CHAPTER 15

Instrumental Variables Estimation and Two Stage Least Squares

In this chapter, we further study the problem of **endogenous explanatory variables** in multiple regression models. In Chapter 3, we derived the bias in the OLS estimators when an important variable is omitted; in Chapter 5, we showed that OLS is generally inconsistent under **omitted variables**. Chapter 9 demonstrated that omitted variables bias can be eliminated (or at least mitigated) when a suitable proxy variable is given for an unobserved explanatory variable. Unfortunately, suitable proxy variables are not always available.

In the previous two chapters, we explained how fixed effects estimation or first differencing can be used with panel data to estimate the effects of time-varying independent variables in the presence of *time-constant* omitted variables. Although such methods are very useful, we do not always have access to panel data. Even if we can obtain panel data, it does us little good if we are interested in the effect of a variable that does not change over time: first differencing or fixed effects estimation eliminates time-constant explanatory variables. In addition, the panel data methods that we have studied so far do not solve the problem of time-varying omitted variables that are correlated with the explanatory variables.

In this chapter, we take a different approach to the endogeneity problem. You will see how the method of instrumental variables (IV) can be used to solve the problem of endogeneity of one or more explanatory variables. The method of two stage least squares (2SLS or TSLS) is second in popularity only to ordinary least squares for estimating linear equations in applied econometrics.

We begin by showing how IV methods can be used to obtain consistent estimators in the presence of omitted variables. IV can also be used to solve the **errors-in-variables** problem, at least under certain assumptions. Chapter 16 will demonstrate how to estimate simultaneous equations models using IV methods.

Our treatment of instrumental variables estimation closely follows our development of ordinary least squares in Part 1, where we assumed that we had a random sample from an underlying population. This is a desirable starting point because, in addition to simplifying the notation, it emphasizes that the important assumptions for IV estimation are stated in terms of the underlying population (just as with OLS). As we showed in Part 2, OLS can be applied to time series data, and the same is true of instrumental variables methods. Section 15-7 discusses some special issues that arise when IV methods are applied to time series data. In Section 15-8, we cover applications to pooled cross sections and panel data.

15-1 Motivation: Omitted Variables in a Simple Regression Model

When faced with the prospect of omitted variables bias (or unobserved heterogeneity), we have so far discussed three options: (1) we can ignore the problem and suffer the consequences of biased and inconsistent estimators; (2) we can try to find and use a suitable proxy variable for the unobserved variable; or (3) we can assume that the omitted variable does not change over time and use the fixed effects or first-differencing methods from Chapters 13 and 14. The first response can be satisfactory if the estimates are coupled with the direction of the biases for the key parameters. For example, if we can say that the estimator of a positive parameter, say, the effect of job training on subsequent wages, is biased toward zero and we have found a statistically significant positive estimate, we have still learned something: job training has a positive effect on wages, and it is likely that we have underestimated the effect. Unfortunately, the opposite case, where our estimates may be too large in magnitude, often occurs, which makes it very difficult for us to draw any useful conclusions.

The proxy variable solution discussed in Section 9-2 can also produce satisfying results, but it is not always possible to find a good proxy. This approach attempts to solve the omitted variable problem by replacing the unobservable with one or more proxy variables.

Another approach leaves the unobserved variable in the error term, but rather than estimating the model by OLS, it uses an estimation method that recognizes the presence of the omitted variable. This is what the method of instrumental variables does.

For illustration, consider the problem of unobserved ability in a wage equation for working adults. A simple model is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e,$$

where e is the error term. In Chapter 9, we showed how, under certain assumptions, a proxy variable such as IQ can be substituted for ability, and then a consistent estimator of β_1 is available from the regression of

$$\log(wage) \text{ on } educ, IQ.$$

Suppose, however, that a proxy variable is not available (or does not have the properties needed to produce a consistent estimator of β_1). Then, we put *abil* into the error term, and we are left with the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{15.1}$$

where u contains *abil*. Of course, if equation (15.1) is estimated by OLS, a biased and inconsistent estimator of β_1 results if *educ* and *abil* are correlated.

It turns out that we can still use equation (15.1) as the basis for estimation, provided we can find an instrumental variable for *educ*. To describe this approach, the simple regression model is written as

$$y = \beta_0 + \beta_1 x + u, \tag{15.2}$$

where we think that x and u are correlated (have nonzero covariance):

$$\mathrm{Cov}(x,u) \neq 0. \tag{15.3}$$

The method of instrumental variables works whether or not x and u are correlated, but, for reasons we will see later, OLS should be used if x is uncorrelated with u.

In order to obtain consistent estimators of β_0 and β_1 when x and u are correlated, we need some additional information. The information comes by way of a new variable that satisfies certain properties. Suppose that we have an observable variable z that satisfies these two assumptions: (1) z is uncorrelated with u, that is,

$$\mathrm{Cov}(z,u) = 0; \tag{15.4}$$

(2) z is correlated with x, that is,

$$\mathrm{Cov}(z,x) \neq 0. \tag{15.5}$$

Then, we call z an **instrumental variable** for x, or sometimes simply an **instrument** for x.

The requirement that the instrument z satisfies (15.4) is summarized by saying "z is exogenous in equation (15.2)," and so we often refer to (15.4) as **instrument exogeneity**. In the context of omitted variables, instrument exogeneity means that z should have no partial effect on y (after x and omitted variables have been controlled for), and z should be uncorrelated with the omitted variables. Equation (15.5) means that z must be related, either positively or negatively, to the endogenous explanatory variable x. This condition is sometimes referred to as **instrument relevance** (as in "z is relevant for explaining variation in x").

There is a very important difference between the two requirements for an instrumental variable. Because (15.4) involves the covariance between z and the unobserved error u, we cannot generally hope to test this assumption: in most cases, we must maintain $\mathrm{Cov}(z,u) = 0$ by appealing to economic behavior or introspection. Sometimes, we might have an observable proxy variable for some factor contained in u, in which case we can check to see if z and the proxy variable are roughly uncorrelated. Of course, if we have a good proxy for an important element of u, we might just add the proxy as an explanatory variable and estimate the expanded equation by ordinary least squares. See Section 9-2.

Some readers may be wondering why we do not attempt to check (15.4) by using the following procedure. Given a sample of size n, obtain the OLS residuals, $\hat{u}_i$, from the regression y_i on x_i. Then, devise a test based on the sample correlation between z_i and $\hat{u}_i$ as a check on whether z_i and the unobserved errors u_i are correlated. A moment's thought reveals the logical problem with this procedure. The entire reason for moving beyond OLS is that we think the OLS estimators of β_0 and β_1 are inconsistent due to correlation between x and u. Therefore, in computing the OLS residuals $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, we are not getting useful estimates of the u_i. Therefore, we can learn nothing by studying the correlation between z_i and $\hat{u}_i$. A related suggestion is to use the OLS regression y_i on x_i, z_i and to conclude z_i satisfies the exogeneity requirement if its coefficient is statistically insignificant. Again, this procedure does not work, regardless of the outcome of the test, because x is allowed to be endogenous. The bottom line is that, in the current setting, we have no way of testing (15.4) unless we use external information.

By contrast, the condition that z is correlated with x (in the population) can be tested, given a random sample from the population. The easiest way to do this is to estimate a simple regression between x and z. In the population, we have

$$x = \pi_0 + \pi_1 z + v. \tag{15.6}$$

Then, because $\pi_1 = \mathrm{Cov}(z,x)/\mathrm{Var}(z)$, assumption (15.5) holds if, and only if, $\pi_1 \neq 0$. Thus, we should be able to *reject* the null hypothesis

$$\mathrm{H}_0\colon \pi_1 = 0 \tag{15.7}$$

against the two-sided alternative H_0: $\pi_1 \neq 0$, at a sufficiently small significance level. If this is the case, then we can be fairly confident that (15.5) holds.

For the log(*wage*) equation in (15.1), an instrumental variable z for *educ* must be (1) uncorrelated with ability (and any other unobserved factors affecting wage) and (2) correlated with education. Something such as the last digit of an individual's Social Security Number almost certainly satisfies the first requirement: it is uncorrelated with ability because it is determined randomly. However, it is precisely because of the randomness of the last digit of the SSN that it is not correlated with education, either; therefore it makes a poor instrumental variable for *educ* because it violates the instrument relevance requirement in equation (15.5).

What we have called a *proxy variable* for the omitted variable makes a poor IV for the opposite reason. For example, in the log(*wage*) example with omitted ability, a proxy variable for *abil* should be as highly correlated as possible with *abil*. An instrumental variable must be *uncorrelated* with *abil*. Therefore, while *IQ* is a good candidate as a proxy variable for *abil*, it is not a good instrumental variable for *educ* because it violates the instrument exogeneity requirement in equation (15.4).

Whether other possible instrumental variable candidates satisfy the exogeneity requirement in (15.4) is less clear-cut. In wage equations, labor economists have used family background variables as IVs for education. For example, mother's education (*motheduc*) is positively correlated with child's education, as can be seen by collecting a sample of data on working people and running a simple regression of *educ* on *motheduc*. Therefore, *motheduc* satisfies equation (15.5). The problem is that mother's education might also be correlated with child's ability (through mother's ability and perhaps quality of nurturing at an early age), in which case (15.4) fails.

Another IV choice for *educ* in (15.1) is number of siblings while growing up (*sibs*). Typically, having more siblings is associated with lower average levels of education. Thus, if number of siblings is uncorrelated with ability, it can act as an instrumental variable for *educ*.

As a second example, consider the problem of estimating the causal effect of skipping classes on final exam score. In a simple regression framework, we have

$$score = \beta_0 + \beta_1 skipped + u, \quad [15.8]$$

where *score* is the final exam score and *skipped* is the total number of lectures missed during the semester. We certainly might be worried that *skipped* is correlated with other factors in u: more able, highly motivated students might miss fewer classes. Thus, a simple regression of *score* on *skipped* may not give us a good estimate of the causal effect of missing classes.

What might be a good IV for *skipped*? We need something that has no direct effect on *score* and is not correlated with student ability and motivation. At the same time, the IV must be correlated with *skipped*. One option is to use distance between living quarters and classrooms. Especially at large universities, some living quarters will be further from a student's classrooms, and this may essentially be a random occurrence. Some students live off campus while others commute long distances. Living further away from classrooms may increase the likelihood of missing lectures due to bad weather, oversleeping, and so on. Thus, *skipped* may be positively correlated with *distance*; this can be checked by regressing *skipped* on *distance* and doing a t test, as described earlier.

Is *distance* uncorrelated with u? In the simple regression model (15.8), some factors in u may be correlated with *distance*. For example, students from low-income families may live off campus; if income affects student performance, this could cause *distance* to be correlated with u. Section 15-2 shows how to use IV in the context of multiple regression, so that other factors affecting *score* can be included directly in the model. Then, *distance* might be a good IV for *skipped*. An IV approach may not be necessary at all if a good proxy exists for student ability, such as cumulative GPA prior to the semester.

There is a final point worth emphasizing before we turn to the mechanics of IV estimation: namely, in using the simple regression in equation (15.6) to test (15.7), it is important to take note of the sign (and even magnitude) of $\hat{\pi}_1$ and not just its statistical significance. Arguments for why a variable z makes a good IV candidate for an endogenous explanatory variable x should include a discussion about the nature of the relationship between x and z. For example, due to genetics and background influences it makes sense that child's education (x) and mother's education (z) are positively correlated. If in your sample of data you find that they are actually negatively correlated—that is, $\hat{\pi}_1 < 0$—then your use of mother's education as an IV for child's education is likely to be unconvincing. [And this has nothing to do with whether condition (15.4) is likely to hold.] In the example of measuring whether skipping classes has an effect on test performance, one should find a positive, statistically significant relationship between *skipped* and *distance* in order to justify using *distance* as an IV for skipped: a negative relationship would be difficult to justify [and would suggest that there are important omitted variables driving a negative correlation—variables that might themselves have to be included in the model (15.8)].

We now demonstrate that the availability of an instrumental variable can be used to estimate consistently the parameters in equation (15.2). In particular, we show that assumptions (15.4) and (15.5) serve to *identify* the parameter β_1. **Identification** of a parameter in this context means that we can write β_1 in terms of population moments that can be estimated using a sample of data. To write β_1 in terms of population covariances, we use equation (15.2): the covariance between z and y is

$$\mathrm{Cov}(z,y) = \beta_1 \mathrm{Cov}(z,x) + \mathrm{Cov}(z,u).$$

Now, under assumption (15.4), $\mathrm{Cov}(z,u) = 0$, and under assumption (15.5), $\mathrm{Cov}(z,x) \neq 0$. Thus, we can solve for β_1 as

$$\beta_1 = \frac{\mathrm{Cov}(z,y)}{\mathrm{Cov}(z,x)}. \qquad [15.9]$$

[Notice how this simple algebra fails if z and x are uncorrelated, that is, if $\mathrm{Cov}(z,x) = 0$.] Equation (15.9) shows that β_1 is the population covariance between z and y divided by the population covariance between z and x, which shows that β_1 is identified. Given a random sample, we estimate the population quantities by the sample analogs. After canceling the sample sizes in the numerator and denominator, we get the **instrumental variables (IV) estimator** of β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}. \qquad [15.10]$$

Given a sample of data on x, y, and z, it is simple to obtain the IV estimator in (15.10). The IV estimator of β_0 is simply $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, which looks just like the OLS intercept estimator except that the slope estimator, $\hat{\beta}_1$, is now the IV estimator.

It is no accident that when $z = x$ we obtain the OLS estimator of β_1. In other words, when x is exogenous, it can be used as its own IV, and the IV estimator is then identical to the OLS estimator.

