17 Sample Selection, Attrition, and Stratified Sampling

17.1 Introduction

Up to this point, with the exception of occasionally touching on cluster samples and independently pooled cross sections, we have assumed the availability of a random sample from the underlying population. This assumption is not always realistic: because of the way some economic data sets are collected, and often because of the behavior of the units being sampled, random samples are not always available.

A **selected sample** is a general term that describes a nonrandom sample. There are a variety of **selection mechanisms** that result in nonrandom samples. Some of these are due to sample design, while others are due to the behavior of the units being sampled, including nonresponse on survey questions and attrition from social programs. Before we launch into specifics, there is an important general point to remember: sample selection can only be an issue once the population of interest has been carefully specified. If we are interested in a subset of a larger population, then the proper approach is to specify a model for that part of the population, obtain a random sample from that part of the population, and proceed with standard econometric methods.

The following are some examples with nonrandomly selected samples.

Example 17.1 (Saving Function): Suppose we wish to estimate a saving function for all families in a given country, and the population saving function is

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 married + \beta_4 kids + u \tag{17.1}$$

where *age* is the age of the household head and the other variables are self-explanatory. However, we only have access to a survey that included families whose household head was 45 years of age or older. This limitation raises a sample selection issue because we are interested in the saving function for all families, but we can obtain a random sample only for a subset of the population.

Example 17.2 (Truncation Based on Wealth): We are interested in estimating the effect of worker eligibility in a particular pension plan [for example, a 401(k) plan] on family wealth. Let the population model be

$$wealth = \beta_0 + \beta_1 plan + \beta_2 educ + \beta_3 age + \beta_4 income + u \tag{17.2}$$

where *plan* is a binary indicator for eligibility in the pension plan. However, we can only sample people with a net wealth less than $200,000, so the sample is selected on the basis of *wealth*. As we will see, sampling based on a response variable is much more serious than sampling based on an exogenous explanatory variable.

In these two examples data were missing on all variables for a subset of the population as a result of survey design. In other cases, units *are* randomly drawn from the population, but data are missing on one or more variables for some units in the sample. Using a subset of a random sample because of missing data can lead to a sample selection problem. As we will see, if the reason the observations are missing is appropriately exogenous, using the subsample has no serious consequences.

Our final example illustrates a more subtle form of a missing data problem.

Example 17.3 (Wage Offer Function): Consider estimating a wage offer equation for people of working age. By definition, this equation is supposed to represent *all* people of working age, whether or not a person is actually working at the time of the survey. Because we can only observe the wage offer for working people, we effectively select our sample on this basis.

This example is not as straightforward as the previous two. We treat it as a sample selection problem because data on a key variable—the wage offer, $wage^o$—are available only for a clearly defined subset of the population. This is sometimes called **incidental truncation** because $wage^o$ is missing as a result of the outcome of another variable, labor force participation.

The incidental truncation in this example has a strong self-selection component: people self-select into employment, so whether or not we observe $wage^o$ depends on an individual's labor supply decision. Whether we call examples like this sample selection or self-selection is largely irrelevant. The important point is that we must account for the nonrandom nature of the sample we have for estimating the wage offer equation.

In the next several sections we cover a variety of sample selection issues, including tests and corrections. Section 17.7 treats sample selection and the related problem of attrition in panel data. Stratified sampling, which arises out of sampling design, is covered in Section 17.8.

17.2 When Can Sample Selection Be Ignored?

In some cases, the fact that we have a nonrandom sample does not affect the way we estimate population parameters; it is important to understand when this is the case.

17.2.1 Linear Models: OLS and 2SLS

We begin by obtaining conditions under which estimation of the population model by 2SLS using the selected sample is consistent for the population parameters. These

results are of interest in their own right, but we will also apply them to several specific models later in the chapter.

We assume that there is a population represented by the random vector $(\mathbf{x}, y, \mathbf{z})$, where $\mathbf{x}$ is a $1 \times K$ vector of explanatory variables, y is the scalar response variable, and $\mathbf{z}$ is a $1 \times L$ vector of instruments.

The population model is the standard single-equation linear model with possibly endogenous explanatory variables:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u \tag{17.3}$$

$$\mathrm{E}(u \mid \mathbf{z}) = 0 \tag{17.4}$$

where we take $x_1 \equiv 1$ for notational simplicity. The sense in which the instruments $\mathbf{z}$ are exogenous, given in assumption (17.4), is stronger than we need for 2SLS to be consistent when using a random sample from the population. With random sampling, the zero correlation condition $\mathrm{E}(\mathbf{z}'u) = \mathbf{0}$ is sufficient. *If* we could obtain a random sample from the population, equation (17.3) could be estimated by 2SLS under the condition $\mathrm{rank}[\mathrm{E}(\mathbf{z}'\mathbf{x})] = K$.

A leading special case is $\mathbf{z} = \mathbf{x}$, so that the explanatory variables are exogenous and equation (17.3) is a model of the conditional expectation $\mathrm{E}(y \mid \mathbf{x})$:

$$\mathrm{E}(y \mid \mathbf{x}) = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K \tag{17.5}$$

But our general treatment allows elements of $\mathbf{x}$ to be correlated with u.

Rather than obtaining a random sample—that is, a sample representative of the population—we only use data points that satisfy certain conditions. Let s be a binary **selection indicator** representing a random draw from the population. By definition, $s = 1$ if we use the draw in the estimation, and $s = 0$ if we do not. Usually, we do not use observations when $s = 0$ because data on at least some elements of $(\mathbf{x}, y, \mathbf{z})$ are unobserved—because of survey design, nonresponse, or incidental truncation.

The key assumption underlying the validity of 2SLS on selected sample is

$$\mathrm{E}(u \mid \mathbf{z}, s) = 0 \tag{17.6}$$

There are some important cases where assumption (17.6) necessarily follows from assumption (17.4). If s is a deterministic function of $\mathbf{z}$, then $\mathrm{E}(u \mid \mathbf{z}, s) = \mathrm{E}(u \mid \mathbf{z})$. Such cases arise when selection is a fixed rule involving only the exogenous variables $\mathbf{z}$. Also, if selection is independent of $(\mathbf{z}, u)$—a sufficient condition is that selection is independent of $(\mathbf{x}, y, \mathbf{z})$—then $\mathrm{E}(u \mid \mathbf{z}, s) = \mathrm{E}(u \mid \mathbf{z})$.

In estimating equation (17.3), we apply 2SLS to the observations for which $s = 1$. To study the properties of the 2SLS estimator on the selected sample, let

$\{(\mathbf{x}_i, y_i, \mathbf{z}_i, s_i)\colon i = 1, 2, \ldots, N\}$ denote a random sample from the *population*. We use observation i if $s_i = 1$, but not if $s_i = 0$. Therefore, we do not actually have N observations to use in the estimation; in fact, we do not even need to know N.

The 2SLS estimator using the selected sample can be expressed as

$$\hat{\boldsymbol{\beta}} = \left[\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)'\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1}$$
$$\times\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)'\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'y_i\right)$$

Substituting $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$ gives

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)'\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1}$$
$$\times\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)'\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'u_i\right) \tag{17.7}$$

By assumption, $\mathrm{E}(u_i \mid \mathbf{z}_i, s_i) = 0$, and so $\mathrm{E}(s_i\mathbf{z}_i'u_i) = \mathbf{0}$ by iterated expectations. [In the case where s is a function of $\mathbf{z}$, this result shows why assumption (17.4) cannot be replaced with $\mathrm{E}(\mathbf{z}'u) = \mathbf{0}$.] Now the law of large numbers applies to show that plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, at least under a modification of the rank condition. We summarize with a theorem:

THEOREM 17.1 (Consistency of 2SLS under Sample Selection): In model (17.3), assume that $\mathrm{E}(u^2) < \infty$, $\mathrm{E}(x_j^2) < \infty$, $j = 1, \ldots, K$, and $\mathrm{E}(z_j^2) < \infty$, $j = 1, \ldots, L$. Maintain assumption (17.6) and, in addition, assume

$$\text{rank } \mathrm{E}(\mathbf{z}'\mathbf{z} \mid s = 1) = L \tag{17.8}$$

$$\text{rank } \mathrm{E}(\mathbf{z}'\mathbf{x} \mid s = 1) = K \tag{17.9}$$

Then the 2SLS estimator using the selected sample is consistent for $\boldsymbol{\beta}$ and $\sqrt{N}$-asymptotically normal. Further, if $\mathrm{E}(u^2 \mid \mathbf{z}, s) = \sigma^2$, then the usual asymptotic variance of the 2SLS estimator is valid.

Equation (17.7) essentially proves the consistency result. Showing that the usual 2SLS asymptotic variance matrix is valid requires two steps. First, under the homo-

skedasticity assumption in the population, the usual iterated expectations argument gives $\mathrm{E}(su^2\mathbf{z}'\mathbf{z}) = \sigma^2\mathrm{E}(s\mathbf{z}'\mathbf{z})$. This equation can be used to show that $\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma^2\{\mathrm{E}(s\mathbf{x}'\mathbf{z})[\mathrm{E}(s\mathbf{z}'\mathbf{z})]^{-1}\mathrm{E}(s\mathbf{z}'\mathbf{x})\}^{-1}$. The second step is to show that the usual 2SLS estimator of σ^2 is consistent. This fact can be seen as follows. Under the homoskedasticity assumption, $\mathrm{E}(su^2) = \mathrm{E}(s)\sigma^2$, where $\mathrm{E}(s)$ is just the fraction of the subpopulation in the overall population. The estimator of σ^2 (without degrees-of-freedom adjustment) is

$$\left(\sum_{i=1}^{N} s_i\right)^{-1} \sum_{i=1}^{N} s_i\hat{u}_i^2 \tag{17.10}$$

since $\sum_{i=1}^{N} s_i$ is simply the number of observations in the selected sample. Removing the "^" from u_i^2 and applying the law of large numbers gives $N^{-1}\sum_{i=1}^{N} s_i \xrightarrow{p} \mathrm{E}(s)$ and $N^{-1}\sum_{i=1}^{N} s_i u_i^2 \xrightarrow{p} \mathrm{E}(su^2) = \mathrm{E}(s)\sigma^2$. Since the N^{-1} terms cancel, expression (17.10) converges in probability to σ^2.

If s is a function only of $\mathbf{z}$, or s is independent of $(\mathbf{z}, u)$, and $\mathrm{E}(u^2 \mid \mathbf{z}) = \sigma^2$—that is, if the homoskedasticity assumption holds in the original population—then $\mathrm{E}(u^2 \mid \mathbf{z}, s) = \sigma^2$. Without the homoskedasticity assumption we would just use the heteroskedasticity-robust standard errors, just as if a random sample were available with heteroskedasticity present in the population model.

When $\mathbf{x}$ is exogenous and we apply OLS on the selected sample, Theorem 17.1 implies that we can select the sample on the basis of the explanatory variables. Selection based on y or on endogenous elements of $\mathbf{x}$ is not allowed because then $\mathrm{E}(u \mid \mathbf{z}, s) \neq \mathrm{E}(u)$.

Example 17.4 (Nonrandomly Missing IQ Scores): As an example of how Theorem 17.1 can be applied, consider the analysis in Griliches, Hall, and Hausman (1978) (GHH). The structural equation of interest is

$$\log(wage) = \mathbf{z}_1\boldsymbol{\delta}_1 + abil + v, \qquad \mathrm{E}(v \mid \mathbf{z}_1, abil, IQ) = 0$$

and we assume that IQ is a valid proxy for $abil$ in the sense that $abil = \theta_1 IQ + e$ and $\mathrm{E}(e \mid \mathbf{z}_1, IQ) = 0$ (see Section 4.3.2). Write

$$\log(wage) = \mathbf{z}_1\boldsymbol{\delta}_1 + \theta_1 IQ + u \tag{17.11}$$

where $u = v + e$. Under the assumptions made, $\mathrm{E}(u \mid \mathbf{z}_1, IQ) = 0$. It follows immediately from Theorem 17.1 that, if we choose the sample excluding all people with IQs below a fixed value, then OLS estimation of equation (17.11) will be consistent. This problem is not quite the one faced by GHH. Instead, GHH noticed that the

probability of IQ missing was higher at lower IQs (because people were reluctant to give permission to obtain IQ scores). A simple way to model this situation is $s = 1$ if $IQ + r \geq 0$, $s = 0$ if $IQ + r < 0$, where r is an unobserved random variable. If r is redundant in the structural equation *and* in the proxy variable equation for IQ, that is, if $\mathrm{E}(v \mid \mathbf{z}_1, abil, IQ, r) = 0$ and $\mathrm{E}(e \mid \mathbf{z}_1, IQ, r) = 0$, then $\mathrm{E}(u \mid \mathbf{z}_1, IQ, r) = 0$. Since s is a function of IQ and r, it follows immediately that $\mathrm{E}(u \mid \mathbf{z}_1, IQ, s) = 0$. Therefore, using OLS on the sample for which IQ is observed yields consistent estimators.

If r is correlated with either v or e, $\mathrm{E}(u \mid \mathbf{z}_1, IQ, s) \neq \mathrm{E}(u)$ in general, and OLS estimation of equation (17.11) using the selected sample would not consistently estimate δ_1 and θ_1. Therefore, even though IQ is exogenous in the population equation (17.11), the sample selection is not exogenous. In Section 17.4.2 we cover a method that can be used to correct for sample selection bias.

Theorem 17.1 has other useful applications. Suppose that $\mathbf{x}$ is exogenous in equation (17.3) and that s is a nonrandom function of $(\mathbf{x}, v)$, where v is a variable not appearing in equation (17.3). If (u, v) is independent of $\mathbf{x}$, then $\mathrm{E}(u \mid \mathbf{x}, v) = \mathrm{E}(u \mid v)$, and so

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

If we make an assumption about the functional form of $\mathrm{E}(u \mid v)$, for example, $\mathrm{E}(u \mid v) = \gamma v$, then we can write

$$y = \mathbf{x}\boldsymbol{\beta} + \gamma v + e, \qquad \mathrm{E}(e \mid \mathbf{x}, v) = 0 \tag{17.12}$$

where $e = u - \mathrm{E}(u \mid v)$. Because s is just a function of $(\mathbf{x}, v)$, $\mathrm{E}(e \mid \mathbf{x}, v, s) = 0$, and so $\boldsymbol{\beta}$ and γ can be estimated consistently by the OLS regression y on $\mathbf{x}$, v, using the selected sample. Effectively, including v in the regression on the selected subsample eliminates the sample selection problem and allows us to consistently estimate $\boldsymbol{\beta}$. [Incidentally, because v is independent of $\mathbf{x}$, we would not have to include it in equation (17.3) to consistently estimate $\boldsymbol{\beta}$ *if* we had a random sample from the population. However, including v would result in an asymptotically more efficient estimator of $\boldsymbol{\beta}$ when $\mathrm{Var}(y \mid \mathbf{x}, v)$ is homoskedastic. See Problem 4.5.] In Section 17.5 we will see how equation (17.12) can be implemented.

17.2.2 Nonlinear Models

Results similar to those in the previous section hold for nonlinear models as well. We will cover explicitly the case of nonlinear regression and maximum likelihood. See Problem 17.8 for the GMM case.

In the nonlinear regression case, if $E(y \mid \mathbf{x}, s) = E(y \mid \mathbf{x})$—so that selection is *ignorable* in the conditional mean sense—then NLS on the selected sample is consistent. Sufficient is that s is a deterministic function of $\mathbf{x}$. The consistency argument is simple: NLS on the selected sample solves

$$\min_{\beta} N^{-1} \sum_{i=1}^{N} s_i [y_i - m(\mathbf{x}_i, \beta)]^2$$

so it suffices to show that β_o in $E(y \mid \mathbf{x}) = m(\mathbf{x}, \beta_o)$ minimizes $E\{s[y - m(\mathbf{x}, \beta)]^2\}$ over β. By iterated expectations,

$$E\{s[y - m(\mathbf{x}, \beta)]^2\} = E(sE\{[y - m(\mathbf{x}, \beta)]^2 \mid \mathbf{x}, s\})$$

Next, write $[y - m(\mathbf{x}, \beta)]^2 = u^2 + 2[m(\mathbf{x}, \beta_o) - m(\mathbf{x}, \beta)]u + [m(\mathbf{x}, \beta_o) - m(\mathbf{x}, \beta)]^2$, where $u = y - m(\mathbf{x}, \beta_o)$. By assumption, $E(u \mid \mathbf{x}, s) = 0$. Therefore,

$$E\{[y - m(\mathbf{x}, \beta)]^2 \mid \mathbf{x}, s\} = E(u^2 \mid \mathbf{x}, s) + [m(\mathbf{x}, \beta_o) - m(\mathbf{x}, \beta)]^2$$

and the second term is clearly minimized at $\beta = \beta_o$. We do have to assume that β_o is the *unique* value of β that makes $E\{s[m(\mathbf{x}, \beta) - m(\mathbf{x}, \beta_o)]^2\}$ zero. This is the identification condition on the subpopulation.

It can also be shown that, if $\mathrm{Var}(y \mid \mathbf{x}, s) = \mathrm{Var}(y \mid \mathbf{x})$ and $\mathrm{Var}(y \mid \mathbf{x}) = \sigma_o^2$, then the usual, nonrobust NLS statistics are valid. If heteroskedasticity exists either in the population or the subpopulation, standard heteroskedasticity-robust inference can be used. The arguments are very similar to those for 2SLS in the previous subsection.

Another important case is the general conditional maximum likelihood setup. Assume that the distribution of $\mathbf{y}$ given $\mathbf{x}$ and s is the same as the distribution of $\mathbf{y}$ given $\mathbf{x}$: $D(\mathbf{y} \mid \mathbf{x}, s) = D(\mathbf{y} \mid \mathbf{x})$. This is a stronger form of ignorability of selection, but it always holds if s is a nonrandom function of $\mathbf{x}$, or if s is independent of $(\mathbf{x}, \mathbf{y})$. In any case, $D(\mathbf{y} \mid \mathbf{x}, s) = D(\mathbf{y} \mid \mathbf{x})$ ensures that the MLE on the selected sample is consistent and that the usual MLE statistics are valid. The analogy argument should be familiar by now. Conditional MLE on the selected sample solves

$$\max_{\theta} N^{-1} \sum_{i=1}^{N} s_i \ell(\mathbf{y}_i, \mathbf{x}_i; \theta) \tag{17.13}$$

where $\ell(\mathbf{y}_i, \mathbf{x}_i; \theta)$ is the log likelihood for observation i. Now for each $\mathbf{x}$, θ_o maximizes $E[\ell(\mathbf{y}, \mathbf{x}; \theta) \mid \mathbf{x}]$ over θ. But $E[s\ell(\mathbf{y}, \mathbf{x}; \theta)] = E\{sE[\ell(\mathbf{y}, \mathbf{x}; \theta) \mid \mathbf{x}, s]\} = E\{sE[\ell(\mathbf{y}, \mathbf{x}; \theta) \mid \mathbf{x}]\}$, since, by assumption, the conditional distribution of $\mathbf{y}$ given $(\mathbf{x}, s)$ does not depend on s. Since $E[\ell(\mathbf{y}, \mathbf{x}; \theta) \mid \mathbf{x}]$ is maximized at θ_o, so is $E\{sE[\ell(\mathbf{y}, \mathbf{x}; \theta) \mid \mathbf{x}]\}$. We must make

the stronger assumption that θ_o is the unique maximum, just as in the previous cases: if the selected subset of the population is too small, we may not be able to identify θ_o. Inference can be carried out using the usual MLE statistics obtained from the selected subsample because the information equality now holds conditional on $\mathbf{x}$ and s under the assumption that $D(\mathbf{y} \mid \mathbf{x}, s) = D(\mathbf{y} \mid \mathbf{x})$. We omit the details.

Problem 17.8 asks you to work through the case of GMM estimation of general nonlinear models based on conditional moment restrictions.

17.3 Selection on the Basis of the Response Variable: Truncated Regression

Let $(\mathbf{x}_i, y_i)$ denote a random draw from a population. In this section we explicitly treat the case where the sample is selected on the basis of y_i.

In applying the following methods it is important to remember that there is an underlying population of interest, often described by a linear conditional expectation: $E(y_i \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$. If we could observe a random sample from the population, then we would just use standard regression analysis. The problem comes about because the sample we can observe is chosen at least partly based on the value of y_i. Unlike in the case where selection is based only on $\mathbf{x}_i$, selection based on y_i causes problems for standard OLS analysis on the selected sample.

A classic example of selection based on y_i is Hausman and Wise's (1977) study of the determinants of earnings. Hausman and Wise recognized that their sample from a negative income tax experiment was truncated because only families with income below 1.5 times the poverty level were allowed to participate in the program; no data were available on families with incomes above the threshold value. The truncation rule was known, and so the effects of truncation could be accounted for.

A similar example is Example 17.2. We do not observe data on families with wealth above $200,000. This case is different from the top coding example we discussed in Chapter 16. Here, we observe *nothing* about families with high wealth: they are entirely excluded from the sample. In the top coding case, we have a random sample of families, and we always observe $\mathbf{x}_i$; the information on $\mathbf{x}_i$ is useful even if wealth is top coded.

We assume that y_i is a continuous random variable and that the selection rule takes the form

$$s_i = 1[a_1 < y_i < a_2]$$

where a_1 and a_2 are known constants such that $a_1 < a_2$. A good way to think of the sample selection is that we draw $(\mathbf{x}_i, y_i)$ randomly from the population. If y_i falls in

the interval (a_1, a_2), then we observe both y_i and $\mathbf{x}_i$. If y_i is outside this interval, then we do not observe y_i or $\mathbf{x}_i$. Thus all we know is that there is some subset of the population that does not enter our data set because of the selection rule. We know how to characterize the part of the population not being sampled because we know the constants a_1 and a_2.

In most applications we are still interested in estimating $\mathrm{E}(y_i \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$. However, because of sample selection based on y_i, we must—at least in a parametric context—specify a full conditional distribution of y_i given $\mathbf{x}_i$. Parameterize the conditional density of y_i given $\mathbf{x}_i$ by $f(\cdot \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma)$, where $\boldsymbol{\beta}$ are the conditional mean parameters and γ is a $G \times 1$ vector of additional parameters. The cdf of y_i given $\mathbf{x}_i$ is $F(\cdot \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma)$.

What we can use in estimation is the density of y_i conditional on $\mathbf{x}_i$ *and* the fact that we observe $(y_i, \mathbf{x}_i)$. In other words, we must condition on $a_1 < y_i < a_2$ or, equivalently, $s_i = 1$. The cdf of y_i conditional on $(\mathbf{x}_i, s_i = 1)$ is simply

$$\mathrm{P}(y_i \leq y \mid \mathbf{x}_i, s_i = 1) = \frac{\mathrm{P}(y_i \leq y, s_i = 1 \mid \mathbf{x}_i)}{\mathrm{P}(s_i = 1 \mid \mathbf{x}_i)}$$

Because y_i is continuously distributed, $\mathrm{P}(s_i = 1 \mid \mathbf{x}_i) = \mathrm{P}(a_1 < y_i < a_2 \mid \mathbf{x}_i) = F(a_2 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) - F(a_1 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) > 0$ for all possible values of $\mathbf{x}_i$. The case $a_2 = \infty$ corresponds to truncation only from below, in which case $F(a_2 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) \equiv 1$. If $a_1 = -\infty$ (truncation only from above), then $F(a_1 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) = 0$. To obtain the numerator when $a_1 < y < a_2$, we have

$$\mathrm{P}(y_i \leq y, s_i = 1 \mid \mathbf{x}_i) = \mathrm{P}(a_1 < y_i \leq y \mid \mathbf{x}_i) = F(y \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) - F(a_1 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma)$$

When we put this equation over $\mathrm{P}(s_i = 1 \mid \mathbf{x}_i)$ and take the derivative with respect to the dummy argument y, we obtain the density of y_i given $(\mathbf{x}_i, s_i = 1)$:

$$p(y \mid \mathbf{x}_i, s_i = 1) = \frac{f(y \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma)}{F(a_2 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma) - F(a_1 \mid \mathbf{x}_i; \boldsymbol{\beta}, \gamma)} \tag{17.14}$$

for $a_1 < y < a_2$.

Given a model for $f(y \mid \mathbf{x}; \boldsymbol{\beta}, \gamma)$, the log-likelihood function for any $(\mathbf{x}_i, y_i)$ in the sample can be obtained by plugging y_i into equation (17.14) and taking the log. The CMLEs of $\boldsymbol{\beta}$ and γ using the selected sample are efficient in the class of estimators that do not use information about the distribution of $\mathbf{x}_i$. Standard errors and test statistics can be computed using the general theory of conditional MLE.

In most applications of truncated samples, the population conditional distribution is assumed to be Normal$(\mathbf{x}\boldsymbol{\beta}, \sigma^2)$, in which case we have the **truncated Tobit model** or **truncated normal regression model**. The truncated Tobit model is related to the censored Tobit model for data-censoring applications (see Chapter 16), but there is a key

difference: in censored regression, we observe the covariates $\mathbf{x}$ for *all* people, even those for whom the response is not known. If we drop observations entirely when the response is not observed, we obtain the truncated regression model. If in Example 16.1 we use the information in the top coded observations, we are in the censored regression case. If we drop all top coded observations, we are in the truncated regression case. (Given a choice, we should use a censored regression analysis, as it uses all of the information in the sample.)

From our analysis of the censored regression model in Chapter 16, it is not surprising that heteroskedasticity or nonnormality in truncated regression results in inconsistent estimators of $\boldsymbol{\beta}$. This outcome is unfortunate because, if not for the sample selection problem, we could consistently estimate $\boldsymbol{\beta}$ under $\mathrm{E}(y \mid \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, without specifying $\mathrm{Var}(y \mid \mathbf{x})$ or the conditional distribution. Distribution-free methods for the truncated regression model have been suggested by Powell (1986) under the assumption of a symmetric error distribution; see Powell (1994) for a recent survey.

Truncating a sample on the basis of y is related to **choice-based sampling**. Traditional choice-based sampling applies when y is a discrete response taking on a finite number of values, where sampling frequencies differ depending on the outcome of y. [In the truncation case, the sampling frequency is one when y falls in the interval (a_1, a_2) and zero when y falls outside of the interval.] We do not cover choice-based sampling here; see Manksi and McFadden (1981), Imbens (1992), and Cosslett (1993). In Section 17.8 we cover some estimation methods for stratified sampling, which can be applied to some choice-based samples.

17.4 A Probit Selection Equation

We now turn to sample selection corrections when selection is determined by a probit model. This setup applies to problems different from those in Section 17.3, where the problem was that a survey or program was designed to intentionally exclude part of the population. We are now interested in selection problems that are due to incidental truncation, attrition in the context of program evalution, and general nonresponse that leads to missing data on the response variable or the explanatory variables.

17.4.1 Exogenous Explanatory Variables

The incidental truncation problem is motivated by Gronau's (1974) model of the wage offer and labor force participation.

Example 17.5 (Labor Force Participation and the Wage Offer): Interest lies in estimating $\mathrm{E}(w_i^o \mid \mathbf{x}_i)$, where w_i^o is the hourly wage offer for a randomly drawn individual

i. If w_i^o were observed for everyone in the (working age) population, we would proceed in a standard regression framework. However, a potential sample selection problem arises because w_i^o is observed only for people who work.

We can cast this problem as a weekly labor supply model:

$$\max_h \ \text{util}_i(w_i^o h + a_i, h) \quad \text{subject to } 0 \leq h \leq 168 \tag{17.15}$$

where h is hours worked per week and a_i is nonwage income of person i. Let $s_i(h) \equiv \text{util}_i(w_i^o h + a_i, h)$, and assume that we can rule out the solution $h_i = 168$. Then the solution can be $h_i = 0$ or $0 < h_i < 168$. If $ds_i/dh \leq 0$ at $h = 0$, then the optimum is $h_i = 0$. Using this condition, straightforward algebra shows that $h_i = 0$ if and only if

$$w_i^o \leq -mu_i^h(a_i, 0)/mu_i^q(a_i, 0) \tag{17.16}$$

where $mu_i^h(\cdot, \cdot)$ is the marginal disutility of working and $mu_i^q(\cdot, \cdot)$ is the marginal utility of income. Gronau (1974) called the right-hand side of equation (17.16) the *reservation wage*, w_i^r, which is assumed to be strictly positive.

We now make the parametric assumptions

$$w_i^o = \exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + u_{i1}), \qquad w_i^r = \exp(\mathbf{x}_{i2}\boldsymbol{\beta}_2 + \gamma_2 a_i + u_{i2}) \tag{17.17}$$

where (u_{i1}, u_{i2}) is independent of $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, a_i)$. Here, $\mathbf{x}_{i1}$ contains productivity characteristics, and possibly demographic characteristics, of individual i, and $\mathbf{x}_{i2}$ contains variables that determine the marginal utility of leisure and income; these may overlap with $\mathbf{x}_{i1}$. From equation (17.17) we have the log wage equation

$$\log w_i^o = \mathbf{x}_{i1}\boldsymbol{\beta}_1 + u_{i1} \tag{17.18}$$

But the wage offer w_i^o is observed only if the person works, that is, only if $w_i^o \geq w_i^r$, or

$$\log w_i^o - \log w_i^r = \mathbf{x}_{i1}\boldsymbol{\beta}_1 - \mathbf{x}_{i2}\boldsymbol{\beta}_2 - \gamma_2 a_i + u_{i1} - u_{i2} \equiv \mathbf{x}_i\boldsymbol{\delta}_2 + v_{i2} > 0$$

This behavior introduces a potential sample selection problem if we use data only on working people to estimate equation (17.18).

This example differs in an important respect from top coding examples. With top coding, the censoring rule is known for each unit in the population. In Gronau's example, we do not know w_i^r, so we cannot use w_i^o in a censored regression analysis. If w_i^r were observed and exogenous and $\mathbf{x}_{i1}$ were always observed, then we would be in the censored regression framework (see Problem 16.3). If w_i^r were observed and exogenous but $\mathbf{x}_{i1}$ were observed only when w_i^o is, we would be in the truncated Tobit framework. But w_i^r is allowed to depend on unobservables, and so we need a new framework.

If we drop the i subscript, let $y_1 \equiv \log w^o$, and let y_2 be the binary labor force participation indicator, Gronau's model can be written for a random draw from the population as

$$y_1 = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \tag{17.19}$$

$$y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0] \tag{17.20}$$

We discuss estimation of this model under the following set of assumptions:

ASSUMPTION 17.1: (a) $(\mathbf{x}, y_2)$ are always observed, y_1 is observed only when $y_2 = 1$; (b) (u_1, v_2) is independent of $\mathbf{x}$ with zero mean; (c) $v_2 \sim \text{Normal}(0, 1)$; and (d) $\text{E}(u_1 \mid v_2) = \gamma_1 v_2$.

Assumption 17.1a emphasizes the sample selection nature of the problem. Part b is a strong, but standard, form of exogeneity of $\mathbf{x}$. We will see that Assumption 17.1c is needed to derive a conditional expectation given the selected sample. It is probably the most restrictive assumption because it is an explicit distributional assumption. Assuming $\text{Var}(v_2) = 1$ is without loss of generality because y_2 is a binary variable.

Assumption 17.1d requires linearity in the population regression of u_1 on v_2. It always holds if (u_1, v_2) is bivariate normal—a standard assumption in these contexts—but Assumption 17.1d holds under weaker assumptions. In particular, we do not need to assume that u_1 itself is normally distributed.

Amemiya (1985) calls equations (17.19) and (17.20) the **type II Tobit model**. This name is fine as a label, but we must understand that it is a model of sample selection, and it has nothing to do with y_1 being a corner solution outcome. Unfortunately, in almost all treatments of this model, y_1 is set to zero when $y_2 = 0$. Setting y_1 to zero (or any value) when $y_2 = 0$ is misleading and can lead to inappropriate use of the model. For example, it makes no sense to set the wage offer to zero just because we do not observe it. As another example, it makes no sense to set the price per dollar of life insurance (y_1) to zero for someone who did not buy life insurance (so $y_2 = 1$ if and only if a person owns a life insurance policy).

We also have some interest in the parameters of the selection equation (17.20); for example, in Gronau's model it is a reduced-form labor force participation equation. In program evaluation with attrition, the selection equation explains the probability of dropping out of the program.

We can allow a little more generality in the model by replacing $\mathbf{x}$ in equation (17.20) with $\mathbf{x}_2$; then, as will become clear, $\mathbf{x}_1$ would only need to be observed whenever y_1 is, whereas $\mathbf{x}_2$ must always be observed. This extension is not especially useful for something like Gronau's model because it implies that $\mathbf{x}_1$ contains elements

that cannot also appear in $\mathbf{x}_2$. Because the selection equation is not typically a structural equation, it is undesirable to impose exclusion restrictions in equation (17.20). If a variable affecting y_1 is observed only along with y_1, the instrumental variables method that we cover in Section 17.4.2 is more attractive.

To derive an estimating equation, let $(y_1, y_2, \mathbf{x}, u_1, v_2)$ denote a random draw from the population. Since y_1 is observed only when $y_2 = 1$, what we can hope to estimate is $\mathrm{E}(y_1 \mid \mathbf{x}, y_2 = 1)$ [along with $\mathrm{P}(y_2 = 1 \mid \mathbf{x})$]. How does $\mathrm{E}(y_1 \mid \mathbf{x}, y_2 = 1)$ depend on the vector of interest, $\boldsymbol{\beta}_1$? First, under Assumption 17.1 and equation (17.19),

$$\mathrm{E}(y_1 \mid \mathbf{x}, v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathrm{E}(u_1 \mid \mathbf{x}, v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathrm{E}(u_1 \mid v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2 \tag{17.21}$$

where the second equality follows because (u_1, v_2) is independent of $\mathbf{x}$. Equation (17.21) is very useful. The first thing to note is that, if $\gamma_1 = 0$—which implies that u_1 and v_2 are uncorrelated—then $\mathrm{E}(y_1 \mid \mathbf{x}, v_2) = \mathrm{E}(y_1 \mid \mathbf{x}) = \mathrm{E}(y_1 \mid \mathbf{x}_1) = \mathbf{x}_1\boldsymbol{\beta}_1$. Because y_2 is a function of $(\mathbf{x}, v_2)$, it follows immediately that $\mathrm{E}(y_1 \mid \mathbf{x}, y_2) = \mathrm{E}(y_1 \mid \mathbf{x}_1)$. In other words, if $\gamma_1 = 0$, then there is no sample selection problem, and $\boldsymbol{\beta}_1$ can be consistently estimated by OLS using the selected sample.

