbol{\gamma})$ for each i. However, from the assumptions we have made, s given z follows a probit model:

$$\mathrm{P}(s = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\gamma}). \tag{17.49}$$

Therefore, we can estimate $\boldsymbol{\gamma}$ by probit of s_i on $\mathbf{z}_i$, using the *entire* sample. In a second step, we can estimate $\boldsymbol{\beta}$. We summarize the procedure, which has recently been dubbed the **Heckit method** in econometrics literature after the work of Heckman (1976).

SAMPLE SELECTION CORRECTION:

(i) Using all n observations, estimate a probit model of s_i on $\mathbf{z}_i$ and obtain the estimates $\hat{\gamma}_h$. Compute the inverse Mills ratio, $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\gamma})$ for each i. (Actually, we only need these for the i with $s_i = 1$.)

(ii) Using the selected sample, that is, the observations for which $s_i = 1$ (say, n_1 of them), run the regression of

$$y_i \text{ on } \mathbf{x}_i, \hat{\lambda}_i. \tag{17.50}$$

The $\hat{\beta}_j$ are consistent and approximately normally distributed.

A simple test of selection bias is available from regression (17.50). Namely, we can use the usual t statistic on $\hat{\lambda}_i$ as a test of H_0: $\rho = 0$. Under H_0, there is no sample selection problem.

When $\rho \neq 0$, the usual OLS standard errors reported from (17.50) are not exactly correct. This is because they do not account for estimation of $\boldsymbol{\gamma}$, which uses the same observations in regression (17.50), and more. Some econometrics packages compute corrected standard errors. (Unfortunately, it is not as simple as a heteroskedasticity adjustment. See Wooldridge [2002, Chapter 6] for further discussion.) In many cases, the adjustments do not lead to important differences, but it is hard to know that beforehand (unless $\hat{\rho}$ is small and insignificant).

We recently mentioned that $\mathbf{x}$ should be a strict subset of $\mathbf{z}$. This has two implications. First, any element that appears as an explanatory variable in (17.46) should also be an explanatory variable in the selection equation. Although in rare cases it makes sense to exclude elements from the selection equation, including all elements of $\mathbf{x}$ in $\mathbf{z}$ is not very costly; excluding them can lead to inconsistency if they are incorrectly excluded.

A second major implication is that we have at least one element of $\mathbf{z}$ that is not also in $\mathbf{x}$. This means that we need a variable that affects selection but does *not* have a partial effect on y. This is not absolutely necessary to apply the procedure—in fact, we can mechanically carry out the two steps when $\mathbf{z} = \mathbf{x}$—but the results are usually less than convincing unless we have an *exclusion restriction* in (17.46). The reason for this is that while the inverse Mills ratio is a nonlinear function of $\mathbf{z}$, it is often well approximated by a linear function. If $\mathbf{z} = \mathbf{x}$, $\hat{\lambda}_i$ can be highly correlated with the elements of $\mathbf{x}_i$. As we know, such multicollinearity can lead to very high standard errors for the $\hat{\beta}_j$. Intuitively, if we do not have a variable that affects selection but not y, it is extremely difficult, if not impossible, to distinguish sample selection from a misspecified functional form in (17.46).

EXAMPLE 17.5

(Wage Offer Equation for Married Women)

We apply the sample selection correction to the data on married women in MROZ.RAW. Recall that of the 753 women in the sample, 428 worked for a wage during the year. The wage offer equation is standard, with log(*wage*) as the dependent variable and *educ*, *exper*, and $exper^2$ as the explanatory variables. In order to test and correct for sample selection bias—

TABLE 17.5
Wage Offer Equation for Married Women

Dependent Variable: log(*wage*)		
Independent Variables	**OLS**	**Heckit**
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
$exper^2$	−.00081 (.00039)	−.00086 (.00044)
constant	−.522 (.199)	−.578 (.307)
$\hat{\lambda}$	—	.032 (.134)
Sample Size *R*-Squared	428 .157	428 .157

due to unobservability of the wage offer for nonworking women—we need to estimate a probit model for labor force participation. In addition to the education and experience variables, we include the factors in Table 17.1: other income, age, number of young children, and number of older children. The fact that these four variables are excluded from the wage offer equation is an *assumption*: we assume that, given the productivity factors, *nwifeinc*, *age*, *kidslt6*, and *kidsge6* have no effect on the wage offer. It is clear from the probit results in Table 17.1 that at least *age* and *kidslt6* have a strong effect on labor force participation.