A simple application of the law of large numbers shows that the IV estimator is consistent for β_1: $\mathrm{plim}(\hat{\beta}_1) = \beta_1$, provided assumptions (15.4) and (15.5) are satisfied. If either assumption fails, the IV estimators are not consistent (more on this later). One feature of the IV estimator is that, when x and u are in fact correlated—so that instrumental variables estimation is actually needed—it is essentially never unbiased. This means that, in small samples, the IV estimator can have a substantial bias, which is one reason why large samples are preferred.

When discussing the application of instrumental variables it is important to be careful with language. Like OLS, IV is an *estimation* method. It makes little sense to refer to "an instrumental variables model"—just as the phrase "OLS model" makes little sense. As we know, a model is an equation such as (15.8), which is a special case of the generic model in equation (15.2). When we have a model such as (15.2), we can choose to estimate the parameters of that model in many different ways. Prior to this chapter we focused primarily on OLS, but, for example, we also know from Chapter 8 that one can use weighted least squares as an alternative estimation method (and there are unlimited possibilities for the weights). If we have an instrumental variable candidate z for x, then we can instead apply instrumental variables estimation. It is certainly true that the estimation method we apply is motivated by the model and assumptions we make about that model. But the estimators are well defined and exist apart from any underlying model or assumptions: remember, an estimator is simply a rule for combining data. The bottom line is that while we probably know what a researcher means when using a phrase such as "I estimated an IV model," such language betrays a lack of understanding about the difference between a model and an estimation method.

15-1a Statistical Inference with the IV Estimator

Given the similar structure of the IV and OLS estimators, it is not surprising that the IV estimator has an approximate normal distribution in large sample sizes. To perform inference on β_1, we need a standard error that can be used to compute t statistics and confidence intervals. The usual approach is to impose a homoskedasticity assumption, just as in the case of OLS. Now, the homoskedasticity assumption is stated conditional on the instrumental variable, z, not the endogenous explanatory variable, x. Along with the previous assumptions on u, x, and z, we add

$$\mathrm{E}(u^2|z) = \sigma^2 = \mathrm{Var}(u). \quad [15.11]$$

It can be shown that, under (15.4), (15.5), and (15.11), the asymptotic variance of $\hat{\beta}_1$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}, \quad [15.12]$$

where σ_x^2 is the population variance of x, σ^2 is the population variance of u, and $\rho_{x,z}^2$ is the square of the population correlation between x and z. This tells us how highly correlated x and z are in the population. As with the OLS estimator, the asymptotic variance of the IV estimator decreases to zero at the rate of $1/n$, where n is the sample size.

Equation (15.12) is interesting for two reasons. First, it provides a way to obtain a standard error for the IV estimator. All quantities in (15.12) can be consistently estimated given a random sample. To estimate σ_x^2, we simply compute the sample variance of x_i; to estimate $\rho_{x,z}^2$, we can run the regression of x_i on z_i to obtain the R-squared, say, $R_{x,z}^2$. Finally, to estimate σ^2, we can use the IV residuals,

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the IV estimates. A consistent estimator of σ^2 looks just like the estimator of σ^2 from a simple OLS regression:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2,$$

where it is standard to use the degrees of freedom correction (even though this has little effect as the sample size grows).

The (asymptotic) standard error of $\hat{\beta}_1$ is the square root of the estimated asymptotic variance, the latter of which is given by

$$\frac{\hat{\sigma}^2}{\mathrm{SST}_x \cdot R_{x,z}^2}, \quad [15.13]$$

where SST_x is the total sum of squares of the x_i. [Recall that the sample variance of x_i is SST_x/n, and so the sample sizes cancel to give us (15.13).] The resulting standard error can be used to construct either t statistics for hypotheses involving β_1 or confidence intervals for β_1. $\hat{\beta}_0$ also has a standard error that we do not present here. Any modern econometrics package computes the standard error after any IV estimation; there is rarely any reason to perform the calculations by hand.

A second reason (15.12) is interesting is that it allows us to compare the asymptotic variances of the IV and the OLS estimators (when x and u are uncorrelated). Under the Gauss-Markov assumptions, the variance of the OLS estimator is σ^2/SST_x, while the comparable formula for the IV estimator is $\sigma^2/(\mathrm{SST}_x \cdot R^2_{x,z})$; they differ only in that $R^2_{x,z}$ appears in the denominator of the IV variance. Because an R-squared is always less than one, the IV variance is always larger than the OLS variance (when OLS is valid). If $R^2_{x,z}$ is small, then the IV variance can be much larger than the OLS variance. Remember, $R^2_{x,z}$ measures the strength of the linear relationship between x and z in the sample. If x and z are only slightly correlated, $R^2_{x,z}$ can be small, and this can translate into a very large sampling variance for the IV estimator. The more highly correlated z is with x, the closer $R^2_{x,z}$ is to one, and the smaller is the variance of the IV estimator. In the case that $z = x$, $R^2_{x,z} = 1$, and we get the OLS variance, as expected.

The previous discussion highlights an important cost of performing IV estimation when x and u are uncorrelated: the asymptotic variance of the IV estimator is always larger, and sometimes much larger, than the asymptotic variance of the OLS estimator.

EXAMPLE 15.1 Estimating the Return to Education for Married Women

We use the data on married working women in MROZ to estimate the return to education in the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u. \qquad [15.14]$$

For comparison, we first obtain the OLS estimates:

$$\begin{aligned} \widehat{\log(wage)} &= -.185 + .109\, educ \\ &\quad\;\; (.185) \quad (.014) \\ n &= 428,\ R^2 = .118. \end{aligned} \qquad [15.15]$$

The estimate for β_1 implies an almost 11% return for another year of education.

Next, we use father's education (*fatheduc*) as an instrumental variable for *educ*. We have to maintain that *fatheduc* is uncorrelated with u. The second requirement is that *educ* and *fatheduc* are correlated. We can check this very easily using a simple regression of *educ* on *fatheduc* (using only the working women in the sample):

$$\begin{aligned} \widehat{educ} &= 10.24 + .269\, fatheduc \\ &\quad\;\; (.28) \quad (.029) \\ n &= 428,\ R^2 = .173. \end{aligned} \qquad [15.16]$$

The t statistic on *fatheduc* is 9.28, which indicates that *educ* and *fatheduc* have a statistically significant positive correlation. (In fact, *fatheduc* explains about 17% of the variation in *educ* in the sample.) Using *fatheduc* as an IV for *educ* gives

$$\begin{aligned} \widehat{\log(wage)} &= .441 + .059\, educ \\ &\quad\;\; (.446) \quad (.035) \\ n &= 428,\ R^2 = .093. \end{aligned} \qquad [15.17]$$

The IV estimate of the return to education is 5.9%, which is barely more than one-half of the OLS estimate. This *suggests* that the OLS estimate is too high and is consistent with omitted ability bias. But we should remember that these are estimates from just one sample: we can never know whether .109 is above the true return to education, or whether .059 is closer to the true return to education. Further, the standard error of the IV estimate is two and one-half times as large as the OLS standard error (this is expected, for the reasons we gave earlier). The 95% confidence interval for β_1 using OLS is much tighter than that using the IV; in fact, the IV confidence interval actually contains the OLS estimate. Therefore, although the differences between (15.15) and (15.17) are practically large, we cannot say whether the difference is *statistically* significant. We will show how to test this in Section 15-5.

In the previous example, the estimated return to education using IV was less than that using OLS, which corresponds to our expectations. But this need not have been the case, as the following example demonstrates.

EXAMPLE 15.2 Estimating the Return to Education for Men

We now use WAGE2 to estimate the return to education for men. We use the variable *sibs* (number of siblings) as an instrument for *educ*. These are negatively correlated, as we can verify from a simple regression:

$$\widehat{educ} = \underset{(.11)}{14.14} - \underset{(.030)}{.228}\, sibs$$
$$n = 935,\ R^2 = .057.$$

This equation implies that every sibling is associated with, on average, about .23 less of a year of education. If we assume that *sibs* is uncorrelated with the error term in (15.14), then the IV estimator is consistent. Estimating equation (15.14) using *sibs* as an IV for *educ* gives

$$\widehat{\log(wage)} = \underset{(.36)}{5.13} + \underset{(.026)}{.122}\, educ$$
$$n = 935.$$

(The R-squared is computed to be negative, so we do not report it. A discussion of R-squared in the context of IV estimation follows.) For comparison, the OLS estimate of β_1 is .059 with a standard error of .006. Unlike in the previous example, the IV estimate is now much higher than the OLS estimate. Although we do not know whether the difference is statistically significant, this does not mesh with the omitted ability bias from OLS. It could be that *sibs* is also correlated with ability: more siblings means, on average, less parental attention, which could result in lower ability. Another interpretation is that the OLS estimator is biased toward zero because of measurement error in *educ*. This is not entirely convincing because, as we discussed in Section 9-3, *educ* is unlikely to satisfy the classical errors-in-variables model.

In the previous examples, the endogenous explanatory variable (*educ*) and the instrumental variables (*fatheduc*, *sibs*) have quantitative meaning. But nothing prevents the explanatory variable or IV from being binary variables. Angrist and Krueger (1991), in their simplest analysis, came up with a clever binary instrumental variable for *educ*, using census data on men in the United States. Let *frstqrt* be equal to one if the man was born in the first quarter of the year, and zero otherwise. It seems that the error term in (15.14)—and, in particular, ability—should be unrelated to quarter of birth. But *frstqrt* also needs to be correlated with *educ*. It turns out that years of education *do* differ systematically in the population based on quarter of birth. Angrist and Krueger argued persuasively that this is due to compulsory school attendance laws in effect in all states. Briefly, students born early in the year typically begin school at an older age. Therefore, they reach the compulsory schooling age (16 in most states) with somewhat less education than students who begin school at a younger age. For students who finish high school, Angrist and Krueger verified that there is no relationship between years of education and quarter of birth.

Because years of education varies only slightly across quarter of birth—which means $R^2_{x,z}$ in (15.13) is very small—Angrist and Krueger needed a very large sample size to get a reasonably precise IV estimate. Using 247,199 men born between 1920 and 1929, the OLS estimate of the return to education was .0801 (standard error .0004), and the IV estimate was .0715 (.0219); these are reported in Table III of Angrist and Krueger's paper. Note how large the t statistic is for the OLS estimate (about 200), whereas the t statistic for the IV estimate is only 3.26. Thus, the IV estimate is statistically different from zero, but its confidence interval is much wider than that based on the OLS estimate.

An interesting finding by Angrist and Krueger is that the IV estimate does not differ much from the OLS estimate. In fact, using men born in the next decade, the IV estimate is somewhat higher than the OLS estimate. One could interpret this as showing that there is no omitted ability bias when wage equations are estimated by OLS. However, the Angrist and Krueger paper has been criticized on econometric grounds. As discussed by Bound, Jaeger, and Baker (1995), it is not obvious that season of birth is unrelated to unobserved factors that affect wage. As we will explain in the next subsection, even a small amount of correlation between z and u can cause serious problems for the IV estimator.

For policy analysis, the endogenous explanatory variable is often a binary variable. For example, Angrist (1990) studied the effect that being a veteran of the Vietnam War had on lifetime earnings. A simple model is

$$\log(earns) = \beta_0 + \beta_1 veteran + u, \tag{15.18}$$

where *veteran* is a binary variable. The problem with estimating this equation by OLS is that there may be a *self-selection* problem, as we mentioned in Chapter 7: perhaps people who get the most out of the military choose to join, or the decision to join is correlated with other characteristics that affect earnings. These will cause *veteran* and u to be correlated.

GOING FURTHER 15.1

If some men who were assigned low draft lottery numbers obtained additional schooling to reduce the probability of being drafted, is lottery number a good instrument for *veteran* in (15.18)?

Angrist pointed out that the Vietnam draft lottery provided a **natural experiment** (see also Chapter 13) that created an instrumental variable for *veteran*. Young men were given lottery numbers that determined whether they would be called to serve in Vietnam. Because the numbers given were (eventually) randomly assigned, it seems plausible that draft lottery number is uncorrelated with the error term u. But those with a low enough number had to serve in Vietnam, so that the probability of being a veteran is correlated with lottery number. If both of these assertions are true, draft lottery number is a good IV candidate for *veteran*.

It is also possible to have a binary endogenous explanatory variable and a binary instrumental variable. See Problem 1 for an example.

15-1b Properties of IV with a Poor Instrumental Variable

We have already seen that, though IV is consistent when z and u are uncorrelated and z and x have any positive or negative correlation, IV estimates can have large standard errors, especially if z and x are only weakly correlated. Weak correlation between z and x can have even more serious consequences: the IV estimator can have a large asymptotic bias even if z and u are only moderately correlated.

We can see this by studying the probability limit of the IV estimator when z and u are possibly correlated. Letting $\hat{\beta}_{1,\,\text{IV}}$ denote the IV estimator, we can write

$$\text{plim}\,\hat{\beta}_{1,\,\text{IV}} = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}, \tag{15.19}$$

where σ_u and σ_x are the standard deviations of u and x in the population, respectively. The interesting part of this equation involves the correlation terms. It shows that, even if $\text{Corr}(z,u)$ is small, the inconsistency in the IV estimator can be very large if $\text{Corr}(z,x)$ is also small. Thus, even if we focus only on consistency, it is not necessarily better to use IV than OLS if the correlation between z and u is smaller than that between x and u. Using the fact that $\text{Corr}(x,u) = \text{Cov}(x,u)/(\sigma_x\sigma_u)$ along with equation (5.3), we can write the plim of the OLS estimator—call it $\hat{\beta}_{1,\,\text{OLS}}$—as

$$\text{plim}\,\hat{\beta}_{1,\,\text{OLS}} = \beta_1 + \text{Corr}(x,u) \cdot \frac{\sigma_u}{\sigma_x}. \tag{15.20}$$

Comparing these formulas shows that it is possible for the directions of the asymptotic biases to be different for IV and OLS. For example, suppose $\text{Corr}(x,u) > 0$, $\text{Corr}(z,x) > 0$, and $\text{Corr}(z,u) < 0$.