What if $\gamma_1 \neq 0$? Using iterated expectations on equation (17.21),

$$\mathrm{E}(y_1 \mid \mathbf{x}, y_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1\mathrm{E}(v_2 \mid \mathbf{x}, y_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 h(\mathbf{x}, y_2)$$

where $h(\mathbf{x}, y_2) = \mathrm{E}(v_2 \mid \mathbf{x}, y_2)$. If we knew $h(\mathbf{x}, y_2)$, then, from Theorem 17.1, we could estimate $\boldsymbol{\beta}_1$ and γ_1 from the regression y_1 on $\mathbf{x}_1$ and $h(\mathbf{x}, y_2)$, using only the selected sample. Because the selected sample has $y_2 = 1$, we need only find $h(\mathbf{x}, 1)$. But $h(\mathbf{x}, 1) = \mathrm{E}(v_2 \mid v_2 > -\mathbf{x}\boldsymbol{\delta}_2) = \lambda(\mathbf{x}\boldsymbol{\delta}_2)$, where $\lambda(\cdot) \equiv \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio, and so we can write

$$\mathrm{E}(y_1 \mid \mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}\boldsymbol{\delta}_2) \tag{17.22}$$

Equation (17.22), which can be found in numerous places (see, for example, Heckman, 1979, and Amemiya, 1985) makes it clear that an OLS regression of y_1 on $\mathbf{x}_1$ using the selected sample omits the term $\lambda(\mathbf{x}\boldsymbol{\delta}_2)$ and generally leads to inconsistent estimation of $\boldsymbol{\beta}_1$. As pointed out by Heckman (1979), the presence of selection bias can be viewed as an omitted variable problem in the selected sample. An interesting point is that, even though only $\mathbf{x}_1$ appears in the population expectation, $\mathrm{E}(y_1 \mid \mathbf{x})$, other elements of $\mathbf{x}$ appear in the expectation on the subpopulation, $\mathrm{E}(y_1 \mid \mathbf{x}, y_2 = 1)$.

Equation (17.22) also suggests a way to consistently estimate $\boldsymbol{\beta}_1$. Following Heckman (1979), we can consistently estimate $\boldsymbol{\beta}_1$ and γ_1 using the selected sample by regressing y_{i1} on $\mathbf{x}_{i1}$, $\lambda(\mathbf{x}_i\boldsymbol{\delta}_2)$. The problem is that $\boldsymbol{\delta}_2$ is unknown, so we cannot compute the additional regressor $\lambda(\mathbf{x}_i\boldsymbol{\delta}_2)$. Nevertheless, a consistent estimator of $\boldsymbol{\delta}_2$ is available from the first-stage probit estimation of the selection equation.

Procedure 17.1: (a) Obtain the probit estimate $\hat{\delta}_2$ from the model

$$P(y_{i2} = 1 \mid \mathbf{x}_i) = \Phi(\mathbf{x}_i\delta_2) \tag{17.23}$$

using all N observations. Then, obtain the estimated inverse Mills ratios $\hat{\lambda}_{i2} \equiv \lambda(\mathbf{x}_i\hat{\delta}_2)$ (at least for $i = 1, \ldots, N_1$).

(b) Obtain $\hat{\beta}_1$ and $\hat{\gamma}_1$ from the OLS regression on the selected sample,

$$y_{i1} \text{ on } \mathbf{x}_{i1}, \hat{\lambda}_{i2}, \qquad i = 1, 2, \ldots, N_1 \tag{17.24}$$

These estimators are consistent and $\sqrt{N}$-asymptotically normal.

The procedure is sometimes called **Heckit** after Heckman (1976) and the tradition of putting "it" on the end of procedures related to probit (such as Tobit).

A very simple test for selection bias is available from regression (17.24). Under the null of no selection bias, H_0: $\gamma_1 = 0$, we have $\text{Var}(y_1 \mid \mathbf{x}, y_2 = 1) = \text{Var}(y_1 \mid \mathbf{x}) = \text{Var}(u_1)$, and so homoskedasticity holds under H_0. Further, from the results on generated regressors in Chapter 6, the asymptotic variance of $\hat{\gamma}_1$ (and $\hat{\beta}_1$) is not affected by $\hat{\delta}_2$ when $\gamma_1 = 0$. Thus, a standard t test on $\hat{\gamma}_1$ is a valid test of the null hypothsesis of no selection bias.

When $\gamma_1 \neq 0$, obtaining a consistent estimate for the asymptotic variance of $\hat{\beta}_1$ is complicated for two reasons. The first is that, if $\gamma_1 \neq 0$, then $\text{Var}(y_1 \mid \mathbf{x}, y_2 = 1)$ is not constant. As we know, heteroskedasticity itself is easy to correct for using the robust standard errors. However, we should also account for the fact that $\hat{\delta}_2$ is an estimator of δ_2. The adjustment to the variance of $(\hat{\beta}_1, \hat{\gamma}_1)$ because of the two-step estimation is cumbersome—it is *not* enough to simply make the standard errors heteroskedasticity-robust. Some statistical packages now have this feature built in.

As a technical point, we do not need $\mathbf{x}_1$ to be a strict subset of $\mathbf{x}$ for β_1 to be indentified, and Procedure 17.1 does carry through when $\mathbf{x}_1 = \mathbf{x}$. However, if $\mathbf{x}_i\hat{\delta}_2$ does not have much variation in the sample, then $\hat{\lambda}_{i2}$ can be approximated well by a linear function of $\mathbf{x}$. If $\mathbf{x} = \mathbf{x}_1$, this correlation can introduce severe collinearity among the regressors in regression (17.24), which can lead to large standard errors of the elements of $\hat{\beta}_1$. When $\mathbf{x}_1 = \mathbf{x}$, β_1 is identified only due to the nonlinearity of the inverse Mills ratio.

The situation is not quite as bad as in Section 9.5.1. There, identification failed for certain values of the structural parameters. Here, we still have identification for any value of β_1 in equation (17.19), but it is unlikely we can estimate β_1 with much precision. Even if we can, we would have to wonder whether a statistically inverse Mills ratio term is due to sample selection or functional form misspecification in the population model (17.19).

Table 17.1
Wage Offer Equation for Married Women

Dependent Variable: log(*wage*)		
Independent Variable	OLS	Heckit
educ	.108	.109
	(.014)	(.016)
exper	.042	.044
	(.012)	(.016)
$exper^2$	−.00081	−.00086
	(.00039)	(.00044)
constant	−.522	−.578
	(.199)	(.307)
$\hat{\lambda}_2$	—	.032
		(.134)
Sample size	428	428
R-squared	.157	.157

Example 17.6 (Wage Offer Equation for Married Women): We use the data in MROZ.RAW to estimate a wage offer function for married women, accounting for potential selectivity bias into the workforce. Of the 753 women, we observe the wage offer for 428 working women. The labor force participation equation contains the variables in Table 15.1, including other income, age, number of young children, and number of older children—in addition to *educ*, *exper*, and $exper^2$. The results of OLS on the selected sample and the Heckit method are given in Table 17.1.

The differences between the OLS and Heckit estimates are practically small, and the inverse Mills ratio term is statistically insignificant. The fact that the intercept estimates differ somewhat is usually unimportant. [The standard errors reported for Heckit are the unadjusted ones from regression (17.24). If $\hat{\lambda}_2$ were statistically significant, we should obtain the corrected standard errors.]

The Heckit results in Table 17.1 use four exclusion restrictions in the structural equation, because *nwifeinc*, *age*, *kidslt6*, and *kidsge6* are all excluded from the wage offer equation. If we allow all variables in the selection equation to also appear in the wage offer equation, the Heckit estimates become very imprecise. The coefficient on *educ* becomes .119 (se = .034), compared with the OLS estimate .100 (se = .015). The coefficient on *kidslt6*—which now appears in the wage offer equation—is −.188 (se = .232) in the Heckit estimation, and −.056 (se = .009) in the OLS estimation. The imprecision of the Heckit estimates is due to the severe collinearity that comes from adding $\hat{\lambda}_2$ to the equation, because $\hat{\lambda}_2$ is now a function only of the explanatory variables in the wage offer equation. In fact, using the selected sample, regressing $\hat{\lambda}_2$ on

the seven explanatory variables gives R-squared = .962. Unfortunately, comparing the OLS and Heckit results does not allow us to resolve some important issues. For example, the OLS results suggest that another young child reduces the wage offer by about 5.6 percent (t statistic ≈ -6.2), other things being equal. Is this effect real, or is it simply due to our inability to adequately correct for sample selection bias? Unless we have a variable that affects labor force participation without affecting the wage offer, we cannot answer this question.

If we replace parts c and d in Assumption 17.1 with the stronger assumption that (u_1, v_2) is bivariate normal with mean zero, $\text{Var}(u_1) = \sigma_1^2$, $\text{Cov}(u_1, v_2) = \sigma_{12}$, and $\text{Var}(v_2) = 1$, then partial maximum likelihood estimation can be used, as described generally in Problem 13.7. Partial MLE will be more efficient than the two-step procedure under joint normality of u_1 and v_2, and it will produce standard errors and likelihood ratio statistics that can be used directly (this conclusion follows from Problem 13.7). The drawbacks are that it is less robust than the two-step procedure and that it is sometimes difficult to get the problem to converge.

The reason we cannot perform full conditional MLE is that y_1 is only observed when $y_2 = 1$. Thus, while we can use the full density of y_2 given $\mathbf{x}$, which is $f(y_2 \mid \mathbf{x}) = [\Phi(\mathbf{x}\boldsymbol{\delta}_2)]^{y_2}[1 - \Phi(\mathbf{x}\boldsymbol{\delta}_2)]^{1-y_2}$, $y_2 = 0, 1$, we can only use the density $f(y_1 \mid y_2, \mathbf{x})$ when $y_2 = 1$. To find $f(y_1 \mid y_2, \mathbf{x})$ at $y_2 = 1$, we can use Bayes' rule to write $f(y_1 \mid y_2, \mathbf{x}) = f(y_2 \mid y_1, \mathbf{x})f(y_1 \mid \mathbf{x})/f(y_2 \mid \mathbf{x})$. Therefore, $f(y_1 \mid y_2 = 1, \mathbf{x}) = \text{P}(y_2 = 1 \mid y_1, \mathbf{x})f(y_1 \mid \mathbf{x})/\text{P}(y_2 = 1 \mid \mathbf{x})$. But $y_1 \mid \mathbf{x} \sim \text{Normal}(\mathbf{x}_1\boldsymbol{\beta}_1, \sigma_1^2)$. Further, $y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + \sigma_{12}\sigma_1^{-2}(y_1 - \mathbf{x}_1\boldsymbol{\beta}_1) + e_2 > 0]$, where e_2 is independent of $(\mathbf{x}, y_1)$ and $e_2 \sim \text{Normal}(0, 1 - \sigma_{12}^2\sigma_1^{-2})$ (this conclusion follows from standard conditional distribution results for joint normal random variables). Therefore,

$$\text{P}(y_2 = 1 \mid y_1, \mathbf{x}) = \Phi\{[\mathbf{x}\boldsymbol{\delta}_2 + \sigma_{12}\sigma_1^{-2}(y_1 - \mathbf{x}_1\boldsymbol{\beta}_1)](1 - \sigma_{12}^2\sigma_1^{-2})^{-1/2}\}$$

Combining all of these pieces [and noting the cancellation of $\text{P}(y_2 = 1 \mid \mathbf{x})$] we get

$$\ell_i(\boldsymbol{\theta}) = (1 - y_{i2}) \log[1 - \Phi(\mathbf{x}_i\boldsymbol{\delta}_2)] + y_{i2}(\log \Phi\{[\mathbf{x}_i\boldsymbol{\delta}_2 + \sigma_{12}\sigma_1^{-2}(y_{i1} - \mathbf{x}_{i1}\boldsymbol{\beta}_1)] \times (1 - \sigma_{12}^2\sigma_1^{-2})^{-1/2}\} + \log \phi[(y_{i1} - \mathbf{x}_{i1}\boldsymbol{\beta}_1)/\sigma_1] - \log(\sigma_1))$$

The partial log likelihood is obtained by summing $\ell_i(\boldsymbol{\theta})$ across *all* observations; $y_{i2} = 1$ picks out when y_{i1} is observed and therefore contains information for estimating $\boldsymbol{\beta}_1$.

Ahn and Powell (1993) show how to consistently estimate $\boldsymbol{\beta}_1$ without making any distributional assumptions; in particular, the selection equation need not have the probit form. Vella (1998) contains a recent survey.

17.4.2 Endogenous Explanatory Variables

We now study the sample selection model when one of the elements of $\mathbf{x}_1$ is thought to be correlated with u_1. Or, all the elements of $\mathbf{x}_1$ are exogenous in the population model but data are missing on an element of $\mathbf{x}_1$, and the reason data are missing might be systematically related to u_1. For simplicity, we focus on the case of a single endogenous explanatory variable.

The model in the population is

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{17.25}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2 \tag{17.26}$$

$$y_3 = 1(\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0) \tag{17.27}$$

The first equation is the structural equation of interest, the second equation is a linear projection for the potentially endogenous or missing variable y_2, and the third equation is the selection equation. We allow arbitrary correlation among u_1, v_2, and v_3.

The setup in equations (17.25)–(17.27) encompasses at least three cases of interest. The first occurs when y_2 is always observed but is endogenous in equation (17.25). An example is seen when y_1 is $\log(wage^o)$ and y_2 is years of schooling: years of schooling is generally available whether or not someone is in the workforce. The model also applies when y_2 is observed only along with y_1, as would happen if $y_1 = \log(wage^o)$ and y_2 is the ratio of the benefits offer to wage offer. As a second example, let y_1 be the percentage of voters supporting the incumbent in a congressional district, and let y_2 be intended campaign expenditures. Then $y_3 = 1$ if the incumbent runs for reelection, and we only observe (y_1, y_2) when $y_3 = 1$. A third application is to missing data only on y_2, as in Example 17.4 where y_2 is IQ score. In the last two cases, y_2 might in fact be exogenous in equation (17.25), but endogenous sample selection effectively makes y_2 endogenous in the selected sample.

If y_1 and y_2 were always observed along with $\mathbf{z}$, we would just estimate equation (17.25) by 2SLS if y_2 is endogenous. We can use the results from Section 17.2.1 to show that 2SLS with the inverse Mills ratio added to the regressors is consistent. Regardless of the data availability on y_1 and y_2, in the second step we use only observations for which both y_1 and y_2 are observed.

ASSUMPTION 17.2: (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) is observed when $y_3 = 1$; (b) (u_1, v_3) is independent of $\mathbf{z}$; (c) $v_3 \sim \text{Normal}(0, 1)$; (d) $\text{E}(u_1 \mid v_3) = \gamma_1 v_3$; and (e) $\text{E}(\mathbf{z}'v_2) = \mathbf{0}$ and, writing $\mathbf{z}\boldsymbol{\delta}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_2\boldsymbol{\delta}_{22}$, $\boldsymbol{\delta}_{22} \neq \mathbf{0}$.

Parts b, c, and d are identical to the corresponding assumptions when all explanatory variables are observed and exogenous. Assumption e is new, resulting from the endogeneity of y_2 in equation (17.25). It is important to see that Assumption 17.2e is identical to the rank condition needed for identifying equation (17.25) in the absence of sample selection. As we will see, stating identification in the population is not always sufficient, but, from a practical point of view, the focus should be on Assumption 17.2e.

To derive an estimating equation, write (in the population)

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{17.28}$$

where $g(\mathbf{z}, y_3) \equiv \mathrm{E}(u_1 \mid \mathbf{z}, y_3)$ and $e_1 \equiv u_1 - \mathrm{E}(u_1 \mid \mathbf{z}, y_3)$. By definition, $\mathrm{E}(e_1 \mid \mathbf{z}, y_3) = 0$. If we knew $g(\mathbf{z}, y_3)$ then, from Theorem 17.1, we could just estimate equation (17.28) by 2SLS on the selected sample ($y_3 = 1$) using instruments $[\mathbf{z}, g(\mathbf{z}, 1)]$. It turns out that we do know $g(\mathbf{z}, 1)$ up to some estimable parameters: $\mathrm{E}(u_1 \mid \mathbf{z}, y_3 = 1) = \gamma_1\lambda(\mathbf{z}\boldsymbol{\delta}_3)$. Since $\boldsymbol{\delta}_3$ can be consistently estimated by probit of y_3 on $\mathbf{z}$ (using the entire sample), we have the following:

Procedure 17.2: (a) Obtain $\hat{\boldsymbol{\delta}}_3$ from probit of y_3 on $\mathbf{z}$ using all observations. Obtain the estimated inverse Mills ratios, $\hat{\lambda}_{i3} = \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$.

(b) Using the selected subsample (for which we observe both y_1 and y_2), estimate the equation

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \gamma_1\hat{\lambda}_{i3} + error_i \tag{17.29}$$

by 2SLS, using instruments $(\mathbf{z}_i, \hat{\lambda}_{i3})$.

The steps in this procedure show that identification actually requires that $\mathbf{z}_2$ appear in the linear projection of y_2 onto $\mathbf{z}_1, \mathbf{z}_2$, *and* $\lambda(\mathbf{z}\boldsymbol{\delta}_3)$ in the selected subpopulation. It would be unusual if this statement were not true when the rank condition 17.2e holds in the population.

The hypothesis-of-no-selection problem (allowing y_2 to be endogenous or not), H_0: $\gamma_1 = 0$, is tested using the usual 2SLS t statistic for $\hat{\gamma}_1$. When $\gamma_1 \neq 0$, standard errors and test statistics should be corrected for the generated regressors problem, as in Chapter 6.

Example 17.7 (Education Endogenous and Sample Selection): In Example 17.6 we now allow *educ* to be endogenous in the wage offer equation, and we test for sample selection bias. Just as if we did not have a sample selection problem, we need IVs for *educ* that do not appear in the wage offer equation. As in Example 5.3, we use parents' education (*motheduc*, *fatheduc*) and husband's education as IVs. In addition,

we need some variables that affect labor force participation but not the wage offer; we use the same four variables as in Example 17.6. Therefore, all variables except *educ* (and, of course, the wage offer) are treated as exogenous.

Unless we have very reliable prior information, *all* exogenous variables should appear in the selection equation, and all should be listed as instruments in estimating equation (17.29) by 2SLS. Dropping some exogenous variables in either the selection equation or in estimating equation (17.29) imposes exclusion restrictions on a reduced-form equation, something that can be dangerous and is unnecessary. Therefore, in the labor force participation equation we include *exper*, $exper^2$, *nwifeinc*, *kidslt6*, *kidsge6*, *motheduc*, *fatheduc*, and *huseduc* (not *educ*). In estimating equation (17.29), the same set of variables, along with $\hat{\lambda}_3$, are used as IVs. The 2SLS coefficient on $\hat{\lambda}_3$ is .040 (se = .133), and so, again, there is little evidence of sample selection bias. The coefficient on *educ* is .088 (se = .021), which is similar to the 2SLS estimate obtained without the sample selection correction (see Example 5.3). Because there is little evidence of sample selection bias, the standard errors are not corrected for first-stage estimation of δ_3.

Importantly, Procedure 17.2 applies to any kind of endogenous variable y_2, including binary and other discrete variables, without any additional assumptions. This statement is true because the reduced form for y_2 is just a linear projection; we do not have to assume, for example, that v_2 is normally distributed or even independent of $\mathbf{z}$. As an example, we might wish to look at the effects of participation in a job training program on the subsequent wage offer, accounting for the fact that not all who participated in the program will be employed in the following period (y_2 is always observed in this case). If participation is voluntary, an instrument for it might be whether the person was randomly chosen as a potential participant.

Even if y_2 is exogenous in the population equation (17.25), when y_2 is sometimes missing we generally need an instrument for y_2 when selection is not ignorable [that is, $\mathrm{E}(u_1 \mid \mathbf{z}_1, y_2, y_3) \neq \mathrm{E}(u_1)$]. In Example 17.4 we could use family background variables and another test score, such as *KWW*, as IVs for *IQ*, assuming these are always observed. We would generally include all such variables in the reduced-form selection equation. Procedure 17.2 works whether we assume *IQ* is a proxy variable for ability or an indicator of ability (see Chapters 4 and 5).

As a practical matter, we should have at least *two* elements of $\mathbf{z}$ that are not also in $\mathbf{z}_1$; that is, we need at least two exclusion restrictions in the structural equation. Intuitively, for the procedure to be convincing, we should have at least one instrument for y_2 and another exogenous variable that determines selection. Suppose that the scalar z_2 is our only exogenous variable excluded from equation (17.25). Then,

under random sampling, the equation would be just identified. When we account for sample selection bias, the Mills ratio term in equation (17.29) is a function of $\mathbf{z}_1$ and z_2. While the nonlinearity of the Mills ratio technically allows us to identify $\boldsymbol{\delta}_1$ and α_1, it is unlikely to work very well in practice because of severe multicollinearity among the IVs. This situation is analogous to using the standard Heckit method when there are no exclusion restrictions in the structural equation (see Section 17.4.1).

If we make stronger assumptions, it is possible to estimate model (17.25)–(17.27) by partial maximum likelihood of the kind discussed in Problem 13.7. One possibility is to assume that (u_1, v_2, v_3) is trivariate normal and independent of $\mathbf{z}$. In addition to ruling out discrete y_2, such a procedure would be computationally difficult. If y_2 is binary, we can model it as $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 > 0]$, where $v_2 \mid \mathbf{z} \sim \text{Normal}(0, 1)$. But maximum likelihood estimation that allows any correlation matrix for (u_1, v_2, v_3) is complicated and less robust than Procedure 17.2.

17.4.3 Binary Response Model with Sample Selection

We can estimate binary response models with sample selection if we assume that the latent errors are bivariate normal and independent of the explanatory variables. Write the model as

$$y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0] \tag{17.30}$$

$$y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0] \tag{17.31}$$

where the second equation is the sample selection equation and y_1 is observed only when $y_2 = 1$; we assume that $\mathbf{x}$ is always observed. For example, suppose y_1 is an employment indicator and $\mathbf{x}_1$ contains a job training binary indicator (which we assume is exogenous), as well as other human capital and family background variables. We might lose track of some people who are eligible to participate in the program; this is an example of sample attrition. If attrition is systematically related to u_1, estimating equation (17.30) on the sample at hand can result in an inconsistent estimator of $\boldsymbol{\beta}_1$.

If we assume that (u_1, v_2) is independent of $\mathbf{x}$ with a zero-mean normal distribution (and unit variances), we can apply partial maximum likelihood. What we need is the density of y_1 conditional on $\mathbf{x}$ and $y_2 = 1$. We have essentially found this density in Chapter 15: in equation (15.55) set $\alpha_1 = 0$, replace $\mathbf{z}$ with $\mathbf{x}$, and replace $\boldsymbol{\delta}_1$ with $\boldsymbol{\beta}_1$. The parameter ρ_1 is still the correlation between u_1 and v_2. A two-step procedure can be applied: first, estimate $\boldsymbol{\delta}_2$ by probit of y_2 on $\mathbf{x}$. Then, estimate $\boldsymbol{\beta}_1$ and ρ_1 in the second stage using equation (15.55) along with $\text{P}(y_1 = 0 \mid \mathbf{x}, y_2 = 1)$.

A convincing analysis requires at least one variable in $\mathbf{x}$—that is, something that determines selection—that is not also in $\mathbf{x}_1$. Otherwise, identification is off of the nonlinearities in the probit models.

Allowing for endogenous explanatory variables in equation (17.30) along with sample selection is difficult, and it could be the focus of future research.

17.5 A Tobit Selection Equation

We now study the case where more information is available on sample selection, primarily in the context of incidental truncation. In particular, we assume that selection is based on the outcome of a Tobit, rather than a probit, equation. The analysis of the models in this section comes from Wooldridge (1998). The model in Section 17.5.1 is a special case of the model studied by Vella (1992) in the context of testing for selectivity bias.

17.5.1 Exogenous Explanatory Variables

We now consider the case where the selection equation is of the censored Tobit form. The population model is

$$y_1 = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \tag{17.32}$$

$$y_2 = \max(0, \mathbf{x}\boldsymbol{\delta}_2 + v_2) \tag{17.33}$$

where $(\mathbf{x}, y_2)$ is always observed in the population but y_1 is observed only when $y_2 > 0$. A standard example occurs when y_1 is the log of the hourly wage offer and y_2 is weekly or annual hours of labor supply.

ASSUMPTION 17.3: (a) $(\mathbf{x}, y_2)$ is always observed in the population, but y_1 is observed only when $y_2 > 0$; (b) (u_1, v_2) is independent of $\mathbf{x}$; (c) $v_2 \sim \text{Normal}(0, \tau_2^2)$; and (d) $\text{E}(u_1 \mid v_2) = \gamma_1 v_2$.

These assumptions are very similar to the assumptions for a probit selection equation. The only difference is that v_2 now has an unknown variance, since y_2 is a censored as opposed to binary variable.

Amemiya (1985) calls equations (17.32) and (17.33) the **type III Tobit model**, but we emphasize that equation (17.32) is the structural population equation of interest and that equation (17.33) simply determines when y_1 is observed. In the labor economics example, we are interested in the wage offer equation, and equation (17.33) is a reduced-form hours equation. It makes no sense to define y_1 to be, say, zero, just because we do not observe y_1.

The starting point is equation (17.21), just as in the probit selection case. Now define the selection indicator as $s_2 = 1$ if $y_2 > 0$, and $s_2 = 0$ otherwise. Since s_2 is a function of $\mathbf{x}$ and v_2, it follows immediately that

$$\mathrm{E}(y_1 \mid \mathbf{x}, v_2, s_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2 \tag{17.34}$$

This equation means that, if we could observe v_2, then an OLS regression of y_1 on $\mathbf{x}_1$, v_2 using the selected subsample would consistently estimate $(\boldsymbol{\beta}_1, \gamma_1)$, as we discussed in Section 17.2.1. While v_2 cannot be observed when $y_2 = 0$ (because when $y_2 = 0$, we only know that $v_2 \leq -\mathbf{x}\boldsymbol{\delta}_2$), for $y_2 > 0$, $v_2 = y_2 - \mathbf{x}\boldsymbol{\delta}_2$. Thus, if we knew $\boldsymbol{\delta}_2$, we would know v_2 whenever $y_2 > 0$. It seems reasonable that, because $\boldsymbol{\delta}_2$ can be consistently estimated by Tobit on the whole sample, we can replace v_2 with consistent estimates.

Procedure 17.3: (a) Estimate equation (17.33) by standard Tobit using all N observations. For $y_{i2} > 0$ (say $i = 1, 2, \ldots, N_1$), define

$$\hat{v}_{i2} = y_{i2} - \mathbf{x}_i\hat{\boldsymbol{\delta}}_2 \tag{17.35}$$

(b) Using observations for which $y_{i2} > 0$, estimate $\boldsymbol{\beta}_1$, γ_1 by the OLS regression

$$y_{i1} \text{ on } \mathbf{x}_{i1}, \hat{v}_{i2} \quad i = 1, 2, \ldots, N_1 \tag{17.36}$$

This regression produces consistent, $\sqrt{N}$-asymptotically normal estimators of $\boldsymbol{\beta}_1$ and γ_1 under Assumption 17.3.

The statistic to test for selectivity bias is just the usual t statistic on $\hat{v}_{i2}$ in regression (17.36). This was suggested by Vella (1992). Wooldridge (1998) showed that this procedure also solves the selection problem when $\gamma_1 \neq 0$.

It seems likely that there is an efficiency gain over Procedure 17.1. If v_2 were known and we could use regression (17.36) for the entire population, there would definitely be an efficiency gain: the error variance is reduced by conditioning on v_2 along with $\mathbf{x}$, and there would be no heteroskedasticity in the population. See Problem 4.5.

Unlike in the probit selection case, $\mathbf{x}_1 = \mathbf{x}$ causes no problems here: v_2 always has separate variation from $\mathbf{x}_1$ because of variation in y_2. We do not need to rely on the nonlinearity of the inverse Mills ratio.

Example 17.8 (Wage Offer Equation for Married Women): We now apply Procedure 17.3 to the wage offer equation for married women in Example 17.6. (We assume education is exogenous.) The only difference is that the first-step estimation is Tobit, rather than probit, and we include the Tobit residuals as the additional

explanatory variables, not the inverse Mills ratio. In regression (17.36), the coefficient on $\hat{v}_2$ is $-.000053$ (se = .000041), which is somewhat more evidence of a sample selection problem, but we still do not reject the null hypothesis H_0: $\gamma_1 = 0$ at even the 15 percent level against a two-sided alternative. Further, the coefficient on *educ* is .103 (se = .015), which is not much different from the OLS and Heckit estimates. (Again, we use the usual OLS standard error.) When we include all exogenous variables in the wage offer equation, the estimates from Procedure 17.3 are much more stable than the Heckit estimates. For example, the coefficient on *educ* becomes .093 (se = .016), which is comparable to the OLS estimates discussed in Example 17.6.

For partial maximum likelihood estimation, we assume that (u_1, v_2) is jointly normal, and we use the density for $f(y_2 \mid \mathbf{x})$ for the entire sample and the conditional density $f(y_1 \mid \mathbf{x}, y_2, s_2 = 1) = f(y_1 \mid \mathbf{x}, y_2)$ for the selected sample. This approach is fairly straightforward because, when $y_2 > 0$, $y_1 \mid \mathbf{x}, y_2 \sim \text{Normal}[\mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1(y_2 - \mathbf{x}\boldsymbol{\delta}_2), \eta_1^2]$, where $\eta_1^2 = \sigma_1^2 - \sigma_{12}^2/\tau_2^2$, $\sigma_1^2 = \text{Var}(u_1)$, and $\sigma_{12} = \text{Cov}(u_1, v_2)$. The log likelihood for observation i is

$$\ell_i(\boldsymbol{\theta}) = s_{i2} \log f(y_{i1} \mid \mathbf{x}_i, y_{i2}; \boldsymbol{\theta}) + \log f(y_{i2} \mid \mathbf{x}_i; \boldsymbol{\delta}_2, \tau_2^2) \tag{17.37}$$

where $f(y_{i1} \mid \mathbf{x}_i, y_{i2}; \boldsymbol{\theta})$ is the Normal$[\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \gamma_1(y_{i2} - \mathbf{x}_i\boldsymbol{\delta}_2), \eta_1^2]$ distribution, evaluated at y_{i1}, and $f(y_{i2} \mid \mathbf{x}_i; \boldsymbol{\delta}_2, \tau_2^2)$ is the standard censored Tobit density [see equation (16.19)]. As shown in Problem 13.7, the usual MLE theory can be used even though the log-likelihood function is not based on a true conditional density.

It is possible to obtain sample selection corrections and tests for various other nonlinear models when the selection rule is of the Tobit form. For example, suppose that the binary variable y_1 given $\mathbf{z}$ follows a probit model, but it is observed only when $y_2 > 0$. A valid test for selection bias is to include the Tobit residuals, $\hat{v}_2$, in a probit of y_1 on $\mathbf{z}$, $\hat{v}_2$ using the selected sample; see Vella (1992). This procedure also produces consistent estimates (up to scale), as can be seen by applying the maximum likelihood results in Section 17.2.2 along with two-step estimation results.

Honoré, Kyriazidou, and Udry (1997) show how to estimate the parameters of the type III Tobit model without making distributional assumptions.

17.5.2 Endogenous Explanatory Variables

We explicitly consider the case of a single endogenous explanatory variable, as in Section 17.4.2. We use equations (17.25) and (17.26), and, in place of equation (17.27), we have a Tobit selection equation:

$$y_3 = \max(0, \mathbf{z}\boldsymbol{\delta}_3 + v_3) \tag{17.38}$$

ASSUMPTION 17.4: (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) is observed when $y_3 > 0$; (b) (u_1, v_3) is independent of $\mathbf{z}$; (c) $v_3 \sim \text{Normal}(0, \tau_3^2)$; (d) $\text{E}(u_1 \mid v_3) = \gamma_1 v_3$; and (e) $\text{E}(\mathbf{z}'v_2) = \mathbf{0}$ and, writing $\mathbf{z}\boldsymbol{\delta}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_2\boldsymbol{\delta}_{22}$, $\boldsymbol{\delta}_{22} \neq \mathbf{0}$.

Again, these assumptions are very similar to those used with a probit selection mechanism.

To derive an estimating equation, write

$$y_1 = \mathbf{z}_1\delta_1 + \alpha_1 y_2 + \gamma_1 v_3 + e_1 \tag{17.39}$$

where $e_1 \equiv u_1 - \text{E}(u_1 \mid v_3)$. Since (e_1, v_3) is independent of $\mathbf{z}$ by Assumption 17.4b, $\text{E}(e_1 \mid \mathbf{z}, v_3) = 0$. From Theorem 17.1, if v_3 were observed, we could estimate equation (17.39) by 2SLS on the selected sample using instruments $(\mathbf{z}, v_3)$. As before, we can estimate v_3 when $y_3 > 0$, since $\boldsymbol{\delta}_3$ can be consistently estimated by Tobit of y_3 on $\mathbf{z}$ (using the entire sample).

Procedure 17.4: (a) Obtain $\hat{\boldsymbol{\delta}}_3$ from Tobit of y_3 on $\mathbf{z}$ using all observations. Obtain the Tobit residuals $\hat{v}_{i3} = y_{i3} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_3$ for $y_{i3} > 0$.

(b) Using the selected subsample, estimate the equation

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \gamma_1 \hat{v}_{i3} + error_i \tag{17.40}$$

by 2SLS, using instruments $(\mathbf{z}_i, \hat{v}_{i3})$. The estimators are $\sqrt{N}$-consistent and asymptotically normal under Assumption 17.4.

Comments similar to those after Procedure 17.2 hold here as well. Strictly speaking, identification really requires that $\mathbf{z}_2$ appear in the linear projection of y_2 onto $\mathbf{z}_1$, $\mathbf{z}_2$, and v_3 in the selected subpopulation. The null of no selection bias is tested using the 2SLS t statistic (or maybe its heteroskedasticity-robust version) on $\hat{v}_{i3}$. When $\gamma_1 \neq 0$, standard errors should be corrected using two-step methods.