Table 17.5 contains the results from OLS and Heckit. [The standard errors reported for the Heckit results are just the usual OLS standard errors from regression (17.50).] There is no evidence of a sample selection problem in estimating the wage offer equation. The coefficient on $\hat{\lambda}$ has a very small *t* statistic (.239), so we fail to reject H_0: $\rho = 0$. Just as importantly, there are no practically large differences in the estimated slope coefficients in Table 17.5. The estimated returns to education differ by only one-tenth of a percentage point.

An alternative to the preceding two-step estimation method is full maximum likelihood estimation. This is more complicated as it requires obtaining the joint distribution of y and s. It often makes sense to test for sample selection using the previous procedure; if

there is no evidence of sample selection, there is no reason to continue. If we detect sample selection bias, we can either use the two-step estimates or estimate the regression and selection equations jointly by MLE. (See Wooldridge [2002, Chapter 17].)

In Example 17.5, we know more than just whether a woman worked during the year: we know how many hours each woman worked. It turns out that we can use this information in an alternative sample selection procedure. In place of the inverse Mills ratio $\hat{\lambda}_i$, we use the Tobit residuals, say, $\hat{v}_i$, which are computed as $\hat{v}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ whenever $y_i > 0$. It can be shown that the regression in (17.50) with $\hat{v}_i$ in place of $\hat{\lambda}_i$ also produces consistent estimates of the β_j, and the standard t statistic on $\hat{v}_i$ is a valid test for sample selection bias. This approach has the advantage of using more information, but it is less widely applicable. (See Wooldridge [2002, Chapter 17].)

There are many more topics concerning sample selection. One worth mentioning is models with endogenous explanatory variables *in addition to* possible sample selection bias. Write a model with a single endogenous explanatory variable as

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1, \tag{17.51}$$

where y_1 is only observed when $s = 1$, and y_2 may only be observed along with y_1. An example is when y_1 is the percentage of votes received by an incumbent, and y_2 is the percentage of total expenditures accounted for by the incumbent. For incumbents who do not run, we cannot observe y_1 or y_2. If we have exogenous factors that affect the decision to run and that are correlated with campaign expenditures, we can consistently estimate α_1 and the elements of $\boldsymbol{\beta}_1$ by instrumental variables. To be convincing, we need *two* exogenous variables that do not appear in (17.51). Effectively, one should affect the selection decision, and one should be correlated with y_2 [the usual requirement for estimating (17.51) by 2SLS]. Briefly, the method is to estimate the selection equation by probit, where *all* exogenous variables appear in the probit equation. Then, we add the inverse Mills ratio to (17.51) and estimate the equation by 2SLS. The inverse Mills ratio acts as its own instrument, as it depends only on exogenous variables. We use all exogenous variables as the other instruments. As before, we can use the t statistic on $\hat{\lambda}_i$ as a test for selection bias. (See Wooldridge [2002, Chapter 17] for further information.)

SUMMARY

In this chapter, we have covered several advanced methods that are often used in applications, especially in microeconomics. Logit and probit models are used for binary response variables. These models have some advantages over the linear probability model: fitted probabilities are between zero and one, and the partial effects diminish. The primary cost to logit and probit is that they are harder to interpret.

The Tobit model is applicable to nonnegative outcomes that pile up at zero but also take on a broad range of positive values. Many individual choice variables, such as labor supply, amount of life insurance, and amount of pension fund invested in stocks, have this feature. As with logit and probit, the expected values of y given $\mathbf{x}$—either conditional on $y > 0$ or unconditionally—depend on $\mathbf{x}$ and $\boldsymbol{\beta}$ in nonlinear ways. We gave the

expressions for these expectations as well as formulas for the partial effects of each x_j on the expectations. These can be estimated after the Tobit model has been estimated by maximum likelihood.

When the dependent variable is a count variable—that is, it takes on nonnegative, integer values—a Poisson regression model is appropriate. The expected value of y given the x_j has an exponential form. This gives the parameter interpretations as semi-elasticities or elasticities, depending on whether x_j is in level or logarithmic form. In short, we can interpret the parameters *as if* they are in a linear model with $\log(y)$ as the dependent variable. The parameters can be estimated by MLE. However, because the Poisson distribution imposes equality of the variance and mean, it is often necessary to compute standard errors and test statistics that allow for over- or underdispersion. These are simple adjustments to the usual MLE standard errors and statistics.

Censored and truncated regression models handle specific kinds of missing data problems. In censored regression, the dependent variable is censored above or below a threshold. We can use information on the censored outcomes because we always observe the explanatory variables, as in duration applications or top coding of observations. A truncated regression model arises when a part of the population is excluded entirely: we observe no information on units that are not covered by the sampling scheme. This is a special case of a sample selection problem.