Then the IV estimator has a downward bias, whereas the OLS estimator has an upward bias (asymptotically). In practice, this situation is probably rare. More problematic is when the direction of the bias is the same and the correlation between z and x is small. For concreteness, suppose x and z are both positively correlated with u and $\text{Corr}(z,x) > 0$. Then the asymptotic bias in the IV estimator is less than that for OLS only if $\text{Corr}(z,u)/\text{Corr}(z,x) < \text{Corr}(x,u)$. If $\text{Corr}(z,x)$ is small, then a seemingly small correlation between z and u can be magnified and make IV worse than OLS, even if we restrict attention to bias. For example, if $\text{Corr}(z,x) = .2$, $\text{Corr}(z,u)$ must be less than one-fifth of $\text{Corr}(z,u)$ before IV has less asymptotic bias than OLS. In many applications, the correlation between the instrument and x is less than .2. Unfortunately, because we rarely have an idea about the relative magnitudes of $\text{Corr}(z,u)$ and $\text{Corr}(x,u)$, we can never know for sure which estimator has the largest asymptotic bias [unless, of course, we assume $\text{Corr}(z,u) = 0$].

In the Angrist and Krueger (1991) example mentioned earlier, where x is years of schooling and z is a binary variable indicating quarter of birth, the correlation between z and x is very small. Bound, Jaeger, and Baker (1995) discussed reasons why quarter of birth and u might be somewhat correlated. From equation (15.19), we see that this can lead to a substantial bias in the IV estimator.

When z and x are not correlated at all, things are especially bad, whether or not z is uncorrelated with u. The following example illustrates why we should always check to see if the endogenous explanatory variable is correlated with the IV candidate.

EXAMPLE 15.3 Estimating the Effect of Smoking on Birth Weight

In Chapter 6, we estimated the effect of cigarette smoking on child birth weight. Without other explanatory variables, the model is

$$\log(bwght) = \beta_0 + \beta_1 packs + u, \tag{15.21}$$

where *packs* is the number of packs smoked by the mother per day. We might worry that *packs* is correlated with other health factors or the availability of good prenatal care, so that *packs* and u might be correlated. A possible instrumental variable for *packs* is the average price of cigarettes in the state of residence, *cigprice*. We will assume that *cigprice* and u are uncorrelated (even though state support for health care could be correlated with cigarette taxes).

If cigarettes are a typical consumption good, basic economic theory suggests that *packs* and *cigprice* are negatively correlated, so that *cigprice* can be used as an IV for *packs*. To check this, we regress *packs* on *cigprice*, using the data in BWGHT:

$$\begin{aligned}\widehat{packs} &= \underset{(.103)}{.067} + \underset{(.0008)}{.0003}\, cigprice \\ n &= 1{,}388,\ R^2 = .0000,\ \bar{R}^2 = -.0006.\end{aligned}$$

This indicates no relationship between smoking during pregnancy and cigarette prices, which is perhaps not too surprising given the addictive nature of cigarette smoking.

Because *packs* and *cigprice* are not correlated, we should not use *cigprice* as an IV for *packs* in (15.21). But what happens if we do? The IV results would be

$$\begin{aligned}\widehat{\log(bwght)} &= \underset{(.91)}{4.45} + \underset{(8.70)}{2.99}\, packs \\ n &= 1{,}388\end{aligned}$$

(the reported R-squared is negative). The coefficient on *packs* is huge and of an unexpected sign. The standard error is also very large, so *packs* is not significant. But the estimates are meaningless because *cigprice* fails the one requirement of an IV that we can always test: assumption (15.5).

The previous example shows that IV estimation can produce strange results when the instrument relevance condition, $\text{Corr}(z,x) \neq 0$, fails. Of practically greater interest is the so-called problem of **weak instruments**, which is loosely defined as the problem of "low" (but not zero) correlation between z and x. In a particular application, it is difficult to define how low is too low, but recent theoretical research, supplemented by simulation studies, has shed considerable light on the issue. Staiger and Stock (1997) formalized the problem of weak instruments by modeling the correlation between z and x as a function of the sample size; in particular, the correlation is assumed to shrink to zero at the rate $1/\sqrt{n}$. Not surprisingly, the asymptotic distribution of the instrumental variables estimator is different compared with the usual asymptotics, where the correlation is assumed to be fixed and nonzero. One of the implications of the Stock–Staiger work is that the usual statistical inference, based on t statistics and the standard normal distribution, can be seriously misleading. We discuss this further in Section 15-3.

15-1c Computing *R*-Squared after IV Estimation

Most regression packages compute an R-squared after IV estimation, using the standard formula: $R^2 = 1 - \text{SSR/SST}$, where SSR is the sum of squared *IV residuals* and SST is the total sum of squares of y. Unlike in the case of OLS, the R-squared from IV estimation can be negative because SSR for IV can actually be larger than SST. Although it does not really hurt to report the R-squared for IV estimation, it is not very useful, either. When x and u are correlated, we cannot decompose the variance of y into $\beta_1^2\text{Var}(x) + \text{Var}(u)$, and so the R-squared has no natural interpretation. In addition, as we will discuss in Section 15-3, these R-squareds *cannot* be used in the usual way to compute F tests of joint restrictions.

If our goal was to produce the largest R-squared, we would always use OLS. IV methods are intended to provide better estimates of the ceteris paribus effect of x on y when x and u are correlated; goodness-of-fit is not a factor. A high R-squared resulting from OLS is of little comfort if we cannot consistently estimate β_1.

15-2 IV Estimation of the Multiple Regression Model

The IV estimator for the simple regression model is easily extended to the multiple regression case. We begin with the case where only one of the explanatory variables is correlated with the error. In fact, consider a standard linear model with two explanatory variables:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1. \tag{15.22}$$

We call this a **structural equation** to emphasize that we are interested in the β_j, which simply means that the equation is supposed to measure a causal relationship. We use a new notation here to distinguish endogenous from **exogenous variables**. The dependent variable y_1 is clearly endogenous, as it is correlated with u_1. The variables y_2 and z_1 are the explanatory variables, and u_1 is the error. As usual, we assume that the expected value of u_1 is zero: $\text{E}(u_1) = 0$. We use z_1 to indicate that this variable is exogenous in (15.22) (z_1 is uncorrelated with u_1). We use y_2 to indicate that this variable is suspected of being correlated with u_1. We do not specify why y_2 and u_1 are correlated, but for now it is best to think of u_1 as containing an omitted variable correlated with y_2. The notation in equation (15.22) originates in simultaneous equations models (which we cover in Chapter 16), but we use it more generally to easily distinguish exogenous from endogenous explanatory variables in a multiple regression model.

An example of (15.22) is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1, \tag{15.23}$$

where $y_1 = \log(wage)$, $y_2 = educ$, and $z_1 = exper$. In other words, we assume that *exper* is exogenous in (15.23), but we allow that *educ*—for the usual reasons—is correlated with u_1.

We know that if (15.22) is estimated by OLS, *all* of the estimators will be biased and inconsistent. Thus, we follow the strategy suggested in the previous section and seek an instrumental variable for y_2. Because z_1 is assumed to be uncorrelated with u_1, can we use z_1 as an instrument for y_2, assuming y_2 and z_1 are correlated? The answer is no. Because z_1 itself appears as an explanatory variable in (15.22), it cannot serve as an instrumental variable for y_2. We need another exogenous variable—call it z_2—that does *not* appear in (15.22). Therefore, key assumptions are that z_1 and z_2 are uncorrelated with u_1; we also assume that u_1 has zero expected value, which is without loss of generality when the equation contains an intercept:

$$\mathrm{E}(u_1) = 0, \mathrm{Cov}(z_1,u_1) = 0, \text{ and } \mathrm{Cov}(z_2,u_1) = 0. \qquad [15.24]$$

Given the zero mean assumption, the latter two assumptions are equivalent to $\mathrm{E}(z_1u_1) = \mathrm{E}(z_2u_1) = 0$, and so the method of moments approach suggests obtaining estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by solving the sample counterparts of (15.24):

$$\begin{aligned} \sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0. \end{aligned} \qquad [15.25]$$

This is a set of three linear equations in the three unknowns $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and it is easily solved given the data on y_1, y_2, z_1, and z_2. The estimators are called *instrumental variables estimators*. If we think y_2 is exogenous and we choose $z_2 = y_2$, equations (15.25) are exactly the first order conditions for the OLS estimators; see equations (3.13).

We still need the instrumental variable z_2 to be correlated with y_2, but the sense in which these two variables must be correlated is complicated by the presence of z_1 in equation (15.22). We now need to state the assumption in terms of *partial* correlation. The easiest way to state the condition is to write the endogenous explanatory variable as a linear function of the exogenous variables and an error term:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2, \qquad [15.26]$$

where, by construction, $\mathrm{E}(v_2) = 0$, $\mathrm{Cov}(z_1,v_2) = 0$, and $\mathrm{Cov}(z_2,v_2) = 0$, and the π_j are unknown parameters. The key identification condition [along with (15.24)] is that

$$\pi_2 \neq 0. \qquad [15.27]$$

In other words, after partialling out z_1, y_2 and z_2 are still correlated. This correlation can be positive or negative, but it cannot be zero. Testing (15.27) is easy: we estimate (15.26) by OLS and use a t test (possibly making it robust to heteroskedasticity). We should always test this assumption. Unfortunately, we cannot test that z_1 and z_2 are uncorrelated with u_1; hopefully, we can make the case based on economic reasoning or introspection.

GOING FURTHER 15.2

Suppose we wish to estimate the effect of marijuana usage on college grade point average. For the population of college seniors at a university, let *daysused* denote the number of days in the past month on which a student smoked marijuana and consider the structural equation

$$colGPA = \beta_0 + \beta_1 daysused + \beta_2 SAT + u.$$

(i) Let *percHS* denote the percentage of a student's high school graduating class that reported regular use of marijuana. If this is an IV candidate for *daysused*, write the reduced form for *daysused*. Do you think (15.27) is likely to be true?

(ii) Do you think *percHS* is truly exogenous in the structural equation? What problems might there be?

Equation (15.26) is an example of a **reduced form equation**, which means that we have written an endogenous variable in terms of exogenous variables. This name comes from simultaneous equations models—which we study in Chapter 16—but it is a useful concept whenever we have an endogenous explanatory variable. The name helps distinguish it from the structural equation (15.22).

Adding more **exogenous explanatory variables** to the model is straightforward. Write the structural model as

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1, \quad [15.28]$$

where y_2 is thought to be correlated with u_1. Let z_k be a variable not in (15.28) that is also exogenous. Therefore, we assume that

$$\mathrm{E}(u_1) = 0,\ \mathrm{Cov}(z_j, u_1) = 0, \quad j = 1, \ldots, k. \quad [15.29]$$

Under (15.29), $z_1, \ldots, z_{k-1}$ are the exogenous variables appearing in (15.28). In effect, these act as their own instrumental variables in estimating the β_j in (15.28). The special case of $k = 2$ is given in the equations in (15.25); along with z_2, z_1 appears in the set of moment conditions used to obtain the IV estimates. More generally, $z_1, \ldots, z_{k-1}$ are used in the moment conditions along with the instrumental variable for y_2, z_k.

The reduced form for y_2 is

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2, \quad [15.30]$$

and we need some partial correlation between z_k and y_2:

$$\pi_k \neq 0. \quad [15.31]$$

Under (15.29) and (15.31), z_k is a valid IV for y_2. [We do not care about the remaining π_j in (15.30); some or all of them could be zero.] A minor additional assumption is that there are no perfect linear relationships among the exogenous variables; this is analogous to the assumption of no perfect collinearity in the context of OLS.

For standard statistical inference, we need to assume homoskedasticity of u_1. We give a careful statement of these assumptions in a more general setting in Section 15-3.

EXAMPLE 15.4 Using College Proximity as an IV for Education

Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He used a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an instrumental variable for education. In a log(*wage*) equation, he included other standard controls: experience, a black dummy variable, dummy variables for living in an SMSA and living in the South, and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966. In order for *nearc4* to be a valid instrument, it must be uncorrelated with the error term in the wage equation—we assume this—and it must be partially correlated with *educ*. To check the latter requirement, we regress *educ* on *nearc4* and all of the exogenous variables appearing in the equation. (That is, we estimate the reduced form for *educ*.) Using the data in CARD, we obtain, in condensed form,

$$\begin{aligned} educ &= \underset{(.24)}{16.64} + \underset{(.088)}{.320}\, nearc4 - \underset{(.034)}{.413}\, exper + \cdots \\ n &= 3{,}010,\ R^2 = .477. \end{aligned}$$

We are interested in the coefficient and t statistic on *nearc4*. The coefficient implies that in 1976, other things being fixed (experience, race, region, and so on), people who lived near a college in 1966 had, on average, about one-third of a year more education than those who did not grow up near a college. The t statistic on *nearc4* is 3.64, which gives a p-value that is zero in the first three decimals.

Therefore, if *nearc4* is uncorrelated with unobserved factors in the error term, we can use *nearc4* as an IV for *educ*.

The OLS and IV estimates are given in Table 15.1. Like the OLS standard errors, the reported IV standard errors employ a degrees-of-freedom adjustment in estimating the error variance. In some statistical packages the degrees-of-freedom adjustment is the default; in others it is not.