As in the case with a probit selection equation, the endogenous variable y_2 can be continuous, discrete, censored, and so on. Extending the method to multiple endogenous explanatory variables is straightforward. The only restriction is the usual one for linear models: we need enough instruments to identify the structural equation. See Problem 17.6 for an application to the Mroz data.

An interesting special case of model (17.25), (17.26), and (17.38) is when $y_2 = y_3$. Actually, because we only use observations for which $y_3 > 0$, $y_2 = y_3^*$ is also allowed, where $y_3^* = \mathbf{z}\boldsymbol{\delta}_3 + v_3$. Either way, the variable that determines selection also appears in the structural equation. This special case could be useful when sample selection is caused by a corner solution outcome on y_3 (in which case $y_2 = y_3$ is

natural) or because y_3^* is subject to data censoring (in which case $y_2 = y_3^*$ is more realistic). An example of the former occurs when y_3 is hours worked and we assume hours appears in the wage offer function. As a data-censoring example, suppose that y_1 is a measure of growth in an infant's weight starting from birth and that we observe y_1 only if the infant is brought into a clinic within three months. Naturally, birth weight depends on age, and so y_3^*—length of time between the first and second measurements, which has quantitiative meaning—appears as an explanatory variable in the equation for y_1. We have a data-censoring problem for y_3^*, which causes a sample selection problem for y_1. In this case, we would estimate a censored regression model for y_3 [or, possibly, $\log(y_3)$] to account for the data censoring. We would include the residuals $\hat{v}_{i3} = y_{i3} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_3$ in equation (17.40) for the noncensored observations. As our extra instrument we might use distance from the child's home to the clinic.

17.6 Estimating Structural Tobit Equations with Sample Selection

We briefly show how a structural Tobit model can be estimated using the methods of the previous section. As an example, consider the structural labor supply model

$$\log(w^o) = \mathbf{z}_1\boldsymbol{\beta}_1 + u_1 \tag{17.41}$$

$$h = \max[0, \mathbf{z}_2\boldsymbol{\beta}_2 + \alpha_2 \log(w^o) + u_2] \tag{17.42}$$

This system involves simultaneity and sample selection because we observe w^o only if $h > 0$.

The general form of the model is

$$y_1 = \mathbf{z}_1\boldsymbol{\beta}_1 + u_1 \tag{17.43}$$

$$y_2 = \max(0, \mathbf{z}_2\boldsymbol{\beta}_2 + \alpha_2 y_1 + u_2) \tag{17.44}$$

ASSUMPTION 17.5: (a) $(\mathbf{z}, y_2)$ is always observed; y_1 is observed when $y_2 > 0$; (b) (u_1, u_2) is independent of $\mathbf{z}$ with a zero-mean bivariate normal distribution; and (c) $\mathbf{z}_1$ contains at least one element whose coefficient is different from zero that is not in $\mathbf{z}_2$.

As always, it is important to see that equations (17.43) and (17.44) constitute a model describing a population. If y_1 were always observed, then equation (17.43) could be estimated by OLS. If, in addition, u_1 and u_2 were uncorrelated, equation (17.44) could be estimated by censored Tobit. Correlation between u_1 and u_2 could be handled by the methods of Section 16.6.2. Now, we require new methods, whether or not u_1 and u_2 are uncorrelated, because y_1 is not observed when $y_2 = 0$.

The restriction in Assumption 17.5c is needed to identify the structural parameters $(\boldsymbol{\beta}_2, \alpha_2)$ ($\boldsymbol{\beta}_1$ is always identified). To see that this condition is needed, and for finding the reduced form for y_2, it is useful to introduce the latent variable

$$y_2^* \equiv \mathbf{z}_2\boldsymbol{\beta}_2 + \alpha_2 y_1 + u_2 \tag{17.45}$$

so that $y_2 = \max(0, y_2^*)$. If equations (17.43) and (17.45) make up the system of interest—that is, if y_1 and y_2^* are always observed—then $\boldsymbol{\beta}_1$ is identified without further restrictions, but identification of α_2 and $\boldsymbol{\beta}_2$ requires exactly Assumption 17.5c. This turns out to be sufficient even when y_2 follows a Tobit model and we have nonrandom sample selection.

The reduced form for y_2^* is $y_2^* = \mathbf{z}\boldsymbol{\delta}_2 + v_2$. Therefore, we can write the reduced form of equation (17.44) as

$$y_2 = \max(0, \mathbf{z}\boldsymbol{\delta}_2 + v_2). \tag{17.46}$$

But then equations (17.43) and (17.46) constitute the model we studied in Section 17.5.1. The vector $\boldsymbol{\delta}_2$ is consistently estimated by Tobit, and $\boldsymbol{\beta}_1$ is estimated as in Procedure 17.3. The only remaining issue is how to estimate the structural parameters of equation (17.44), α_2 and $\boldsymbol{\beta}_2$. In the labor supply case, these are the labor supply parameters.

Assuming identification, estimation of $(\alpha_2, \boldsymbol{\beta}_2)$ is fairly straightforward after having estimated $\boldsymbol{\beta}_1$. To see this point, write the reduced form of y_2 in terms of the structural parameters as

$$y_2 = \max[0, \mathbf{z}_2\boldsymbol{\beta}_2 + \alpha_2(\mathbf{z}_1\boldsymbol{\beta}_1) + v_2] \tag{17.47}$$

Under joint normality of u_1 and u_2, v_2 is normally distributed. Therefore, if $\boldsymbol{\beta}_1$ were *known*, $\boldsymbol{\beta}_2$ and α_2 could be estimated by standard Tobit using $\mathbf{z}_2$ and $\mathbf{z}_1\boldsymbol{\beta}_1$ as regressors. Operationalizing this procedure requires replacing $\boldsymbol{\beta}_1$ with its consistent estimator. Thus, using all observations, $\boldsymbol{\beta}_2$ and α_2 are estimated from the Tobit equation

$$y_{i2} = \max[0, \mathbf{z}_{i2}\boldsymbol{\beta}_2 + \alpha_2(\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_1) + error_i] \tag{17.48}$$

To summarize, we have the following:

Procedure 17.5: (a) Use Procedure 17.3 to obtain $\hat{\boldsymbol{\beta}}_1$.
(b) Obtain $\hat{\boldsymbol{\beta}}_2$ and $\hat{\alpha}_2$ from the Tobit in equation (17.48).

In applying this procedure, it is important to note that the explanatory variable in equation (17.48) is $\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_1$ for all i. These are *not* the fitted values from regression (17.36), which depend on $\hat{v}_{i2}$. Also, it may be tempting to use y_{i1} in place of $\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_1$ for

that part of the sample for which y_{i1} is observed. This approach is not a good idea: the estimators are inconsistent in this case.

The estimation in equation (17.48) makes it clear that the procedure fails if $\mathbf{z}_1$ does not contain at least one variable not in $\mathbf{z}_2$. If $\mathbf{z}_1$ is a subset of $\mathbf{z}_2$, then $\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_1$ is a linear combination of $\mathbf{z}_{i2}$, and so perfect multicollinearity will exist in equation (17.48).

Estimating $\text{Avar}(\hat{\alpha}_2, \hat{\boldsymbol{\beta}}_2)$ is even messier than estimating $\text{Avar}(\hat{\boldsymbol{\beta}}_1)$, since $(\hat{\alpha}_2, \hat{\boldsymbol{\beta}}_2)$ comes from a three-step procedure. Often just the usual Tobit standard errors and test statistics reported from equation (17.48) are used, even though these are not strictly valid. By setting the problem up as a large GMM problem, as illustrated in Chapter 14, correct standard errors and test statistics can be obtained.

Under Assumption 17.5, a full maximum likelihood approach is possible. In fact, the log-likelihood function can be constructed from equations (17.43) and (17.47), and it has a form very similar to equation (17.37). The only difference is that nonlinear restrictions are imposed automatically on the structural parameters. In addition to making it easy to obtain valid standard errors, MLE is desirable because it allows us to estimate $\sigma_2^2 = \text{Var}(u_2)$, which is needed to estimate average partial effects in equation (17.44).

In examples such as labor supply, it is not clear where the elements of $\mathbf{z}_1$ that are not in $\mathbf{z}_2$ might come from. One possibility is a union binary variable, if we believe that union membership increases wages (other factors accounted for) but has no effect on labor supply once wage and other factors have been controlled for. This approach would require knowing union status for people whether or not they are working in the period covered by the survey. In some studies past experience is assumed to affect wage—which it certainly does—and is assumed not to appear in the labor supply function, a tenuous assumption.

17.7 Sample Selection and Attrition in Linear Panel Data Models

In our treatment of panel data models we have assumed that a balanced panel is available—each cross section unit has the same time periods available. Often, some time periods are missing for some units in the population of interest, and we are left with an **unbalanced panel**. Unbalanced panels can arise for several reasons. First, the survey design may simply rotate people or firms out of the sample based on prespecified rules. For example, if a survey of individuals begins at time $t = 1$, at time $t = 2$ some of the original people may be dropped and new people added. At $t = 3$ some additional people might be dropped and others added; and so on. This is an example of a **rotating panel**.

Provided the decision to rotate units out of a panel is made randomly, unbalanced panels are fairly easy to deal with, as we will see shortly. A more complicated problem arises when attrition from a panel is due to units electing to drop out. If this decision is based on factors that are systematically related to the response variable, even after we condition on explanatory variables, a sample selection problem can result—just as in the cross section case. Nevertheless, a panel data set provides us with the means to handle, in a simple fashion, attrition that is based on a time-constant, unobserved effect, provided we use first-differencing methods; we show this in Section 17.7.3.

A different kind of sample selection problem occurs when people do not disappear from the panel but certain variables are unobserved for at least some time periods. This is the incidental truncation problem discussed in Section 17.4. A leading case is estimating a wage offer equation using a panel of individuals. Even if the population of interest is people who are employed in the initial year, some people will become unemployed in subsequent years. For those people we cannot observe a wage offer, just as in the cross-sectional case. This situation is different from the attrition problem where people leave the sample entirely and, usually, do not reappear in later years. In the incidental truncation case we observe some variables on everyone in each time period.

17.7.1 Fixed Effects Estimation with Unbalanced Panels

We begin by studying assumptions under which the usual fixed effects estimator on the unbalanced panel is consistent. The model is the usual linear, unobserved effects model under random sampling in the cross section: for any i,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \qquad t = 1, \ldots, T \tag{17.49}$$

where $\mathbf{x}_{it}$ is $1 \times K$ and $\boldsymbol{\beta}$ is the $K \times 1$ vector of interest. As before, we assume that N cross section observations are available and the asymptotic analysis is as $N \to \infty$. We explicitly cover the case where c_i is allowed to be correlated with $\mathbf{x}_{it}$, so that all elements of $\mathbf{x}_{it}$ are time varying. A random effects analysis is also possible under stronger assumptions; see, for example, Verbeek and Nijman (1992, 1996).

We covered the case where all T time periods are available in Chapters 10 and 11. Now we consider the case where some time periods might be missing for some of the cross section draws. Think of $t = 1$ as the first time period for which data on anyone in the population are available, and $t = T$ as the last possible time period. For a random draw i from the population, let $\mathbf{s}_i \equiv (s_{i1}, \ldots, s_{iT})'$ denote the $T \times 1$ vector of selection indicators: $s_{it} = 1$ if $(\mathbf{x}_{it}, y_{it})$ is observed, and zero otherwise. Generally, we

have an unbalanced panel. We can treat $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i): i = 1, 2, \ldots, N\}$ as a random sample from the population; the selection indicators tell us which time periods are missing for each i.

We can easily find assumptions under which the fixed effects estimator on the unbalanced panel is consistent by writing it as

$$\hat{\boldsymbol{\beta}} = \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right)$$

$$= \boldsymbol{\beta} + \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'u_{it}\right) \tag{17.50}$$

where we define

$$\ddot{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}, \qquad \ddot{y}_{it} \equiv y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}, \qquad \text{and} \qquad T_i \equiv \sum_{t=1}^{T} s_{it}$$

That is, T_i is the number of time periods observed for cross section i, and we apply the within transformation on the available time periods.

If fixed effects on the unbalanced panel is to be consistent, we should have $\mathrm{E}(s_{it}\ddot{\mathbf{x}}_{it}'u_{it}) = 0$ for all t. Now, since $\ddot{\mathbf{x}}_{it}$ depends on all of $\mathbf{x}_i$ and $\mathbf{s}_i$, a form of strict exogeneity is needed.

ASSUMPTION 17.6: (a) $\mathrm{E}(u_{it} \mid \mathbf{x}_i, \mathbf{s}_i, c_i) = 0$, $t = 1, 2, \ldots, T$; (b) $\sum_{t=1}^{T} \mathrm{E}(s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it})$ is nonsingular; and (c) $\mathrm{E}(\mathbf{u}_i\mathbf{u}_i' \mid \mathbf{x}_i, \mathbf{s}_i, c_i) = \sigma_u^2\mathbf{I}_T$.

Under Assumption 17.6a, $\mathrm{E}(s_{it}\ddot{\mathbf{x}}_{it}'u_{it}) = \mathbf{0}$ from the law of iterated expectations [because $s_{it}\ddot{\mathbf{x}}_{it}$ is a function of $(\mathbf{x}_i, \mathbf{s}_i)$]. The second assumption is the rank condition on the expected outer product matrix, after accounting for sample selection; naturally, it rules out time-constant elements in $\mathbf{x}_{it}$. These first two assumptions ensure consistency of FE on the unbalanced panel.

In the case of a randomly rotating panel, and in other cases where selection is entirely random, $\mathbf{s}_i$ is independent of $(\mathbf{u}_i, \mathbf{x}_i, c_i)$, in which case Assumption 17.6a follows under the standard fixed effects assumption $\mathrm{E}(u_{it} \mid \mathbf{x}_i, c_i) = 0$ for all t. In this case, the natural assumptions on the population model imply consistency and asympotic normality on the unbalanced panel. Assumption 17.6a also holds under much weaker conditions. In particular, it does not assume anything about the relationship between $\mathbf{s}_i$ and $(\mathbf{x}_i, c_i)$. Therefore, if we think selection in all time periods is correlated with c_i or $\mathbf{x}_i$, but that u_{it} is mean independent of $\mathbf{s}_i$ given $(\mathbf{x}_i, c_i)$ for all t, then FE on the

unbalanced panel is consistent and asymptotically normal. This conclusion may be a reasonable approximation, especially for short panels. What Assumption 17.6a rules out is selection that is partially correlated with the idiosyncratic errors, u_{it}.

A random effects analysis on the unbalanced panel requires much stronger assumptions: it effectively requires $\mathbf{s}_i$ and c_i to be independent. Random effects will be inconsistent if, say, in a wage offer equation, less able people are more likely to disappear from the sample. This conclusion is true even if $\mathrm{E}(c_i \mid \mathbf{x}_i) = 0$ (Assumption RE.1 from Chapter 10) holds in the underlying population; see Wooldridge (1995a) for further discussion.

When we add Assumption 17.6c, standard inference procedures based on FE are valid. In particular, under Assumptions 17.6a and 17.6c,

$$\mathrm{Var}\left(\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'u_{it}\right) = \sigma_u^2\left[\sum_{t=1}^{T}\mathrm{E}(s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it})\right]$$

Therefore, the asymptotic variance of the fixed effects estimator is estimated as

$$\hat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1} \tag{17.51}$$

The estimator $\hat{\sigma}_u^2$ can be derived from

$$\mathrm{E}\left(\sum_{t=1}^{T} s_{it}\ddot{u}_{it}^2\right) = \mathrm{E}\left[\sum_{t=1}^{T} s_{it}\mathrm{E}(\ddot{u}_{it}^2 \mid \mathbf{s}_i)\right] = \mathrm{E}\{T_i[\sigma_u^2(1 - 1/T_i)]\} = \sigma_u^2\mathrm{E}[(T_i - 1)]$$

Now, define the FE residuals as $\hat{u}_{it} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}$ when $s_{it} = 1$. Then, because $N^{-1}\sum_{i=1}^{N}(T_i - 1) \xrightarrow{p} \mathrm{E}(T_i - 1)$,

$$\hat{\sigma}_u^2 = \left[N^{-1}\sum_{i=1}^{N}(T_i - 1)\right]^{-1} N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{u}_{it}^2 = \left[\sum_{i=1}^{N}(T_i - 1)\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{u}_{it}^2$$

is consistent for σ_u^2 as $N \to \infty$. Standard software packages also make a degrees-of-freedom adjustment by subtracting K from $\sum_{i=1}^{N}(T_i - 1)$. It follows that all of the usual test statistics based on an unbalanced fixed effects analysis are valid. In particular, the dummy variable regression discussed in Chapter 10 produces asymptotically valid statistics.

Because the FE estimator uses time demeaning, any unit i for which $T_i = 1$ drops out of the fixed effects estimator. To use these observations we would need to add more assumptions, such as the random effects assumption $\mathrm{E}(c_i \mid \mathbf{x}_i, \mathbf{s}_i) = 0$.

Relaxing Assumption 17.6c is easy: just apply the robust variance matrix estimator in equation (10.59) to the unbalanced panel. The only changes are that the rows of $\ddot{\mathbf{X}}_i$ are $s_{it}\ddot{\mathbf{x}}_{it}$ and the elements of $\hat{\mathbf{u}}_i$ are $s_{it}\hat{u}_{it}$, $t = 1, \ldots, T$.

Under Assumption 17.6, it is also valid to used a standard fixed effects analysis on any balanced subset of the unbalanced panel; in fact, we can condition on any outcomes of the s_{it}. For example, if we use unit i only when observations are available in all time periods, we are conditioning on $s_{it} = 1$ for all t.

Using similar arguments, it can be shown that any kind of differencing method on any subset of the observed panel is consistent. For example, with $T = 3$, we observe cross section units with data for one, two, or three time periods. Those units with $T_i = 1$ drop out, but any other combinations of differences can be used in a pooled OLS analysis. The analogues of Assumption 17.6 for first differencing—for example, Assumption 17.6c is replaced with $\mathrm{E}(\Delta\mathbf{u}_i\Delta\mathbf{u}_i' \mid \mathbf{x}_i, \mathbf{s}_i, c_i) = \sigma_e^2\mathbf{I}_{T-1}$—ensure that the usual statistics from pooled OLS on the unbalanced first differences are asymptotically valid.

17.7.2 Testing and Correcting for Sample Selection Bias

The results in the previous subsection imply that sample selection in a fixed effects context is only a problem when selection is related to the idiosyncratic errors, u_{it}. Therefore, any test for selection bias should test only this assumption. A simple test was suggested by Nijman and Verbeek (1992) in the context of random effects estimation, but it works for fixed effects as well: add, say, the lagged selection indicator, $s_{i,t-1}$, to the equation, estimate the model by fixed effects (on the unbalanced panel), and do a t test (perhaps making it fully robust) for the significance of $s_{i,t-1}$. (This method loses the first time period for *all* observations.) Under the null hypothesis, u_{it} is uncorrelated with s_{ir} for all r, and so selection in the previous time period should not be significant in the equation at time t. (Incidentally, it never makes sense to put s_{it} in the equation at time t because $s_{it} = 1$ for all i and t in the selected subsample.)

Putting $s_{i,t-1}$ does not work if $s_{i,t-1}$ is unity whenever s_{it} is unity because then there is no variation in $s_{i,t-1}$ in the selected sample. This is the case in attrition problems if (say) a person can only appear in period t if he or she appeared in $t - 1$. An alternative is to include a lead of the selection indicator, $s_{i,t+1}$. For observations i that are in the sample every time period, $s_{i,t+1}$ is always zero. But for attriters, $s_{i,t+1}$ switches from zero to one in the period just before attrition. If we use fixed effects or first differencing, we need $T > 2$ time periods to carry out the test.

For incidental truncation problems it makes sense to extend Heckman's (1976) test to the unobserved effects panel data context. This is done in Wooldridge (1995a). Write the equation of interest as

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + u_{it1}, \qquad t = 1, \ldots, T \tag{17.52}$$

Initially, suppose that y_{it1} is observed only if the binary selection indicator, s_{it2}, is unity. Let $\mathbf{x}_{it}$ denote the set of all exogenous variables at time t; we assume that these are observed in every time period, and $\mathbf{x}_{it1}$ is a subset of $\mathbf{x}_{it}$. Suppose that, for each t, s_{it2} is determined by the probit equation

$$s_{it2} = 1[\mathbf{x}_i\boldsymbol{\psi}_{t2} + v_{it2} > 0], \qquad v_{it2} \mid \mathbf{x}_i \sim \text{Normal}(0, 1) \tag{17.53}$$

where $\mathbf{x}_i$ contains unity. This is best viewed as a reduced-form selection equation: we let the explanatory variables in all time periods appear in the selection equation at time t to allow for general selection models, including those with unobserved effect and the Chamberlain (1980) device discussed in Section 15.8.2, as well as certain dynamic models of selection. A Mundlak (1978) approach would replace $\mathbf{x}_i$ with $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ at time t and assume that coefficients are constant across time. [See equation (15.68).] Then the parameters can be estimated by pooled probit, greatly conserving on degrees of freedom. Such conservation may be important for small N. For testing purposes, under the null hypothesis it does not matter whether equation (17.53) is the proper model of sample selection, but we will need to assume equation (17.53), or a Mundlak version of it, when correcting for sample selection.

Under the null hypothesis in Assumption 17.6a (with the obvious notational changes), the inverse Mills ratio obtained from the sample selection probit should not be significant in the equation estimated by fixed effects. Thus, let $\hat{\lambda}_{it2}$ be the estimated Mills ratios from estimating equation (17.53) by pooled probit across i and t. Then a valid test of the null hypothesis is a t statistic on $\hat{\lambda}_{it2}$ in the FE estimation on the unbalanced panel. Under Assumption 17.6c the usual t statistic is valid, but the approach works whether or not the u_{it1} are homoskedastic and serially uncorrelated: just compute the robust standard error. Wooldridge (1995a) shows formally that the first-stage estimation of $\boldsymbol{\psi}_2$ does not affect the limiting distribution of the t statistic under H_0. This conclusion also follows from the results in Chapter 12 on M-estimation.

Correcting for sample selection requires much more care. Unfortunately, under any assumptions that actually allow for an unobserved effect in the underlying selection equation, adding $\hat{\lambda}_{it2}$ to equation (17.52) and using FE does not produce consistent estimators. To see why, suppose

$$s_{it2} = 1[\mathbf{x}_{it}\boldsymbol{\delta}_2 + c_{i2} + a_{it2} > 0], \qquad a_{it2} \mid (\mathbf{x}_i, c_{i1}, c_{i2}) \sim \text{Normal}(0, 1) \tag{17.54}$$

Then, to get equation (17.53), v_{it2} depends on a_{it2} and, at least partially, on c_{i2}. Now, suppose we make the strong assumption $\text{E}(u_{it1} \mid \mathbf{x}_i, c_{i1}, c_{i2}, \mathbf{v}_{i2}) = g_{i1} + \rho_1 v_{it2}$, which would hold under the assumption that the (u_{it1}, a_{it2}) are independent across t condi-

tional on $(\mathbf{x}_i, c_{i1}, c_{i2})$. Then we have

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \rho_1 \mathrm{E}(v_{it2} \mid \mathbf{x}_i, \mathbf{s}_{i2}) + (c_{i1} + g_{i1}) + e_{it1} + \rho_1[v_{it2} - \mathrm{E}(v_{it2} \mid \mathbf{x}_i, \mathbf{s}_{i2})]$$

The composite error, $e_{it1} + \rho_1[v_{it2} - \mathrm{E}(v_{it2} \mid \mathbf{x}_i, \mathbf{s}_{i2})]$, is uncorrelated with any function of $(\mathbf{x}_i, \mathbf{s}_{i2})$. The problem is that $\mathrm{E}(v_{it2} \mid \mathbf{x}_i, \mathbf{s}_{i2})$ depends on all elements in $\mathbf{s}_{i2}$, and this expectation is complicated for even small T.

A method that does work is available using Chamberlain's approach to panel data models, but we need some linearity assumptions on the expected values of u_{it1} and c_{i1} given $\mathbf{x}_i$ and v_{it2}.

ASSUMPTION 17.7: (a) The selection equation is given by equation (17.53); (b) $\mathrm{E}(u_{it1} \mid \mathbf{x}_i, v_{it2}) = \mathrm{E}(u_{it1} \mid v_{it2}) = \rho_{t1} v_{it2}$, $t = 1, \ldots, T$; and (c) $\mathrm{E}(c_{i1} \mid \mathbf{x}_i, v_{it2}) = \mathrm{L}(c_{i1} \mid 1, \mathbf{x}_i, v_{it2})$

The second assumption is standard and follows under joint normality of (u_{it1}, v_{it2}) when this vector is independent of $\mathbf{x}_i$. Assumption 17.7c implies that

$$\mathrm{E}(c_{i1} \mid \mathbf{x}_i, v_{it2}) = \mathbf{x}_i\boldsymbol{\pi}_1 + \phi_{t1} v_{it2}$$

where, by equation (17.53) and iterated expectations, $\mathrm{E}(c_{i1} \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\pi}_1 + \mathrm{E}(v_{it2} \mid \mathbf{x}_{it}) = \mathbf{x}_i\boldsymbol{\pi}_1$. These assumptions place no restrictions on the serial dependence in (u_{it1}, v_{it2}). They do imply that

$$\mathrm{E}(y_{it1} \mid \mathbf{x}_i, v_{it2}) = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \mathbf{x}_i\boldsymbol{\pi}_1 + \gamma_{t1} v_{it2} \tag{17.55}$$

where $\gamma_{t1} \equiv \rho_{t1} + \phi_{t1}$
Conditioning on $s_{it2} = 1$ gives

$$\mathrm{E}(y_{it1} \mid \mathbf{x}_i, s_{it2} = 1) = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \mathbf{x}_i\boldsymbol{\pi}_1 + \gamma_{t1}\lambda(\mathbf{x}_i\boldsymbol{\psi}_{t2})$$

Therefore, we can consistently estimate $\boldsymbol{\beta}_1$ by first estimating a probit of s_{it2} on $\mathbf{x}_i$ for each t and then saving the inverse Mills ratio, $\hat{\lambda}_{it2}$, all i and t. Next, run the pooled OLS regression using the selected sample:

$$y_{it1} \text{ on } \mathbf{x}_{it1}, \mathbf{x}_i, \hat{\lambda}_{it2}, d2_t\hat{\lambda}_{it2}, \ldots, dT_t\hat{\lambda}_{it2} \quad \text{for all } s_{it2} = 1 \tag{17.56}$$

where $d2_t$ through dT_t are time dummies. If γ_{t1} in equation (17.55) is constant across t, simply include $\hat{\lambda}_{it2}$ by itself in equation (17.56).

The asymptotic variance of $\hat{\boldsymbol{\beta}}_1$ needs to be corrected for general heteroskedasticity and serial correlation, as well as first-stage estimation of the $\boldsymbol{\psi}_{t2}$. These corrections can be made using the formulas for two-step M-estimation from Chapter 12; Wooldridge (1995a) contains the formulas.

If the selection equation is of the Tobit form, we have somewhat more flexibility. Write the selection equation now as

$$y_{it2} = \max(0, \mathbf{x}_i\boldsymbol{\psi}_{t2} + v_{it2}), \qquad v_{it2} \mid \mathbf{x}_i \sim \text{Normal}(0, \sigma_{t2}^2) \tag{17.57}$$

where y_{it1} is observed if $y_{it2} > 0$. Then, under Assumption 17.6, with the Tobit selection equation in place of equation (17.53), consistent estimation follows from the pooled regression (17.56) where $\hat{\lambda}_{it2}$ is replaced by the Tobit residuals, $\hat{v}_{it2}$ when $y_{it2} > 0$ ($s_{it2} = 1$). The Tobit residuals are obtained from the T cross section Tobits in equation (17.57); alternatively, especially with small N, we can use a Mundlak-type approach and use pooled Tobit with $\mathbf{x}_i\boldsymbol{\psi}_{t2}$ replaced with $\mathbf{x}_{it}\boldsymbol{\delta}_2 + \bar{\mathbf{x}}_i\boldsymbol{\pi}_2$; see equation (16.52).

It is easy to see that we can add $\alpha_1 y_{it2}$ to the structural equation (17.52), provided we make an explicit exclusion restriction in Assumption 17.7. In particular, we must assume that $\mathrm{E}(c_{i1} \mid \mathbf{x}_i, v_{it2}) = \mathbf{x}_{i1}\boldsymbol{\pi}_1 + \phi_{t1}v_{it2}$, and that $\mathbf{x}_{it1}$ is a strict subset of $\mathbf{x}_{it}$. Then, because y_{it2} is a function of $(\mathbf{x}_i, v_{it2})$, we can write $\mathrm{E}(y_{it1} \mid \mathbf{x}_i, v_{it2}) = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \alpha_1 y_{it2} + \mathbf{x}_{i1}\boldsymbol{\pi}_1 + \gamma_{t1}v_{it2}$. We obtain the Tobit residuals, $\hat{v}_{it2}$ for each t, and then run the regression y_{it1} on $\mathbf{x}_{it1}, y_{it2}, \mathbf{x}_{i1}$, and $\hat{v}_{it2}$ (possibly interacted with time dummies) for the selected sample. If we do not have an exclusion restriction, this regression suffers from perfect multicollinearity. As an example, we can easily include hours worked in a wage offer function for panel data, provided we have a variable affecting labor supply (such as the number of young children) but not the wage offer.

A pure fixed effects approach is more fruitful when the selection equation is of the Tobit form. The following assumption comes from Wooldridge (1995a):

ASSUMPTION 17.8: (a) The selection equation is equation (17.57). (b) For some unobserved effect g_{i1}, $\mathrm{E}(u_{it1} \mid \mathbf{x}_i, c_{i1}, g_{i1}, \mathbf{v}_{i2}) = \mathrm{E}(u_{it1} \mid g_{i1}, v_{it2}) = g_{i1} + \rho_1 v_{it2}$.

Under part b of this assumption,

$$\mathrm{E}(y_{it1} \mid \mathbf{x}_i, \mathbf{v}_{i2}, c_{i1}, g_{i1}) = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \rho_1 v_{it2} + f_{i1} \tag{17.58}$$

where $f_{i1} = c_{i1} + g_{i1}$. The same expectation holds when we also condition on $\mathbf{s}_{i2}$ (since $\mathbf{s}_{i2}$ is a function of $\mathbf{x}_i$, $\mathbf{v}_{i2}$). Therefore, estimating equation (17.58) by fixed effects on the unbalanced panel would consistently estimate $\boldsymbol{\beta}_1$ and ρ_1. As usual, we replace v_{it2} with the Tobit residuals $\hat{v}_{it2}$ whenever $y_{it2} > 0$. A t test of H_0: $\rho_1 = 0$ is valid very generally as a test of the null hypothesis of no sample selection. If the $\{u_{it1}\}$ satisfy the standard homoskedasticity and serial uncorrelatedness assumptions, then the usual t statistic is valid. A fully robust test may be warranted. (Again, with an exclusion restriction, we can add y_{it2} as an additional explanatory variable.)

Wooldridge (1995a) discusses an important case where Assumption 17.8b holds: in the Tobit version of equation (17.54) with $(\mathbf{u}_{i1}, \mathbf{a}_{i2})$ independent of $(\mathbf{x}_i, c_{i1}, c_{i2})$ and $\mathrm{E}(u_{it1} \mid \mathbf{a}_{i2}) = \mathrm{E}(u_{it1} \mid a_{it2}) = \rho_1 a_{it2}$. The second-to-last equality holds under the common assumption that $\{(u_{it1}, a_{it2}): t = 1, \ldots, T\}$ is serially independent.

The preceding methods assume normality of the errors in the selection equation and, implicitly, the unobserved heterogeneity. Kyriazidou (1997) and Honoré and Kyriazidou (2000b) have proposed methods that do not require distributional assumptions. Dustmann and Rochina-Barrachina (2000) apply Wooldridge's (1995a) and Kyriazidou's (1997) methods to the problem of estimating a wage offer equation with selection into the work force.

17.7.3 Attrition

We now turn specifically to testing and correcting for attrition in a linear, unobserved effects panel data model. General attrition, where units may reenter the sample after leaving, is complicated. We analyze a common special case. At $t = 1$ a random sample is obtained from the relevant population—people, for concreteness. In $t = 2$ and beyond, some people drop out of the sample for reasons that may not be entirely random. We assume that, once a person drops out, he or she is out forever: attrition is an *absorbing* state. Any panel data set with attrition can be set up in this way by ignoring any subsequent observations on units after they initially leave the sample. In Section 17.7.2 we discussed one way to test for attrition bias when we assume that attrition is an absorbing state: include $s_{i,t+1}$ as an additional explanatory variable in a fixed effects analysis.

One method for correcting for attrition bias is closely related to the corrections for incidental truncation covered in the previous subsection. Write the model for a random draw from the population as in equation (17.49), where we assume that $(\mathbf{x}_{it}, y_{it})$ is observed for all i when $t = 1$. Let s_{it} denote the selection indicator for each time period, where $s_{it} = 1$ if $(\mathbf{x}_{it}, y_{it})$ are observed. Because we ignore units once they initially leave the sample, $s_{it} = 1$ implies $s_{ir} = 1$ for $r < t$.

The sequential nature of attrition makes first differencing a natural choice to remove the unobserved effect:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \qquad t = 2, \ldots, T$$

Conditional on $s_{i,t-1} = 1$, write a (reduced-form) selection equation for $t \geq 2$ as

$$s_{it} = 1[\mathbf{w}_{it}\boldsymbol{\delta}_t + v_{it} > 0], \qquad v_{it} \mid \{\Delta \mathbf{x}_{it}, \mathbf{w}_{it}, s_{i,t-1} = 1\} \sim \text{Normal}(0, 1) \tag{17.59}$$

where $\mathbf{w}_{it}$ must contain variables observed at time t for all units with $s_{i,t-1} = 1$. Good candidates for $\mathbf{w}_{it}$ include the variables in $\mathbf{x}_{i,t-1}$ and any variables in $\mathbf{x}_{it}$ that are observed at time t when $s_{i,t-1} = 1$ (for example, if $\mathbf{x}_{it}$ contains lags of variables or a variable such as age). In general, the dimension of $\mathbf{w}_{it}$ can grow with t. For example, if equation (17.49) is dynamically complete, then $y_{i,t-2}$ is orthogonal to Δu_{it}, and so it can be an element of $\mathbf{w}_{it}$. Since $y_{i,t-1}$ is correlated with $u_{i,t-1}$, it should not be included in $\mathbf{w}_{it}$.