Section 17.5 gave a systematic treatment of nonrandom sample selection. We showed that exogenous sample selection does not affect consistency of OLS when it is applied to the subsample, but endogenous sample selection does. We showed how to test and correct for sample selection bias for the general problem of incidental truncation, where observations are missing on y due to the outcome of another variable (such as labor force participation). Heckman's method is relatively easy to implement in these situations.

KEY TERMS

Average Partial Effect
Binary Response Models
Censored Normal Regression Model
Censored Regression Model
Corner Solution Response
Count Variable
Duration Analysis
Exogenous Sample Selection
Heckit Method
Incidental Truncation
Inverse Mills Ratio
Latent Variable Model
Likelihood Ratio Statistic
Limited Dependent Variable (LDV)
Logit Model
Log-Likelihood Function
Maximum Likelihood Estimation (MLE)
Nonrandom Sample Selection
Overdispersion
Percent Correctly Predicted
Poisson Distribution
Poisson Regression Model
Probit Model
Pseudo R-Squared
Quasi-Likelihood Ratio Statistic
Quasi-Maximum Likelihood Estimation (QMLE)
Response Probability
Selected Sample
Tobit Model
Top Coding
Truncated Normal Regression Model
Truncated Regression Model
Wald Statistic

PROBLEMS

17.1 (i) For a binary response y, let $\bar{y}$ be the proportion of ones in the sample (which is equal to the sample average of the y_i). Let $\hat{q}_0$ be the percent correctly predicted for the outcome $y = 0$ and let $\hat{q}_1$ be the percent correctly predicted for the outcome $y = 1$. If $\hat{p}$ is the overall percent correctly predicted, show that $\hat{p}$ is a weighted average of $\hat{q}_0$ and $\hat{q}_1$:

$$\hat{p} = (1 - \bar{y})\,\hat{q}_0 + \bar{y}\hat{q}_1.$$

(ii) In a sample of 300, suppose that $\bar{y} = .70$, so that there are 210 outcomes with $y_i = 1$ and 90 with $y_i = 0$. Suppose that the percent correctly predicted when $y = 0$ is 80, and the percent correctly predicted when $y = 1$ is 40. Find the overall percent correctly predicted.

17.2 Let *grad* be a dummy variable for whether a student-athlete at a large university graduates in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following logit model is obtained:

$$\widehat{\mathrm{P}(grad = 1|hsGPA,SAT,study)} = \Lambda(-1.17 + .24\,hsGPA + .00058\,SAT + .073\,study),$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logit function. Holding *hsGPA* fixed at 3.0 and *SAT* fixed at 1,200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

17.3 (Requires calculus)

(i) Suppose in the Tobit model that $x_1 = \log(z_1)$, and this is the only place z_1 appears in $\mathbf{x}$. Show that

$$\frac{\partial \mathrm{E}(y|y > 0,\mathbf{x})}{\partial z_1} = (\beta_1/z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}, \quad (17.52)$$

where β_1 is the coefficient on $\log(z_1)$.

(ii) If $x_1 = z_1$, and $x_2 = z_1^2$, show that

$$\frac{\partial \mathrm{E}(y|y > 0,\mathbf{x})}{\partial z_1} = (\beta_1 + 2\beta_2 z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\},$$

where β_1 is the coefficient on z_1 and β_2 is the coefficient on z_1^2.

17.4 Let mvp_i be the marginal value product for worker i, which is the price of a firm's good multiplied by the marginal product of the worker. Assume that

$$\log(mvp_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

$$wage_i = \max(mvp_i, minwage_i),$$

where the explanatory variables include education, experience, and so on, and $minwage_i$ is the minimum wage relevant for person i. Write $\log(wage_i)$ in terms of $\log(mvp_i)$ and $\log(minwage_i)$.

17.5 (Requires calculus) Let *patents* be the number of patents applied for by a firm during a given year. Assume that the conditional expectation of *patents* given *sales* and *RD* is

$$E(patents|sales,RD) = \exp[\beta_0 + \beta_1 \log(sales) + \beta_2 RD + \beta_3 RD^2],$$

where *sales* is annual firm sales and *RD* is total spending on research and development over the past 10 years.

(i) How would you estimate the β_j? Justify your answer by discussing the nature of *patents*.

(ii) How do you interpret β_1?

(iii) Find the partial effect of *RD* on $E(patents|sales,RD)$.

17.6 Consider a family saving function for the population of all families in the United States:

$$sav = \beta_0 + \beta_1 inc + \beta_2 hhsize + \beta_3 educ + \beta_4 age + u,$$

where *hhsize* is household size, *educ* is years of education of the household head, and *age* is age of the household head. Assume that $E(u|inc,hhsize,educ,age) = 0$.

(i) Suppose that the sample includes only families whose head is over 25 years old. If we use OLS on such a sample, do we get unbiased estimators of the β_j? Explain.