Interestingly, the IV estimate of the return to education is almost twice as large as the OLS estimate, but the standard error of the IV estimate is over 18 times larger than the OLS standard error. The 95% confidence interval for the IV estimate is between .024 and .239, which is a very wide range. The presence of larger confidence intervals is a price we must pay to get a consistent estimator of the return to education when we think *educ* is endogenous.

TABLE 15.1 Dependent Variable: log(*wage*)

Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations *R*-squared	3,010 .300	3,010 .238
Other controls: *smsa66, reg662, …, reg669*		

As discussed earlier, we should not make anything of the smaller *R*-squared in the IV estimation: by definition, the OLS *R*-squared will always be larger because OLS minimizes the sum of squared residuals.

It is worth noting, especially for studying the effects of policy interventions, that a reduced form equation exists for y_1, too. In the context of equation (15.28) with z_k an IV for y_2, the reduced form for y_1 always has the form

$$y_1 = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_k + e_1, \qquad [15.32]$$

where $\gamma_j = \beta_j + \beta_1\pi_j$ for $j < k$, $\gamma_k = \beta_1\pi_k$, and $e_1 = u_1 + \beta_1 v_2$—as can be verified by plugging (15.30) into (15.28) and rearranging. Because the z_j are exogenous in (15.32), the γ_j can be consistently estimated by OLS. In other words, we regress y_1 on all of the exogenous variables, including z_k, the IV for y_2. Only if we want to estimate β_1 in (15.28) do we need to apply IV.

When y_2 is a zero-one variable denoting participation and z_k is a zero-one variable representing *eligibility* for program participation—which is, hopefully, either randomized across individuals or, at most, a function of the other exogenous variables $z_1, \ldots, z_{k-1}$ (such as income)—the coefficient γ_k has an interesting interpretation. Rather than an estimate of the effect of the program itself, it is an estimate of the effect of *offering* the program. Unlike β_1 in (15.28)—which measures the effect of the program itself—γ_k accounts for the possibility that some units made eligible will choose not to participate. In the program evaluation literature, γ_k is an example of an *intention-to-treat* parameter: it measures the effect of being made *eligible* and not the effect of actual participation. The intention-to-treat coefficient, $\gamma_k = \beta_1\pi_k$, depends on the effect of participating, β_1, and the change (typically, increase) in the probability of participating due to being eligible, π_k. [When y_2 is binary, equation (15.30) is a linear probability model, and therefore π_k measures the ceteris paribus change in probability that $y_2 = 1$ as z_k switches from zero to one.]

15-3 Two Stage Least Squares

In the previous section, we assumed that we had a single endogenous explanatory variable (y_2), along with one instrumental variable for y_2. It often happens that we have more than one exogenous variable that is excluded from the structural model and might be correlated with y_2, which means they are valid IVs for y_2. In this section, we discuss how to use multiple instrumental variables.

15-3a A Single Endogenous Explanatory Variable

Consider again the structural model (15.22), which has one endogenous and one exogenous explanatory variable. Suppose now that we have *two* exogenous variables excluded from (15.22): z_2 and z_3. Our assumptions that z_2 and z_3 do not appear in (15.22) and are uncorrelated with the error u_1 are known as **exclusion restrictions**.

If z_2 and z_3 are both correlated with y_2, we could just use each as an IV, as in the previous section. But then we would have two IV estimators, and neither of these would, in general, be efficient. Since each of z_1, z_2, and z_3 is uncorrelated with u_1, any linear combination is also uncorrelated with u_1, and therefore any linear combination of the exogenous variables is a valid IV. To find the best IV, we choose the linear combination that is most highly correlated with y_2. This turns out to be given by the reduced form equation for y_2. Write

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2, \quad [15.33]$$

where

$$\mathrm{E}(v_2) = 0,\ \mathrm{Cov}(z_1,v_2) = 0,\ \mathrm{Cov}(z_2,v_2) = 0,\ \text{and}\ \mathrm{Cov}(z_3,v_2) = 0.$$

Then, the best IV for y_2 (under the assumptions given in the chapter appendix) is the linear combination of the z_j in (15.33), which we call y_2^*:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3. \quad [15.34]$$

For this IV not to be perfectly correlated with z_1 we need at least one of π_2 or π_3 to be different from zero:

$$\pi_2 \neq 0 \text{ or } \pi_3 \neq 0. \quad [15.35]$$

This is the key identification assumption, once we assume the z_j are all exogenous. (The value of π_1 is irrelevant.) The structural equation (15.22) is not identified if $\pi_2 = 0$ and $\pi_3 = 0$. We can test H_0: $\pi_2 = 0$ and $\pi_3 = 0$ against (15.35) using an F statistic.

A useful way to think of (15.33) is that it breaks y_2 into two pieces. The first is y_2^*; this is the part of y_2 that is uncorrelated with the error term, u_1. The second piece is v_2, and this part is possibly correlated with u_1—which is why y_2 is possibly endogenous.

Given data on the z_j, we can compute y_2^* for each observation, provided we know the population parameters π_j. This is never true in practice. Nevertheless, as we saw in the previous section, we can always estimate the reduced form by OLS. Thus, using the sample, we regress y_2 on z_1, z_2, and z_3 and obtain the fitted values:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \qquad [15.36]$$

(that is, we have $\hat{y}_{i2}$ for each i). At this point, we should verify that z_2 and z_3 are jointly significant in (15.33) at a reasonably small significance level (no larger than 5%). If z_2 and z_3 are not jointly significant in (15.33), then we are wasting our time with IV estimation.

Once we have $\hat{y}_2$, we can use it as the IV for y_2. The three equations for estimating β_0, β_1, and β_2 are the first two equations of (15.25), with the third replaced by

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0. \qquad [15.37]$$

Solving the three equations in three unknowns gives us the IV estimators.

With multiple instruments, the IV estimator using $\hat{y}_{i2}$ as the instrument is also called the **two stage least squares (2SLS) estimator**. The reason is simple. Using the algebra of OLS, it can be shown that when we use $\hat{y}_2$ as the IV for y_2, the IV estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are *identical* to the OLS estimates from the regression of

$$y_1 \text{ on } \hat{y}_2 \text{ and } z_1. \qquad [15.38]$$

In other words, we can obtain the 2SLS estimator in two stages. The **first stage** is to run the regression in (15.36), where we obtain the fitted values $\hat{y}_2$. The second stage is the OLS regression (15.38). Because we use $\hat{y}_2$ in place of y_2, the 2SLS estimates can differ substantially from the OLS estimates.

Some economists like to interpret the regression in (15.38) as follows. The fitted value, $\hat{y}_2$, is the estimated version of y_2^*, and y_2^* is uncorrelated with u_1. Therefore, 2SLS first "purges" y_2 of its correlation with u_1 before doing the OLS regression in (15.38). We can show this by plugging $y_2 = y_2^* + v_2$ into (15.22):

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2. \qquad [15.39]$$

Now, the composite error $u_1 + \beta_1 v_2$ has zero mean and is uncorrelated with y_2^* and z_1, which is why the OLS regression in (15.38) works.

Most econometrics packages have special commands for 2SLS, so there is no need to perform the two stages explicitly. In fact, in most cases you should avoid doing the second stage manually, as the standard errors and test statistics obtained in this way are *not* valid. [The reason is that the error term in (15.39) includes v_2, but the standard errors involve the variance of u_1 only.] Any regression software that supports 2SLS asks for the dependent variable, the list of explanatory variables (both exogenous and endogenous), and the entire list of instrumental variables (that is, all exogenous variables). The output is typically quite similar to that for OLS.

In model (15.28) with a single IV for y_2, the IV estimator from Section 15-2 is identical to the 2SLS estimator. Therefore, when we have one IV for each endogenous explanatory variable, we can call the estimation method IV or 2SLS.

Adding more exogenous variables changes very little. For example, suppose the wage equation is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1, \qquad [15.40]$$

where u_1 is uncorrelated with both *exper* and $exper^2$. Suppose that we also think mother's and father's educations are uncorrelated with u_1. Then, we can use both of these as IVs for *educ*. The reduced form (or first stage equation) equation for *educ* is

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2, \qquad [15.41]$$

and identification requires that $\pi_3 \neq 0$ or $\pi_4 \neq 0$ (or both, of course).

EXAMPLE 15.5 Return to Education for Working Women

We estimate equation (15.40) using the data in MROZ. First, we test H_0: $\pi_3 = 0$, $\pi_4 = 0$ in (15.41) using an F test. The result is $F = 124.76$, and p-value $= .0000$. As expected, *educ* is (partially) correlated with parents' education.

When we estimate (15.40) by 2SLS, we obtain, in equation form,

$$\widehat{\log(wage)} = \underset{(.400)}{.048} + \underset{(.031)}{.061}\, educ + \underset{(.013)}{.044}\, exper - \underset{(.0004)}{.0009}\, exper^2$$
$$n = 428,\ R^2 = .136.$$

The estimated return to education is about 6.1%, compared with an OLS estimate of about 10.8%. Because of its relatively large standard error, the 2SLS estimate is barely statistically significant at the 5% level against a two-sided alternative.

The assumptions needed for 2SLS to have the desired large sample properties are given in the chapter appendix, but it is useful to briefly summarize them here. If we write the structural equation as in (15.28),

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1, \qquad [15.42]$$

then we assume each z_j to be uncorrelated with u_1. In addition, we need at least one exogenous variable *not* in (15.42) that is partially correlated with y_2. This ensures consistency. For the usual 2SLS standard errors and t statistics to be asymptotically valid, we also need a homoskedasticity assumption: the variance of the structural error, u_1, cannot depend on any of the exogenous variables. For time series applications, we need more assumptions, as we will see in Section 15-7.

15-3b Multicollinearity and 2SLS

In Chapter 3, we introduced the problem of multicollinearity and showed how correlation among regressors can lead to large standard errors for the OLS estimates. Multicollinearity can be even more serious with 2SLS. To see why, the (asymptotic) variance of the 2SLS estimator of β_1 can be approximated as

$$\sigma^2/[\widehat{\text{SST}}_2(1 - \hat{R}_2^2)], \qquad [15.43]$$

where $\sigma^2 = \text{Var}(u_1)$, $\widehat{\text{SST}}_2$ is the total variation in $\hat{y}_2$, and $\hat{R}_2^2$ is the R-squared from a regression of $\hat{y}_2$ on all other exogenous variables appearing in the structural equation. There are two reasons why the variance of the 2SLS estimator is larger than that for OLS. First, $\hat{y}_2$, by construction, has less variation than y_2. (Remember: Total sum of squares = explained sum of squares + residual sum of squares; the variation in y_2 is the total sum of squares, while the variation in $\hat{y}_2$ is the explained sum of squares from the first stage regression.) Second, the correlation between $\hat{y}_2$ and the exogenous variables in (15.42) is often much higher than the correlation between y_2 and these variables. This essentially defines the multicollinearity problem in 2SLS.

As an illustration, consider Example 15.4. When *educ* is regressed on the exogenous variables in Table 15.1 (not including *nearc4*), R-squared $= .475$; this is a moderate degree of multicollinearity, but the important thing is that the OLS standard error on $\hat{\beta}_{educ}$ is quite small. When we obtain the first stage fitted values, $\widehat{educ}$, and regress these on the exogenous variables in Table 15.1, R-squared $= .995$, which indicates a very high degree of multicollinearity between $\widehat{educ}$ and the remaining exogenous variables in the table. (This high R-squared is not too surprising because $\widehat{educ}$ is a function of all the exogenous variables in Table 15.1, plus *nearc4*.) Equation (15.43) shows that an $\hat{R}_2^2$ close to one can result in a very large standard error for the 2SLS estimator. But as with OLS, a large sample size can help offset a large $\hat{R}_2^2$.

15-3c Detecting Weak Instruments

In Section 15-1 we briefly discussed the problem of weak instruments. We focused on equation (15.19), which demonstrates how a small correlation between the instrument and error can lead to very large inconsistency (and therefore bias) if the instrument, z, also has little correlation with the explanatory variable, x. The same problem can arise in the context of the multiple equation model in equation (15.42), whether we have one instrument for y_2 or more instruments than we need.

We also mentioned the findings of Staiger and Stock (1997), and we now discuss the practical implications of this research in a bit more depth. Importantly, Staiger and Stock study the case of where all instrumental variables are exogenous. With the exogeneity requirement satisfied by the instruments, they focus on the case where the instruments are weakly correlated with y_2, and they study the validity of standard errors, confidence intervals, and t statistics involving the coefficient β_1 on y_2. The mechanism they used to model weak correlation led to an important finding: even with very large sample sizes the 2SLS estimator can be biased and a distribution that is very different from standard normal.

Building on Staiger and Stock (1997), Stock and Yogo (2005) (SY for short) proposed methods for detecting situations where weak instruments will lead to substantial bias and distorted statistical inference. Conveniently, Stock and Yogo obtained rules concerning the size of the t statistic (with one instrument) or the F statistic (with more than one instrument) from the first-stage regression. The theory is much too involved to pursue here. Instead, we describe some simple rules of thumb proposed by Stock and Yogo that are easy to implement.

The key implication of the SY work is that one needs more than just a statistical rejection of the null hypothesis in the first stage regression at the usual significance levels. For example, in equation (15.6), it is not enough to reject the null hypothesis stated in (15.7) at the 5% significance level. Using bias calculations for the instrumental variables estimator, SY recommend that one can proceed with the usual IV inference if the first-stage t statistic has absolute value larger than $\sqrt{10} \approx 3.2$. Readers will recognize this value as being well above the 95$^{\text{th}}$ percentile of the standard normal distribution, 1.96, which is what we would use for a standard 5% significance level. This same rule of thumb applies in the multiple regression model with a single endogenous explanatory variable, y_2, and a single instrumental variable, z_k. In particular, the t statistic in testing hypothesis (15.31) should be at least 3.2 in absolute value.