If the $\mathbf{x}_{it}$ are strictly exogenous *and* selection does not depend on $\Delta\mathbf{x}_{it}$ once $\mathbf{w}_{it}$ has been controlled for, a reasonable assumption (say, under joint normality of Δu_{it} and v_{it}) is

$$\mathrm{E}(\Delta u_{it} \mid \Delta\mathbf{x}_{it}, \mathbf{w}_{it}, v_{it}, s_{i,t-1} = 1) = \mathrm{E}(\Delta u_{it} \mid v_{it}, s_{i,t-1} = 1) = \rho_t v_{it} \tag{17.60}$$

Then

$$\mathrm{E}(\Delta y_{it} \mid \Delta\mathbf{x}_{it}, \mathbf{w}_{it}, s_{it} = 1) = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \rho_t\lambda(\mathbf{w}_{it}\boldsymbol{\delta}_t), \qquad t = 2, \ldots, T \tag{17.61}$$

Notice how, because $s_{i,t-1} = 1$ when $s_{it} = 1$, we do not have to condition on $s_{i,t-1}$ in equation (17.61). It now follows from equation (17.61) that pooled OLS of Δy_{it} on $\Delta\mathbf{x}_{it}, d2_t\hat{\lambda}_{it}, \ldots, dT_t\hat{\lambda}_{it}$, $t = 2, \ldots, T$, where the $\hat{\lambda}_{it}$ are from the $T - 1$ cross section probits in equation (17.59), is consistent for $\boldsymbol{\beta}_1$ and the ρ_t. A joint test of H_0: $\rho_t = 0$, $t = 2, \ldots, T$, is a fairly simple test for attrition bias, although nothing guarantees serial independence of the errors.

There are two potential problems with this approach. For one, the first equality in equation (17.60) is restrictive because it means that $\mathbf{x}_{it}$ does not affect attrition once the elements in $\mathbf{w}_{it}$ have been controlled for. Second, we have assumed strict exogeneity of $\mathbf{x}_{it}$. Both these restrictions can be relaxed by using an IV procedure.

Let $\mathbf{z}_{it}$ be a vector of variables such that $\mathbf{z}_{it}$ is redundant in the selection equation (possibly because $\mathbf{w}_{it}$ contains $\mathbf{z}_{it}$) and that $\mathbf{z}_{it}$ is exogenous in the sense that equation (17.58) holds with $\mathbf{z}_{it}$ in place of $\Delta\mathbf{x}_{it}$; for example, $\mathbf{z}_{it}$ should contain $\mathbf{x}_{ir}$ for $r < t$. Now, using an argument similar to the cross section case in Section 17.4.2, we can estimate the equation

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \rho_2\, d2_t\hat{\lambda}_{it} + \cdots + \rho_T\, dT_t\hat{\lambda}_{it} + error_{it} \tag{17.62}$$

by instrumental variables with instruments $(\mathbf{z}_{it}, d2_t\hat{\lambda}_{it}, \ldots, dT_t\hat{\lambda}_{it})$, using the selected sample. For example, the pooled 2SLS estimator on the selected sample is consistent and asymptotically normal, and attrition bias can be tested by a joint test of H_0: $\rho_t = 0$, $t = 2, \ldots, T$. Under H_0, only serial correlation and heteroskedasticity adjustments are possibly needed. If H_0 fails we have the usual generated regressors problem for estimating the asymptotic variance. Other IV procedures, such as GMM, can also be used, but they too must account for the generated regressors problem.

Example 17.9 (Dynamic Model with Attrition): Consider the model

$$y_{it} = \mathbf{g}_{it}\boldsymbol{\gamma} + \eta_1 y_{i,t-1} + c_i + u_{it}, \qquad t = 1, \ldots, T \tag{17.63}$$

where we assume that $(y_{i0}, \mathbf{g}_{i1}, y_{i1})$ are all observed for a random sample from the population. Assume that $\mathrm{E}(u_{it} \mid \mathbf{g}_i, y_{i,t-1}, \ldots, y_{i0}, c_i) = 0$, so that $\mathbf{g}_{it}$ is strictly exoge-

nous. Then the explanatory variables in the probit at time t, $\mathbf{w}_{it}$, can include $\mathbf{g}_{i,t-1}$, $y_{i,t-2}$, and further lags of these. After estimating the selection probit for each t, and differencing, we can estimate

$$\Delta y_{it} = \Delta \mathbf{g}_{it}\boldsymbol{\beta} + \eta_1 \Delta y_{i,t-1} + \rho_3 \, d3_t \hat{\lambda}_{it} + \cdots + \rho_T \, dT_t \hat{\lambda}_{it} + error_{it}$$

by pooled 2SLS on the selected sample starting at $t = 3$, using instruments $(\mathbf{g}_{i,t-1}, \mathbf{g}_{i,t-2}, y_{i,t-2}, y_{i,t-3})$. As usual, there are other possibilities for the instruments.

Although the focus in this section has been on pure attrition, where units disappear entirely from the sample, the methods can also be used in the context of incidental truncation without strictly exogenous explanatory variables. For example, suppose we are interested in the population of men who are employed at $t = 0$ and $t = 1$, and we would like to estimate a dynamic wage equation with an unobserved effect. Problems arise if men become unemployed in future periods. Such events can be treated as an attrition problem if all subsequent time periods are dropped once a man first becomes unemployed. This approach loses information but makes the econometrics relatively straightforward, especially because, in the preceding general model, $\mathbf{x}_{it}$ will always be observed at time t and so can be included in the labor force participation probit (assuming that men do not leave the sample entirely). Things become much more complicated if we are interested in the wage offer for all working age men at $t = 1$ because we have to deal with the sample selection problem into employment at $t = 0$ and $t = 1$.

The methods for attrition and selection just described apply only to linear models, and it is difficult to extend them to general nonlinear models. An alternative approach is based on **inverse probability weighting (IPW)**, which can be applied to general M-estimation, at least under certain assumptions.

Moffitt, Fitzgerald, and Gottschalk (1999) (MFG) propose inverse probability weighting to estimate linear panel data models under possibly nonrandom attrition. [MFG propose a different set of weights, analogous to those studied by Horowitz and Manski (1998), to solve missing data problems. The weights we use require estimation of only one attrition model, rather than two as in MFG.] IPW must be used with care to solve the attrition problem. As before, we assume that we have a random sample from the population at $t = 1$. We are interested in some feature, such as the conditional mean, or maybe the entire conditional distribution, of y_{it} given $\mathbf{x}_{it}$. Ideally, at each t we would observe $(y_{it}, \mathbf{x}_{it})$ for any unit that was in the random sample at $t = 1$. Instead, we observe $(y_{it}, \mathbf{x}_{it})$ only if $s_{it} = 1$. We can easily solve the attrition problem if we assume that, conditional on observables in the first time period, say, $\mathbf{z}_{i1}$, $(y_{it}, \mathbf{x}_{it})$ is independent of s_{it}:

$$P(s_{it} = 1 \mid y_{it}, \mathbf{x}_{it}, \mathbf{z}_{i1}) = P(s_{it} = 1 \mid \mathbf{z}_{i1}), \qquad t = 2, \ldots, T \tag{17.64}$$

Assumption (17.64) has been called **selection on observables** because we assume that $\mathbf{z}_{i1}$ is a strong enough predictor of selection in each time period so that the distribution of s_{it} given $[\mathbf{z}_{i1}, (y_{it}, \mathbf{x}_{it})]$ does not depend on $(y_{it}, \mathbf{x}_{it})$. In the statistics literature, selection on observables is also called **ignorability of selection** conditional on $\mathbf{z}_{i1}$. [The more standard approach, where selection is given by equation (17.59) and Δu_{it} is correlated with v_{it}, is sometimes called **selection on unobservables**. These categorizations are not strictly correct, as selection in both cases depends on observables and unobservables, but they serve as useful shorthand.]

Inverse probability weighting involves two steps. First, for each t, we estimate a probit or logit of s_{it} on $\mathbf{z}_{i1}$. (A crucial point is that the same cross section units—namely, all units appearing in the first time period—are used in the probit or logit for each time period.) Let $\hat{p}_{it}$ be the fitted probabilities, $t = 2, \ldots, T$, $i = 1, \ldots, N$. In the second step, the objective function for (i, t) is weighted by $1/\hat{p}_{it}$. For general M-estimation, the objective function is

$$\sum_{i=1}^{N} \sum_{t=1}^{T} (s_{it}/\hat{p}_{it}) q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) \tag{17.65}$$

where $\mathbf{w}_{it} \equiv (y_{it}, \mathbf{x}_{it})$ and $q_t(\mathbf{w}_{it}, \boldsymbol{\theta})$ is the objective function in each time period. As usual, the selection indicator s_{it} chooses the observations where we actually observe data. (For $t = 1$, $s_{it} = \hat{p}_{it} = 1$ for all i.) For least squares, $q_t(\mathbf{w}_{it}, \boldsymbol{\theta})$ is simply the squared residual function; for partial MLE, $q_t(\mathbf{w}_{it}, \boldsymbol{\theta})$ is the log-likelihood function.

The argument for why IPW works is rather simple. Let $\boldsymbol{\theta}_o$ denote the value of $\boldsymbol{\theta}$ that solves the population problem $\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^{T} \mathrm{E}[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})]$. Let $\boldsymbol{\delta}_t^o$ denote the true values of the selection response parameters in each time period, so that $P(s_{it} = 1 \mid \mathbf{z}_{i1}) = p_t(\mathbf{z}_{i1}, \boldsymbol{\delta}_t^o) \equiv p_{it}^o$. Now, under standard regularity conditions, we can replace p_{it}^o with $\hat{p}_{it} \equiv p_t(\mathbf{z}_{i1}, \hat{\boldsymbol{\delta}}_t)$ without affecting the consistency argument. So, apart from regularity conditions, it is sufficient to show that $\boldsymbol{\theta}_o$ minimizes $\sum_{t=1}^{T} \mathrm{E}[(s_{it}/p_{it}^o) q_t(\mathbf{w}_{it}, \boldsymbol{\theta})]$ over $\boldsymbol{\Theta}$. But, from iterated expectations,

$$\begin{aligned} \mathrm{E}[(s_{it}/p_{it}^o) q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] &= \mathrm{E}\{\mathrm{E}[(s_{it}/p_{it}^o) q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) \mid \mathbf{w}_{it}, \mathbf{z}_{i1}]\} \\ &= \mathrm{E}\{[\mathrm{E}(s_{it} \mid \mathbf{w}_{it}, \mathbf{z}_{i1})/p_{it}^o] q_t(\mathbf{w}_{it}, \boldsymbol{\theta})\} = \mathrm{E}[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] \end{aligned}$$

because $\mathrm{E}(s_{it} \mid \mathbf{w}_{it}, \mathbf{z}_{i1}) = P(s_{it} = 1 \mid \mathbf{z}_{i1})$ by assumption (17.64). Therefore, the probability limit of the weighted objective function is identical to that of the unweighted function *if* we had no attrition problem. Using this simple analogy argument, Wooldridge (2000d) shows that the inverse probability weighting produces a consistent,

$\sqrt{N}$-asymptotically normal estimator. The methods for adjusting the asymptotic variance matrix of two step M-estimators—described in Subsection 12.5.2—can be applied to the IPW M-estimator from (17.65). For reasons we will see, a sequential method of estimating attrition probabilities can be more attractive.

MFG propose an IPW scheme where the conditioning variables in the attrition probits change across time. In particular, at time t an attrition probit is estimated restricting attention to those units still in the sample at time $t-1$. (Out of this group, some are lost to attrition at time t, and some are not.) If we assume that attrition is an absorbing state, we can include in the conditioning variables, $\mathbf{z}_{it}$, all values of y and $\mathbf{x}$ dated at time $t-1$ and earlier (as well as other variables observed for all units in the sample at $t-1$). This approach is appealing because the ignorability assumption is much more plausible if we can condition on both recent responses and covariates. [That is, $\mathrm{P}(s_{it}=1 \mid \mathbf{w}_{it}, \mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1}, s_{i,t-1}=1) = \mathrm{P}(s_{it}=1 \mid \mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1}, s_{i,t-1}=1)$ is more likely than assumption (17.64).] Unfortunately, obtaining the fitted probabilities in this way and using them in an IPW procedure does not generally produce consistent estimators. The problem is that the selection models at each time period are not representative of the population that was originally sampled at $t=1$. Letting $p_{it}^{o} = \mathrm{P}(s_{it}=1 \mid \mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1}, s_{i,t-1}=1)$, we can no longer use the iterated expectations argument to conclude that $\mathrm{E}[(s_{it}/p_{it}^{o})q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] = \mathrm{E}[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})]$. Only if $\mathrm{E}[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] = \mathrm{E}[q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) \mid s_{i,t-1}=1]$ for all $\boldsymbol{\theta}$ does the argument work, but this assumption essentially requires that $\mathbf{w}_{it}$ be independent of $s_{i,t-1}$.

It is possible to allow the covariates in the selection probabilities to increase in richness over time, but the MFG procedure must be modified. For the case where attrition is an absorbing state, Wooldridge (2000d), building on work for regression models by Robins, Rotnitzky, and Zhao (1995) (RRZ), shows that the following probabilities can be used in the IPW procedure:

$$p_{it}(\boldsymbol{\delta}_t^{o}) \equiv \pi_{i2}(\boldsymbol{\gamma}_2^{o})\pi_{i3}(\boldsymbol{\gamma}_3^{o}) \cdots \pi_{it}(\boldsymbol{\gamma}_t^{o}), \qquad t=2, \ldots, T \tag{17.66}$$

where

$$\pi_{it}(\boldsymbol{\gamma}_t^{o}) \equiv \mathrm{P}(s_{it}=1 \mid \mathbf{z}_{it}, s_{i,t-1}=1) \tag{17.67}$$

In other words, as in the MFG procedure, we estimate probit models at each time t, restricted to units that are in the sample at $t-1$. The covariates in the probit are essentially everything we can observe for units in the sample at time $t-1$ that might affect attrition. For $t=2, \ldots, T$, let $\hat{\pi}_{it}$ denote the fitted selection probabilities. Then we construct the probability weights as the product $\hat{p}_{it} \equiv \hat{\pi}_{i2}\hat{\pi}_{i3} \cdots \hat{\pi}_{it}$ and use the objective function (17.65). Naturally, this method only works under certain assumptions. The key ignorability condition can be stated as

$$P(s_{it} = 1 \mid \mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}, s_{i,t-1} = 1) = P(s_{it} = 1 \mid \mathbf{z}_{it}, s_{i,t-1} = 1) \tag{17.68}$$

where $\mathbf{v}_{it} \equiv (\mathbf{w}_{it}, \mathbf{z}_{it})$. Now, we must include future values of $\mathbf{w}_{it}$ *and* $\mathbf{z}_{it}$ in the conditioning set on the left-hand side. Assumption (17.68) is fairly strong, but it does allow for attrition to be strongly related to past outcomes on y and $\mathbf{x}$ (which can be included in $\mathbf{z}_{it}$).

A convenient feature of the sequential method described above is that ignoring the first-stage estimation of the probabilities actually leads to *conservative* inference concerning $\boldsymbol{\theta}_0$: the (correct) asymptotic variance that adjusts for the first-stage estimation is actually smaller than the one that does not. See Wooldridge (2000d) for the general case and RRZ (1995) for the nonlinear regression case. See Wooldridge (2000d) for more on the pros and cons of using inverse probability weighting to reduce attrition bias.

17.8 Stratified Sampling

Nonrandom samples also come in the form of **stratified samples**, where different subsets of the population are sampled with different frequencies. For example, certain surveys are designed to learn primarily about a particular subset of the population, in which case that group is usually overrepresented in the sample. Stratification can be based on exogenous variables or endogenous variables (which are known once a model and assumptions have been specified), or some combination of these. As in the case of sample selection problems, it is important to know which is the case.

As mentioned in Section 17.3, choice-based sampling occurs when the stratification is based entirely on a discrete response variable. Various methods have been proposed for estimating discrete response models from choice-based samples under different assumptions; most of these are variations of maximum likelihood. Manski and McFadden (1981) and Cosslett (1993) contain general treatments, with the latter being a very useful survey. For a class of discrete response models, Cosslett (1981) proposed an efficient estimator, and Imbens (1992) obtained a computationally simple method of moments estimator that also achieves the efficiency bound. Imbens and Lancaster (1996) allow for general response variables in a maximum likelihood setting. Here, we focus on a simple, albeit often inefficient, method for estimating models in the context of two kinds of stratified sampling.

17.8.1 Standard Stratified Sampling and Variable Probability Sampling

The two most common kinds of stratification used in obtaining data sets in the social sciences are **standard stratified sampling (SS sampling)** and **variable probability sam-**

pling (VP sampling). In SS sampling, the population is first partitioned into J groups, $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_J$, which we assume are nonoverlapping and exhaustive. We let $\mathbf{w}$ denote the random variable representing the population of interest.

STANDARD STRATIFIED SAMPLING: For $j = 1, \ldots, J$, draw a random sample of size N_j from stratum j. For each j, denote this random sample by $\{\mathbf{w}_{ij}: i = 1, 2, \ldots, N_j\}$.

The strata sample sizes N_j are nonrandom. Therefore, the total sample size, $N = N_1 + \cdots + N_J$, is also nonrandom. A randomly drawn observation from stratum j, $\mathbf{w}_{ij}$, has distribution $D(\mathbf{w} \mid \mathbf{w} \in \mathcal{W}_j)$. Therefore, while observations within a stratum are identically distributed, observations across strata are not. A scheme that is similar in nature to SS sampling is called **multinomial sampling**, where a stratum is first picked at random and then an observation is randomly drawn from the stratum. This *does* result in i.i.d. observations, but it does not correspond to how stratified samples are obtained in practice. It also leads to the same estimators as under SS sampling, so we do not discuss it further; see Cosslett (1993) or Wooldridge (1999b) for further discussion.

Variable probability samples are obtained using a different scheme. First, an observation is drawn at random from the population. If the observation falls into stratum j, it is kept with probability p_j. Thus, random draws from the population are discarded with varying frequencies depending on which stratum they fall into. This kind of sampling is appropriate when information on the variable or variables that determine the strata is relatively easy to obtain compared with the rest of the information. Survey data sets, including initial interviews to collect panel or longitudinal data, are good examples. Suppose we want to oversample individuals from, say, lower income classes. We can first ask an individual her or his income. If the response is in income class j, this person is kept in the sample with probability p_j, and then the remaining information, such as education, work history, family background, and so on can be collected; otherwise, the person is dropped without further interviewing.

A key feature of VP sampling is that observations within a stratum are discarded randomly. As discussed by Wooldridge (1999b), VP sampling is equivalent to the following:

VARIABLE PROBABILITY SAMPLING: Repeat the following steps N times:

1. Draw an observation $\mathbf{w}_i$ at random from the population.
2. If $\mathbf{w}_i$ is in stratum j, toss a (biased) coin with probability p_j of turning up heads. Let $h_{ij} = 1$ if the coin turns up heads and zero otherwise.
3. Keep observation i if $h_{ij} = 1$; otherwise, omit it from the sample.

The number of observations falling into stratum j is denoted N_j, and the number of data points we actually have for estimation is $N_0 = N_1 + N_2 + \cdots + N_J$. Notice that if N—the number of times the population is sampled—is fixed, then N_0 is a random variable: we do not know what each N_j will be prior to sampling. Also, we will not use information on the number of discarded observations in each stratum, so that N is not required to be known.

The assumption that the probability of the coin turning up heads in step 2 depends only on the stratum ensures that sampling is random within each stratum. This roughly reflects how samples are obtained for certain large cross-sectional and panel data sets used in economics, including the panel study of income dynamics and the national longitudinal survey.

To see that a VP sample can be analyzed as a random sample, we construct a population that incorporates the stratification. The VP sampling scheme is equivalent to first tossing all J coins before actually observing which stratum $\mathbf{w}_i$ falls into; this gives $(h_{i1}, \ldots, h_{iJ})$. Next, $\mathbf{w}_i$ is observed to fall into one of the strata. Finally, the outcome is kept or not depending on the coin flip for that stratum. The result is that the vector $(\mathbf{w}_i, \mathbf{h}_i)$, where $\mathbf{h}_i$ is the J-vector of binary indicators h_{ij}, is a random sample from a new population with sample space $\mathcal{W} \times \mathcal{H}$, where $\mathcal{W}$ is the original sample space and $\mathcal{H}$ denotes the sample space associated with outcomes from flipping J coins. Under this alternative way of viewing the sampling scheme, $\mathbf{h}_i$ is independent of $\mathbf{w}_i$. Treating $(\mathbf{w}_i, \mathbf{h}_i)$ as a random draw from the new population is not at odds with the fact that our estimators are based on a nonrandom sample from the original population: we simply use the vector $\mathbf{h}_i$ to determine which observations are kept in the estimation procedure.

17.8.2 Weighted Estimators to Account for Stratification

With variable probability sampling, it is easy to construct weighted objective functions that produce consistent and asymptotically normal estimators of the population parameters. It is useful to define a set of binary variables that indicate whether a random draw $\mathbf{w}_i$ is kept in the sample and, if so, which stratum it falls into:

$$r_{ij} = h_{ij}s_{ij} \tag{17.69}$$

By definition, $r_{ij} = 1$ for at most one j. If $h_{ij} = 1$ then $r_{ij} = s_{ij}$. If $r_{ij} = 0$ for all $j = 1, 2, \ldots, J$, then the random draw $\mathbf{w}_i$ does not appear in the sample (and we do not know which stratum it belonged to).

With these definitions, we can define the **weighted M-estimator**, $\hat{\theta}_w$, as the solution to

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{j=1}^{J} p_j^{-1} r_{ij} q(\mathbf{w}_i, \theta) \tag{17.70}$$

where $q(\mathbf{w}, \boldsymbol{\theta})$ is the objective function that is chosen to identify the population parameters $\boldsymbol{\theta}_o$. Note how the outer summation is over all *potential* observations, that is, the observations that *would* appear in a random sample. The indicators r_{ij} simply pick out the observations that actually appear in the available sample, and these indicators also attach each observed data point to its stratum. The objective function (17.70) weights each observed data point in the sample by the inverse of the sampling probability. For implementation it is useful to write the objective function as

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N_0} p_{j_i}^{-1} q(\mathbf{w}_i, \boldsymbol{\theta}) \tag{17.71}$$

where, without loss of generality, the data points actually observed are ordered $i = 1, \ldots, N_0$. Since j_i is the stratum for observation i, $p_{j_i}^{-1}$ is the weight attached to observation i in the estimation. In practice, the $p_{j_i}^{-1}$ are the **sampling weights** reported with other variables in stratified samples.

The objective function $q(\mathbf{w}, \boldsymbol{\theta})$ contains all of the M-estimator examples we have covered so far in the book, including least squares (linear and nonlinear), conditional maximum likelihood, and partial maximum likelihood. In panel data applications, the probability weights are from sampling in an initial year. Weights for later years are intended to reflect both stratification (if any) and possible attrition, as discussed in Section 17.7.3 and in Wooldridge (2000d).

Wooldridge (1999b) shows that, under the same assumptions as Theorem 12.2 and the assumption that each sampling probability is strictly positive, the weighted M-estimator consistently estimates $\boldsymbol{\theta}_o$, which is assumed to uniquely minimize $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$. To see that the weighted objective function identifies $\boldsymbol{\theta}_o$, we use the fact that h_j is independent of $\mathbf{w}$ [and therefore of $(\mathbf{w}, s_j)$ for each j], and so

$$\begin{aligned} \mathrm{E}\left[\sum_{j=1}^{J} p_j^{-1} h_j s_j q(\mathbf{w}, \boldsymbol{\theta})\right] &= \sum_{j=1}^{J} p_j^{-1} \mathrm{E}(h_j) \mathrm{E}[s_j q(\mathbf{w}, \boldsymbol{\theta})] \\ &= \sum_{j=1}^{J} p_j^{-1} p_j \mathrm{E}[s_j q(\mathbf{w}, \boldsymbol{\theta})] = \mathrm{E}\left[\left(\sum_{j=1}^{J} s_j\right) q(\mathbf{w}, \boldsymbol{\theta})\right] = \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})] \end{aligned} \tag{17.72}$$

where the final equality follows because the s_j sum to unity. Therefore, the expected value of the weighted objective function [over the distribution of $(\mathbf{w}, \mathbf{h})$] equals the expected value of $q(\mathbf{w}, \boldsymbol{\theta})$ (over the distribution of $\mathbf{w}$). Consistency of the weighted M-estimator follows under the regularity conditions in Theorem 12.2.

Asymptotic normality also follows under the same regularity conditions as in Chapter 12. Wooldridge (1999b) shows that a valid estimator of the asymptotic

variance of $\hat{\theta}_w$ is

$$\left[\sum_{i=1}^{N_0} p_{j_i}^{-1}\nabla_\theta^2 q_i(\hat{\theta}_w)\right]^{-1}\left[\sum_{i=1}^{N_0} p_{j_i}^{-2}\nabla_\theta q_i(\hat{\theta}_w)'\nabla_\theta q_i(\hat{\theta}_w)\right]\left[\sum_{i=1}^{N_0} p_{j_i}^{-1}\nabla_\theta^2 q_i(\hat{\theta}_w)\right]^{-1} \tag{17.73}$$

which looks like the standard formula for a robust variance matrix estimator except for the presence of the sampling probabilities p_{j_i}.

When $\mathbf{w}$ partitions as $(\mathbf{x}, \mathbf{y})$, an alternative estimator replaces the Hessian $\nabla_\theta^2 q_i(\hat{\theta}_w)$ in expression (17.73) with $\mathbf{A}(\mathbf{x}_i, \hat{\theta}_w)$, where $\mathbf{A}(\mathbf{x}_i, \theta_o) \equiv \mathrm{E}[\nabla_\theta^2 q(\mathbf{w}_i, \theta_o) \mid \mathbf{x}_i]$, as in Chapter 12. Asymptotic standard errors and Wald statistics can be obtained using either estimate of the asymptotic variance.

Example 17.10 (Linear Model under Stratified Sampling): In estimating the linear model

$$y = \mathbf{x}\beta_o + u, \qquad \mathrm{E}(\mathbf{x}'u) = \mathbf{0} \tag{17.74}$$

by weighted least squares, the asymptotic variance matrix estimator is

$$\left(\sum_{i=1}^{N_0} p_{j_i}^{-1}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N_0} p_{j_i}^{-2}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\sum_{i=1}^{N_0} p_{j_i}^{-1}\mathbf{x}_i'\mathbf{x}_i\right)^{-1} \tag{17.75}$$

where $\hat{u}_i = y_i - \mathbf{x}_i\hat{\beta}_w$ is the residual after WLS estimation. Interestingly, this is simply the White (1980b) heteroskedasticity-consistent covariance matrix estimator applied to the stratified sample, where all variables for observation i are weighted by $p_{j_i}^{-1/2}$ before performing the regression. This estimator has been suggested by, among others, Hausman and Wise (1981). Hausman and Wise use maximum likelihood to obtain more efficient estimators in the context of the normal linear regression model, that is, $u \mid \mathbf{x} \sim \mathrm{Normal}(\mathbf{x}\beta_o, \sigma_o^2)$. Because of stratification, MLE is not generally robust to failure of the homoskedastic normality assumption.

It is important to remember that the form of expression (17.75) in this example is not due to potential heteroskedasticity in the underlying population model. Even if $\mathrm{E}(u^2 \mid \mathbf{x}) = \sigma_o^2$, the estimator (17.75) is generally needed because of the stratified sampling. This estimator also works in the presence of heteroskedasticity of arbitrary and unknown form in the population, and it is routinely computed by many regression packages.

Example 17.11 (Conditional MLE under Stratified Sampling): When $f(\mathbf{y} \mid \mathbf{x}; \theta)$ is a correctly specified model for the density of $\mathbf{y}_i$ given $\mathbf{x}_i$ in the population, the inverse-probability-weighted MLE is obtained with $q_i(\theta) \equiv -\log[f(\mathbf{y}_i \mid \mathbf{x}_i; \theta)]$. This estimator

is consistent and asymptotically normal, with asymptotic variance estimator given by expression (17.73) [or, preferably, the form that uses $\mathbf{A}(\mathbf{x}_i, \hat{\theta}_w)$].

A weighting scheme is also available in the standard stratified sampling case, but the weights are different from the VP sampling case. To derive them, let $Q_j = \mathrm{P}(\mathbf{w} \in \mathcal{W}_j)$ denote the population frequency for stratum j; we assume that the Q_j are *known*. By the law of iterated expectations,

$$\mathrm{E}[q(\mathbf{w}, \theta)] = Q_1 \mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{w} \in \mathcal{W}_1] + \cdots + Q_J \mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{w} \in \mathcal{W}_J] \tag{17.76}$$

for any θ. For each j, $\mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{w} \in \mathcal{W}_j]$ can be consistently estimated using a random sample obtained from stratum j. This scheme leads to the sample objective function

$$Q_1 \left[N_1^{-1} \sum_{i=1}^{N_1} q(\mathbf{w}_{i1}, \theta) \right] + \cdots + Q_J \left[N_J^{-1} \sum_{i=1}^{N_J} q(\mathbf{w}_{iJ}, \theta) \right]$$

where $\mathbf{w}_{ij}$ denotes a random draw i from stratum j and N_j is the nonrandom sample size for stratum j. We can apply the uniform law of large numbers to each term, so that the sum converges uniformly to equation (17.76) under the regularity conditions in Chapter 12. By multiplying and dividing each term by the total number of observations $N = N_1 + \cdots + N_J$, we can write the sample objective function more simply as

$$N^{-1} \sum_{i=1}^{N} (Q_{j_i}/H_{j_i}) q(\mathbf{w}_i, \theta) \tag{17.77}$$

where j_i denotes the stratum for observation i and $H_j \equiv N_j/N$ denotes the fraction of observations in stratum j. Because we have the stratum indicator j_i, we can drop the j subscript on $\mathbf{w}_i$. When we omit the division by N, equation (17.77) has the same form as equation (17.71), but the weights are (Q_{j_i}/H_{j_i}) rather than $p_{j_i}^{-1}$ (and the arguments for why each weighting works are very different). Also, in general, the formula for the asymptotic variance is different in the SS sampling case. In addition to the minor notational change of replacing N_0 with N, the middle matrix in equation (17.73) becomes

$$\sum_{j=1}^{J} (Q_j^2/H_j^2) \left[\sum_{i=1}^{N_j} (\nabla_\theta \hat{q}_{ij} - \overline{\nabla_\theta q_j})' (\nabla_\theta \hat{q}_{ij} - \overline{\nabla_\theta q_j}) \right]$$

where $\nabla_\theta \hat{q}_{ij} \equiv \nabla_\theta q(\mathbf{w}_{ij}, \hat{\theta}_w)$ and $\overline{\nabla_\theta q_j} \equiv N_j^{-1} \sum_{i=1}^{N_j} \nabla_\theta \hat{q}_{ij}$ (the within-stratum sample average). This approach requires us to explicitly partition observations into their respective strata. See Wooldridge (2001) for a detailed derivation. [If in the VP

sampling case the population frequencies Q_j are known, it is better to use as weights $Q_j/(N_j/N_0)$ rather than p_j^{-1}, which makes the analysis look just like the SS sampling case. See Wooldridge (1999b) for details.]

If in Example 17.11 we have standard stratified sampling rather than VP sampling, the weighted MLE is typically called the **weighted exogenous sample MLE (WESMLE)**; this estimator was suggested by Manski and Lerman (1977) in the context of choice-based sampling in discrete response models. [Actually, Manski and Lerman (1977) use multinomial sampling where H_j is the probability of picking stratum j. But Cosslett (1981) showed that a more efficient estimator is obtained by using N_j/N, as one always does in the case of SS sampling; see Wooldridge (1999b) for an extension of Cosslett's result to the M-estimator case.]

Provided that the sampling weights Q_{j_i}/H_{j_i} or $p_{j_i}^{-1}$ are given (along with the stratum), analysis with the weighted M-estimator under SS or VP sampling is fairly straightforward, but it is not likely to be efficient. In the conditional maximum likelihood case it is certainly possible to do better. See Imbens and Lancaster (1996) for a careful treatment.

17.8.3 Stratification Based on Exogenous Variables

When $\mathbf{w}$ partitions as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ is exogenous in a sense to be made precise, and stratification is based entirely on $\mathbf{x}$, the standard unweighted estimator on the stratified sample is consistent and asymptotically normal. The sense in which $\mathbf{x}$ must be exogenous is that $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x}] \tag{17.78}$$

for each possible outcome $\mathbf{x}$. This assumption holds in a variety of contexts with conditioning variables and correctly specified models. For example, as we discussed in Chapter 12, this holds for nonlinear regression when the conditional mean is correctly specified and $\boldsymbol{\theta}_o$ is the vector of conditional mean parameters; in Chapter 13 we showed that this holds for conditional maximum likelihood when the density of $\mathbf{y}$ given $\mathbf{x}$ is correct. It also holds in other cases, including quasi-maximum likelihood, which we cover in Chapter 19. One interesting observation is that, in the linear regression model (17.74), the exogeneity of $\mathbf{x}$ must be strengthened to $\mathrm{E}(u \mid \mathbf{x}) = 0$.