(ii) Now, suppose our sample includes only married couples without children. Can we estimate all of the parameters in the saving equation? Which ones can we estimate?

(iii) Suppose we exclude from our sample families that save more than $25,000 per year. Does OLS produce consistent estimators of the β_j?

17.7 Suppose you are hired by a university to study the factors that determine whether students admitted to the university actually come to the university. You are given a large random sample of students who were admitted the previous year. You have information on whether each student chose to attend, high school performance, family income, financial aid offered, race, and geographic variables. Someone says to you, "Any analysis of that data will lead to biased results because it is not a random sample of all college applicants, but only those who apply to this university." What do you think of this criticism?

COMPUTER EXERCISES

C17.1 Use the data in PNTSPRD.RAW for this exercise.

(i) The variable *favwin* is a binary variable if the team favored by the Las Vegas point spread wins. A linear probability model to estimate the probability that the favored team wins is

$$P(favwin = 1|spread) = \beta_0 + \beta_1 spread.$$

by ordinary least squares. Report the results in the usual form. Do there appear to be significant wage differences by race and ethnicity?

(iii) Estimate a probit model for *inlf* that includes the explanatory variables in the wage equation from part (ii) as well as *nwifeinc* and *kidlt6*. Do these last two variables have coefficients of the expected sign? Are they statistically significant?

(iv) Explain why, for the purposes of testing and, possibly, correcting the wage equation for selection into the workforce, it is important for *nwifeinc* and *kidlt6* to help explain *inlf*. What must you assume about *nwifeinc* and *kidlt6* in the wage equation?

(v) Compute the inverse Mills ratio (for each observation) and add it as an additional regressor to the wage equation from part (ii). What is its two-sided p-value? Do you think this is particularly small with 3,286 observations?

(vi) Does adding the inverse Mills ratio change the coefficients in the wage regression in important ways? Explain.

APPENDIX 17A

Asymptotic Standard Errors in Limited Dependent Variable Models

Derivations of the asymptotic standard errors for the models and methods introduced in this chapter are well beyond the scope of this text. Not only do the derivations require matrix algebra, but they also require advanced asymptotic theory of nonlinear estimation. The background needed for a careful analysis of these methods and several derivations are given in Wooldridge (2002).

It is instructive to see the formulas for obtaining the asymptotic standard errors for at least some of the methods. Given the binary response model $P(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})$, where $G(\cdot)$ is the logit or probit function, and $\boldsymbol{\beta}$ is the $k \times 1$ vector of parameters, the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is estimated as

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \frac{[g(\mathbf{x}_i\hat{\boldsymbol{\beta}})]^2 \mathbf{x}_i'\mathbf{x}_i}{G(\mathbf{x}_i\hat{\boldsymbol{\beta}})[1 - G(\mathbf{x}_i\hat{\boldsymbol{\beta}})]} \right)^{-1}, \tag{17.53}$$

which is a $k \times k$ matrix. (See Appendix D for a summary of matrix algebra.) Without the terms involving $g(\cdot)$ and $G(\cdot)$, this formula looks a lot like the estimated variance matrix for the OLS estimator, minus the term $\hat{\sigma}^2$. The expression in (17.53) accounts for the nonlinear nature of the response probability—that is, the nonlinear nature of $G(\cdot)$—as well as the particular form of heteroskedasticity in a binary response model: $\text{Var}(y|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})[1 - G(\mathbf{x}\boldsymbol{\beta})]$.

The square roots of the diagonal elements of (17.53) are the asymptotic standard errors of the $\hat{\beta}_j$, and they are routinely reported by econometrics software that supports logit and probit analysis. Once we have these, (asymptotic) t statistics and confidence intervals are obtained in the usual ways.

The matrix in (17.53) is also the basis for Wald tests of multiple restrictions on $\boldsymbol{\beta}$ (see Wooldridge [2002, Chapter 15]).

The asymptotic variance matrix for Tobit is more complicated but has a similar structure. Note that we can obtain a standard error for $\hat{\sigma}$ as well. The asymptotic variance for Poisson regression, allowing for $\sigma^2 \neq 1$ in (17.35), has a form much like (17.53):

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left(\sum_{i=1}^{n} \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}_i' \mathbf{x}_i \right)^{-1}.$$

The square roots of the diagonal elements of this matrix are the asymptotic standard errors. If the Poisson assumption holds, we can drop $\hat{\sigma}^2$ from the formula (because $\sigma^2 = 1$).

Asymptotic standard errors for censored regression, truncated regression, and the Heckit sample selection correction are more complicated, although they share features with the previous formulas. See Wooldridge (2002) for details.