SY cover the case of 2SLS, too. In this case, we must focus on the first-stage F statistic for exclusion of the instrumental variables for y_2, and the SY rule is $F > 10$. (Notice this is the same rule based on the t statistic when there is only one instrument, as $t^2 = F$.) For example, consider equation (15.34), where we have two instruments for y_2, z_2 and z_3. Then the F statistic for the null hypothesis

$$H_0\colon \pi_2 = 0, \pi_3 = 0$$

should have $F > 10$. Remember, this is not the overall F statistic for all of the exogenous variables in (15.34). We test only the coefficients on the proposed IVs for y_2, that is, the exogenous variables that do not appear in (15.22). In Example 15.5 the relevant F statistic is 124.76, which is well above 10, implying that we do not have to worry about weak instruments. (Of course, the exogeneity of the parents' education variables is in doubt.)

The rule of thumb of requiring the F statistic to be larger than 10 tends to work well and is easy to remember. However, like all rules of thumb involving statistical inference, it makes no sense to use 10 as a knife-edge cutoff. For example, one can probably proceed if $F = 9.94$, as it is pretty close to 10. The rule of thumb should be used as a guideline. SY have more detailed suggestions for cases where there are many instruments for y_2, say five or more.

A more complicated issue is what happens if there is heteroskedasticity in either the equation of interest, (15.28), or the reduced form (first stage) for the endogenous explanatory variables, (15.30). Stock and Yogo (2005) did not allow for heteroskedasticity in either equation (or, in a time series or panel context, serial correlation). It makes sense that the requirements for the first-stage t or F statistic would be more stringent. Work by Olea and Pflueger (2013) suggests this is the case: the first-stage F might need to be more like 20 rather than 10 in order to ensure the instruments are sufficiently strong. This is an ongoing area of research.

15-3d Multiple Endogenous Explanatory Variables

Two stage least squares can also be used in models with more than one endogenous explanatory variable. For example, consider the model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1, \quad [15.44]$$

where $\mathrm{E}(u_1) = 0$ and u_1 is uncorrelated with z_1, z_2, and z_3. The variables y_2 and y_3 are endogenous explanatory variables: each may be correlated with u_1.

To estimate (15.44) by 2SLS, we need *at least two* exogenous variables that do not appear in (15.44) but that are correlated with y_2 and y_3. Suppose we have two excluded exogenous variables, say z_4 and z_5. Then, from our analysis of a single endogenous explanatory variable, we need either z_4 or z_5 to appear in each reduced form for y_2 and y_3. (As before, we can use F statistics to test this.) Although this is necessary for identification, unfortunately, it is not sufficient. Suppose that z_4 appears in each reduced form, but z_5 appears in neither. Then, we do not really have two exogenous variables partially correlated with y_2 and y_3. Two stage least squares will not produce consistent estimators of the β_j.

Generally, when we have more than one endogenous explanatory variable in a regression model, identification can fail in several complicated ways. But we can easily state a necessary condition for identification, which is called the **order condition**.

GOING FURTHER 15.3

The following model explains violent crime rates, at the city level, in terms of a binary variable for whether gun control laws exist and other controls:

$$\begin{aligned} violent = {} & \beta_0 + \beta_1 guncontrol + \beta_2 unem \\ & + \beta_3 popul + \beta_4 percblck \\ & + \beta_5 age18\_21 + \dots . \end{aligned}$$

Some researchers have estimated similar equations using variables such as the number of National Rifle Association members in the city and the number of subscribers to gun magazines as instrumental variables for *guncontrol* [see, for example, Kleck and Patterson (1993)]. Are these convincing instruments?

Order Condition for Identification of an Equation. We need at least as many excluded exogenous variables as there are included endogenous explanatory variables in the structural equation. The order condition is simple to check, as it only involves counting endogenous and exogenous variables. The sufficient condition for identification is called the **rank condition**. We have seen special cases of the rank condition before—for example, in the discussion surrounding equation (15.35). A general statement of the rank condition requires matrix algebra and is beyond the scope of this text. [See Wooldridge (2010, Chapter 5).] It is even more difficult to obtain diagnostics for weak instruments.

15-3e Testing Multiple Hypotheses after 2SLS Estimation

We must be careful when testing multiple hypotheses in a model estimated by 2SLS. It is tempting to use either the sum of squared residuals or the R-squared form of the F statistic, as we learned with OLS in Chapter 4. The fact that the R-squared in 2SLS can be negative suggests that the usual way of computing F statistics might not be appropriate; this is the case. In fact, if we use the 2SLS residuals to compute the SSRs for both the restricted and unrestricted models, there is no guarantee that $\mathrm{SSR}_r \geq \mathrm{SSR}_{ur}$; if the reverse is true, the F statistic would be negative.

It is possible to combine the sum of squared residuals from the second stage regression [such as (15.38)] with SSR_{ur} to obtain a statistic with an approximate F distribution in large samples. Because many econometrics packages have simple-to-use test commands that can be used to test multiple hypotheses after 2SLS estimation, we omit the details. Davidson and MacKinnon (1993) and Wooldridge (2010, Chapter 5) contain discussions of how to compute F-type statistics for 2SLS.

15-4 IV Solutions to Errors-in-Variables Problems

In the previous sections, we presented the use of instrumental variables as a way to solve the omitted variables problem, but they can also be used to deal with the measurement error problem. As an illustration, consider the model

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u, \quad [15.45]$$

where y and x_2 are observed but x_1^* is not. Let x_1 be an observed measurement of x_1^*: $x_1 = x_1^* + e_1$, where e_1 is the measurement error. In Chapter 9, we showed that correlation between x_1 and e_1 causes OLS, where x_1 is used in place of x_1^*, to be biased and inconsistent. We can see this by writing

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1). \quad [15.46]$$

If the classical errors-in-variables (CEV) assumptions hold, the bias in the OLS estimator of β_1 is toward zero. Without further assumptions, we can do nothing about this.

In some cases, we can use an IV procedure to solve the measurement error problem. In (15.45), we assume that u is uncorrelated with x_1^*, x_1, and x_2; in the CEV case, we assume that e_1 is uncorrelated with x_1^* and x_2. These imply that x_2 is exogenous in (15.46), but that x_1 is correlated with e_1. What we need is an IV for x_1. Such an IV must be correlated with x_1, uncorrelated with u—so that it can be excluded from (15.45)—and uncorrelated with the measurement error, e_1.

One possibility is to obtain a second measurement on x_1^*, say, z_1. Because it is x_1^* that affects y, it is only natural to assume that z_1 is uncorrelated with u. If we write $z_1 = x_1^* + a_1$, where a_1 is the measurement error in z_1, then we must assume that a_1 and e_1 are uncorrelated. In other words, x_1 and z_1 both mismeasure x_1^*, but their measurement errors are uncorrelated. Certainly, x_1 and z_1 are correlated through their dependence on x_1^*, so we can use z_1 as an IV for x_1.

Where might we get two measurements on a variable? Sometimes, when a group of workers is asked for their annual salary, their employers can provide a second measure. For married couples, each spouse can independently report the level of savings or family income. In the Ashenfelter and Krueger (1994) study cited in Section 14-3, each twin was asked about his or her sibling's years of education; this gives a second measure that can be used as an IV for self-reported education in a wage equation. (Ashenfelter and Krueger combined differencing and IV to account for the omitted ability problem as well; more on this in Section 15-8.) Generally, though, having two measures of an explanatory variable is rare.

An alternative is to use other exogenous variables as IVs for a potentially mismeasured variable. For example, our use of *motheduc* and *fatheduc* as IVs for *educ* in Example 15.5 can serve this purpose. If we think that $educ = educ^* + e_1$, then the IV estimates in Example 15.5 do not suffer from measurement error if *motheduc* and *fatheduc* are uncorrelated with the measurement error, e_1. This is probably more reasonable than assuming *motheduc* and *fatheduc* are uncorrelated with ability, which is contained in u in (15.45).

IV methods can also be adopted when using things like test scores to control for unobserved characteristics. In Section 9-2, we showed that, under certain assumptions, proxy variables can be used to solve the omitted variables problem. In Example 9.3, we used IQ as a proxy variable for unobserved ability. This simply entails adding IQ to the model and performing an OLS regression. But there is an alternative that works when IQ does not fully satisfy the proxy variable assumptions. To illustrate, write a wage equation as

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \quad [15.47]$$

where we again have the omitted ability problem. But we have two test scores that are *indicators* of ability. We assume that the scores can be written as

$$test_1 = \gamma_1 abil + e_1$$

and

$$test_2 = \delta_1 abil + e_2,$$

where $\gamma_1 > 0$, $\delta_1 > 0$. Since it is ability that affects wage, we can assume that $test_1$ and $test_2$ are uncorrelated with u. If we write *abil* in terms of the first test score and plug the result into (15.47), we get

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test_1 + (u - \alpha_1 e_1), \qquad [15.48]$$

where $\alpha_1 = 1/\gamma_1$. Now, if we assume that e_1 is uncorrelated with all the explanatory variables in (15.47), including *abil*, then e_1 and $test_1$ *must* be correlated. [Notice that *educ* is *not* endogenous in (15.48); however, $test_1$ is.] This means that estimating (15.48) by OLS will produce inconsistent estimators of the β_j (and α_1). Under the assumptions we have made, $test_1$ does not satisfy the proxy variable assumptions.

If we assume that e_2 is also uncorrelated with all the explanatory variables in (15.47) *and* that e_1 and e_2 are uncorrelated, then e_1 is uncorrelated with the second test score, $test_2$. Therefore, $test_2$ can be used as an IV for $test_1$.

EXAMPLE 15.6 Using Two Test Scores as Indicators of Ability

We use the data in WAGE2 to implement the preceding procedure, where *IQ* plays the role of the first test score and *KWW* (knowledge of the world of work) is the second test score. The explanatory variables are the same as in Example 9.3: *educ*, *exper*, *tenure*, *married*, *south*, *urban*, and *black*. Rather than adding *IQ* and doing OLS, as in column (2) of Table 9.2, we add *IQ* and use *KWW* as its instrument. The coefficient on *educ* is .025 (se = .017). This is a low estimate, and it is not statistically different from zero. This is a puzzling finding, and it suggests that one of our assumptions fails; perhaps e_1 and e_2 are correlated.

15-5 Testing for Endogeneity and Testing Overidentifying Restrictions

In this section, we describe two important tests in the context of instrumental variables estimation.

15-5a Testing for Endogeneity

The 2SLS estimator is less efficient than OLS when the explanatory variables are exogenous; as we have seen, the 2SLS estimates can have very large standard errors. Therefore, it is useful to have a test for endogeneity of an explanatory variable that shows whether 2SLS is even necessary. Obtaining such a test is rather simple.

To illustrate, suppose we have a single suspected endogenous variable,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1, \qquad [15.49]$$

where z_1 and z_2 are exogenous. We have two additional exogenous variables, z_3 and z_4, which do not appear in (15.49). If y_2 is uncorrelated with u_1, we should estimate (15.49) by OLS. How can we test this? Hausman (1978) suggested directly comparing the OLS and 2SLS estimates and determining whether the differences are statistically significant. After all, both OLS and 2SLS are consistent if all variables are exogenous. If 2SLS and OLS differ significantly, we conclude that y_2 must be endogenous (maintaining that the z_j are exogenous).

It is a good idea to compute OLS and 2SLS to see if the estimates are practically different. To determine whether the differences are statistically significant, it is easier to use a regression test. This is based on estimating the reduced form for y_2, which in this case is

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2. \qquad [15.50]$$

Now, since each z_j is uncorrelated with u_1, y_2 is uncorrelated with u_1 if, and only if, v_2 is uncorrelated with u_1; this is what we wish to test. Write $u_1 = \delta_1 v_2 + e_1$, where e_1 is uncorrelated with v_2 and has zero mean. Then, u_1 and v_2 are uncorrelated if, and only if, $\delta_1 = 0$. The easiest way to test this is to include v_2 as an additional regressor in (15.49) and to do a t test. There is only one problem with implementing this: v_2 is not observed, because it is the error term in (15.50). Because we can estimate the reduced form for y_2 by OLS, we can obtain the reduced form residuals, $\hat{v}_2$. Therefore, we estimate

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error \tag{15.51}$$

by OLS and test H_0: $\delta_1 = 0$ using a t statistic. If we reject H_0 at a small significance level, we conclude that y_2 is endogenous because v_2 and u_1 are correlated.

Testing for Endogeneity of a Single Explanatory Variable:

(i) Estimate the reduced form for y_2 by regressing it on *all* exogenous variables (including those in the structural equation and the additional IVs). Obtain the residuals, $\hat{v}_2$.

(ii) Add $\hat{v}_2$ to the structural equation (which includes y_2) and test for significance of $\hat{v}_2$ using an OLS regression. If the coefficient on $\hat{v}_2$ is statistically different from zero, we conclude that y_2 is indeed endogenous. We might want to use a heteroskedasticity-robust t test.

EXAMPLE 15.7 Return to Education for Working Women

We can test for endogeneity of *educ* in (15.40) by obtaining the residuals $\hat{v}_2$ from estimating the reduced form (15.41)—using only working women—and including these in (15.40). When we do this, the coefficient on $\hat{v}_2$ is $\hat{\delta}_1 = .058$, and $t = 1.67$. This is moderate evidence of positive correlation between u_1 and v_2. It is probably a good idea to report both estimates because the 2SLS estimate of the return to education (6.1%) is well below the OLS estimate (10.8%).