In the case of VP sampling, selection on the basis of $\mathbf{x}$ means that each selection indicator s_j is a deterministic function of $\mathbf{x}$. The unweighted M-estimator on the stratified sample, $\hat{\boldsymbol{\theta}}_u$, minimizes

$$\sum_{i=1}^{N} \sum_{j=1}^{J} h_{ij} s_{ij} q(\mathbf{w}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N_0} q(\mathbf{w}_i, \boldsymbol{\theta})$$

Consistency follows from standard M-estimation results if we can show that θ_o uniquely solves

$$\min_{\theta \in \Theta} \sum_{j=1}^{J} \mathrm{E}[h_j s_j q(\mathbf{w}, \theta)] \tag{17.79}$$

Since s_j is a function of $\mathbf{x}$ and h_j is independent of $\mathbf{w}$ (and therefore $\mathbf{x}$), $\mathrm{E}[h_j s_j q(\mathbf{w}, \theta) \mid \mathbf{x}] = \mathrm{E}(h_j \mid \mathbf{x}) s_j \mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{x}] = p_j s_j \mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{x}]$ for each j. By assumption, $\mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{x}]$ is minimized at θ_o for all $\mathbf{x}$, and therefore so is $p_j s_j \mathrm{E}[q(\mathbf{w}, \theta) \mid \mathbf{x}]$ (but probably not uniquely). By iterated expectations it follows that θ_o is a solution to equation (17.79). Unlike in the case of the weighted estimator, it no longer suffices to assume that θ_o uniquely minimizes $\mathrm{E}[q(\mathbf{w}, \theta)]$; we must directly assume θ_o is the unique solution to problem (17.79). This assumption could fail if, for example, $p_j = 0$ for some j—so that we do not observe part of the population at all. (Unlike in the case of the weighted estimator, $p_j = 0$ for at least some j is allowed for the unweighted estimator, subject to identification holding.) For example, in the context of linear wage regression, we could not identify the return to education if we only sample those with exactly a high school education.

Wooldridge (1999b) shows that the usual asymptotic variance estimators (see Section 12.5) are valid when stratification is based on $\mathbf{x}$ and we ignore the stratification problem. For example, the usual conditional maximum likelihood analysis holds. In the case of regression, we can use the usual heteroskedasticity-robust variance matrix estimator. Or, if we assume homoskedasticity in the population, the nonrobust form [see equation (12.58)] is valid with the usual estimator of the error variance.

When a generalized conditional information matrix equality holds, and stratification is based on $\mathbf{x}$, Wooldridge (1999b) shows that the unweighted estimator is more efficient than the weighted estimator. The key assumption is

$$\mathrm{E}[\nabla_\theta q(\mathbf{w}, \theta_o)' \nabla_\theta q(\mathbf{w}, \theta_o) \mid \mathbf{x}] = \sigma_o^2 \mathrm{E}[\nabla_\theta^2 q(\mathbf{w}, \theta_o) \mid \mathbf{x}] \tag{17.80}$$

for some $\sigma_o^2 > 0$. When assumption (17.80) holds and θ_o solves equation (17.79), the asymptotic variance of the unweighted M-estimator is smaller than that for the weighted M-estimator. This generalization includes conditional maximum likelihood (with $\sigma_o^2 = 1$) and nonlinear regression under homoskedasticity.

Very similar conclusions hold for standard stratified sampling. One useful fact is that, when stratification is based on $\mathbf{x}$, the estimator (17.73) is valid with $p_j = H_j / Q_j$ (and $N_0 = N$); therefore, we need not compute within-strata variation in the estimated score. The unweighted estimator is consistent when stratification is based on $\mathbf{x}$ and the usual asymptotic variance matrix estimators are valid. The unweighted

estimator is also more efficient when assumption (17.80) holds. See Wooldridge (2001) for statements of assumptions and proofs of theorems.

Problems

17.1. a. Suppose you are hired to explain fire damage to buildings in terms of building and neighborhood characteristics. If you use cross section data on reported fires, is there a sample selection problem due to the fact that most buildings do not catch fire during the year?

b. If you want to estimate the relationship between contributions to a 401(k) plan and the match rate of the plan—the rate at which the employer matches employee contributions—is there a sample selection problem if you only use a sample of workers already enrolled in a 401(k) plan?

17.2. In Example 17.4, suppose that *IQ* is an indicator of *abil*, and *KWW* is another indicator (see Section 5.3.2). Find assumptions under which IV on the selected sample is valid.

17.3. Let $f(\cdot \mid \mathbf{x}_i; \boldsymbol{\theta})$ denote the density of y_i given $\mathbf{x}_i$ for a random draw from the population. Find the conditional density of y_i given $(\mathbf{x}_i, s_i = 1)$ when the selection rule is $s_i = 1[a_1(\mathbf{x}_i) < y_i < a_2(\mathbf{x}_i)]$, where $a_1(\mathbf{x})$ and $a_2(\mathbf{x})$ are known functions of $\mathbf{x}$. In the Hausman and Wise (1977) example, $a_2(\mathbf{x})$ was a function of family size because the poverty income level depends on family size.

17.4. Suppose in Section 17.4.1 we replace Assumption 17.1d with

$$\mathrm{E}(u_1 \mid v_2) = \gamma_1 v_2 + \gamma_2(v_2^2 - 1)$$

(We subtract unity from v_2^2 to ensure that the second term has zero expectation.)

a. Using the fact that $\mathrm{Var}(v_2 \mid v_2 > -a) = 1 - \lambda(a)[\lambda(a) + a]$, show that

$$\mathrm{E}(y_1 \mid \mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}\boldsymbol{\delta}_2) - \gamma_2\lambda(\mathbf{x}\boldsymbol{\delta}_2)\mathbf{x}\boldsymbol{\delta}_2$$

[Hint: Take $a = \mathbf{x}\boldsymbol{\delta}_2$ and use the fact that $\mathrm{E}(v_2^2 \mid v_2 > -a) = \mathrm{Var}(v_2 \mid v_2 > -a) + [\mathrm{E}(v_2 \mid v_2 > -a)]^2$.]

b. Explain how to correct for sample selection in this case.

c. How would you test for the presence of sample selection bias?

17.5. Consider the following alternative to Procedure 17.2. First, run the OLS regression of y_2 on $\mathbf{z}$ and obtain the fitted values, $\hat{y}_2$. Next, get the inverse Mills ratio,

$\hat{\lambda}_3$, from the probit of y_3 on $\mathbf{z}$. Finally, run the OLS regression y_1 on $\mathbf{z}_1, \hat{y}_2, \hat{\lambda}_3$ using the selected sample.

a. Find a set of sufficient conditions that imply consistency of the proposed procedure. (Do not worry about regularity conditions.)

b. Show that the assumptions from part a are more restrictive than those in Procedure 17.2, and give some examples that are covered by Procedure 17.2 but not by the alternative procedure.

17.6. Apply Procedure 17.4 to the data in MROZ.RAW. Use a constant, *exper*, and $exper^2$ as elements of $\mathbf{z}_1$; take $y_2 = educ$. The other elements of $\mathbf{z}$ should include *age*, *kidslt6*, *kidsge6*, *nwifeinc*, *motheduc*, *fatheduc*, and *huseduc*.

17.7. Consider the model

$$y_1 = \mathbf{z}\boldsymbol{\delta}_1 + v_1$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$$

$$y_3 = \max(0, \alpha_{31} y_1 + \alpha_{32} y_2 + \mathbf{z}_3\boldsymbol{\delta}_3 + u_3)$$

where $(\mathbf{z}, y_2, y_3)$ are always observed and y_1 is observed when $y_3 > 0$. The first two equations are reduced-form equations, and the third equation is of primary interest. For example, take $y_1 = \log(wage^o)$, $y_2 = educ$, and $y_3 = hours$, and then education and $\log(wage^o)$ are possibly endogenous in the labor supply function. Assume that (v_1, v_2, u_3) are jointly zero-mean normal and independent of $\mathbf{z}$.

a. Find a simple way to consistently estimate the parameters in the third equation allowing for arbitrary correlations among (v_1, v_2, u_3). Be sure to state any identification assumptions needed.

b. Now suppose that y_2 is observed only when $y_3 > 0$; for example, $y_1 = \log(wage^o)$, $y_2 = \log(benefits^o)$, $y_3 = hours$. Now derive a multistep procedure for estimating the third equation under the same assumptions as in part a.

c. How can we estimate the average partial effects?

17.8. Consider the following conditional moment restrictions problem with a selected sample. In the population, $\mathrm{E}[\mathbf{r}(\mathbf{w}, \boldsymbol{\theta}_o) \mid \mathbf{x}] = \mathbf{0}$. Let s be the selection indicator, and assume that

$$\mathrm{E}[\mathbf{r}(\mathbf{w}, \boldsymbol{\theta}_o) \mid \mathbf{x}, s] = \mathbf{0}$$

Sufficient is that $s = f(\mathbf{x})$ for a nonrandom function f.

a. Let $\mathbf{Z}_i$ be a $G \times L$ matrix of functions of $\mathbf{x}_i$. Show that $\boldsymbol{\theta}_o$ satisfies

$$\mathrm{E}[s_i \mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = \mathbf{0}$$

b. Write down the objective function for the system nonlinear 2SLS estimator based on the selected sample. Argue that, under the appropriate rank condition, the estimator is consistent and $\sqrt{N}$-asymptotically normal.

c. Write down the objective function for a minimum chi-square estimator using the selected sample. Use the estimates from part b to estimate the weighting matrix. Argue that the estimator is consistent and $\sqrt{N}$-asymptotically normal.

17.9. Consider the problem of standard stratified sampling. Argue that when $\boldsymbol{\theta}_o$ solves equation (17.78) for each $\mathbf{x}$, $\boldsymbol{\theta}_o$ is identified in the population, stratification is based on $\mathbf{x}$, and $\bar{H}_j > 0$ for $j = 1, \ldots, J$, the unweighted estimator is consistent. {Hint: Write the objective function for the unweighted estimator as

$$\sum_{j=1}^{J} H_j \left[N_j^{-1} \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta}) \right] \tag{17.81}$$

and assume that $H_j \to \bar{H}_j > 0$ as $N \to \infty$. If the strata are $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_J$, argue that equation (17.81) converges uniformly to

$$\bar{H}_1 \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathscr{X}_1] + \cdots + \bar{H}_J \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathscr{X}_J] \tag{17.82}$$

Why does $\boldsymbol{\theta}_o$ necessarily minimize expression (17.82)? Identification follows when you show that $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}) \mid \mathbf{x} \in \mathscr{X}_j]$ is uniquely minimized at $\boldsymbol{\theta}_o$ for at least one j.}

17.10. Consider model (17.25), where selection is ignorable in the sense that $\mathrm{E}(u_1 \mid \mathbf{z}, u_3) = 0$. However, data are missing on y_2 when $y_3 = 0$, and $\mathrm{E}(y_2 \mid \mathbf{z}, y_3) \neq \mathrm{E}(y_2 \mid \mathbf{z})$.

a. Find $\mathrm{E}(y_1 \mid \mathbf{z}, y_3)$.

b. If, in addition to Assumption 17.2, (v_2, v_3) is independent of $\mathbf{z}$ and $\mathrm{E}(v_2 \mid v_3) = \gamma_2 v_3$, find $\mathrm{E}(y_1 \mid \mathbf{z}, y_3 = 1)$.

c. Suggest a two-step method for consistently estimating $\boldsymbol{\delta}_1$ and α_1.

d. Does this method generally work if $\mathrm{E}(u_1 \mid \mathbf{z}, y_3) \neq 0$?

e. Would you bother with the method from part c if $\mathrm{E}(u_1 \mid \mathbf{z}, y_2, y_3) = 0$? Explain.

17.11. In Section 16.7 we discussed two-part models for a corner solution outcome, say, y. These models have sometimes been studied in the context of incidental truncation.

a. Suppose you have a parametric model for the distribution of y conditional on $\mathbf{x}$ and $y > 0$. (Cragg's model and the lognormal model from Section 16.7 are examples.) If you estimate the parameters of this model by conditional MLE, using only the observations for which $y_i > 0$, do the parameter estimates suffer from sample selection bias? Explain.

b. If instead you specify only $\mathrm{E}(y \mid \mathbf{x}, y > 0) = \exp(\mathbf{x}\boldsymbol{\beta})$ and estimate $\boldsymbol{\beta}$ by nonlinear least squares using observations for which $y_i > 0$, do the estimates suffer from sample selection bias?

c. In addition to the specification from part b, suppose that $\mathrm{P}(y = 0 \mid \mathbf{x}) = 1 - \Phi(\mathbf{x}\gamma)$. How would you estimate γ?

d. Given the assumptions in parts b and c, how would you estimate $\mathrm{E}(y \mid \mathbf{x})$?

e. Given your answers to the first four parts, do you think viewing estimation of two-part models as an incidental truncation problem is appropriate?

17.12. Consider Theorem 17.1. Suppose that we relax assumption (17.6) to $\mathrm{E}(u \mid \mathbf{z}, s) = \mathrm{E}(u \mid s) = (1 - s)\alpha_0 + s\alpha_1$. The first equality is the assumption; the second is unrestrictive, as it simply allows the mean of u to differ in the selected and unselected subpopulations.

a. Show that 2SLS estimation using the selected subsample consistently estimates the slope parameters, $\beta_2, \ldots, \beta_K$. What is the plim of the intercept estimator? [Hint: Replace u with $(1 - s)\alpha_0 + s\alpha_1 + e$, where $\mathrm{E}(e \mid \mathbf{z}, s) = 0$.]

b. Show that $\mathrm{E}(u \mid \mathbf{z}, s) = \mathrm{E}(u \mid s)$ if (u, s) is independent of $\mathbf{z}$. Does independence of s and $\mathbf{z}$ seem reasonable?

17.13. Suppose that y given $\mathbf{x}$ follows a standard censored Tobit, where y is a corner solution response. However, there is at least one element of $\mathbf{x}$ that we can observe only when $y > 0$. (An example is seen when y is quantity demanded of a good or service, and one element of $\mathbf{x}$ is price, derived as total expenditure on the good divided by y whenever $y > 0$.)

a. Explain why we cannot use standard censored Tobit maximum likelihood estimation to estimate $\boldsymbol{\beta}$ and σ^2. What method can we use instead?

b. How is it that we can still estimate $\mathrm{E}(y \mid \mathbf{x})$, even though we do not observe some elements of $\mathbf{x}$ when $y = 0$?

18 Estimating Average Treatment Effects

18.1 Introduction

In this chapter we explicitly study the problem of estimating an **average treatment effect (ATE)**. An average treatment effect is a special case of an average partial effect: an ATE is an average partial effect for a binary explanatory variable.

Estimating ATEs has become important in the program evaluation literature, such as in the evaluation of job training programs. Originally, the binary indicators represented medical treatment or program participation, but the methods are applicable when the explanatory variable of interest is any binary variable.

We begin by introducing a counterfactual framework pioneered by Rubin (1974) and since adopted by many in both statistics and econometrics, including Rosenbaum and Rubin (1983), Heckman (1992, 1997), Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996), Manski (1996), Heckman, Ichimura, and Todd (1997), and Angrist (1998). The counterfactual framework allows us to define various treatment effects that may be of interest. Once we define the different treatment effects, we can study ways to consistently estimate these effects. We will not provide a comprehensive treatment of this rapidly growing literature, but we will show that, under certain assumptions, estimators that we are already familiar with consistently estimate average treatment effects. We will also study some extensions that consistently estimate ATEs under weaker assumptions.

Broadly, most estimators of ATEs fit into one of two categories. The first set exploits assumptions concerning *ignorability* of the treatment conditional on a set of covariates. As we will see in Section 18.3, this approach is analogous to the proxy variable solution to the omitted variables problem that we discussed in Chapter 4, and in some cases reduces exactly to an OLS regression with many controls. A second set of estimators relies on the availability of one or more instrumental variables that are redundant in the response equations but help determine participation. Different IV estimators are available depending on functional form assumptions concerning how unobserved heterogeneity affects the responses. We study IV estimators in Section 18.4.

In Section 18.5 we briefly discuss some further topics, including special considerations for binary and corner solution responses, using panel data to estimate treatment effects, and nonbinary treatments.

18.2 A Counterfactual Setting and the Self-Selection Problem

The modern literature on treatment effects begins with a counterfactual, where each individual (or other agent) has an outcome with and without treatment (where

"treatment" is interpreted very broadly). This section draws heavily on Heckman (1992, 1997), Imbens and Angrist (1994), and Angrist, Imbens, and Rubin (1996) (hereafter AIR). Let y_1 denote the outcome with treatment and y_0 the outcome without treatment. Because an individual cannot be in both states, we cannot observe both y_0 and y_1; in effect, the problem we face is one of missing data.

It is important to see that we have made no assumptions about the distributions of y_0 and y_1. In many cases these may be roughly continuously distributed (such as salary), but often y_0 and y_1 are binary outcomes (such as a welfare participation indicator), or even corner solution outcomes (such as married women's labor supply). However, some of the assumptions we make will be less plausible for discontinuous random variables, something we discuss after introducing the assumptions.

The following discussion assumes that we have an independent, identically distributed sample from the population. This assumption rules out cases where the treatment of one unit affects another's outcome (possibly through general equilibrium effects, as in Heckman, Lochner, and Taber, 1998). The assumption that treatment of unit i affects only the outcome of unit i is called the **stable unit treatment value assumption (SUTVA)** in the treatment literature (see, for example, AIR). We are making a stronger assumption because random sampling implies SUTVA.

Let the variable w be a binary treatment indicator, where $w = 1$ denotes treatment and $w = 0$ otherwise. The triple (y_0, y_1, w) represents a random vector from the underlying population of interest. For a random draw i from the population, we write (y_{i0}, y_{i1}, w_i). However, as we have throughout, we state assumptions in terms of the population.

To measure the effect of treatment, we are interested in the difference in the outcomes with and without treatment, $y_1 - y_0$. Because this is a random variable (that is, it is individual specific), we must be clear about what feature of its distribution we want to estimate. Several possibilities have been suggested in the literature. In Rosenbaum and Rubin (1983), the quantity of interest is the **average treatment effect (ATE)**,

$$ATE \equiv \mathrm{E}(y_1 - y_0) \tag{18.1}$$

ATE is the expected effect of treatment on a randomly drawn person from the population. Some have criticized this measure as not being especially relevant for policy purposes: because it averages across the entire population, it includes in the average units who would never be eligible for treatment. Heckman (1997) gives the example of a job training program, where we would not want to include millionaires in computing the average effect of a job training program. This criticism is somewhat misleading, as we can—and would—exclude people from the population who would never be eligible. For example, in evaluating a job training program, we might re-

strict attention to people whose pretraining income is below a certain threshold; wealthy people would be excluded precisely because we have no interest in how job training affects the wealthy. In evaluating the benefits of a program such as Head Start, we could restrict the population to those who are actually eligible for the program or are likely to be eligible in the future. In evaluating the effectiveness of enterprise zones, we could restrict our analysis to block groups whose unemployment rates are above a certain threshold or whose per capita incomes are below a certain level.

A second quantity of interest, and one that has received much recent attention, is the **average treatment effect on the treated**, which we denote ATE_1:

$$ATE_1 \equiv \mathrm{E}(y_1 - y_0 \mid w = 1) \tag{18.2}$$

That is, ATE_1 is the mean effect for those who actually participated in the program. As we will see, in some special cases equations (18.1) and (18.2) are equivalent, but generally they differ.

Imbens and Angrist (1994) define another treatment effect, which they call a **local average treatment effect (LATE)**. LATE has the advantage of being estimable using instrumental variables under very weak conditions. It has two potential drawbacks: (1) it measures the effect of treatment on a generally unidentifiable subpopulation; and (2) the definition of LATE depends on the particular instrumental variable that we have available. We will discuss LATE in the simplest setting in Section 18.4.2.

We can expand the definition of both treatment effects by conditioning on covariates. If x is an observed covariate, the ATE conditional on x is simply $\mathrm{E}(y_1 - y_0 \mid x)$; similarly, equation (18.2) becomes $\mathrm{E}(y_1 - y_0 \mid x, w = 1)$. By choosing x appropriately, we can define ATEs for various subsets of the population. For example, x can be pretraining income or a binary variable indicating poverty status, race, or gender. For the most part, we will focus on ATE and ATE_1 without conditioning on covariates.

As noted previously, the difficulty in estimating equation (18.1) or (18.2) is that we observe only y_0 or y_1, not both, for each person. More precisely, along with w, the observed outcome is

$$y = (1 - w)y_0 + wy_1 = y_0 + w(y_1 - y_0) \tag{18.3}$$

Therefore, the question is, How can we estimate equation (18.1) or (18.2) with a random sample on y and w (and usually some observed covariates)?

First, suppose that the treatment indicator w is statistically independent of (y_0, y_1), as would occur when treatment is *randomized* across agents. One implication of independence between treatment status and the potential outcomes is that ATE and ATE_1 are identical: $\mathrm{E}(y_1 - y_0 \mid w = 1) = \mathrm{E}(y_1 - y_0)$. Furthermore, estimation of

ATE is simple. Using equation (18.3), we have

$$\mathrm{E}(y \mid w = 1) = \mathrm{E}(y_1 \mid w = 1) = \mathrm{E}(y_1)$$

where the last equality follows because y_1 and w are independent. Similarly,

$$\mathrm{E}(y \mid w = 0) = \mathrm{E}(y_0 \mid w = 0) = \mathrm{E}(y_0)$$

It follows that

$$ATE = ATE_1 = \mathrm{E}(y \mid w = 1) - \mathrm{E}(y \mid w = 0) \tag{18.4}$$

The right-hand side is easily estimated by a difference in sample means: the sample average of y for the treated units minus the sample average of y for the untreated units. Thus, randomized treatment guarantees that the difference-in-means estimator from basic statistics is unbiased, consistent, and asymptotically normal. In fact, these properties are preserved under the weaker assumption of **mean independence**: $\mathrm{E}(y_0 \mid w) = \mathrm{E}(y_0)$ and $\mathrm{E}(y_1 \mid w) = \mathrm{E}(y_1)$.

Randomization of treatment is often infeasible in program evaluation (although randomization of *eligibility* often is feasible; more on this topic later). In most cases, individuals at least partly determine whether they receive treatment, and their decisions may be related to the benefits of treatment, $y_1 - y_0$. In other words, there is **self-selection** into treatment.

It turns out that ATE_1 can be consistently estimated as a difference in means under the weaker assumption that w is independent of y_0, without placing any restriction on the relationship between w and y_1. To see this point, note that we can always write

$$\begin{aligned}\mathrm{E}(y \mid w = 1) - \mathrm{E}(y \mid w = 0) &= \mathrm{E}(y_0 \mid w = 1) - \mathrm{E}(y_0 \mid w = 0) + \mathrm{E}(y_1 - y_0 \mid w = 1) \\ &= [\mathrm{E}(y_0 \mid w = 1) - \mathrm{E}(y_0 \mid w = 0)] + ATE_1 \end{aligned} \tag{18.5}$$

If y_0 is mean independent of w, that is,

$$\mathrm{E}(y_0 \mid w) = \mathrm{E}(y_0) \tag{18.6}$$

then the first term in equation (18.5) disappears, and so the difference in means estimator is an unbiased estimator of ATE_1. Unfortunately, condition (18.6) is a strong assumption. For example, suppose that people are randomly made eligible for a voluntary job training program. Condition (18.6) effectively implies that the participation decision is unrelated to what people would earn in the absence of the program.

A useful expression relating ATE_1 and ATE is obtained by writing $y_0 = \mu_0 + v_0$ and $y_1 = \mu_1 + v_1$, where $\mu_g = \mathrm{E}(y_g)$, $g = 0, 1$. Then

$$y_1 - y_0 = (\mu_1 - \mu_0) + (v_1 - v_0) = ATE + (v_1 - v_0)$$

Taking the expectation of this equation conditional on $w = 1$ gives

$$ATE_1 = ATE + \mathrm{E}(v_1 - v_0 \mid w = 1)$$

We can think of $v_1 - v_0$ as the person-specific gain from participation, and so ATE_1 differs from ATE by the expected person-specific gain for those who participated. If $y_1 - y_0$ is not mean independent of w, ATE_1 and ATE generally differ.

Fortunately, we can estimate ATE and ATE_1 under assumptions less restrictive than independence of (y_0, y_1) and w. In most cases, we can collect data on individual characteristics and relevant pretreatment outcomes—sometimes a substantial amount of data. If, in an appropriate sense, treatment depends on the observables and not on the unobservables determining (y_0, y_1), then we can estimate average treatment effects quite generally, as we show in the next section.

18.3 Methods Assuming Ignorability of Treatment

We adopt the framework of the previous section, and, in addition, we let $\mathbf{x}$ denote a vector of observed covariates. Therefore, the population is described by $(y_0, y_1, w, \mathbf{x})$, and we observe y, w, and $\mathbf{x}$, where y is given by equation (18.3). When w and (y_0, y_1) are allowed to be correlated, we need an assumption in order to identify treatment effects. Rosenbaum and Rubin (1983) introduced the following assumption, which they called **ignorability of treatment** (given observed covariates $\mathbf{x}$):

ASSUMPTION ATE.1: Conditional on $\mathbf{x}$, w and (y_0, y_1) are independent.

For many purposes, it suffices to assume ignorability in a **conditional mean independence** sense:

ASSUMPTION ATE.1′: (a) $\mathrm{E}(y_0 \mid \mathbf{x}, w) = \mathrm{E}(y_0 \mid \mathbf{x})$; and (b) $\mathrm{E}(y_1 \mid \mathbf{x}, w) = \mathrm{E}(y_1 \mid \mathbf{x})$.

Naturally, Assumption ATE.1 implies Assumption ATE.1′. In practice, Assumption ATE.1′ might not afford much generality, although it does allow $\mathrm{Var}(y_0 \mid \mathbf{x}, w)$ and $\mathrm{Var}(y_1 \mid \mathbf{x}, w)$ to depend on w. The idea underlying Assumption ATE.1′ is this: if we can observe enough information (contained in $\mathbf{x}$) that determines treatment, then (y_0, y_1) might be mean independent of w, conditional on $\mathbf{x}$. Loosely, even though (y_0, y_1) and w might be correlated, they are uncorrelated once we partial out $\mathbf{x}$.

Assumption ATE.1 certainly holds if w is a deterministic function of $\mathbf{x}$, which has prompted some authors in econometrics to call assumptions like ATE.1 **selection on observables**; see, for example, Barnow, Cain, and Goldberger (1980, 1981), Heckman and Robb (1985), and Moffitt (1996). (We discussed a similar assumption in Section

17.7.3 in the context of attrition in panel data.) The name is fine as a label, but we must realize that Assumption ATE.1 does allow w to depend on unobservables, albeit in a restricted fashion. If $w = g(\mathbf{x}, a)$, where a is an unobservable random variable independent of $(\mathbf{x}, y_0, y_1)$, then Assumption ATE.1 holds. But a cannot be arbitrarily correlated with y_0 and y_1.

An important fact is that, under Assumption ATE.1′, the average treatment effect conditional on $\mathbf{x}$ and the average treatment effect of the treated, conditional on $\mathbf{x}$, are identical:

$$ATE_1(\mathbf{x}) \equiv \mathrm{E}(y_1 - y_0 \mid \mathbf{x}, w = 1) = \mathrm{E}(y_1 - y_0 \mid \mathbf{x}) = ATE(\mathbf{x})$$

because $\mathrm{E}(y_g \mid \mathbf{x}, w) = \mathrm{E}(y_g \mid \mathbf{x})$, $g = 0, 1$. However, the unconditional versions of the treatment effects are not generally equal. For clarity, define $r(\mathbf{x}) = \mathrm{E}(y_1 - y_0 \mid \mathbf{x}) = ATE(\mathbf{x})$. Then ATE is the expected value of $r(\mathbf{x})$ across the entire population, whereas ATE_1 is the expected value of $r(\mathbf{x})$ in the treated subpopulation. Mathematically,

$$ATE = \mathrm{E}[r(\mathbf{x})] \qquad \text{and} \qquad ATE_1 = \mathrm{E}[r(\mathbf{x}) \mid w = 1]$$

If we can estimate $r(\cdot)$, then ATE can be estimated by averaging across the entire random sample from the population, whereas ATE_1 would be estimated by averaging across the part of the sample with $w_i = 1$. We will discuss specific estimation strategies in the next subsection.

An interesting feature of Assumptions ATE.1 and ATE.1′—and one that is perhaps foreign to economists—is that they are stated without imposing any kind of model on joint or conditional distributions. It turns out that no more structure is needed in order to identify either of the treatment effects. We first show how the ignorability assumption relates to standard regression analysis.

18.3.1 Regression Methods

We can use equation (18.3), along with Assumption ATE.1′, to obtain estimators of $ATE(\mathbf{x})$, which can then be used to estimate ATE and ATE_1. First,

$$\begin{aligned}\mathrm{E}(y \mid \mathbf{x}, w) &= \mathrm{E}(y_0 \mid \mathbf{x}, w) + w[\mathrm{E}(y_1 \mid \mathbf{x}, w) - \mathrm{E}(y_0 \mid \mathbf{x}, w)] \\ &= \mathrm{E}(y_0 \mid \mathbf{x}) + w[\mathrm{E}(y_1 \mid \mathbf{x}) - \mathrm{E}(y_0 \mid \mathbf{x})]\end{aligned}$$

where the first equality follows from equation (18.3) and the second follows from Assumption ATE.1′. Therefore, under Assumption ATE.1′,

$$\mathrm{E}(y \mid \mathbf{x}, w = 1) - \mathrm{E}(y \mid \mathbf{x}, w = 0) = \mathrm{E}(y_1 \mid \mathbf{x}) - \mathrm{E}(y_0 \mid \mathbf{x}) = ATE(\mathbf{x}) \tag{18.7}$$

Because we have a random sample on $(y, w, \mathbf{x})$ from the relevant population, $r_1(\mathbf{x}) \equiv \mathrm{E}(y \mid \mathbf{x}, w = 1)$ and $r_0(\mathbf{x}) \equiv \mathrm{E}(y \mid \mathbf{x}, w = 0)$ are **nonparametrically identified**. That is, these are conditional expectations that depend entirely on observables, and so they can be consistently estimated quite generally. (See Härdle and Linton, 1994, for assumptions and methods.) For the purposes of identification, we can just assume $r_1(\mathbf{x})$ and $r_0(\mathbf{x})$ are known, and the fact that they are known means that $ATE(\mathbf{x})$ is identified. If $\hat{r}_1(\mathbf{x})$ and $\hat{r}_0(\mathbf{x})$ are consistent estimators (in an appropriate sense), using the random sample of size N, a consistent estimator of ATE under fairly weak assumptions is

$$\hat{ATE} = N^{-1} \sum_{i=1}^{N} [\hat{r}_1(\mathbf{x}_i) - \hat{r}_0(\mathbf{x}_i)]$$

while a consistent estimator of ATE_1 is

$$\hat{ATE}_1 = \left(\sum_{i=1}^{N} w_i \right)^{-1} \left\{ \sum_{i=1}^{N} w_i [\hat{r}_1(\mathbf{x}_i) - \hat{r}_0(\mathbf{x}_i)] \right\}$$

The formula for $\hat{ATE}_1$ simply averages $[\hat{r}_1(\mathbf{x}_i) - \hat{r}_0(\mathbf{x}_i)]$ over the subsample with $w_i = 1$.

There are several implementation issues that arise in computing and using $\hat{ATE}$ and $\hat{ATE}_1$. The most obvious of these is obtaining $\hat{r}_1(\cdot)$ and $\hat{r}_0(\cdot)$. To be as flexible as possible, we could use nonparametric estimators, such as a **kernel estimator** (see Härdle and Linton, 1994). Obtaining reliable standard errors when we use nonparametric estimates can be difficult. An alternative is to use flexible parametric models, such as low-order polynomials that include interaction terms. [Presumably, we would also account for the nature of y in estimating $\mathrm{E}(y \mid \mathbf{x}, w = 1)$ and $\mathrm{E}(y \mid \mathbf{x}, w = 0)$. For example, if y is binary, we would use a flexible logit or probit; if y is a corner solution, we might use a flexible Tobit or a flexible exponential regression function.]

With plenty of data, a third possibility is to list all possible values that $\mathbf{x}$ can take, say, $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M$, and to estimate $\mathrm{E}(y \mid \mathbf{x} = \mathbf{c}_m, w = 1)$ by averaging the y_i over all i with $\mathbf{x}_i = \mathbf{c}_m$ and $w_i = 1$; $\mathrm{E}(y \mid \mathbf{x} = \mathbf{c}_m, w = 0)$ is estimated similarly. For each m and $w = 0$ or 1, this method is just estimation of a mean using a sample average. Typically, M is large because $\mathbf{x}$ takes on many values, and many of the cells may have only a small number of observations.

Regardless of how $\hat{r}_1(\cdot)$ and $\hat{r}_0(\cdot)$ are obtained, to use the estimated treatment effects we need to obtain asymptotically valid standard errors. Generally, this task can be very difficult, especially if nonparametric methods are used in estimation.

Nevertheless, we will show how a linear regression model involving level effects and interactions can be used to obtain good estimates of the treatment effects as well as reliable standard errors.

Before we turn to standard regression models, we need to discuss a problem that can arise in the evaluation of programs, especially when flexible estimation of $E(y \mid \mathbf{x}, w = 1)$ and $E(y \mid \mathbf{x}, w = 0)$ is desirable. To illustrate the problem, suppose there is only one binary covariate, x, and Assumption ATE.1′ holds; for concreteness, x could be an indicator for whether pretraining earnings are below a certain threshold. Suppose that everyone in the relevant population with $x = 1$ participates in the program. Then, while we can estimate $E(y \mid x = 1, w = 1)$ with a random sample from the population, we cannot estimate $E(y \mid x = 1, w = 0)$ because we have no data on the subpopulation with $x = 1$ and $w = 0$. Intuitively, we only observe the counterfactual y_1 when $x = 1$; we never observe y_0 for any members of the population with $x = 1$. Therefore, $ATE(x)$ is not identified at $x = 1$.

If some people with $x = 0$ participate while others do not, we can estimate $E(y \mid x = 0, w = 1) - E(y \mid x = 0, w = 0)$ using a simple difference in averages over the group with $x = 0$, and so $ATE(x)$ is identified at $x = 0$. But if we cannot estimate $ATE(1)$, we cannot estimate the unconditional ATE because $ATE = P(x = 0) \cdot ATE(0) + P(x = 1) \cdot ATE(1)$. In effect, we can only estimate the ATE over the subpopulation with $x = 0$, which means that we must redefine the population of interest. This limitation is unfortunate: presumably we would be very interested in the program's effects on the group that always participates.