An interesting feature of the regression from step (ii) of the test for endogeneity is that the coefficient estimates on all explanatory variables (except, of course, $\hat{v}_2$) are identical to the 2SLS estimates. For example, estimating (15.51) by OLS produces the same $\hat{\beta}_j$ as estimating (15.49) by 2SLS. One benefit of this equivalence is that it provides an easy check on whether you have done the proper regression in testing for endogeneity. But it also gives a different, useful interpretation of 2SLS: adding $\hat{v}_2$ to the original equation as an explanatory variable, and applying OLS, clears up the endogeneity of y_2. So, when we start by estimating (15.49) by OLS, we can quantify the importance of allowing y_2 to be endogenous by seeing how much $\hat{\beta}_1$ changes when $\hat{v}_2$ is added to the equation. Irrespective of the outcome of the statistical tests, we can see whether the change in $\hat{\beta}_1$ is expected and is practically significant.

If, in the end, the 2SLS estimates are chosen, one should obtain the standard errors using built-in 2SLS routines rather than those from regression (15.51). The standard errors obtained from the OLS regression (15.51) are valid only under the null hypothesis $\delta_1 = 0$.

We can also test for endogeneity of multiple explanatory variables. For each suspected endogenous variable, we obtain the reduced form residuals, as in part (i). Then, we test for joint significance of these residuals in the structural equation, using an F test. Joint significance indicates that at least one suspected explanatory variable is endogenous. The number of exclusion restrictions tested is the number of suspected endogenous explanatory variables.

15-5b Testing Overidentification Restrictions

When we introduced the simple instrumental variables estimator in Section 15-1, we emphasized that the instrument must satisfy two requirements: it must be uncorrelated with the error (exogeneity) and correlated with the endogenous explanatory variable (relevance). We have now seen that, even in models with additional explanatory variables, the second requirement can be tested using a t test (with just one instrument) or an F test (when there are multiple instruments). In the context of the simple IV estimator, we noted that the exogeneity requirement cannot be tested. However, if we have more instruments than we need, we can effectively test whether some of them are uncorrelated with the structural error.

As a specific example, again consider equation (15.49) with two instrumental variables for y_2, z_3, and z_4. Remember, z_1 and z_2 essentially act as their own instruments. Because we have two instruments for y_2, we can estimate (15.49) using, say, only z_3 as an IV for y_2; let $\check{\beta}_1$ be the resulting IV estimator of β_1. Then, we can estimate (15.49) using only z_4 as an IV for y_2; call this IV estimator $\tilde{\beta}_1$. If all z_j are exogenous, and if z_3 and z_4 are each partially correlated with y_2, then $\check{\beta}_1$ and $\tilde{\beta}_1$ are both consistent for β_1. Therefore, if our logic for choosing the instruments is sound, $\check{\beta}_1$ and $\tilde{\beta}_1$ should differ only by sampling error. Hausman (1978) proposed basing a test of whether z_3 and z_4 are both exogenous on the difference, $\check{\beta}_1 - \tilde{\beta}_1$. Shortly, we will provide a simpler way to obtain a valid test, but, before doing so, we should understand how to interpret the outcome of the test.

If we conclude that $\check{\beta}_1$ and $\tilde{\beta}_1$ are statistically different from one another, then we have no choice but to conclude that either z_3, z_4, or both fail the exogeneity requirement. Unfortunately, we cannot know which is the case (unless we simply assert from the beginning that, say, z_3 is exogenous). For example, if y_2 denotes years of schooling in a log wage equation, z_3 is mother's education, and z_4 is father's education, a statistically significant difference in the two IV estimators implies that one or both of the parents' education variables are correlated with u_1 in (15.54).

Certainly, rejecting that one's instruments are exogenous is serious and requires a new approach. But the more serious, and subtle, problem in comparing IV estimates is that they may be similar even though both instruments fail the exogeneity requirement. In the previous example, it seems likely that if mother's education is positively correlated with u_1, then so is father's education. Therefore, the two IV estimates may be similar even though each is inconsistent. In effect, because the IVs in this example are chosen using similar reasoning, their separate use in IV procedures may very well lead to similar estimates that are nevertheless both inconsistent. The point is that we should not feel especially comfortable if our IV procedures pass the Hausman test.

Another problem with comparing two IV estimates is that often they may seem practically different yet, statistically, we cannot reject the null hypothesis that they are consistent for the same population parameter. For example, in estimating (15.40) by IV using *motheduc* as the only instrument, the coefficient on *educ* is .049 (.037). If we use only *fatheduc* as the IV for *educ*, the coefficient on *educ* is .070 (.034). [Perhaps not surprisingly, the estimate using both parents' education as IVs is in between these two, .061 (.031).] For policy purposes, the difference between 5% and 7% for the estimated return to a year of schooling is substantial. Yet, as shown in Example 15.8, the difference is not statistically significant.

The procedure of comparing different IV estimates of the same parameter is an example of testing **overidentifying restrictions**. The general idea is that we have more instruments than we need to estimate the parameters consistently. In the previous example, we had one more instrument than we need, and this results in one overidentifying restriction that can be tested. In the general case, suppose that we have q more instruments than we need. For example, with one endogenous explanatory variable, y_2, and three proposed instruments for y_2, we have $q = 3 - 1 = 2$ overidentifying restrictions. When q is two or more, comparing several IV estimates is cumbersome. Instead, we can easily compute a test statistic based on the 2SLS residuals. The idea is that, if all instruments are exogenous, the 2SLS residuals should be uncorrelated with the instruments, up to sampling error. But if there are $k + 1$ parameters and $k + 1 + q$ instruments, the 2SLS residuals have a zero mean and are identically uncorrelated with k linear combinations of the instruments. (This algebraic fact contains, as a special case, the fact that the OLS residuals have a zero mean and are uncorrelated with the k explanatory variables.) Therefore, the test checks whether the 2SLS residuals are correlated with q linear functions of the instruments, and we need not decide on the functions; the test does that for us automatically.

The following regression-based test is valid when the homoskedasticity assumption, listed as Assumption 2SLS.5 in the chapter appendix, holds.

Testing Overidentifying Restrictions:

(i) Estimate the structural equation by 2SLS and obtain the 2SLS residuals, $\hat{u}_1$.

(ii) Regress $\hat{u}_1$ on *all exogenous* variables. Obtain the R-squared, say, R_1^2.

(iii) Under the null hypothesis that all IVs are uncorrelated with u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. If nR_1^2 exceeds (say) the 5% critical value in the χ_q^2 distribution, we reject H_0 and conclude that at least some of the IVs are not exogenous.

EXAMPLE 15.8 Return to Education for Working Women

When we use *motheduc* and *fatheduc* as IVs for *educ* in (15.40), we have a single overidentifying restriction. Regressing the 2SLS residuals $\hat{u}_1$ on *exper*, *exper*^{2}, *motheduc*, and *fatheduc* produces $R_1^2 = .0009$. Therefore, $nR_1^2 = 428(.0009) = .3852$, which is a very small value in a χ_1^2 distribution (p-value $= .535$). Therefore, the parents' education variables pass the overidentification test. When we add husband's education to the IV list, we get two overidentifying restrictions, and $nR_1^2 = 1.11$ (p-value $= .574$). Subject to the preceding cautions, it seems reasonable to add *huseduc* to the IV list, as this reduces the standard error of the 2SLS estimate: the 2SLS estimate on *educ* using all three instruments is .080 ($se = .022$), so this makes *educ* much more significant than when *huseduc* is not used as an IV ($\hat{\beta}_{educ} = .061$, se $= .031$).

When $q = 1$, a natural question is: How does the test obtained from the regression-based procedure compare with a test based on directly comparing the estimates? In fact, the two procedures are asymptotically the same. As a practical matter, it makes sense to compute the two IV estimates to see how they differ. More generally, when $q \geq 2$, one can compare the 2SLS estimates using all IVs to the IV estimates using single instruments. By doing so, one can see if the various IV estimates are practically different, whether or not the overidentification test rejects or fails to reject.

In the previous example, we alluded to a general fact about 2SLS: under the standard 2SLS assumptions, adding instruments to the list improves the asymptotic efficiency of the 2SLS. But this requires that any new instruments are in fact exogenous—otherwise, 2SLS will not even be consistent—and it is only an asymptotic result. With the typical sample sizes available, adding too many instruments—that is, increasing the number of overidentifying restrictions—can cause severe biases in 2SLS. A detailed discussion would take us too far afield. A nice illustration is given by Bound, Jaeger, and Baker (1995), who argue that the 2SLS estimates of the return to education obtained by Angrist and Krueger (1991), using many instrumental variables, are likely to be seriously biased (even with hundreds of thousands of observations!).

The overidentification test can be used whenever we have more instruments than we need. If we have just enough instruments, the model is said to be *just identified*, and the R-squared in part (ii) will be identically zero. As we mentioned earlier, we cannot test exogeneity of the instruments in the just identified case.

The test can be made robust to heteroskedasticity of arbitrary form; for details, see Wooldridge (2010, Chapter 5).

15-6 2SLS with Heteroskedasticity

Heteroskedasticity in the context of 2SLS raises essentially the same issues as with OLS. Most importantly, it is possible to obtain standard errors and test statistics that are (asymptotically) robust to heteroskedasticity of arbitrary and unknown form. In fact, expression (8.4) continues to be valid if the $\hat{r}_{ij}$ are obtained as the residuals from regressing $\hat{x}_{ij}$ on the other $\hat{x}_{ih}$, where the "ˆ" denotes fitted values from the first stage regressions (for endogenous explanatory variables). Wooldridge (2010, Chapter 5) contains more details. Some software packages do this routinely.

We can also test for heteroskedasticity, using an analog of the Breusch-Pagan test that we covered in Chapter 8. Let $\hat{u}$ denote the 2SLS residuals and let $z_1, z_2, \ldots, z_m$ denote all the exogenous variables (including those used as IVs for the endogenous explanatory variables). Then, under reasonable assumptions [spelled out, for example, in Wooldridge (2010, Chapter 5)], an asymptotically valid statistic is the usual F statistic for joint significance in a regression of $\hat{u}^2$ on $z_1, z_2, \ldots, z_m$. The null hypothesis of homoskedasticity is rejected if the z_j are jointly significant.

If we apply this test to Example 15.8, using *motheduc*, *fatheduc*, and *huseduc* as instruments for *educ*, we obtain $F_{5,422} = 2.53$ and p-value $= .029$. This is evidence of heteroskedasticity at the 5% level. We might want to compute heteroskedasticity-robust standard errors to account for this.

If we know how the error variance depends on the exogenous variables, we can use a weighted 2SLS procedure, essentially the same as in Section 8-4. After estimating a model for $\text{Var}(u|z_1, z_2, \ldots, z_m)$, we divide the dependent variable, the explanatory variables, and all the instrumental variables for observation i by $\sqrt{\hat{h}_i}$, where $\hat{h}_i$ denotes the estimated variance. (The constant, which is both an explanatory variable and an IV, is divided by $\sqrt{\hat{h}_i}$; see Section 8-4.) Then, we apply 2SLS on the transformed equation using the transformed instruments.

15-7 Applying 2SLS to Time Series Equations

When we apply 2SLS to time series data, many of the considerations that arose for OLS in Chapters 10, 11, and 12 are relevant. Write the structural equation for each time period as

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \tag{15.52}$$

where one or more of the explanatory variables x_{tj} might be correlated with u_t. Denote the set of exogenous variables by $z_{t1}, \ldots, z_{tm}$:

$$\text{E}(u_t) = 0, \text{Cov}(z_{tj}, u_t) = 0, \quad j = 1, \ldots, m.$$

Any exogenous explanatory variable is also a z_{tj}. For identification, it is necessary that $m \geq k$ (we have as many exogenous variables as explanatory variables).

The mechanics of 2SLS are identical for time series or cross-sectional data, but for time series data the statistical properties of 2SLS depend on the trending and correlation properties of the underlying sequences. In particular, we must be careful to include trends if we have trending dependent or explanatory variables. Since a time trend is exogenous, it can always serve as its own instrumental variable. The same is true of seasonal dummy variables, if monthly or quarterly data are used.

Series that have strong persistence (have unit roots) must be used with care, just as with OLS. Often, differencing the equation is warranted before estimation, and this applies to the instruments as well.

GOING FURTHER 15.4

A model to test the effect of growth in government spending on growth in output is

$$gGDP_t = \beta_0 + \beta_1 gGOV_t + \beta_2 INVRAT_t + \beta_3 gLAB_t + u_t,$$

where g indicates growth, *GDP* is real gross domestic product, GOV is real government spending, *INVRAT* is the ratio of gross domestic investment to GDP, and *LAB* is the size of the labor force. [See equation (6) in Ram (1986).] Under what assumptions would a dummy variable indicating whether the president in year $t - 1$ is a Republican be a suitable IV for $gGOV_t$?

Under analogs of the assumptions in Chapter 11 for the asymptotic properties of OLS, 2SLS using time series data is consistent and asymptotically normally distributed. In fact, if we replace the explanatory variables with the instrumental variables in stating the assumptions, we only need to add the identification assumptions for 2SLS. For example, the homoskedasticity assumption is stated as

$$\text{E}(u_t^2|z_{t1}, \ldots, z_{tm}) = \sigma^2, \tag{15.53}$$

and the no serial correlation assumption is stated as

$$\mathrm{E}(u_t u_s | \mathbf{z}_t, \mathbf{z}_s) = 0 \quad \text{for all } t \neq s, \tag{15.54}$$

where $\mathbf{z}_t$ denotes all exogenous variables at time t. A full statement of the assumptions is given in the chapter appendix. We will provide examples of 2SLS for time series problems in Chapter 16; see also Computer Exercise C4.

As in the case of OLS, the no serial correlation assumption can often be violated with time series data. Fortunately, it is very easy to test for AR(1) serial correlation. If we write $u_t = \rho u_{t-1} + e_t$ and plug this into equation (15.52), we get

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \rho u_{t-1} + e_t, t \geq 2. \tag{15.55}$$

To test H_0: $\rho_1 = 0$, we must replace u_{t-1} with the 2SLS residuals, $\hat{u}_{t-1}$. Further, if x_{tj} is endogenous in (15.52), then it is endogenous in (15.55), so we still need to use an IV. Because e_t is uncorrelated with all past values of u_t, $\hat{u}_{t-1}$ can be used as its own instrument.

Testing for AR(1) Serial Correlation after 2SLS:

(i) Estimate (15.52) by 2SLS and obtain the 2SLS residuals, $\hat{u}_t$.

(ii) Estimate

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \rho \hat{u}_{t-1} + error_t, \quad t = 2, \ldots, n$$

by 2SLS, using the same instruments from part (i), in addition to $\hat{u}_{t-1}$. Use the t statistic on $\hat{\rho}$ to test H_0: $\rho = 0$.

As with the OLS version of this test from Chapter 12, the t statistic only has asymptotic justification, but it tends to work well in practice. A heteroskedasticity-robust version can be used to guard against heteroskedasticity. Further, lagged residuals can be added to the equation to test for higher forms of serial correlation using a joint F test.

What happens if we detect serial correlation? Some econometrics packages will compute standard errors that are robust to fairly general forms of serial correlation and heteroskedasticity. This is a nice, simple way to go if your econometrics package does this. The computations are very similar to those in Section 12-5 for OLS. [See Wooldridge (1995) for formulas and other computational methods.]

An alternative is to use the AR(1) model and correct for serial correlation. The procedure is similar to that for OLS and places additional restrictions on the instrumental variables. The quasi-differenced equation is the same as in equation (12.32):

$$\tilde{y}_t = \beta_0(1 - \rho) + \beta_1 \tilde{x}_{t1} + \cdots + \beta_k \tilde{x}_{tk} + e_t, \quad t \geq 2, \tag{15.56}$$

where $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,j}$. (We can use the $t = 1$ observation just as in Section 12-3, but we omit that for simplicity here.) The question is: What can we use as instrumental variables? It seems natural to use the quasi-differenced instruments, $\tilde{z}_{tj} = z_{tj} - \rho z_{t-1,j}$. This only works, however, if in (15.52) the original error u_t is uncorrelated with the instruments at times t, $t - 1$, and $t + 1$. That is, the instrumental variables must be strictly exogenous in (15.52). This rules out lagged dependent variables as IVs, for example. It also eliminates cases where future movements in the IVs react to current and past changes in the error, u_t.

2SLS with AR(1) Errors:

(i) Estimate (15.52) by 2SLS and obtain the 2SLS residuals, $\hat{u}_t$, $t = 1, 2, \ldots, n$.

(ii) Obtain $\hat{\rho}$ from the regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ $t = 2, \ldots, n$ and construct the quasi-differenced variables $\tilde{y}_t = y_t - \hat{\rho} y_{t-1}$, $\tilde{x}_{tj} = x_{tj} - \hat{\rho} x_{t-1,j}$, and $\tilde{z}_{tj} = z_{tj} - \hat{\rho} z_{t-1,j}$ for $t \geq 2$. (Remember, in most cases, some of the IVs will also be explanatory variables.)

(iii) Estimate (15.56) (where ρ is replaced with $\hat{\rho}$) by 2SLS, using the $\tilde{z}_{tj}$ as the instruments. Assuming that (15.56) satisfies the 2SLS assumptions in the chapter appendix, the usual 2SLS test statistics are asymptotically valid.

We can also use the first time period as in Prais-Winsten estimation of the model with exogenous explanatory variables. The transformed variables in the first time period—the dependent variable, explanatory variables, and instrumental variables—are obtained simply by multiplying all first-period values by $(1 - \hat{\rho})^{1/2}$. (See also Section 12-3.)

15-8 Applying 2SLS to Pooled Cross Sections and Panel Data

Applying instrumental variables methods to independently pooled cross sections raises no new difficulties. As with models estimated by OLS, we should often include time period dummy variables to allow for aggregate time effects. These dummy variables are exogenous—because the passage of time is exogenous—and so they act as their own instruments.

EXAMPLE 15.9 Effect of Education on Fertility

In Example 13.1, we used the pooled cross section in FERTIL1 to estimate the effect of education on women's fertility, controlling for various other factors. As in Sander (1992), we allow for the possibility that *educ* is endogenous in the equation. As instrumental variables for *educ*, we use mother's and father's education levels (*meduc*, *feduc*). The 2SLS estimate of β_{educ} is $-.153$ $(\text{se} = .039)$, compared with the OLS estimate $-.128$ $(\text{se} = .018)$. The 2SLS estimate shows a somewhat larger effect of education on fertility, but the 2SLS standard error is over twice as large as the OLS standard error. (In fact, the 95% confidence interval based on 2SLS easily contains the OLS estimate.) The OLS and 2SLS estimates of β_{educ} are *not statistically* different, as can be seen by testing for endogeneity of *educ* as in Section 15-5: when the reduced form residual, $\hat{v}_2$, is included with the other regressors in Table 13.1 (including *educ*), its *t* statistic is .702, which is not significant at any reasonable level. Therefore, in this case, we conclude that the difference between 2SLS and OLS could be entirely due to sampling error.

Instrumental variables estimation can be combined with panel data methods, particularly first differencing, to estimate parameters consistently in the presence of unobserved effects and endogeneity in one or more time-varying explanatory variables. The following simple example illustrates this combination of methods.

EXAMPLE 15.10 Job Training and Worker Productivity

Suppose we want to estimate the effect of another hour of job training on worker productivity. For the two years 1987 and 1988, consider the simple panel data model

$$\log(scrap_{it}) = \beta_0 + \delta_0 d88_t + \beta_1 hrsemp_{it} + a_i + u_{it},\ t = 1, 2,$$

where $scrap_{it}$ is firm i's scrap rate in year t and $hrsemp_{it}$ is hours of job training per employee. As usual, we allow different year intercepts and a constant, unobserved firm effect, a_i.

For the reasons discussed in Section 13-2, we might be concerned that $hrsemp_{it}$ is correlated with a_i, the latter of which contains unmeasured worker ability. As before, we difference to remove a_i:

$$\Delta\log(scrap_i) = \delta_0 + \beta_1 \Delta hrsemp_i + \Delta u_i. \qquad [15.57]$$

Normally, we would estimate this equation by OLS. But what if Δu_i is correlated with $\Delta hrsemp_i$? For example, a firm might hire more skilled workers, while at the same time reducing the level of job training. In this case, we need an instrumental variable for $\Delta hrsemp_i$. Generally, such an IV would be hard to find, but we can exploit the fact that some firms received job training grants in 1988. If we assume that grant designation is uncorrelated with Δu_i—something that is reasonable, because the grants were given at the beginning of 1988—then $\Delta grant_i$ is valid as an IV, provided $\Delta hrsemp$ and $\Delta grant$ are correlated. Using the data in JTRAIN differenced between 1987 and 1988, the first stage regression is

$$\widehat{\Delta hrsemp} = \underset{(1.56)}{.51} + \underset{(3.13)}{27.88}\,\Delta grant$$
$$n = 45, R^2 = .392.$$

This confirms that the change in hours of job training per employee is strongly positively related to receiving a job training grant in 1988. In fact, receiving a job training grant increased per-employee training by almost 28 hours, and grant designation accounted for almost 40% of the variation in $\Delta hrsemp$. Two stage least squares estimation of (15.57) gives

$$\Delta\log(scrap) = \underset{(.127)}{-.033} - \underset{(.008)}{.014}\,\Delta hrsemp$$
$$n = 45, R^2 = .016.$$

This means that 10 more hours of job training per worker are estimated to reduce the scrap rate by about 14%. For the firms in the sample, the average amount of job training in 1988 was about 17 hours per worker, with a minimum of zero and a maximum of 88.

For comparison, OLS estimation of (15.57) gives $\hat{\beta}_1 = -.0076$ $(\text{se} = .0045)$, so the 2SLS estimate of β_1 is almost twice as large in magnitude and is slightly more statistically significant.

When $T \geq 3$, the differenced equation may contain serial correlation. The same test and correction for AR(1) serial correlation from Section 15-7 can be used, where all regressions are pooled across i as well as t. Because we do not want to lose an entire time period, the Prais-Winsten transformation should be used for the initial time period.

Unobserved effects models containing lagged dependent variables also require IV methods for consistent estimation. The reason is that, after differencing, $\Delta y_{i,t-1}$ is correlated with Δu_{it} because $y_{i,t-1}$ and $u_{i,t-1}$ are correlated. We can use two or more lags of y as IVs for $\Delta y_{i,t-1}$. [See Wooldridge (2010, Chapter 11) for details.]

Instrumental variables after differencing can be used on matched pairs samples as well. Ashenfelter and Krueger (1994) differenced the wage equation across twins to eliminate unobserved ability:

$$\log(wage_2) - \log(wage_1) = \delta_0 + \beta_1(educ_{2,2} - educ_{1,1}) + (u_2 - u_1),$$

where $educ_{1,1}$ is years of schooling for the first twin as reported by the first twin and $educ_{2,2}$ is years of schooling for the second twin as reported by the second twin. To account for possible measurement error in the self-reported schooling measures, Ashenfelter and Krueger used $(educ_{2,1} - educ_{1,2})$ as an IV for $(educ_{2,2} - educ_{1,1})$, where $educ_{2,1}$ is years of schooling for the second twin as reported by the first twin and $educ_{1,2}$ is years of schooling for the first twin as reported by the second twin. The IV estimate of β_1 is $.167(t = 3.88)$, compared with the OLS estimate on the first differences of $.092(t = 3.83)$ [see Ashenfelter and Krueger (1994, Table 3)].

Summary

In Chapter 15, we have introduced the method of instrumental variables as a way to estimate the parameters in a linear model consistently when one or more explanatory variables are endogenous. An instrumental variable must have two properties: (1) it must be exogenous, that is, uncorrelated with the error term of the structural equation; (2) it must be partially correlated with the endogenous explanatory variable. Finding a variable with these two properties is usually challenging.

The method of two stage least squares, which allows for more instrumental variables than we have explanatory variables, is used routinely in the empirical social sciences. When used properly, it can allow us to estimate ceteris paribus effects in the presence of endogenous explanatory variables. This is true in cross-sectional, time series, and panel data applications. But when instruments are poor—which means they are correlated with the error term, only weakly correlated with the endogenous explanatory variable, or both—then 2SLS can be worse than OLS.

When we have valid instrumental variables, we can test whether an explanatory variable is endogenous, using the test in Section 15-5. In addition, though we can never test whether all IVs are exogenous, we can test that at least some of them are—assuming that we have more instruments than we need for consistent estimation (that is, the model is overidentified). Heteroskedasticity and serial correlation can be tested for and dealt with using methods similar to the case of models with exogenous explanatory variables.

In this chapter, we used omitted variables and measurement error to illustrate the method of instrumental variables. IV methods are also indispensable for simultaneous equations models, which we will cover in Chapter 16.

Key Terms

Endogenous Explanatory Variables
Errors-in-Variables
Exclusion Restrictions
Exogenous Explanatory Variables
Exogenous Variables
First Stage
Identification
Instrument
Instrumental Variable
Instrumental Variables (IV) Estimator
Instrument Exogeneity
Instrument Relevance
Natural Experiment
Omitted Variables
Order Condition
Overidentifying Restrictions
Rank Condition
Reduced Form Equation
Structural Equation
Two Stage Least Squares (2SLS) Estimator
Weak Instruments

Problems

1 Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u,$$

where *PC* is a binary variable indicating PC ownership.

(i) Why might PC ownership be correlated with u?

(ii) Explain why *PC* is likely to be related to parents' annual income. Does this mean parental income is a good IV for *PC*? Why or why not?

(iii) Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for *PC*.

2 Suppose that you wish to estimate the effect of class attendance on student performance, as in Example 6.3. A basic model is

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u,$$

where the variables are defined as in Chapter 6.

(i) Let *dist* be the distance from the students' living quarters to the lecture hall. Do you think *dist* is uncorrelated with u?

(ii) Assuming that *dist* and u are uncorrelated, what other assumption must *dist* satisfy to be a valid IV for *atndrte*?

(iii) Suppose, as in equation (6.18), we add the interaction term *priGPA·atndrte*:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \cdot atndrte + u.$$

If *atndrte* is correlated with u, then, in general, so is *priGPA·atndrte*. What might be a good IV for *priGPA·atndrte*? [*Hint*: If $E(u|priGPA, ACT, dist) = 0$, as happens when *priGPA*, *ACT*, and *dist* are all exogenous, then any function of *priGPA* and *dist* is uncorrelated with u.]

3 Consider the simple regression model

$$y = \beta_0 + \beta_1 x + u$$

and let z be a *binary* instrumental variable for x. Use (15.10) to show that the IV estimator $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0),$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of y_i and x_i over the part of the sample with $z_i = 0$, and where $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of y_i and x_i over the part of the sample with $z_i = 1$. This estimator, known as a *grouping estimator*, was first suggested by Wald (1940).