A similar conclusion holds if the group with $x = 0$ never participates in the program. Then $ATE(0)$ is not estimable because $E(y \mid x = 0, w = 1)$ is not estimable. If some people with $x = 1$ participated while others did not, $ATE(1)$ would be identified, and then we would view the population of interest as the subgroup with $x = 1$. There is one important difference between this situation and the one where the $x = 1$ group always receives treatment: it seems perfectly natural to exclude from the population people who have no chance of treatment based on observed covariates. This observation is related to the issue we discussed in Section 18.2 concerning the relevant population for defining ATE. If, for example, people with very high preprogram earnings ($x = 0$) have no chance of participating in a job training program, then we would not want to average together $ATE(0)$ and $ATE(1)$; $ATE(1)$ by itself is much more interesting.

Although the previous example is extreme, its consequences can arise in more plausible settings. Suppose that $\mathbf{x}$ is a vector of binary indicators for pretraining income intervals. For most of the intervals, the probability of participating is strictly between zero and one. If the participation probability is zero at the highest income

level, we simply exclude the high-income group from the relevant population. Unfortunately, if participation is certain at low income levels, we must exclude low-income groups as well.

As a practical matter, we often determine whether the probability of participation is one or zero by looking at the random sample. If we list the possible values of the explanatory variables, $\mathbf{c}_1, \ldots, \mathbf{c}_M$, as described earlier, the problem arises when there is a value, say $\mathbf{c}_m$, where all units with $\mathbf{x}_i = \mathbf{c}_m$ participate in the program. Because we cannot estimate $\mathrm{E}(y \mid \mathbf{x} = \mathbf{c}_m, w = 0)$, the subpopulation with $\mathbf{x} = \mathbf{c}_m$ must be excluded from the analysis.

We now turn to standard parametric regression methods for estimating ATE, and then briefly discuss estimating ATE_1. It is useful to decompose the counterfactual outcomes into their means and a stochastic part with zero mean, as we did at the end of Section 18.2:

$$y_0 = \mu_0 + v_0, \qquad \mathrm{E}(v_0) = 0 \tag{18.8}$$

$$y_1 = \mu_1 + v_1, \qquad \mathrm{E}(v_1) = 0 \tag{18.9}$$

Plugging these into equation (18.3) gives

$$y = \mu_0 + (\mu_1 - \mu_0)w + v_0 + w(v_1 - v_0) \tag{18.10}$$

This is a simple example of a **switching regression model**, where the outcome equations depend on the regime (treatment status in this case).

If we assume that $v_1 - v_0$ has zero mean conditional on $\mathbf{x}$, we obtain a standard regression model under Assumption ATE.1′.

PROPOSITION 18.1: Under Assumption ATE.1′, assume, in addition, that

$$\mathrm{E}(v_1 \mid \mathbf{x}) = \mathrm{E}(v_0 \mid \mathbf{x}) \tag{18.11}$$

Then $ATE_1 = ATE$, and

$$\mathrm{E}(y \mid w, \mathbf{x}) = \mu_0 + \alpha w + g_0(\mathbf{x}) \tag{18.12}$$

where $\alpha \equiv ATE$ and $g_0(\mathbf{x}) = \mathrm{E}(v_0 \mid \mathbf{x})$. If, in addition, $\mathrm{E}(v_0 \mid \mathbf{x}) = \eta_0 + \mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0$ for some vector function $\mathbf{h}_0(\mathbf{x})$, then

$$\mathrm{E}(y \mid w, \mathbf{x}) = \gamma_0 + \alpha w + \mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0 \tag{18.13}$$

where $\gamma_0 = \mu_0 + \eta_0$.

Proof: Under Assumption ATE.1′, $\mathrm{E}(y_1 \mid w, \mathbf{x}) = \mu_1 + \mathrm{E}(v_1 \mid \mathbf{x})$ and $\mathrm{E}(y_0 \mid w, \mathbf{x}) = \mu_0 + \mathrm{E}(v_0 \mid \mathbf{x})$. Under assumption (18.11), $\mathrm{E}(y_1 \mid w, \mathbf{x}) - \mathrm{E}(y_0 \mid w, \mathbf{x}) = \mu_1 - \mu_0$.

Therefore, by iterated expectations, $E(y_1 \mid w) - E(y_0 \mid w) = \mu_1 - \mu_0$, which implies that $ATE_1 = ATE$. The proof of equation (18.12) follows by taking the expectation of equation (18.10) given w, $\mathbf{x}$ and using Assumption ATE.1′ and assumption (18.11).

This proposition shows that when the predicted person-specific gain given $\mathbf{x}$ is zero—that is, when $E(v_1 - v_0 \mid \mathbf{x}) = 0$—$E(y \mid w, \mathbf{x})$ is additive in w and a function of $\mathbf{x}$, and the coefficient on w is the average treatment effect. It follows that standard regression methods can be used to estimate ATE. While nonlinear regression methods can be used if $E(v_0 \mid \mathbf{x})$ is assumed to be nonlinear in parameters, typically we would use an assumption such as equation (18.13). Then, regressing y on an intercept, w, and $\mathbf{h}_0(\mathbf{x})$ consistently estimates the ATE. By putting enough controls in $\mathbf{x}$, we have arranged it so that w and unobservables affecting (y_0, y_1) are appropriately unrelated. In effect, $\mathbf{x}$ proxies for the unobservables. Using flexible functional forms for the elements of $\mathbf{h}_0(\mathbf{x})$ should provide a good approximation to $E(v_0 \mid \mathbf{x})$.

The function $\mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0$ in equation (18.13) is an example of a **control function**: when added to the regression of y on 1, w, it controls for possible self-selection bias. Of course, this statement is only true under the assumptions in Proposition 18.1.

Given Assumption ATE.1′, the additively separable form of equation (18.12) hinges crucially on assumption (18.11). Though assumption (18.11) might be reasonable in some cases, it need not generally hold. [A sufficient, but not necessary, condition for assumption (18.11) is $v_1 = v_0$ or $y_1 = \alpha + y_0$, which means the effect of treatment is the same for everyone in the population.] If we relax assumption (18.11), then we no longer have equality of ATE and ATE_1. Nevertheless, a regression formulation can be used to estimate ATE:

PROPOSITION 18.2: Under Assumption ATE.1′,

$$E(y \mid w, \mathbf{x}) = \mu_0 + \alpha w + g_0(\mathbf{x}) + w[g_1(\mathbf{x}) - g_0(\mathbf{x})] \tag{18.14}$$

where $\alpha = ATE$, $g_0(\mathbf{x}) \equiv E(v_0 \mid \mathbf{x})$, and $g_1(\mathbf{x}) \equiv E(v_1 \mid \mathbf{x})$.

The proof of Proposition 18.2 is immediate by taking the expectation of equation (18.10) given $(w, \mathbf{x})$. Equation (18.14) is interesting because it shows that, under Assumption ATE.1′ only, $E(y \mid w, \mathbf{x})$ is additive in w, a function of $\mathbf{x}$, and an interaction between w and another function of $\mathbf{x}$. The coefficient on w is the average treatment effect (but not generally ATE_1). To operationalize equation (18.14) in a parametric framework, we would replace $g_0(\cdot)$ and $g_1(\cdot)$ with parametric functions of $\mathbf{x}$; typically, these would be linear in parameters, say $\eta_0 + \mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0$ and $\eta_1 + \mathbf{h}_1(\mathbf{x})\boldsymbol{\beta}_1$. For notational simplicity, assume that these are both linear in $\mathbf{x}$. Then we can write

$$\mathrm{E}(y \mid w, \mathbf{x}) = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} \tag{18.15}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\delta}$ are vectors of unknown parameters and $\boldsymbol{\psi} \equiv \mathrm{E}(\mathbf{x})$. Subtracting the mean from $\mathbf{x}$ ensures that ATE is the coefficient on w. In practice, either we would subtract off the known population mean from each element of $\mathbf{x}$, or, more likely, we would demean each element of $\mathbf{x}$ using the sample average. Therefore, under equation (18.15), we would estimate α as the coefficient on w in the regression

$$y_i \text{ on } 1, w_i, \mathbf{x}_i, w_i(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad i = 1, 2, \ldots, N \tag{18.16}$$

where $\bar{\mathbf{x}}$ is the vector of sample averages. (Subtracting the sample averages rather than population averages introduces a generated regressor problem. However, as argued in Problem 6.10, the adjustments to the standard errors typically have minor effects.) The control functions in this case involve not just the $\mathbf{x}_i$, but also interactions of the covariates with the treatment variable. If desired, we can be selective about which elements of $(\mathbf{x}_i - \bar{\mathbf{x}})$ we interact with w_i.

Adding functions of $\mathbf{x}$, such as squares or logarithms, as both level terms and interactions, is simple, provided we demean any functions before constructing the interactions.

Because regression (18.16) consistently estimates $\boldsymbol{\delta}$, we can also study how the ATE given $\mathbf{x}$, that is, $ATE(\mathbf{x}) = \mathrm{E}(y_1 - y_0 \mid \mathbf{x})$, changes with elements of $\mathbf{x}$. In particular, for any $\mathbf{x}$ in the valid range,

$$\hat{ATE}(\mathbf{x}) = \hat{\alpha} + (\mathbf{x} - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}}$$

We can then average this equation over interesting values of $\mathbf{x}$ to obtain the ATE for a subset of the population. For example, if $\mathbf{x}$ contains pretraining earnings or indicators for earnings groups, we can estimate how the ATE changes for various levels of pretraining earnings.

If the functions of $\mathbf{x}$ appearing in the regression are very flexible, problems with estimating $ATE(\mathbf{x})$ at certain values of $\mathbf{x}$ can arise. In the extreme case, we define dummy variables for each possible outcome on $\mathbf{x}$ and use these in place of $\mathbf{x}$. This approach results in what is known as a **saturated model**. We will not be able to include dummy variables for groups that are always treated or never treated, with the result that our estimator of ATE is for the population that excludes these groups.

To estimate ATE_1, write $ATE_1 = \alpha + [\mathrm{E}(\mathbf{x} \mid w = 1) - \boldsymbol{\psi}]\boldsymbol{\delta}$, and so a consistent estimator is

$$\hat{ATE}_1 = \hat{\alpha} + \left(\sum_{i=1}^{N} w_i\right)^{-1}\left[\sum_{i=1}^{N} w_i(\mathbf{x}_i - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}}\right]$$

Obtaining a standard error for this estimator is somewhat complicated, but it can be done using the delta method or bootstrapping.

Example 18.1 (Effects of Enterprise Zones on Economic Development): Consider evaluating the effects of enterprise zone (EZ) designation on employment growth, for block groups in a particular state. Suppose that we have 1980 and 1990 census data, and that the EZ designation originated in the early 1980s. To account for the fact that zone designation is likely to depend on prior economic performance, and perhaps other block characteristics, we can estimate a model such as the following:

$$\begin{aligned} gemp = {} & \mu_0 + \alpha ez + \beta_1 \log(emp80) + \beta_2 \log(pop80) + \beta_3 percmanf80 \\ & + \beta_4 \log(housval80) + \beta_5 ez \cdot [\log(emp80) - m_1] + \beta_6 ez \cdot [\log(pop80) - m_2] \\ & + \beta_7 ez \cdot [percmanf80 - m_3] + \beta_8 ez \cdot [\log(housval80) - m_4] + error \end{aligned}$$

where the right-hand-side variables are a dummy variable for EZ designation, employment, population, percent of employment in manufacturing, and median housing value, all in 1980, and where the m_j are the sample averages.

The regression estimator (18.16), especially with flexible functions of the covariates, applies directly to what are called **regression discontinuity designs**. In this case, treatment is determined as a *nonstochastic* function of a covariate, say $w = f(s)$, where s is an element of $\mathbf{x}$ that has sufficient variation. The key is that f is a discontinuous function of s, typically a step function, $w = 1[s \leq s_0]$, where s_0 is a known threshold. The idea is that once s, which could be income level or class size, reaches a certain threshold, a policy automatically kicks in. (See, for example, Angrist and Lavy, 1999.) Because s is a nonrandom function of $\mathbf{x}$, the conditional independence assumption in Assumption ATE.1 must hold. The key is obtaining flexible functional forms for $g_0(\cdot)$ and $g_1(\cdot)$. Generally, we can identify α only if we are willing to assume that $g_0(\cdot)$ and $g_1(\cdot)$ are smooth functions of $\mathbf{x}$ (which is almost always the case when we estimate parametric or nonparametric regression functions). If we allow $g_0(\cdot)$ to be discontinuous in s—that is, with jumps—we could never distinguish between changes in y due to a change in s or a change in treatment status.

18.3.2 Methods Based on the Propensity Score

Rosenbaum and Rubin (1983) use the ignorability-of-treatment assumption differently in estimating ATE. Regression (18.16) makes functional form assumptions about $\mathrm{E}(v_0 \mid \mathbf{x})$ and $\mathrm{E}(v_1 \mid \mathbf{x})$, where v_0 and v_1 are unobserved. Alternatively, it turns out that ATE and ATE_1 can both be estimated by modeling

$$p(\mathbf{x}) \equiv \mathrm{P}(w = 1 \mid \mathbf{x}) \tag{18.17}$$

which is the probability of treatment given the covariates. The function $p(\mathbf{x})$, which is simply the response probability for treatment, is called the **propensity score** in the evaluation literature. Interestingly, ATE and ATE_1 can be written in terms of the propensity score.

PROPOSITION 18.3: Under Assumption ATE.1′, assume in addition that

$$0 < p(\mathbf{x}) < 1, \qquad \text{all } \mathbf{x} \tag{18.18}$$

Then

$$ATE = \mathrm{E}([w - p(\mathbf{x})]y/\{p(\mathbf{x})[1 - p(\mathbf{x})]\}) \tag{18.19}$$

and

$$ATE_1 = \mathrm{E}\{[w - p(\mathbf{x})]y/[1 - p(\mathbf{x})]\}/\mathrm{P}(w = 1) \tag{18.20}$$

Proof: Plugging equation (18.3) into the numerator inside the expectation in equation (18.19) gives

$$\begin{aligned}[w - p(\mathbf{x})]y &= [w - p(\mathbf{x})][(1 - w)y_0 + wy_1] \\ &= wy_1 - p(\mathbf{x})(1 - w)y_0 - p(\mathbf{x})wy_1\end{aligned}$$

Taking the expectation of this equation conditional on $(w, \mathbf{x})$ and using Assumption ATE.1′ gives

$$wm_1(\mathbf{x}) - p(\mathbf{x})(1 - w)m_0(\mathbf{x}) - p(\mathbf{x})wm_1(\mathbf{x})$$

where $m_j(\mathbf{x}) \equiv \mathrm{E}(y_j \mid \mathbf{x})$, $j = 0, 1$. Taking the expectation conditional on $\mathbf{x}$ gives

$$p(\mathbf{x})m_1(\mathbf{x}) - p(\mathbf{x})[1 - p(\mathbf{x})]m_0(\mathbf{x}) - [p(\mathbf{x})]^2 m_1(\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})][m_1(\mathbf{x}) - m_0(\mathbf{x})]$$

because $p(\mathbf{x}) = \mathrm{E}(w \mid \mathbf{x})$. Therefore, the expected value of the term in equation (18.19) conditional on $\mathbf{x}$ is simply $[m_1(\mathbf{x}) - m_0(\mathbf{x})]$; iterated expectations implies that the right-hand side of equation (18.19) is $\mu_1 - \mu_0$.

Very similar reasoning shows that

$$\mathrm{E}\{[w - p(\mathbf{x})]y/[1 - p(\mathbf{x})] \mid \mathbf{x}\} = p(\mathbf{x})[m_1(\mathbf{x}) - m_0(\mathbf{x})]$$

Next, by iterated expectations,

$$\mathrm{E}\{p(\mathbf{x})[m_1(\mathbf{x}) - m_0(\mathbf{x})]\} = \mathrm{E}\{w[m_1(\mathbf{x}) - m_0(\mathbf{x})]\} = \mathrm{E}[w(y_1 - y_0)]$$

where the last equality follows from Assumption ATE.1′. But

$$\begin{aligned}\mathrm{E}[w(y_1 - y_0)] &= \mathrm{P}(w = 1)\mathrm{E}[w(y_1 - y_0) \mid w = 1] + \mathrm{P}(w = 0)\mathrm{E}[w(y_1 - y_0) \mid w = 0] \\ &= \mathrm{P}(w = 1)\mathrm{E}(y_1 - y_0 \mid w = 1)\end{aligned}$$

Therefore, the right-hand side of equation (18.20) is

$$\{\mathrm{P}(w = 1)\mathrm{E}(y_1 - y_0 \mid w = 1)\}/\mathrm{P}(w = 1) = ATE_1.$$

Rosenbaum and Rubin (1983) call Assumption ATE.1 *plus* condition (18.18) **strong ignorability of treatment** (given covariates $\mathbf{x}$). Proposition 18.3 shows, in a different way from Section 18.3.1, that ATE and ATE_1 are nonparametrically identified under strong ignorability of treatment: the response probability, $\mathrm{P}(y = 1 \mid \mathbf{x})$, can be assumed known for the purposes of identification analysis. Wooldridge (1999c) obtained equation (18.19) in the more general setting of a random coefficient model (see Section 18.5.3), while equation (18.20) is essentially due to Dehejia and Wahba (1999, Proposition 4), who make the stronger assumption ATE.1.

Condition (18.18) is precisely the restriction on the response probability that arose in Section 18.3.1 for identifying ATE. Equation (18.20) shows that ATE_1 is still identified if $p(\mathbf{x}) = 0$ for some $\mathbf{x}$, but this finding has little practical value because we probably want to exclude units that have no chance of being treated, anyway. Importantly, in estimating ATE or ATE_1, we rule out $p(\mathbf{x}) = 1$: we cannot estimate ATE or ATE_1 by including in the population units that are treated with certainty, conditional on $\mathbf{x}$.

Of course, to estimate ATE and ATE_1, we need an estimator of $p(\cdot)$. Rosenbaum and Rubin (1983) suggest using a flexible logit model, where $\mathbf{x}$ and various functions of $\mathbf{x}$—for example, quadratics and interactions—are included. [In this case there is no danger of $\hat{p}(\mathbf{x}) = 0$ or 1 because logit fitted values are strictly in the unit interval, but this functional form restriction might simply mask the problem in the population.] The propensity score can also be estimated using fully nonparametric methods—see, for example, Powell (1994) and Heckman, Ichimura, and Todd (1997). Here, we focus on flexible parametric methods. If $\hat{p}(\mathbf{x}) \equiv F(\mathbf{x}; \hat{\gamma})$ is such an estimator, where $\hat{\gamma}$ is obtained in a first-stage binary response estimation of w on $\mathbf{x}$, then a consistent estimator of ATE is

$$\hat{ATE} = N^{-1} \sum_{i=1}^{N} [w_i - \hat{p}(\mathbf{x}_i)] y_i / \{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]\} \tag{18.21}$$

Interestingly, after simple algebra this estimator can be shown to be identical to an estimator due to Horvitz and Thompson (1952) for handling nonrandom sampling. Consistency under standard regularity conditions follows from Lemma 12.1. Simi-

larly, a consistent estimator of ATE_1 is

$$\hat{ATE}_1 = \left(N^{-1} \sum_{i=1}^{N} w_i \right)^{-1} \left\{ N^{-1} \sum_{i=1}^{N} [w_i - \hat{p}_i(\mathbf{x}_i)] y_i / [1 - \hat{p}(\mathbf{x}_i)] \right\} \tag{18.22}$$

Notice that $N^{-1} \sum_{i=1}^{N} w_i$ is a consistent estimator of $P(w = 1)$. Obtaining valid asymptotic standard errors using the delta method is somewhat complicated, as we need a first-order representation for $\sqrt{N}(\hat{\gamma} - \gamma)$—see Section 12.5.2. Notice that only the predicted probabilities appear in equations (18.21) and (18.22). Therefore, different methods of estimating $p(\mathbf{x})$ that lead to similar predicted values $\hat{p}(\mathbf{x}_i)$ will tend to produce similar treatment effect estimates.

It turns out that the estimators in equations (18.21) and (18.22) not only are convenient, but also can be made to have the smallest asymptotic variances among estimators that are based only on Assumption ATE.1 and condition (18.18) (as well as several regularity conditions). Hirano, Imbens, and Ridder (2000) (HIR) have recently shown that (18.21) and (18.22) achieve the semiparametric efficiency bound obtained by Hahn (1998). In order to achieve the bound, HIR assume that $\hat{p}(\cdot)$ is a series estimator, so that the conditions in Newey (1994) can be verified. As a practical matter, series estimation is not ideal, because, for a binary response, it is identical to a linear probability model in functions of $\mathbf{x}$. Plus, it is difficult to estimate the asymptotic variance of the resulting estimators, $\hat{ATE}$ and $\hat{ATE}_1$. Probably little is lost by using a flexible logit or probit and then obtaining the standard errors by the usual delta method.

A simple, popular estimator in program evaluation is obtained from an OLS regression that simply includes the estimated propensity score, $\hat{p}(\mathbf{x})$, as an additional regressor:

$$y_i \text{ on } 1, w_i, \hat{p}(\mathbf{x}_i), \qquad i = 1, 2, \ldots, N \tag{18.23}$$

where the coefficient on w_i is the estimate of the treatment effect. In other words, the estimated propensity score plays the role of the control function. The idea is that the estimated propensity score should contain all the information in the covariates that is relevant for estimating the treatment effect. The question is, When does regression (18.23) consistently estimate the average treatment effect? The following is a special case of Wooldridge (1999c, Proposition 3.2):

PROPOSITION 18.4: In addition to Assumption ATE.1′, assume that $E(y_1 - y_0 \mid \mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x})$ is *uncorrelated* with $\text{Var}(w \mid \mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$. If the parametric estimator $\hat{p}(\cdot)$ is consistent and $\sqrt{N}$-asymptotically normal, then the OLS coefficient on

w from regression (18.23) is consistent and $\sqrt{N}$-asymptotically normal for the average treatment effect, ATE.

The assumption that $m_1(\mathbf{x}) - m_0(\mathbf{x})$ is uncorrelated with $\text{Var}(w \mid \mathbf{x})$ may appear unlikely, as both are functions of $\mathbf{x}$. However, remember that correlation is a linear measure of dependence. The conditional variance $\text{Var}(w \mid \mathbf{x})$ is a nonmonotonic quadratic in $p(\mathbf{x})$, while $m_1(\mathbf{x}) - m_0(\mathbf{x})$ is likely to be monotonic in many elements of $\mathbf{x}$; zero correlation might hold approximately. (This observation is analogous to the fact that if z is a standard normal random variable, then z and z^2 are uncorrelated.)

Using different auxiliary assumptions, Rosenbaum and Rubin (1983, Corollary 4.3) suggest a more general version of regression (18.23) for estimating ATE:

$$y_i \text{ on } 1, w_i, \hat{p}_i, w_i(\hat{p}_i - \hat{\mu}_p), \qquad i = 1, 2, \ldots, N \tag{18.24}$$

where $\hat{\mu}_p$ is the sample average of $\hat{p}_i$, $i = 1, 2, \ldots, N$.

PROPOSITION 18.5: Under Assumption ATE.1, assume in addition that $\text{E}[y_0 \mid p(\mathbf{x})]$ and $\text{E}[y_1 \mid p(\mathbf{x})]$ are linear in $p(\mathbf{x})$. Then the coefficient on w_i in regression (18.24) consistently estimates ATE.

Proof: Rosenbaum and Rubin (1983, Theorem 3) show that, under Assumption ATE.1, (y_0, y_1) and w are independent conditional on $p(\mathbf{x})$. For completeness, we present the argument. It suffices to show $\text{P}[w = 1 \mid y_0, y_1, p(\mathbf{x})] = \text{P}[w = 1 \mid p(\mathbf{x})]$ or $\text{E}[w \mid y_0, y_1, p(\mathbf{x})] = \text{E}[w \mid p(\mathbf{x})]$. But, under Assumption ATE.1, $\text{E}(w \mid y_0, y_1, \mathbf{x}) = \text{E}(w \mid \mathbf{x}) = p(\mathbf{x})$. By iterated expectations,

$$\text{E}[w \mid y_0, y_1, p(\mathbf{x})] = \text{E}[\text{E}(w \mid y_0, y_1, \mathbf{x}) \mid y_0, y_1, p(\mathbf{x})] = \text{E}[p(\mathbf{x}) \mid y_0, y_1, p(\mathbf{x})] = p(\mathbf{x})$$

We can now use this equation to obtain $\text{E}[y \mid w, p(\mathbf{x})]$. Write $y = y_0 + (\mu_1 - \mu_0)w + w(v_1 - v_0)$. We just showed that (y_0, y_1) and w are independent given $p(\mathbf{x})$, and so

$$\begin{aligned}\text{E}[y \mid w, p(\mathbf{x})] &= \text{E}[y_0 \mid p(\mathbf{x})] + (\mu_1 - \mu_0)w + w\{\text{E}[v_1 \mid p(\mathbf{x})] - \text{E}[v_0 \mid p(\mathbf{x})]\} \\ &= \delta_0 + \delta_1 p(\mathbf{x}) + (\mu_1 - \mu_0)w + \delta_2 w[p(\mathbf{x}) - \mu_p]\end{aligned}$$

under the linearity assumptions, where $\mu_p \equiv \text{E}[p(\mathbf{x})]$. {Remember, as v_1 and v_0 have zero means, the linear function of $p(\mathbf{x})$ must have a zero mean, too; we can always write it as $\delta_2[p(\mathbf{x}) - \mu_p]$.} This step completes the proof, as replacing μ_p with its sample average in the regression does not affect consistency (or asymptotic normality).

The linearity assumptions for $\text{E}[y_0 \mid p(\mathbf{x})]$ and $\text{E}[y_1 \mid p(\mathbf{x})]$ are probably too restrictive in many applications. As $p(\mathbf{x})$ is bounded between zero and one, $\text{E}[y_0 \mid p(\mathbf{x})]$ and $\text{E}[y_1 \mid p(\mathbf{x})]$ are necessarily bounded under linearity, which might be a poor as-

sumption if the y_g have a wide support. If y is binary, linearity of these expected values is also questionable, but it could be a reasonable approximation. Of course, it is a simple matter to replace $\hat{p}_i$ with a low-order polynomial in $\hat{p}_i$, being sure to demean any term before constructing its interaction with w_i.

Example 18.2 (Effects of Job Training on Earnings): The data in JTRAIN2.RAW are from a job training experiment in the 1970s. The response variable is real earnings in 1978, measured in thousands of dollars. Real earnings are zero for men who did not work during the year. Training began up to two years prior to 1978. We use regressions (18.23) and (18.24) to estimate the average treatment effect. The elements of **x** are real earnings in 1974 and 1975, age (in quadratic form), a binary high school degree indicator (*nodegree*), marital status, and binary variables for black and Hispanic. In the first-stage probit of *train* on **x**, only *nodegree* is statistically significant at the 5 percent level. Once we have the fitted propensity scores, we can run regression (18.23). This gives $\hat{\alpha} = 1.626$ (se = .644), where the standard error is not adjusted for the probit first-stage estimation. Job training is estimated to increase earnings by about \$1,626. Interestingly, this estimate is very close to the regression *re78* on 1, *train*, **x**: $\hat{\alpha} = 1.625$ (se = .640); both are somewhat smaller than the simple comparison-of-means estimate, which is 1.794 (se = .633).

Adding the interaction term in regression (18.24), with $\hat{\mu}_p = .416$, lowers the estimate somewhat: $\hat{\alpha} = 1.560$ (se = .642). The interaction term (again, based on the usual OLS standard error) is marginally significant.

Regressions (18.23) and (18.24) are attractive because they account for possibly nonrandom assignment of treatment by including a single function of the covariates, the estimated propensity score. Compared with the regressions that include the full set of covariates, in flexible ways, possibly interacted with the treatment [as in equation (18.16)], the propensity score approach seems much more parsimonious. However, this parsimony is somewhat illusory. Remember, the propensity score is estimated by a first-stage probit or logit, where the treatment is the dependent variable and flexible functions of the elements of **x** are the explanatory variables. It is not obvious that estimating a flexible binary response model in the first stage is somehow better than the kitchen sink regression (18.16). In fact, if the propensity score were estimated using a linear probability model, regression (18.23) and regression (18.16) without the interaction terms would produce identical estimates of α. Also, using regression (18.23) or (18.24) makes it tempting to ignore the first-stage estimation of the propensity score in obtaining the standard error of the treatment effect (as we did in Example 18.2). At least in regression (18.16) we know that the standard error of $\hat{\alpha}$ is reliable; at worst, we must make the standard error robust to heteroskedasticity. In

Example 18.2, we have no way of knowing how much the sampling variation in the first-stage probit estimates would affect a properly computed standard error short of actually doing the calculations. Because the propensity score approach and the standard regression approach require different assumptions for consistency, neither generally dominates the other. [If anything, the linearity assumptions on $\mathrm{E}[y_0 \mid p(\mathbf{x})]$ and $\mathrm{E}[y_1 \mid p(\mathbf{x})]$ are less palatable than the linearity assumptions underlying equation (18.15).]

If we use the propensity score as in equation (18.21) then we need not make auxiliary assumptions as required by regressions (18.23) and (18.24). But we should still adjust the standard error of $\hat{ATE}$ to account for first-stage estimation of the propensity score. Apparently, not much work has been done comparing regression methods that use the propensity score with standard kitchen sink–type regressions, let alone comparing these procedures with the estimator from equation (18.21).

All the previous estimates of ATEs that use the estimated propensity score involve either regressions or formulas that appear similar to regressions in the sense that the propensity score is included in a sample average [see equations (18.21) and (18.22)]. Estimates of the propensity score are also used in a very different way in the treatment effect literature. Various **matching estimators** have been proposed, and asymptotic distributions are available in many cases. The matching approach suggested by Rosenbaum and Rubin (1983) is motivated by the following thought experiment. Suppose we choose a propensity score, $p(\mathbf{x})$, at random from the population. Then, we select two agents from the population sharing the chosen propensity score, where one agent receives treatment and the other does not. Under Assumption ATE.1, the expected difference in the observed outcomes for these agents is

$$\mathrm{E}[y \mid w = 1, p(\mathbf{x})] - \mathrm{E}[y \mid w = 0, p(\mathbf{x})] = \mathrm{E}[y_1 - y_0 \mid p(\mathbf{x})]$$

which is the ATE conditional on $p(\mathbf{x})$. By iterated expectations, averaging across the distribution of propensity scores gives $ATE = \mathrm{E}(y_1 - y_0)$.

An estimation strategy requires estimating the propensity scores, estimating the response differences for pairs matched on the basis of the estimated propensity scores, and then averaging over all such pairs. Because getting identical predicted probabilities is often unlikely, grouping into cells or local averaging is used instead. Effectively, agents with similar propensity scores are considered a match. Heckman, Ichimura, and Todd (1997) (HIT), Angrist (1998), and Dehejia and Wahba (1999) provide recent treatments of matching methods.

As with the regression methods discussed in Section 18.3.1, a practical problem with matching on the propensity score is that it can be hard to find treated and untreated agents with similar estimated propensity scores. HIT discuss trimming

strategies in a nonparametric context and derive asymptotically valid standard errors. Similarly, the practice of grouping on the basis of the estimated propensity scores, and then ignoring the sampling variation in both the estimated propensity scores and the grouping when constructing standard errors and confidence intervals, may be misleading. HIT show how to obtain valid inference.

18.4 Instrumental Variables Methods

We now turn to instrumental variables estimation of average treatment effects when we suspect failure of the ignorability-of-treatment assumption (ATE.1 or ATE.1′). IV methods for estimating ATEs can be very effective if a good instrument for treatment is available. We need the instrument to predict treatment (after partialing out any controls). As we discussed in Section 5.3.1, the instrument should be redundant in a certain conditional expectation and unrelated to unobserved heterogeneity; we give precise assumptions in the following subsections.

Our primary focus in this section is on the average treatment effect defined in equation (18.1), although we touch on estimating ATE_1. In Section 18.4.2, we briefly discuss estimating the local average treatment effect.

18.4.1 Estimating the ATE Using IV

In studying IV procedures, it is useful to write the observed outcome y as in equation (18.10):

$$y = \mu_0 + (\mu_1 - \mu_0)w + v_0 + w(v_1 - v_0) \tag{18.25}$$

However, unlike in Section 18.3, we do not assume that v_0 and v_1 are mean independent of w, given $\mathbf{x}$. Instead, we assume the availability of instruments, which we collect in the vector $\mathbf{z}$. (Here we separate the extra instruments from the covariates, so that $\mathbf{x}$ and $\mathbf{z}$ do not overlap. In many cases $\mathbf{z}$ is a scalar, but the analysis is no easier in that case.)

If we assume that the stochastic parts of y_1 and y_0 are the same, that is, $v_1 = v_0$, then the interaction term disappears (and $ATE = ATE_1$). Without the interaction term we can use standard IV methods under weak assumptions.

ASSUMPTION ATE.2: (a) In equation (18.25), $v_1 = v_0$; (b) $\mathrm{L}(v_0 \mid \mathbf{x}, \mathbf{z}) = \mathrm{L}(v_0 \mid \mathbf{x})$; and (c) $\mathrm{L}(w \mid \mathbf{x}, \mathbf{z}) \neq \mathrm{L}(w \mid \mathbf{x})$.

All linear projections in this chapter contain unity, which we suppress for notational simplicity.

Under parts a and b of Assumption ATE.2, we can write

$$y = \delta_0 + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + u_0 \tag{18.26}$$

where $\alpha = ATE$ and $u_0 \equiv v_0 - \mathrm{L}(v_0 \mid \mathbf{x}, \mathbf{z})$. By definition, u_0 has zero mean and is uncorrelated with $(\mathbf{x}, \mathbf{z})$, but w and u_0 are generally correlated, which makes OLS estimation of equation (18.26) inconsistent. The redundancy of $\mathbf{z}$ in the linear projection $\mathrm{L}(v_0 \mid \mathbf{x}, \mathbf{z})$ means that $\mathbf{z}$ is appropriately excluded from equation (18.26); this is the part of identification that we cannot test (except indirectly using the overidentification test from Chapter 6). Part c means that $\mathbf{z}$ has predictive power in the linear projection of treatment on $(\mathbf{x}, \mathbf{z})$; this is the standard rank condition for identification from Chapter 5, and we can test it using a first-stage regression and heteroskedasticity-robust tests of exclusion restrictions. Under Assumption ATE.2, α [and the other parameters in equation (18.26)] are identified, and they can be consistently estimated by 2SLS. Because the only endogenous explanatory variable in equation (18.26) is binary, equation (18.25) is called a **dummy endogenous variable model** (Heckman, 1978). As we discussed in Chapter 5, there are no special considerations in estimating equation (18.26) by 2SLS when the endogenous explanatory variable is binary.

Assumption ATE.2b holds if the instruments $\mathbf{z}$ are independent of $(y_0, \mathbf{x})$. For example, suppose z is a scalar determining eligibility in a job training program or some other social program. Actual participation, w, might be correlated with v_0, which could contain unobserved ability. If eligibility is randomly assigned, it is often reasonable to assume that z is independent of $(y_0, \mathbf{x})$. Eligibility would positively influence participation, and so Assumption ATE.2c should hold.

Random assignment of eligibility is no guarantee that eligibility is a valid instrument for participation. The outcome of z could affect other behavior, which could feed back into u_0 in equation (18.26). For example, consider Angrist's (1990) draft lottery application, where draft lottery number is used as an instrument for enlisting. Lottery number clearly affected enlistment, so Assumption ATE.2c is satisfied. Assumption ATE.2b is also satisfied *if* men did not change behavior in unobserved ways that affect wage, based on their lottery number. One concern is that men with low lottery numbers may get more education as a way of avoiding service through a deferment. Including years of education in $\mathbf{x}$ effectively solves this problem. But what if men with high draft lottery numbers received more job training because employers did not fear losing them? If a measure of job training status cannot be included in $\mathbf{x}$, lottery number would generally be correlated with u_0. See AIR and Heckman (1997) for additional discussion.

As the previous discussion implies, the redundancy condition in Assumption ATE.2b allows the instruments $\mathbf{z}$ to be correlated with elements of $\mathbf{x}$. For example, in the population of high school graduates, if w is a college degree indicator and the instrument z is distance to the nearest college while attending high school, then z is allowed to be correlated with other controls in the wage equation, such as geographic indicators.

Under $v_1 = v_0$ and the key assumptions on the instruments, 2SLS on equation (18.26) is consistent and asymptotically normal. But if we make stronger assumptions, we can find a more efficient IV estimator.

ASSUMPTION ATE.2′: (a) In equation (18.25), $v_1 = v_0$; (b) $\mathrm{E}(v_0 \mid \mathbf{x}, \mathbf{z}) = \mathrm{L}(v_0 \mid \mathbf{x})$; (c) $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z}) \neq \mathrm{P}(w = 1 \mid \mathbf{x})$ and $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \gamma)$ is a known parametric form (usually probit or logit); and (d) $\mathrm{Var}(v_0 \mid \mathbf{x}, \mathbf{z}) = \sigma_0^2$.

Part b assumes that $\mathrm{E}(v_0 \mid \mathbf{x})$ is linear in $\mathbf{x}$, and so it is more restrictive than Assumption ATE.2b. It does not usually hold for discrete response variables y, although it may be a reasonable approximation in some cases. Under parts a and b, the error u_0 in equation (18.26) has a zero conditional mean:

$$\mathrm{E}(u_0 \mid \mathbf{x}, \mathbf{z}) = 0 \tag{18.27}$$

Part d implies that $\mathrm{Var}(u_0 \mid \mathbf{x}, \mathbf{z})$ is constant. From the results on efficient choice of instruments in Section 14.5.3, the optimal IV for w is $\mathrm{E}(w \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \gamma)$. Therefore, we can use a two-step IV method:

Procedure 18.1 (Under Assumption ATE.2′): (a) Estimate the binary response model $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \gamma)$ by maximum likelihood. Obtain the fitted probabilities, $\hat{G}_i$. The leading case occurs when $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$ follows a probit model.

(b) Estimate equation (18.26) by IV using instruments 1, $\hat{G}_i$, and $\mathbf{x}_i$.

There are several nice features of this IV estimator. First, it can be shown that the conditions sufficient to ignore the estimation of γ in the first stage hold; see Section 6.1.2. Therefore, the usual 2SLS standard errors and test statistics are asymptotically valid. Second, under Assumption ATE.2′, the IV estimator from step b is asymptotically efficient in the class of estimators where the IVs are functions of $(\mathbf{x}_i, \mathbf{z}_i)$; see Problem 8.11. If Assumption ATE.2d does not hold, all statistics should be made robust to heteroskedasticity, and we no longer have the efficient IV estimator.

Procedure 18.1 has an important robustness property. Because we are using $\hat{G}_i$ as an instrument for w_i, the model for $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$ does *not* have to be correctly specified. For example, if we specify a probit model for $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$, we do not need the probit

model to be correct. Generally, what we need is that the linear projection of w onto $[\mathbf{x}, G(\mathbf{x}, \mathbf{z}; \gamma^*)]$ actually depends on $G(\mathbf{x}, \mathbf{z}; \gamma^*)$, where we use γ^* to denote the plim of the maximum likelihood estimator when the model is misspecified (see White, 1982a). These requirements are fairly weak when $\mathbf{z}$ is partially correlated with w.

Technically, α and $\boldsymbol{\beta}$ are identified even if we do not have extra exogenous variables excluded from $\mathbf{x}$. But we can rarely justify the estimator in this case. For concreteness, suppose that w given $\mathbf{x}$ follows a probit model [and we have no $\mathbf{z}$, or $\mathbf{z}$ does not appear in $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$]. Because $G(\mathbf{x}, \gamma) \equiv \Phi(\gamma_0 + \mathbf{x}\gamma_1)$ is a nonlinear function of $\mathbf{x}$, it is not perfectly correlated with $\mathbf{x}$, so it can be used as an IV for w. This situation is very similar to the one discussed in Section 17.4.1: while identification holds for all values of α and $\boldsymbol{\beta}$ if $\gamma_1 \neq \mathbf{0}$, we are achieving identification off of the nonlinearity of $\mathrm{P}(w = 1 \mid \mathbf{x})$. Further, $\Phi(\gamma_0 + \mathbf{x}\gamma_1)$ and $\mathbf{x}$ are typically highly correlated. As we discussed in Section 5.2.6, severe multicollinearity among the IVs can result in very imprecise IV estimators. In fact, if $\mathrm{P}(w = 1 \mid \mathbf{x})$ followed a linear probability model, α would not be identified. See Problem 18.5 for an illustration.

Example 18.3 (Estimating the Effects of Education on Fertility): We use the data in FERTIL2.RAW to estimate the effect of attaining at least seven years of education on fertility. The data are for women of childbearing age in Botswana. Seven years of education is, by far, the modal amount of positive education. (About 21 percent of women report zero years of education. For the subsample with positive education, about 33 percent report seven years of education.) Let $y = children$, the number of living children, and let $w = educ7$ be a binary indicator for at least seven years of education. The elements of $\mathbf{x}$ are *age*, age^2, *evermarr* (ever married), *urban* (lives in an urban area), *electric* (has electricity), and *tv* (has a television).

The OLS estimate of ATE is $-.394$ (se $= .050$). We also use the variable *frsthalf*, a binary variable equal to one if the woman was born in the first half of the year, as an IV for *educ7*. It is easily shown that *educ7* and *frsthalf* are significantly negatively related. The usual IV estimate is much larger in magnitude than the OLS estimate, but only marginally significant: -1.131 (se $= .619$). The estimate from Procedure 18.1 is even bigger in magnitude, and very significant: -1.975 (se $= .332$). The standard error that is robust to arbitrary heteroskedasticity is even smaller. Therefore, using the probit fitted values as an IV, rather than the usual linear projection, produces a more precise estimate (and one notably larger in magnitude).

The IV estimate of education effect seems very large. One possible problem is that, because *children* is a nonnegative integer that piles up at zero, the assumptions underlying Procedure 18.1—namely, Assumptions ATE.2′a and ATE.2′b—might not be met. In Chapter 19 we will discuss other methods for handling integer responses.

In principle, it is important to recognize that Procedure 18.1 is *not* the same as using $\hat{G}$ as a *regressor* in place of w. That is, IV estimation of equation (18.26) is *not* the same as the OLS estimator from

$$y_i \text{ on } 1, \hat{G}_i, \mathbf{x}_i \tag{18.28}$$

Consistency of the OLS estimators from regression (18.28) relies on having the model for $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$ correctly specified. If the first three parts of Assumption ATE.2′ hold, then

$$\mathrm{E}(y \mid \mathbf{x}, \mathbf{z}) = \delta_0 + \alpha G(\mathbf{x}, \mathbf{z}; \gamma) + \mathbf{x}\boldsymbol{\beta}$$

and, from the results on generated regressors in Chapter 6, the estimators from regression (18.28) are generally consistent. Procedure 18.1 is more robust because it does not require Assumption ATE.2′c for consistency.

Another problem with regression (18.28) is that the usual OLS standard errors and test statistics are not valid, for two reasons. First, if $\mathrm{Var}(u_0 \mid \mathbf{x}, \mathbf{z})$ is constant, $\mathrm{Var}(y \mid \mathbf{x}, \mathbf{z})$ cannot be constant because $\mathrm{Var}(w \mid \mathbf{x}, \mathbf{z})$ is not constant. By itself this is a minor nuisance because heteroskedasticity-robust standard errors and test statistics are easy to obtain. [However, it does call into question the efficiency of the estimator from regression (18.28).] A more serious problem is that the asymptotic variance of the estimator from regression (18.28) depends on the asymptotic variance of $\hat{\gamma}$ unless $\alpha = 0$, and the heteroskedasticity-robust standard errors do not correct for this.

In summary, using fitted probabilities from a first-stage binary response model, such as probit or logit, as an instrument for w is a nice way to exploit the binary nature of the endogenous explanatory variable. In addition, the asymptotic inference is always standard. Using $\hat{G}_i$ as an instrument does require the assumption that $\mathrm{E}(v_0 \mid \mathbf{x}, \mathbf{z})$ depends only on $\mathbf{x}$ and is linear in $\mathbf{x}$, which can be more restrictive than Assumption ATE.2b.

Allowing for the interaction $w(v_1 - v_0)$ in equation (18.25) is notably harder. In general, when $v_1 \neq v_0$, the IV estimator (using $\mathbf{z}$ or $\hat{G}$ as IVs for w) does not consistently estimate ATE (or ATE_1). Nevertheless, it is useful to find assumptions under which IV estimation does consistently estimate ATE. This problem has been studied by Angrist (1991), Heckman (1997), and Wooldridge (1997b), and we synthesize results from these papers.

Under the conditional mean redundancy assumptions

$$\mathrm{E}(v_0 \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(v_0 \mid \mathbf{x}) \qquad \text{and} \qquad \mathrm{E}(v_1 \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(v_1 \mid \mathbf{x}) \tag{18.29}$$

we can always write equation (18.25) as

$$y = \mu_0 + \alpha w + g_0(\mathbf{x}) + w[g_1(\mathbf{x}) - g_0(\mathbf{x})] + e_0 + w(e_1 - e_0) \tag{18.30}$$

where α is the ATE and

$$v_0 = g_0(\mathbf{x}) + e_0, \qquad \mathrm{E}(e_0 \mid \mathbf{x}, \mathbf{z}) = 0 \tag{18.31}$$

$$v_1 = g_1(\mathbf{x}) + e_1, \qquad \mathrm{E}(e_1 \mid \mathbf{x}, \mathbf{z}) = 0 \tag{18.32}$$

Given functional form assumptions for g_0 and g_1—which would typically be linear in parameters—we can estimate equation (18.30) by IV, where the error term is $e_0 + w(e_1 - e_0)$. For concreteness, suppose that

$$g_0(\mathbf{x}) = \eta_0 + \mathbf{x}\boldsymbol{\beta}_0, \qquad g_1(\mathbf{x}) - g_0(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} \tag{18.33}$$

where $\boldsymbol{\psi} = \mathrm{E}(\mathbf{x})$. If we plug these equations into equation (18.30), we need instruments for w *and* $w(\mathbf{x} - \boldsymbol{\psi})$ (note that $\mathbf{x}$ does not contain a constant here). If $q \equiv q(\mathbf{x}, \mathbf{z})$ is the instrument for w (such as the response probability in Procedure 18.1), the natural instrument for $w \cdot \mathbf{x}$ is $q \cdot \mathbf{x}$. (And, if q is the efficient IV for w, $q \cdot \mathbf{x}$ is the efficient instrument for $w \cdot \mathbf{x}$.) When will applying IV to

$$y = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + e_0 + w(e_1 - e_0) \tag{18.34}$$

be consistent? If the last term disappears, and, in particular, if

$$e_1 = e_0 \tag{18.35}$$

then the error e_0 has zero mean given $(\mathbf{x}, \mathbf{z})$; this result means that IV estimation of equation (18.34) produces consistent, asymptotically normal estimators.

ASSUMPTION ATE.3: With y expressed as in equation (18.25), conditions (18.29), (18.33), and (18.35) hold. In addition, Assumption ATE.2′c holds.

We have the following extension of Procedure 18.1:

Procedure 18.2 (Under Assumption ATE.3): (a) Same as Procedure 18.1.
(b) Estimate the equation

$$y_i = \gamma + \alpha w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + [w_i(\mathbf{x}_i - \bar{\mathbf{x}})]\boldsymbol{\delta} + error_i \tag{18.36}$$

by IV, using instruments 1, $\hat{G}_i$, $\mathbf{x}_i$, and $\hat{G}_i(\mathbf{x}_i - \bar{\mathbf{x}})$.

If we add Assumption ATE.2′d, Procedure 18.2 produces the efficient IV estimator [when we ignore estimation of $\mathrm{E}(\mathbf{x})$]. As with Procedure 18.1, we do not actually need the binary response model to be correctly specified for identification. As an alternative, we can use $\mathbf{z}_i$ and interactions between $\mathbf{z}_i$ and $\mathbf{x}_i$ as instruments, which generally results in testable overidentifying restrictions.

Technically, the fact that $\bar{\mathbf{x}}$ is an estimator of $\mathrm{E}(\mathbf{x})$ should be accounted for in computing the standard errors of the IV estimators. But, as shown in Problem 6.10,

the adjustments for estimating $\mathrm{E}(\mathbf{x})$ can be expected to have a trivial effect on the standard errors; in practice, we can just use the usual or heteroskedasticity-robust standard errors.

Example 18.4 (An IV Approach to Evaluating Job Training): To evaluate the effects of a job training program on subsequent wages, suppose that $\mathbf{x}$ includes education, experience, and the square of experience. If z indicates eligibility in the program, we would estimate the equation

$$\begin{aligned}\log(wage) = \mu_0 + \alpha\, jobtrain + \beta_{01} educ + \beta_{02} exper + \beta_{03} exper^2 \\ + \delta_1 jobtrain \cdot (educ - \overline{educ}) + \delta_2 jobtrain \cdot (exper - \overline{exper}) \\ + \delta_3 jobtrain \cdot (exper^2 - \overline{exper^2}) + error\end{aligned}$$

by IV, using instruments 1, z, *educ*, *exper*, $exper^2$, and interactions of z with all demeaned covariates. Notice that for the last interaction, we subtract off the average of $exper^2$. Alternatively, we could use in place of z the fitted values from a probit of *jobtrain* on $(\mathbf{x}, z)$.

Procedure 18.2 is easy to carry out, but its consistency generally hinges on condition (18.35), not to mention the functional form assumptions in equation (18.33). We can relax condition (18.35) to

$$\mathrm{E}[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] = \mathrm{E}[w(e_1 - e_0)] \tag{18.37}$$

We do *not* need $w(e_1 - e_0)$ to have zero mean, as a nonzero mean only affects the intercept. It is important to see that correlation between w and $(e_1 - e_0)$ does *not* invalidate the IV estimator of α from Procedure 18.2. However, we must assume that the covariance conditional on $(\mathbf{x}, \mathbf{z})$ is constant. Even if this assumption is not exactly true, it might be approximately true.

It is easy to see why, along with conditions (18.29) and (18.33), condition (18.37) implies consistency of the IV estimator. We can write equation (18.34) as

$$y = \xi + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + e_0 + r \tag{18.38}$$

where $r = w(e_1 - e_0) - \mathrm{E}[w(e_1 - e_0)]$ and $\xi = \gamma + \mathrm{E}[w(e_1 - e_0)]$. Under condition (18.37), $\mathrm{E}(r \mid \mathbf{x}, \mathbf{z}) = 0$, and so the composite error $e_0 + r$ has zero mean conditional on $(\mathbf{x}, \mathbf{z})$. Therefore, any function of $(\mathbf{x}, \mathbf{z})$ can be used as instruments in equation (18.38). Under the following modification of Assumption ATE.3, Procedure 18.2 is still consistent:

ASSUMPTION ATE.3′: With y expressed as in equation (18.25), conditions (18.29), (18.33), and (18.37) hold. In addition, Assumption ATE.2′c holds.

Even if Assumption ATE.2′d holds in addition to Assumption ATE.2′c, the IV estimator is generally not efficient because $\text{Var}(r \mid \mathbf{x}, \mathbf{z})$ would typically be heteroskedastic.

Angrist (1991) provided primitive conditions for assumption (18.37) in the case where $\mathbf{z}$ is independent of $(y_0, y_1, \mathbf{x})$. Then, the covariates can be dropped entirely from the analysis (leading to IV estimation of the simple regression equation $y = \xi + \alpha w + error$). We can extend those conditions here to allow $\mathbf{z}$ and $\mathbf{x}$ to be correlated. Assume that

$$\text{E}(w \mid \mathbf{x}, \mathbf{z}, e_1 - e_0) = h(\mathbf{x}, \mathbf{z}) + k(e_1 - e_0) \tag{18.39}$$

for some functions $h(\cdot)$ and $k(\cdot)$ and that

$$e_1 - e_0 \text{ is independent of } (\mathbf{x}, \mathbf{z}) \tag{18.40}$$

Under these two assumptions,

$$\begin{aligned} \text{E}[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] &= h(\mathbf{x}, \mathbf{z})\text{E}(e_1 - e_0 \mid \mathbf{x}, \mathbf{z}) + \text{E}[(e_1 - e_0)k(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] \\ &= h(\mathbf{x}, \mathbf{z}) \cdot 0 + \text{E}[(e_1 - e_0)k(e_1 - e_0)] \\ &= \text{E}[(e_1 - e_0)k(e_1 - e_0)] \end{aligned} \tag{18.41}$$

which is just an unconditional moment in the distribution of $e_1 - e_0$. We have used the fact that $\text{E}(e_1 - e_0 \mid \mathbf{x}, \mathbf{z}) = 0$ and that any function of $e_1 - e_0$ is independent of $(\mathbf{x}, \mathbf{z})$ under assumption (18.40). If we assume that $k(\cdot)$ is the identity function (as in Wooldridge, 1997b), then equation (18.41) is $\text{Var}(e_1 - e_0)$.

Assumption (18.40) is reasonable for continuously distributed responses, but it would not generally be reasonable when y is a discrete response or corner solution outcome. Further, even if assumption (18.40) holds, assumption (18.39) is violated when w given $\mathbf{x}$, $\mathbf{z}$, and $(e_1 - e_0)$ follows a standard binary response model. For example, a probit model would have

$$\text{P}(w = 1 \mid \mathbf{x}, \mathbf{z}, e_1 - e_0) = \Phi[\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + \rho(e_1 - e_0)] \tag{18.42}$$

which is not separable in $(\mathbf{x}, \mathbf{z})$ and $(e_1 - e_0)$. Nevertheless, assumption (18.39) might be a reasonable approximation in some cases. Without covariates, Angrist (1991) presents simulation evidence that suggests the simple IV estimator does quite well for estimating the ATE even when assumption (18.39) is violated.

Rather than assuming (18.39), different approaches are available, but they require different assumptions. We first consider a solution that involves adding a nonlinear function of $(\mathbf{x}, \mathbf{z})$ to equation (18.38) and estimating the resulting equation by 2SLS. We add to assumptions (18.40) and (18.42) a normality assumption,

$$e_1 - e_0 \sim \text{Normal}(0, \tau^2) \tag{18.43}$$

Under assumptions (18.40), (18.42), and (18.43) we can derive an estimating equation to show that *ATE* is usually identified.

To derive an estimating equation, note that conditions (18.40), (18.42), and (18.43) imply that

$$\text{P}(w = 1 \mid \mathbf{x}, \mathbf{z}) = \Phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2) \tag{18.44}$$

where each theta is the corresponding pi multiplied by $[1 + \rho^2\tau^2]^{-1/2}$. If we let a denote the latent error underlying equation (18.44) (with a standard normal distribution), and define $c \equiv e_1 - e_0$, then conditions (18.40), (18.42), and (18.43) imply that (a, c) has a zero-mean bivariate normal distribution that is independent of $(\mathbf{x}, \mathbf{z})$. Therefore, $\text{E}(c \mid a, \mathbf{x}, \mathbf{z}) = \text{E}(c \mid a) = \xi a$ for some parameter ξ, and

$$\text{E}(wc \mid \mathbf{x}, \mathbf{z}) = \text{E}[w\text{E}(c \mid a, \mathbf{x}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}] = \xi\text{E}(wa \mid \mathbf{x}, \mathbf{z}).$$

Using the fact that $a \sim \text{Normal}(0, 1)$ and is independent of $(\mathbf{x}, \mathbf{z})$, we have

$$\begin{aligned}\text{E}(wa \mid \mathbf{x}, \mathbf{z}) &= \int_{-\infty}^{\infty} 1[\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2 + a \geq 0]a\phi(a)\,da \\ &= \phi(-\{\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2\}) = \phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2)\end{aligned} \tag{18.45}$$

where $\phi(\cdot)$ is the standard normal density. Therefore, we can now write

$$y = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta} + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + \xi\phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2) + e_0 + r \tag{18.46}$$

where $r = wc - \text{E}(wc \mid \mathbf{x}, \mathbf{z})$. The composite error in (18.46) has zero mean conditional on $(\mathbf{x}, \mathbf{z})$, and so we can estimate the parameters using IV methods. One catch is the nonlinear function $\phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2)$. We could use nonlinear two stage least squares, as described in Chapter 14. But a two-step approach is easier. First, we gather together the assumptions:

ASSUMPTION ATE.4: With y written as in equation (18.25), maintain assumptions (18.29), (18.33), (18.40), (18.42) (with $\boldsymbol{\pi}_2 \neq \mathbf{0}$), and (18.43).

Procedure 18.3 (Under Assumption ATE.4): (a) Estimate θ_0, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ from a probit of w on $(1, \mathbf{x}, \mathbf{z})$. Form the predicted probabilities, $\hat{\Phi}_i$, along with $\hat{\phi}_i = \phi(\hat{\theta}_0 + \mathbf{x}_i\hat{\boldsymbol{\theta}}_1 + \mathbf{z}_i\hat{\boldsymbol{\theta}}_2)$, $i = 1, 2, \ldots, N$.

(b) Estimate the equation

$$y_i = \gamma + \alpha w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + \xi\hat{\phi}_i + error_i \tag{18.47}$$

by IV, using instruments $[1, \hat{\Phi}_i, \mathbf{x}_i, \hat{\Phi}_i(\mathbf{x}_i - \bar{\mathbf{x}}), \hat{\phi}_i]$.

The term $\hat{\phi}_i \equiv \phi(\hat{\theta}_0 + \mathbf{x}_i\hat{\theta}_1 + \mathbf{z}_i\hat{\theta}_2)$ in equation (18.47) is another example of a control function, although, unlike in Section 18.3, it is obtained from instrumental variables assumptions, rather than ignorability of treatment assumptions.

Even if $\xi \neq 0$, the effect of adding $\hat{\phi}_i$ to the estimate of α can be small. Consider the version of (18.46) without covariates $\mathbf{x}$ and with a scalar instrument, z:

$$y = \gamma + \alpha w + \xi\phi(\theta_0 + \theta_1 z) + u, \qquad \mathrm{E}(u \mid z) = 0 \tag{18.48}$$

This equation holds, for example, if the instrument z is independent of $(\mathbf{x}, v_0, v_1)$. The simple IV estimator of α is obtained by omitting $\phi(\theta_0 + \theta_1 z)$. If we use z as an IV for w, the simple IV estimator is consistent provided z and $\phi(\theta_0 + \theta_1 z)$ are uncorrelated. (Remember, having an omitted variable that is uncorrelated with the IV does not cause inconsistency of the IV estimator.) Even though $\phi(\theta_0 + \theta_1 z)$ is a function of z, these two variables might have small correlation because z is monotic while $\phi(\theta_0 + \theta_1 z)$ is symmetric about $-(\theta_0/\theta_1)$. This discussion shows that condition (18.37) is not necessary for IV to consistently estimate the ATE: It could be that while $\mathrm{E}[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}]$ is not constant, it is roughly uncorrelated with $\mathbf{x}$ (or the functions of $\mathbf{x}$) that appear in (18.38), as well as with the functions of $\mathbf{z}$ used as instruments.

Equation (18.48) illustrates another important point: If $\xi \neq 0$ and the single instrument z is binary, α is not identified. Lack of identification occurs because $\phi(\theta_0 + \theta_1 z)$ takes on only two values, which means it is perfectly linearly related to z. So long as z takes on more than two values, α is generally identified, although the identification is due to the fact that $\phi(\cdot)$ is a different nonlinear function than $\Phi(\cdot)$. With $\mathbf{x}$ in the model $\hat{\phi}_i$ and $\hat{\Phi}_i$ might be collinear, resulting in imprecise IV estimates.

Because r in (18.46) is heteroskedastic, the instruments below (18.47) are not optimal, and so we might simply use $\mathbf{z}_i$ along with interactions of $\mathbf{z}_i$ with $(\mathbf{x}_i - \bar{\mathbf{x}})$ and $\hat{\phi}_i$ as IVs. If $\mathbf{z}_i$ has dimension greater than one, then we can test the overidentifying restrictions as a partial test of instrument selection and the normality assumptions. Of course, we could use the results of Chapter 14 to characterize and estimate the optimal instruments, but this is fairly involved [see, for example, Newey and McFadden (1994)].

A different approach to estimating the ATE when assumption (18.39) fails is to compute the expected value of y given the endogenous treatment and all exogenous variables: $\mathrm{E}(y \mid w, \mathbf{x}, \mathbf{z})$. Finding this expectation requires somewhat more by way of assumptions, but it also has some advantages, which we discuss later. For completeness, we list a set of assumptions:

ASSUMPTION ATE.4′: With y written as in equation (18.25), maintain assumptions (18.29) and (18.33). Furthermore, the treatment can be written as $w = 1[\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 +$

$\mathbf{z}\boldsymbol{\theta}_2 + a \geq 0]$, where (a, e_0, e_1) is independent of $(\mathbf{x}, \mathbf{z})$ with a trivariate normal distribution; in particular, $a \sim \text{Normal}(0, 1)$.

Under Assumption ATE.4′, we can use calculations very similar to those used in Section 17.4.1 to obtain $\text{E}(y \mid w, \mathbf{x}, \mathbf{z})$. In particular,

$$\text{E}(y \mid w, \mathbf{x}, \mathbf{z}) = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + \rho_1 w[\phi(\mathbf{q}\boldsymbol{\theta})/\Phi(\mathbf{q}\boldsymbol{\theta})] + \rho_2(1 - w)\{\phi(\mathbf{q}\boldsymbol{\theta})/[1 - \Phi(\mathbf{q}\boldsymbol{\theta})]\} \tag{18.49}$$

where $\mathbf{q}\boldsymbol{\theta} \equiv \theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2$ and ρ_1 and ρ_2 are additional parameters. Heckman (1978) used this expectation to obtain two-step estimators of the switching regression model. [See Vella and Verbeek (1999) for a recent discussion of the switching regression model in the context of treatment effects.] Not surprisingly, (18.49) suggests a simple two-step procedure, where the first step is identical to that in Procedure 18.3:

Procedure 18.4 (Under Assumption ATE.4′): (a) Estimate θ_0, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ from a probit of w on $(1, \mathbf{x}, \mathbf{z})$. Form the predicted probabilities, $\hat{\Phi}_i$, along with $\hat{\phi}_i = \phi(\hat{\theta}_0 + \mathbf{x}_i\hat{\boldsymbol{\theta}}_1 + \mathbf{z}_i\hat{\boldsymbol{\theta}}_2)$, $i = 1, 2, \ldots, N$.

(b) Run the OLS regression

$$y_i \text{ on } 1, w_i, \mathbf{x}_i, w_i(\mathbf{x}_i - \bar{\mathbf{x}}), w_i(\hat{\phi}_i/\hat{\Phi}_i), (1 - w_i)[\hat{\phi}_i/(1 - \hat{\Phi}_i)] \tag{18.50}$$

using all of the observations. The coefficient on w_i is a consistent estimator of α, the ATE.

When we restrict attention to the $w_i = 1$ subsample, thereby dropping w_i and $w_i(\mathbf{x}_i - \bar{\mathbf{x}})$, we obtain the sample selection correction from Section 17.4.1; see equation (17.24). (The treatment w_i becomes the sample selection indicator.) But the goal of sample selection corrections is very different from estimating an average treatment effect. For the sample selection problem, the goal is to estimate $\boldsymbol{\beta}_0$, which indexes $\text{E}(y \mid \mathbf{x})$ in the population. By contrast, in estimating an ATE we are interested in the causal effect that w has on y.

It makes sense to check for joint significance of the last two regressors in regression (18.50) as a test of endogeneity of w. Because the coefficients ρ_1 and ρ_2 are zero under H_0, we can use the results from Chapter 6 to justify the usual Wald test (perhaps made robust to heteroskedasticity). If these terms are jointly insignificant at a sufficiently high level, we can justify the usual OLS regression without unobserved heterogeneity. If we reject H_0, we must deal with the generated regressors problem in obtaining a valid standard error for $\hat{\alpha}$.

Technically, Procedure 18.3 is more robust than Procedure 18.4 because the former does not require a trivariate normality assumption. Linear conditional expectations,

along with the assumption that w given $(\mathbf{x}, \mathbf{z})$ follows a probit, suffice. In addition, Procedure 18.3 allows us to separate the issues of endogeneity of w and nonconstant treatment effect: if we ignore the estimation error involved with demeaning $\mathbf{x}_i$ in the interaction term—which generally seems reasonable—then a standard t-test (perhaps made robut to heteroskedasticity) for H_0: $\xi = 0$ is valid for testing the presence of $w(e_1 - e_0)$, even when w is endogenous.

Practically, the extra assumption in Procedure 18.4 is that e_0 is independent of $(\mathbf{x}, \mathbf{z})$ with a normal distribution. We may be willing to make this assumption, especially if the estimates from Procedure 18.3 are too imprecise to be useful. The efficiency issue is a difficult one because of the two-step estimation involved, but, intuitively, Procedure 18.4 is likely to be more efficient because it is based on $E(y \mid w, \mathbf{x}, \mathbf{z})$. Procedure 18.3 involves replacing the unobserved composite error with its expectation conditional only on $(\mathbf{x}, \mathbf{z})$. In at least one case, Procedure 18.4 gives results when Procedure 18.3 cannot: when $\mathbf{x}$ is not in the equation and there is a single binary instrument.

Under a variant of Assumption ATE.3′, we can consistently estimate ATE_1 by IV. As before, we express y as in equation (18.25). First, we show how to consistently estimate $ATE_1(\mathbf{x})$, which can be written as

$$ATE_1(\mathbf{x}) = E(y_1 - y_0 \mid \mathbf{x}, w = 1) = (\mu_1 - \mu_0) + E(v_1 - v_0 \mid \mathbf{x}, w = 1)$$

The following assumption identifies $ATE_1(\mathbf{x})$:

ASSUMPTION ATE.3″: (a) With y expressed as in equation (18.25), the first part of assumption (18.29) holds, that is, $E(v_0 \mid \mathbf{x}, \mathbf{z}) = E(v_0 \mid \mathbf{x})$; (b) $E(v_1 - v_0 \mid \mathbf{x}, \mathbf{z}, w = 1) = E(v_1 - v_0 \mid \mathbf{x}, w = 1)$; and (c) Assumption ATE.2′c holds.

We discussed part a of this assumption earlier, as it also appears in Assumption ATE.3′. It can be violated if agents change their behavior based on $\mathbf{z}$. Part b deserves some discussion. Recall that $v_1 - v_0$ is the person-specific gain from participation or treatment. Assumption ATE.3″ requires that for those in the treatment group, the gain is not predictable given $\mathbf{z}$, once $\mathbf{x}$ is controlled for. Heckman (1997) discusses Angrist's (1990) draft lottery example, where z (a scalar) is draft lottery number. Men who had a large z were virtually certain to escape the draft. But some men with large draft numbers chose to serve anyway. Even with good controls in $\mathbf{x}$, it seems plausible that, for those who chose to serve, a higher z is associated with a higher gain to military service. In other words, for those who chose to serve, $v_1 - v_0$ and z are positively correlated, even after controlling for $\mathbf{x}$. This argument directly applies to estimation of ATE_1; the effect on estimation of ATE is less clear.

Assumption ATE.3″b is plausible when z is a binary indicator for eligibility in a program, which is randomly determined and does not induce changes in behavior other than whether or not to participate.

To see how Assumption ATE.3″ identifies $ATE_1(\mathbf{x})$, rewrite equation (18.25) as

$$\begin{aligned} y &= \mu_0 + g_0(\mathbf{x}) + w[(\mu_1 - \mu_0) + \mathrm{E}(v_1 - v_0 \mid \mathbf{x}, w = 1)] \\ &\quad + w[(v_1 - v_0) - \mathrm{E}(v_1 - v_0 \mid \mathbf{x}, w = 1)] + e_0 \\ &= \mu_0 + g_0(\mathbf{x}) + w \cdot ATE_1(\mathbf{x}) + a + e_0 \end{aligned} \tag{18.51}$$

where $a \equiv w[(v_1 - v_0) - \mathrm{E}(v_1 - v_0 \mid \mathbf{x}, w = 1)]$ and e_0 is defined in equation (18.31). Under Assumption ATE.3″a, $\mathrm{E}(e_0 \mid \mathbf{x}, \mathbf{z}) = 0$. The hard part is dealing with the term a. When $w = 0$, $a = 0$. Therefore, to show that $\mathrm{E}(a \mid \mathbf{x}, \mathbf{z}) = 0$, it suffices to show that $\mathrm{E}(a \mid \mathbf{x}, \mathbf{z}, w = 1) = 0$. [Remember, $\mathrm{E}(a \mid \mathbf{x}, \mathbf{z}) = \mathrm{P}(w = 0) \cdot \mathrm{E}(a \mid \mathbf{x}, \mathbf{z}, w = 0) + \mathrm{P}(w = 1) \cdot \mathrm{E}(a \mid \mathbf{x}, \mathbf{z}, w = 1)$.] But this result follows under Assumption ATE.3″b:

$$\mathrm{E}(a \mid \mathbf{x}, \mathbf{z}, w = 1) = \mathrm{E}(v_1 - v_0 \mid \mathbf{x}, \mathbf{z}, w = 1) - \mathrm{E}(v_1 - v_0 \mid \mathbf{x}, w = 1) = 0$$

Now, letting $r \equiv a + e_0$ and assuming that $g_0(\mathbf{x}) = \eta_0 + \mathbf{h}(\mathbf{x})\boldsymbol{\beta}_0$ and $ATE_1(\mathbf{x}) = \tau + \mathbf{f}(\mathbf{x})\boldsymbol{\delta}$ for some row vector of functions $\mathbf{h}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$, we can write

$$y = \gamma_0 + \mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0 + \tau w + [w \cdot \mathbf{f}(\mathbf{x})]\boldsymbol{\delta} + r, \qquad \mathrm{E}(r \mid \mathbf{x}, \mathbf{z}) = 0$$

All the parameters of this equation can be consistently estimated by IV, using any functions of $(\mathbf{x}, \mathbf{z})$ as IVs. [These would include include 1, $\mathbf{h}_0(\mathbf{x})$, $G(\mathbf{x}, \mathbf{z}; \hat{\gamma})$—the fitted treatment probabilities—and $G(\mathbf{x}, \mathbf{z}; \hat{\gamma}) \cdot \mathbf{f}(\mathbf{x})$.] The average treatment effect on the treated for any $\mathbf{x}$ is estimated as $\hat{\tau} + \mathbf{f}(\mathbf{x})\hat{\boldsymbol{\delta}}$. Averaging over the observations with $w_i = 1$ gives a consistent estimator of ATE_1.

18.4.2 Estimating the Local Average Treatment Effect by IV

We now discuss estimation of an evaluation parameter introduced by Imbens and Angrist (1994), the local average treatment effect (LATE), in the simplest possible setting. This requires a slightly more complicated notation. (More general cases require even more complicated notation, as in AIR.) As before, we let w be the observed treatment indicator (taking on zero or one), and let the counterfactual outcomes be y_1 with treatment and y_0 without treatment. The observed outcome y can be written as in equation (18.3).

To define $LATE$, we need to have an instrumental variable, z. In the simplest case z is a binary variable, and we focus attention on that case here. For each unit i in a random draw from the population, z_i is zero or one. Associated with the two possible outcomes on z are counterfactual treatments, w_0 and w_1. These are the treatment

statuses we would observe if $z = 0$ and $z = 1$, respectively. For each unit, we observe only one of these. For example, z can denote whether a person is eligible for a particular program, while w denotes actual participation in the program.

Write the observed treatment status as

$$w = (1 - z)w_0 + zw_1 = w_0 + z(w_1 - w_0) \tag{18.52}$$

When we plug this equation into $y = y_0 + w(y_1 - y_0)$ we get

$$y = y_0 + w_0(y_1 - y_0) + z(w_1 - w_0)(y_1 - y_0)$$

A key assumption is

$$z \text{ is independent of } (y_0, y_1, w_0, w_1) \tag{18.53}$$

Under assumption (18.53), all expectations involving functions of (y_0, y_1, w_0, w_1), conditional on z, do not depend on z. Therefore,

$$\mathrm{E}(y \mid z = 1) = \mathrm{E}(y_0) + \mathrm{E}[w_0(y_1 - y_0)] + \mathrm{E}[(w_1 - w_0)(y_1 - y_0)]$$

and

$$\mathrm{E}(y \mid z = 0) = \mathrm{E}(y_0) + \mathrm{E}[w_0(y_1 - y_0)]$$

Subtracting the second equation from the first gives

$$\mathrm{E}(y \mid z = 1) - \mathrm{E}(y \mid z = 0) = \mathrm{E}[(w_1 - w_0)(y_1 - y_0)] \tag{18.54}$$

which can be written [see equation (2.49)] as

$$\begin{aligned} 1 \cdot \mathrm{E}(y_1 - y_0 \mid w_1 - w_0 = 1)\mathrm{P}(w_1 - w_0 = 1) \\ + (-1)\mathrm{E}(y_1 - y_0 \mid w_1 - w_0 = -1)\mathrm{P}(w_1 - w_0 = -1) \\ + 0 \cdot \mathrm{E}(y_1 - y_0 \mid w_1 - w_0 = 0)\mathrm{P}(w_1 - w_0 = 0) \\ = \mathrm{E}(y_1 - y_0 \mid w_1 - w_0 = 1)\mathrm{P}(w_1 - w_0 = 1) \\ - \mathrm{E}(y_1 - y_0 \mid w_1 - w_0 = -1)\mathrm{P}(w_1 - w_0 = -1) \end{aligned}$$

To get further, we introduce another important assumption, called *monotonicity* by Imbens and Angrist:

$$w_1 \geq w_0 \tag{18.55}$$

In other words, we are ruling out $w_1 = 0$ and $w_0 = 1$. This assumption has a simple interpretation when z is a dummy variable representing assignment to the treatment group: anyone in the population who would be in the treatment group in the absence

of assignment (or eligibility) would be in the treatment group if assigned to the treatment group. Units of the population who do not satisfy monotonicity are called *defiers*. In many applications, this assumption seems very reasonable. For example, if z denotes randomly assigned eligibility in a job training program, assumption (18.55) simply requires that people who would participate without being eligible would also participate if eligible.

Under assumption (18.55), $P(w_1 - w_0 = -1) = 0$, so assumptions (18.53) and (18.55) imply

$$E(y \mid z = 1) - E(y \mid z = 0) = E(y_1 - y_0 \mid w_1 - w_0 = 1)P(w_1 - w_0 = 1) \tag{18.56}$$

In this setup, Imbens and Angrist (1994) define *LATE* to be

$$LATE = E(y_1 - y_0 \mid w_1 - w_0 = 1) \tag{18.57}$$

Because $w_1 - w_0 = 1$ is equivalent to $w_1 = 1$, $w_0 = 0$, *LATE* has the following interpretation: it is the average treatment effect for those who would be induced to participate by changing z from zero to one. There are two things about *LATE* that make it different from the other treatment parameters. First, it depends on the instrument, z. If we use a different instrument, then *LATE* generally changes. The parameters *ATE* and ATE_1 are defined without reference to an IV, but only with reference to a population. Second, because we cannot observe both w_1 and w_0, we cannot identify the subpopulation with $w_1 - w_0 = 1$. By contrast, *ATE* averages over the entire population, while ATE_1 is the average for those who are actually treated.

Example 18.5 (LATE for Attending a Catholic High School): Suppose that y is a standardized test score, w is an indicator for attending a Catholic high school, and z is an indicator for whether the student is Catholic. Then, generally, *LATE* is the mean effect on test scores for those individuals who choose a Catholic high school because they are Catholic. Evans and Schwab (1995) use a high school graduation indicator for y, and they estimate a probit model with an endogenous binary explanatory variable, as described in Section 15.7.3. Under the probit assumptions, it is possible to estimate *ATE*, whereas the simple IV estimator identifies *LATE* under weaker assumptions.

Because $E(y \mid z = 1)$ and $E(y \mid z = 0)$ are easily estimated using a random sample, *LATE* is identified if $P(w_1 - w_0 = 1)$ is estimable *and* nonzero. Importantly, from the monotonicity assumption, $w_1 - w_0$ is a binary variable because $P(w_1 - w_0 = -1) = 0$. Therefore,

$$\begin{aligned} P(w_1 - w_0 = 1) &= E(w_1 - w_0) = E(w_1) - E(w_0) = E(w \mid z = 1) - E(w \mid z = 0) \\ &= P(w = 1 \mid z = 1) - P(w = 1 \mid z = 0) \end{aligned}$$

where the second-to-last equality follows from equations (18.52) and (18.53). Each conditional probability can be consistently estimated given a random sample on (w, z). Therefore, the final assumption is

$$\mathrm{P}(w = 1 \mid z = 1) \neq \mathrm{P}(w = 1 \mid z = 0) \tag{18.58}$$

To summarize, under assumptions (18.53), (18.55), and (18.58),

$$LATE = [\mathrm{E}(y \mid z = 1) - \mathrm{E}(y \mid z = 0)]/[\mathrm{P}(w = 1 \mid z = 1) - \mathrm{P}(w = 1 \mid z = 0)] \tag{18.59}$$

Therefore, a consistent estimator is $\hat{LATE} = (\bar{y}_1 - \bar{y}_0)/(\bar{w}_1 - \bar{w}_0)$, where $\bar{y}_1$ is the sample average of y_i over that part of the sample where $z_i = 1$ and $\bar{y}_0$ is the sample average over $z_i = 0$, and similarly for $\bar{w}_1$ and $\bar{w}_0$ (which are sample proportions). From Problem 5.13b, we know that $\hat{LATE}$ is identical to the IV estimator of α in the simple equation $y = \delta_0 + \alpha w + error$, where z is the IV for w.

Our conclusion is that, in the simple case of a binary instrument for the binary treatment, the usual IV estimator consistently estimates $LATE$ under weak assumptions. See Angrist, Imbens, and Rubin (1996) and the discussants' comments for much more.

18.5 Further Issues

As we have seen in Sections 18.3 and 18.4, under certain assumptions, OLS or IV can be used to estimate average treatment effects. Therefore, at least in some cases, problems such as attrition or other forms of sample selection can be easily handled using the methods in Chapter 17. For example, in equation (18.34) under assumption (18.35), it is reasonable to assume that the assumptions of Procedure 17.2 hold, with the straightforward extension that the interaction between w_i (which plays the role of y_{i2}) and $(\mathbf{x}_i - \bar{\mathbf{x}})$ is added, along with the appropriate IVs. If the problem is attrition, we need some exogenous elements that affect attrition but do not appear in $\mathbf{x}$ or $\mathbf{z}$.

Other situations may require special attention, and we now briefly discuss some of these.

18.5.1 Special Considerations for Binary and Corner Solution Responses

The definitions of ATE and ATE_1, as well as $LATE$, are valid for any kind of response variable. ATE is simply $\mathrm{E}(y_1) - \mathrm{E}(y_0)$, and for this to be well defined we only need to assume that the expected values exist. If y_0 and y_1 are binary—such as employment indicators—the expected values are probabilities of success. If y_0 and y_1 are corner solution outcomes—such as labor supply—ATE and ATE_1 estimate the effect of treatment on the so-called unconditional expectation rather than, say,

$E(y_1 - y_0 \mid y_0 > 0)$. Still, ATE and ATE_1 are often of interest for corner solution outcomes.

If the average treatment effects as we defined them in Section 18.2 are still of interest, why do we need to consider alternative methods for estimating treatment effects? The answer is that some of the assumptions we have discussed are unrealistic for discrete or corner solution outcomes. For example, to arrive at equation (18.15), we assumed that $E(y_g \mid \mathbf{x})$ is linear in some functions of $\mathbf{x}$. Though these assumptions can be relaxed, the computation of valid standard errors is no longer straightforward because a linear regression no longer generally estimates ATE. [Expression (18.7) is general and does not impose any functional forms, and so it can be used as the basis for estimating ATE. We would simply estimate $E(y \mid \mathbf{x}, w = 1)$ and $E(y \mid \mathbf{x}, w = 0)$ in a way that is consistent with the features of y.]

Under ignorability of treatment, the propensity score approach is attractive because it requires no modeling of expectations involving y_0 or y_1. Only the propensity score needs to be modeled, and this is always a binary response probability.

When we cannot assume ignorability of treatment and must resort to IV methods, allowing for discrete and corner solution responses is theoretically harder. As we discussed in Section 18.4.1, conditions such as equation (18.33) cannot be literally true for binary and Tobit-like responses, and this condition appears in all of the assumptions for IV estimation. It is not easy to relax this assumption because if, say, y_g is a corner solution outcome, a reasonable model for $E(v_1 - v_0 \mid \mathbf{x})$ is not obvious.

Of course, it could be that, even if the assumptions in Section 18.4 cannot be exactly true, the IV methods may nevertheless produce reasonable estimates of ATE and ATE_1. Angrist's (1991) simulation evidence is compelling for binary responses, but he only studies the case without covariates.

As an alternative to the various treatment effect estimators covered in this chapter, we can use probit and Tobit models with a binary endogenous explanatory variable. The maximum likelihood estimator described in Section 15.7.3 requires a strong set of assumptions, but it delivers estimates of the exact average treatment effect, conditional on the exogenous variables, if the probit assumptions hold. Similarly, a Tobit model with a binary endogenous variable can be estimated by maximum likelihood (see Problem 16.6); again, estimates of ATE can be obtained directly.

18.5.2 Panel Data

The availability of panel data allows us to consistently estimate treatment effects without assuming ignorability of treatment and without an instrumental variable, provided the treatment varies over time and is uncorrelated with time-varying unobservables that affect the response.

If the treatment is assumed to have the same effect for each unit and if the effect is constant over time, fixed effects or first-differencing methods can be used, as described in Chapter 10. This approach works well when the treatment and control groups are designated based on time-constant variables and when treatment status is not constant across time. Of course, we must observe the responses and other controls for each cross section unit in at least two different time periods. A more complicated model allows the treatment effect to interact with observable variables and unobserved heterogeneity. For example, consider the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_1 w_{it} + c_i + w_{it}h_i + u_{it}$$

where w_{it} is a binary treatment indicator of some training program, y_{it} is the response variable, and c_i and h_i are unobserved heterogeneity. This is a special case of the model studied in Section 11.2.2. The average treatment effect is $\alpha_1 + \mathrm{E}(h_i)$, and we can use the methods of Section 11.2.2 to estimate α_1 and $\mathrm{E}(h_i)$.

The problem of attrition can be handled as in Section 17.7, provided the treatment effect has an additive form. If attrition is determined solely by whether the participant was not selected for the program, then no adjustments are needed if w_{it} is orthogonal to the idiosyncratic error, u_{it}: this is just attrition on the basis of exogenous explanatory variables.

18.5.3 Nonbinary Treatments

So far, we have restricted attention to the case where w is a binary variable. But we can also estimate average treatment effects when w takes on more than two values. The definitions of ATE, ATE_1, and $LATE$ are more complicated in this case because the counterfactual is more complicated; see Angrist and Imbens (1995) and Heckman (1997). Here, we focus on a **random coefficient model** for the observed outcome, as in Garen (1984), Heckman and Vytlacil (1998), and Wooldridge (1997b, 1999c). The average treatment effect is easy to define in this context, as it is just an average partial effect.

As in the case of binary treatment, two approaches can be used to identify ATE: we can assume ignorability of treatment, conditional on a set of covariates, or we can use an instrumental variables approach. In either case, the model is the same:

$$\mathrm{E}(y \mid w, \mathbf{c}) = a + bw \tag{18.60}$$

where $\mathbf{c} = (a, b)$ and a and b may both depend on observable covariates as well as unobserved heterogeneity. A more traditional approach would introduce observables and unobservables into the equation separately in a parametric fashion—usually,

linear in a set of parameters—but this step is unnecessary when we are interested in estimating $\beta \equiv \mathrm{E}(b)$, which is the average partial effect of w on $\mathrm{E}(y \mid w, \mathbf{c})$.

It is important to see that, unlike in the binary treatment case, equation (18.60) imposes a functional form assumption. This is not as restrictive as it might seem, because a and b are allowed to depend on individual-specific observables and unobservables. Nevertheless, as we know, linear models can have drawbacks for binary and corner solution responses (unless w is binary).

When w is binary, equation (18.60) encompasses the counterfactual setup in Section 18.2, which we analyzed in Sections 18.3 and 18.4. Equation (18.3) shows this result immediately, where we take $a = y_0$ and $b = y_1 - y_0$.

We now establish identification of β under ignorability conditions. The assumptions are collected together as follows:

ASSUMPTION ATE.5: (a) Equation (18.60) holds. For a set of covariates $\mathbf{x}$, the following redundancy assumptions hold: (b) $\mathrm{E}(y \mid w, \mathbf{c}, \mathbf{x}) = \mathrm{E}(y \mid w, \mathbf{c})$; and (c) Conditional on $\mathbf{x}$, $\mathbf{c}$ is redundant in the first two conditional moments of w: $\mathrm{E}(w \mid \mathbf{x}, \mathbf{c}) = \mathrm{E}(w \mid \mathbf{x})$ and $\mathrm{Var}(w \mid \mathbf{x}, \mathbf{c}) = \mathrm{Var}(w \mid \mathbf{x})$.

Given the functional form assumption (18.60), Assumption ATE.5b is not very controversial because a and b can depend in an arbitrary way on $\mathbf{x}$. In effect, a and b already capture any dependence of $\mathrm{E}(y \mid w, \mathbf{c})$ on $\mathbf{x}$. Assumption ATE.5c is much more restrictive, but it is the analogue of the ignorability-of-treatment Assumption ATE.1′. In fact, when w is binary, Assumption ATE.1′ implies Assumption ATE.5c. For general w, Assumption ATE.5c is slightly less restrictive than assuming $(w, \mathbf{c})$ are independent given $\mathbf{x}$. The following is from Wooldridge (1999c, Proposition 3.1):

PROPOSITION 18.6: Under Assumption ATE.5, assume, in addition, that $\mathrm{Var}(w \mid \mathbf{x}) > 0$ for all $\mathbf{x}$ in the relevant population. Then

$$\beta = \mathrm{E}[\mathrm{Cov}(w, y \mid \mathbf{x})/\mathrm{Var}(w \mid \mathbf{x})] \tag{18.61}$$

Because $\mathrm{Var}(w \mid \mathbf{x})$ and $\mathrm{Cov}(w, y \mid \mathbf{x})$ can be estimated generally, equation (18.61) shows that β is identified. If $\hat{m}(\cdot)$ and $\hat{h}(\cdot)$ are consistent estimators of $\mathrm{E}(w \mid \mathbf{x})$ and $\mathrm{Var}(w \mid \mathbf{x})$, respectively, a consistent estimator of β, under fairly weak assumptions, is $N^{-1} \sum_{i=1}^{N} [w_i - \hat{m}(\mathbf{x}_i)] y_i / \hat{h}(\mathbf{x}_i)$; this is the extension of equation (18.21) to the case of nonbinary treatments.

Estimating $m(\cdot)$ and $h(\cdot)$ is easily done using flexible parametric models that should reflect the nature of w. When w is binary, we simply estimate the propensity score. When w is a roughly continuous variable over a broad range, $\mathrm{E}(w \mid \mathbf{x})$ linear in

functions of $\mathbf{x}$ and $\text{Var}(w \mid \mathbf{x})$ constant might be reasonable, in which case, as shown in Wooldridge (1999c), a "kitchen sink" regression of y on w and functions of $\mathbf{x}$ can consistently estimate β. Wooldridge (1999c) discusses additional examples, including when w is a count variable or a fractional variable (both of which we discuss in Chapter 19), and contains an example.

As a computational device, it is useful to see that consistent estimators of β can be computed using an instrumental variables approach. As shown in Wooldridge (1999c), under Assumption ATE.5 we can write

$$y = \beta w + \mathbf{g}(\mathbf{x})\boldsymbol{\theta} + v \tag{18.62}$$

where $\mathbf{g}(\cdot)$ is *any* vector function of $\mathbf{x}$ and $\text{E}[\mathbf{g}(\mathbf{x})'v] = \mathbf{0}$. Typically, we would include levels, squares, and cross products, or logarithms, as elements of $\mathbf{g}(\cdot)$. Adding $\mathbf{g}(\mathbf{x})$ is intended to effectively reduce the error variance. Further, if we define $r = [w - m(\mathbf{x})]/h(\mathbf{x})$, it can be shown that $\text{E}(rv) = 0$. Because r and w are highly correlated, we can use $[r, \mathbf{g}(\mathbf{x})]$ as IVs in equation (18.62). In practice, we replace each unknown r_i with $\hat{r}_i = [w_i - \hat{m}(\mathbf{x}_i)]/\hat{h}(\mathbf{x}_i)$, and use $(\hat{r}_i, \mathbf{g}_i)$ as the IVs for $(w_i, \mathbf{g}_i)$. This estimator is consistent and $\sqrt{N}$-asymptotically normal. Unfortunately, the sufficient conditions for ignoring estimation of the IVs in the first stage—see Section 6.1.2—are not always met in this application.

An alternative approach assumes that w is ignorable in $\text{E}(a \mid \mathbf{x}, w)$ and $\text{E}(b \mid \mathbf{x}, w)$. Under additional linearity assumptions, this leads directly to equation (18.15), regardless of the nature of w.

The previous methods assume some kind of ignorability of treatment. The IV approach also begins with equation (18.60). As in the binary treatment case, we separate the covariates ($\mathbf{x}$) and the IVs ($\mathbf{z}$). We assume both are redundant in equation (18.60):

$$\text{E}(y \mid w, \mathbf{c}, \mathbf{x}, \mathbf{z}) = \text{E}(y \mid w, \mathbf{c}) \tag{18.63}$$

Again, this assumption is noncontroversial once we specify the functional form in equation (18.60). Assumption (18.63) holds trivially in the counterfactual framework with binary treatment.

The difference between $\mathbf{x}$ and $\mathbf{z}$ is that a and b may have conditional means that depend on $\mathbf{x}$, but not on $\mathbf{z}$. For example, if w is a measure of class attendance, $\mathbf{x}$ might contain measures of student ability and motivation. By contrast, we assume $\mathbf{z}$ is redundant for explaining a and b, given $\mathbf{x}$. In the class attendance example, $\mathbf{z}$ might be indicators for different living situations or distances from residence to lecture halls. In effect, the distinction between $\mathbf{x}$ and $\mathbf{z}$ is the kind of distinction we make in struc-

tural models, where some "exogenous" variables ($\mathbf{x}$) are allowed to appear in the structural equation and others ($\mathbf{z}$) are not. Mathematically, we have

$$\mathrm{E}(a \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(a \mid \mathbf{x}), \qquad \mathrm{E}(b \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(b \mid \mathbf{x}). \tag{18.64}$$

Assumption (18.64) is completely analogous to assumption (18.29). For simplicity, we also assume the expectations are linear in $\mathbf{x}$:

$$\mathrm{E}(a \mid \mathbf{x}) = \gamma_0 + \mathbf{x}\gamma, \qquad \mathrm{E}(b \mid \mathbf{x}) = \delta_0 + \mathbf{x}\boldsymbol{\delta} \tag{18.65}$$

Then we can write

$$y = \eta_0 + \mathbf{x}\gamma + \beta w + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + u + w \cdot v + e \tag{18.66}$$

where $\boldsymbol{\psi} = \mathrm{E}(\mathbf{x})$, $u = a - \mathrm{E}(a \mid \mathbf{x}, \mathbf{z})$, $v = b - \mathrm{E}(b \mid \mathbf{x}, \mathbf{z})$, e is the error implied by equation (18.60), and so $\mathrm{E}(e \mid w, \mathbf{c}, \mathbf{x}, \mathbf{z}) = 0$. This equation is basically the same as equation (18.34), except that now w need not be binary. To apply IV to equation (18.66), it suffices that the composite error, $u + w \cdot v + e$, has a constant mean given $(\mathbf{x}, \mathbf{z})$. But $\mathrm{E}(u + e \mid \mathbf{x}, \mathbf{z}) = 0$, and so it suffices to assume

$$\mathrm{E}(w \cdot v \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(w \cdot v) \tag{18.67}$$

which is the same as $\mathrm{Cov}(w, v \mid \mathbf{x}, \mathbf{z}) = \mathrm{Cov}(w, v)$ because $\mathrm{E}(v \mid \mathbf{x}, \mathbf{z}) = 0$. When assumption (18.67) holds along with conditions (18.60), (18.63), (18.64), and (18.65) and an appropriate rank condition—essentially, w is partially correlated with $\mathbf{z}$—2SLS estimation of equation (18.66) consistently estimates all parameters except the intercept. The IVs would be $(1, \mathbf{x}, \mathbf{z}, z_1\mathbf{x}, \ldots, z_L\mathbf{x})$, or we could use $\hat{\mathrm{E}}(w \mid \mathbf{x}, \mathbf{z})$ and $\hat{\mathrm{E}}(w \mid \mathbf{x}, \mathbf{z}) \cdot \mathbf{x}$ as IVs for $[w, w(\mathbf{x} - \boldsymbol{\psi})]$, where $\hat{\mathrm{E}}(w \mid \mathbf{x}, \mathbf{z})$ is an estimate of $\mathrm{E}(w \mid \mathbf{x}, \mathbf{z})$. As before, in practice $\boldsymbol{\psi}$ would be replaced with $\bar{\mathbf{x}}$. The 2SLS estimator is $\sqrt{N}$-consistent and asymptotically normal. Generally, the error in equation (18.66) is heteroskedastic.

Condition (18.67) is the same one we used in the binary treatment case to justify the usual IV estimator. As we discussed in Section 18.4.1, condition (18.67) does not hold when $\mathrm{P}(w = 1 \mid \mathbf{x}, \mathbf{z})$ satisfies logit or probit binary response models. If w is a continuous treatment, condition (18.67) is more reasonable. For example, if $\mathrm{E}(w \mid \mathbf{x}, \mathbf{z}, v)$ is additive in v, then condition (18.67) holds when $\mathrm{Var}(v \mid \mathbf{x}, \mathbf{z})$ is constant, in which case $\mathrm{Cov}(w, v \mid \mathbf{x}, \mathbf{z}) = \sigma_v^2$. See Wooldridge (1997b, 2000f) for further discussion. Wooldridge (2000f) also covers more general cases when condition (18.67) is not true and the treatment is not binary.

Heckman and Vytlacil (1998) use similar assumptions in a general random coefficient model to arrive at a related estimation method. In their simplest approach,

Heckman and Vytlacil suggest a two-step estimator, where $\mathrm{E}(w \mid \mathbf{x}, \dot{\mathbf{z}})$ is estimated using a linear model in the first stage and the fitted values are used in a second-stage regression. The preceding analysis shows that a linear functional form for $\mathrm{E}(w \mid \mathbf{x}, \mathbf{z})$ is not needed for the IV estimator to be consistent, although condition (18.67) generally is.

18.5.4 Multiple Treatments

Sometimes the treatment variable is not simply a scalar. For example, for the population of working high school graduates, w_1 could be credit hours at two-year colleges and w_2 credit hours at four-year colleges. If we make ignorability assumptions of the kind in Section 18.3.1, equation (18.15) extends in a natural way: each treatment variable appears by itself and interacted with the (demeaned) covariates. This approach does not put any restrictions on the nature of the treatments. Alternatively, as in Wooldridge (1999c), Assumption 18.5 extends to a vector $\mathbf{w}$, which leads to an extension of condition (18.60) for multiple treatments.

Wooldridge (2000f) shows how the IV methods in Section 18.4.1 extend easily to multiple treatments, binary or otherwise. For multiple binary treatments, a reduced-form probit is estimated for each treatment, and then terms $w_{ij}(\mathbf{x}_i - \bar{\mathbf{x}})$ and $\hat{\phi}_{ij}$ for each treatment j are added to equation (18.47). See Wooldridge (2000f) for further discussion. An approach based on finding $\mathrm{E}(y \mid w_1, \ldots, w_M, \mathbf{x}, \mathbf{z})$, for M treatments, is difficult but perhaps tractable in some cases.

Problems

18.1. Consider the difference-in-means estimator, $\bar{d} = \bar{y}_1 - \bar{y}_0$, where $\bar{y}_g$ is the sample average of the y_i with $w_i = g$, $g = 0, 1$.

a. Show that, as an estimator of ATE_1, the bias in $\bar{y}_1 - \bar{y}_0$ is $\mathrm{E}(y_0 \mid w = 1) - \mathrm{E}(y_0 \mid w = 0)$.

b. Let y_0 be the earnings someone would earn in the absence of job training, and let $w = 1$ denote the job training indicator. Explain the meaning of $\mathrm{E}(y_0 \mid w = 1) < \mathrm{E}(y_0 \mid w = 0)$. Intuitively, does it make sense that $\mathrm{E}(\bar{d}) < ATE_1$?

18.2. Show that $ATE_1(\mathbf{x})$ is identified under Assumption ATE.1′a; Assumption ATE.1′b is not needed.

18.3. Using the data in JTRAIN2.RAW, repeat the analysis in Example 18.2, using *unem78* as the response variable. For comparison, use the same $\mathbf{x}$ as in Example 18.2.

Compare the estimates from regressions (18.23) and (18.24), along with the estimate of ATE from linear regression *unem78* on 1, *train*, $\mathbf{x}$.

18.4. Carefully derive equation (18.45).

18.5. Use the data in JTRAIN2.RAW for this question.

a. As in Example 18.2, run a probit of *train* on 1, $\mathbf{x}$, where $\mathbf{x}$ contains the covariates from Example 18.2. Obtain the probit fitted values, say $\hat{\Phi}_i$.

b. Estimate the equation $re78_i = \gamma_0 + \alpha\, train_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i$ by IV, using instruments $(1, \hat{\Phi}_i, \mathbf{x}_i)$. Comment on the estimate of α and its standard error.

c. Regress $\hat{\Phi}_i$ on $\mathbf{x}_i$ to obtain the R-squared. What do you make of this result?

d. Does the nonlinearity of the probit model for *train* allow us to estimate α when we do not have an additional instrument? Explain.

18.6. In Procedure 18.2, explain why it is better to estimate equation (18.36) by IV rather than to run the OLS regression y_i on 1, $\hat{G}_i$, $\mathbf{x}_i$, $\hat{G}_i(\mathbf{x}_i - \bar{\mathbf{x}})$, $i = 1, \ldots, N$.

18.7. Use the data in JTRAIN2.RAW for this question.

a. In the ignorability setup of Section 18.5.3, let $w = mostrn$, the number of months spent in job training. Assume that $\mathrm{E}(w \mid \mathbf{x}) = \exp(\gamma_0 + \mathbf{x}\boldsymbol{\gamma})$, where $\mathbf{x}$ contains the same covariates as in Example 18.2. Estimate the parameters by nonlinear least squares, and let $\hat{m}_i$ be the fitted values. Which elements of $\mathbf{x}$ are significant? (You may use the usual NLS standard errors.)

b. Suppose that $\mathrm{Var}(w \mid \mathbf{x}) = h(\mathbf{x}) = \delta_0 + \delta_1 \mathrm{E}(w \mid \mathbf{x}) + \delta_2[\mathrm{E}(w \mid \mathbf{x})]^2$. Use the estimates from part a to estimate the δ_j. (Hint: Regress the squared NLS residuals on a quadratic in the NLS fitted values.) Are any of $\hat{h}_i$—the estimated variances—negative?

c. Form $\hat{r}_i = (w_i - \hat{m}_i)/\hat{h}_i$. Estimate equation (18.62) using $\hat{r}_i$ as an IV for w_i, where $\mathbf{g}(\mathbf{x}) = (1, \mathbf{x})$ and $y = re78$. Compare $\hat{\beta}$ with the OLS estimate of β.

18.8. In the IV setup of Section 18.5.3, suppose that $b = \beta$, and therefore we can write

$$y = a + \beta w + e, \qquad \mathrm{E}(e \mid a, \mathbf{x}, \mathbf{z}) = 0$$

Assume that conditions (18.64) and (18.65) hold for a.

a. Suppose w is a corner solution outcome, such as hours spent in a job training program. If $\mathbf{z}$ is used as IVs for w in $y = \gamma_0 + \beta w + \mathbf{x}\boldsymbol{\gamma} + r$, what is the identification condition?

b. If w given $(\mathbf{x}, \mathbf{z})$ follows a standard Tobit model, propose an IV estimator that uses the Tobit fitted values for w.

c. If $\text{Var}(e \mid a, \mathbf{x}, \mathbf{z}) = \sigma_e^2$ and $\text{Var}(a \mid \mathbf{x}, \mathbf{z}) = \sigma_a^2$, argue that the IV estimator from part b is asymptotically efficient.

d. What is an alternative to IV estimation that would use the Tobit fitted values for w? Which method do you prefer?

e. If $b \neq \beta$, but assumptions (18.64) and (18.65) hold, how would you estimate β?

18.9. Consider the IV approach in Section 18.5.3, under assumptions (18.60), (18.63), (18.64), and (18.65). In place of assumption (18.67), assume that $\text{E}(w \mid \mathbf{x}, \mathbf{z}, v) = \exp(\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + \pi_3 v)$, where v is independent of $(\mathbf{x}, \mathbf{z})$ with $\text{E}[\exp(\pi_3 v)] = 1$. (Therefore, w is some nonnegative treatment.)

a. Show that we can write

$$y = \eta_0 + \mathbf{x}\boldsymbol{\gamma} + \beta w + w \cdot (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + \xi \text{E}(w \mid \mathbf{x}, \mathbf{z}) + r$$

where $\text{E}(w \mid \mathbf{x}, \mathbf{z}) = \exp(\boldsymbol{\pi}_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2)$, and $\text{E}(r \mid \mathbf{x}, \mathbf{z}) = 0$.

b. Use part a to show that β is not identified. {Hint: Let $q \equiv \text{E}(w \mid \mathbf{x}, \mathbf{z})$, and let h be any other function of $(\mathbf{x}, \mathbf{z})$. Does the linear projection of w on $[1, \mathbf{x}, h, h \cdot (\mathbf{x} - \boldsymbol{\psi}), q]$ depend on h?}

c. For $w > 0$ (strictly positive treatment), add the assumption that $E(u \mid v, x, z) = \rho v$. Find $E(y \mid w, x, z) = E(y \mid v, x, z)$ and propose a two-step estimator of β.