4 Suppose that, for a given state in the United States, you wish to use annual time series data to estimate the effect of the state-level minimum wage on the employment of those 18 to 25 years old (*EMP*). A simple model is

$$gEMP_t = \beta_0 + \beta_1 gMIN_t + \beta_2 gPOP_t + \beta_3 gGSP_t + \beta_4 gGDP_t + u_t,$$

where MIN_t is the minimum wage, in real dollars; POP_t is the population from 18 to 25 years old; GSP_t is gross state product; and GDP_t is U.S. gross domestic product. The g prefix indicates the growth rate from year $t - 1$ to year t, which would typically be approximated by the difference in the logs.

(i) If we are worried that the state chooses its minimum wage partly based on unobserved (to us) factors that affect youth employment, what is the problem with OLS estimation?

(ii) Let $USMIN_t$ be the U.S. minimum wage, which is also measured in real terms. Do you think $gUSMIN_t$ is uncorrelated with u_t?

(iii) By law, any state's minimum wage must be at least as large as the U.S. minimum. Explain why this makes $gUSMIN_t$ a potential IV candidate for $gMIN_t$.

5 Refer to equations (15.19) and (15.20). Assume that $\sigma_u = \sigma_x$, so that the population variation in the error term is the same as it is in x. Suppose that the instrumental variable, z, is slightly correlated with u: $\text{Corr}(z, u) = .1$. Suppose also that z and x have a somewhat stronger correlation: $\text{Corr}(z, x) = .2$.

(i) What is the asymptotic bias in the IV estimator?

(ii) How much correlation would have to exist between x and u before OLS has more asymptotic bias than 2SLS?

6 (i) In the model with one endogenous explanatory variable, one exogenous explanatory variable, and one extra exogenous variable, take the reduced form for y_2 (15.26), and plug it into the structural equation (15.22). This gives the reduced form for y_1:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1.$$

Find the α_j in terms of the β_j and the π_j.

(ii) Find the reduced form error, v_1, in terms of u_1, v_2, and the parameters.

(iii) How would you consistently estimate the a_j?

7 The following is a simple model to measure the effect of a school choice program on standardized test performance [see Rouse (1998) for motivation and Computer Exercise C11 for an analysis of a subset of Rouse's data]:

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

where *score* is the score on a statewide test, *choice* is a binary variable indicating whether a student attended a choice school in the last year, and *faminc* is family income. The IV for *choice* is *grant*, the dollar amount granted to students to use for tuition at choice schools. The grant amount differed by family income level, which is why we control for *faminc* in the equation.

(i) Even with *faminc* in the equation, why might *choice* be correlated with u_1?

(ii) If within each income class, the grant amounts were assigned randomly, is *grant* uncorrelated with u_1?

(iii) Write the reduced form equation for *choice*. What is needed for *grant* to be partially correlated with *choice*?

(iv) Write the reduced form equation for *score*. Explain why this is useful. (*Hint*: How do you interpret the coefficient on *grant*?)

8 Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coed schools. You have a random sample of senior high school girls from a state in the United States, and *score* is the score on a standardized math test. Let *girlhs* be a dummy variable indicating whether a student attends a girls' high school.

(i) What other factors would you control for in the equation? (You should be able to reasonably collect data on these factors.)

(ii) Write an equation relating *score* to *girlhs* and the other factors you listed in part (i).

(iii) Suppose that parental support and motivation are unmeasured factors in the error term in part (ii). Are these likely to be correlated with *girlhs*? Explain.

(iv) Discuss the assumptions needed for the number of girls' high schools within a 20-mile radius of a girl's home to be a valid IV for *girlhs*.

(v) Suppose that, when you estimate the reduced form for *girlshs*, you find that the coefficient on *numghs* (the number of girls' high schools within a 20-mile radius) is negative and statistically significant. Would you feel comfortable proceeding with IV estimation where *numghs* is used as an IV for *girlshs*? Explain.

9 Suppose that, in equation (15.8), you do not have a good instrumental variable candidate for *skipped*. But you have two other pieces of information on students: combined SAT score and cumulative GPA prior to the semester. What would you do instead of IV estimation?

10 In a recent article, Evans and Schwab (1995) studied the effects of attending a Catholic high school on the probability of attending college. For concreteness, let *college* be a binary variable equal to unity if a student attends college, and zero otherwise. Let *CathHS* be a binary variable equal to one if the student attends a Catholic high school. A linear probability model is

$$college = \beta_0 + \beta_1 CathHS + other\ factors + u,$$

where the other factors include gender, race, family income, and parental education.

(i) Why might *CathHS* be correlated with u?

(ii) Evans and Schwab have data on a standardized test score taken when each student was a sophomore. What can be done with this variable to improve the ceteris paribus estimate of attending a Catholic high school?

(iii) Let *CathRel* be a binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid IV for *CathHS* in the preceding equation. Which of these can be tested?

(iv) Not surprisingly, being Catholic has a significant positive effect on attending a Catholic high school. Do you think *CathRel* is a convincing instrument for *CathHS*?

11 Consider a simple time series model where the explanatory variable has classical measurement error:

$$y_t = \beta_0 + \beta_1 x_t^* + u_t \qquad [15.58]$$
$$x_t = x_t^* + e_t,$$

where u_t has zero mean and is uncorrelated with x_t^* and e_t. We observe y_t and x_t only. Assume that e_t has zero mean and is uncorrelated with x_t^* and that x_t^* also has a zero mean (this last assumption is only to simplify the algebra).

(i) Write $x_t^* = x_t - e_t$ and plug this into (15.58). Show that the error term in the new equation, say, v_t, is negatively correlated with x_t if $\beta_1 > 0$. What does this imply about the OLS estimator of β_1 from the regression of y_t on x_t?

(ii) In addition to the previous assumptions, assume that u_t and e_t are uncorrelated with all past values of x_t^* and e_t; in particular, with x^*_{t-1} and e_{t-1}. Show that $\mathrm{E}(x_{t-1}v_t) = 0$ where v_t is the error term in the model from part (i).

(iii) Are x_t and x_{t-1} likely to be correlated? Explain.

(iv) What do parts (ii) and (iii) suggest as a useful strategy for consistently estimating β_0 and β_1?

Computer Exercises

C1 Use the data in WAGE2 for this exercise.

(i) In Example 15.2, if *sibs* is used as an instrument for *educ*, the IV estimate of the return to education is .122. To convince yourself that using *sibs* as an IV for *educ* is *not* the same as just plugging *sibs* in for *educ* and running an OLS regression, run the regression of log(*wage*) on *sibs* and explain your findings.

(ii) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on). Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.

(iii) Use *brthord* as an IV for *educ* in equation (15.1). Report and interpret the results.

(iv) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u.$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v.$$

State and test the identification assumption.

(v) Estimate the equation from part (iv) using *brthord* as an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for $\hat{\beta}_{educ}$ and $\hat{\beta}_{sibs}$.

(vi) Using the fitted values from part (iv), $\widehat{educ}$, compute the correlation between $\widehat{educ}$ and *sibs*. Use this result to explain your findings from part (v).

C2 The data in FERTIL2 include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

(i) Estimate the model

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

by OLS and interpret the estimates. In particular, holding *age* fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

(ii) The variable *frsthalf* is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that *frsthalf* is uncorrelated with the error term from part (i), show that *frsthalf* is a reasonable IV candidate for *educ*. (*Hint*: You need to do a regression.)

(iii) Estimate the model from part (i) by using *frsthalf* as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from part (i).

(iv) Add the binary variables *electric*, *tv*, and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

C3 Use the data in CARD for this exercise.

(i) The equation we estimated in Example 15.4 can be written as

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \cdots + u,$$

where the other explanatory variables are listed in Table 15.1. In order for IV to be consistent, the IV for *educ*, *nearc4*, must be uncorrelated with u. Could *nearc4* be correlated with things in the error term, such as unobserved ability? Explain.

(ii) For a subsample of the men in the data set, an IQ score is available. Regress *IQ* on *nearc4* to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

(iii) Now, regress *IQ* on *nearc4*, *smsa66*, and the 1966 regional dummy variables *reg662*, . . . , *reg669*. Are *IQ* and *nearc4* related after the geographic dummy variables have been partialled out? Reconcile this with your findings from part (ii).

(iv) From parts (ii) and (iii), what do you conclude about the importance of controlling for *smsa66* and the 1966 regional dummies in the log(*wage*) equation?

C4 Use the data in INTDEF for this exercise. A simple equation relating the three-month T-bill rate to the inflation rate (constructed from the Consumer Price Index) is

$$i3_t = \beta_0 + \beta_1 inf_t + u_t.$$

(i) Estimate this equation by OLS, omitting the first time period for later comparisons. Report the results in the usual form.

(ii) Some economists feel that the Consumer Price Index mismeasures the true rate of inflation, so that the OLS from part (i) suffers from measurement error bias. Reestimate the equation from part (i), using inf_{t-1} as an IV for inf_t. How does the IV estimate of β_1 compare with the OLS estimate?

(iii) Now, first difference the equation:

$$\Delta i3_t = \beta_0 + \beta_1 \Delta inf_t + \Delta u_t.$$

Estimate this by OLS and compare the estimate of β_1 with the previous estimates.

(iv) Can you use Δinf_{t-1} as an IV for Δinf_t in the differenced equation in part (iii)? Explain. (*Hint*: Are Δinf_t and Δinf_{t-1} sufficiently correlated?)

C5 Use the data in CARD for this exercise.

(i) In Table 15.1, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, $\hat{v}_2$, from the reduced form regression *educ* on *nearc*4, *exper*, *exper*^{2}, *black*, *smsa*, *south*, *smsa*66, *reg*662, …, *reg*669—see Table 15.1. Use these to test whether *educ* is exogenous; that is, determine if the difference between OLS and IV is *statistically* significant.

(ii) Estimate the equation by 2SLS, adding *nearc2* as an instrument. Does the coefficient on *educ* change much?

(iii) Test the single overidentifying restriction from part (ii).

C6 Use the data in MURDER for this exercise. The variable *mrdrte* is the murder rate, that is, the number of murders per 100,000 people. The variable *exec* is the total number of prisoners executed for the current and prior two years; *unem* is the state unemployment rate.

(i) How many states executed at least one prisoner in 1991, 1992, or 1993? Which state had the most executions?

(ii) Using the two years 1990 and 1993, do a pooled regression of *mrdrte* on *d93*, *exec*, and *unem*. What do you make of the coefficient on *exec*?

(iii) Using the changes from 1990 to 1993 only (for a total of 51 observations), estimate the equation

$$\Delta mrdrte = \delta_0 + \beta_1 \Delta exec + \beta_2 \Delta unem + \Delta u$$

by OLS and report the results in the usual form. Now, does capital punishment appear to have a deterrent effect?

(iv) The change in executions may be at least partly related to changes in the expected murder rate, so that $\Delta exec$ is correlated with Δu in part (iii). It might be reasonable to assume that $\Delta exec_{-1}$ is uncorrelated with Δu. (After all, $\Delta exec_{-1}$ depends on executions that occurred three or more years ago.) Regress $\Delta exec$ on $\Delta exec_{-1}$ to see if they are sufficiently correlated; interpret the coefficient on $\Delta exec_{-1}$.

(v) Reestimate the equation from part (iii), using $\Delta exec_{-1}$ as an IV for $\Delta exec$. Assume that $\Delta unem$ is exogenous. How do your conclusions change from part (iii)?

C7 Use the data in PHILLIPS for this exercise.

(i) In Example 11.5, we estimated an expectations augmented Phillips curve of the form

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t,$$

where $\Delta inf_t = inf_t - inf_{t-1}$. In estimating this equation by OLS, we assumed that the supply shock, e_t, was uncorrelated with $unem_t$. If this is false, what can be said about the OLS estimator of β_1?

(ii) Suppose that e_t is unpredictable given all past information: $\mathrm{E}(e_t | inf_{t-1}, unem_{t-1}, \ldots) = 0$. Explain why this makes $unem_{t-1}$ a good IV candidate for $unem_t$.

(iii) Regress $unem_t$ on $unem_{t-1}$. Are $unem_t$ and $unem_{t-1}$ significantly correlated?

(iv) Estimate the expectations augmented Phillips curve by IV. Report the results in the usual form and compare them with the OLS estimates from Example 11.5.

C8 Use the data in 401KSUBS for this exercise. The equation of interest is a linear probability model:

$$pira = \beta_0 + \beta_1 p401k + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u.$$

The goal is to test whether there is a tradeoff between participating in a 401(k) plan and having an individual retirement account (IRA). Therefore, we want to estimate β_1.

(i) Estimate the equation by OLS and discuss the estimated effect of *p401k*.

(ii) For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be a problem with ordinary least squares?

(iii) The variable *e401k* is a binary variable equal to one if a worker is *eligible* to participate in a 401(k) plan. Explain what is required for *e401k* to be a valid IV for *p401k*. Do these assumptions seem reasonable?

(iv) Estimate the reduced form for *p401k* and verify that *e401k* has significant partial correlation with *p401k*. Since the reduced form is also a linear probability model, use a heteroskedasticity-robust standard error.

(v) Now, estimate the structural equation by IV and compare the estimate of β_1 with the OLS estimate. Again, you should obtain heteroskedasticity-robust standard errors.

(vi) Test the null hypothesis that *p401k* is in fact exogenous, using a heteroskedasticity-robust test.

C9 The purpose of this exercise is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Use the data file WAGE2.

(i) Use a 2SLS routine to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u,$$

where *sibs* is the IV for *educ*. Report the results in the usual form.

(ii) Now, manually carry out 2SLS. That is, first regress $educ_i$ on $sibs_i$, $exper_i$, $tenure_i$, and $black_i$ and obtain the fitted values, $\widehat{educ}_i$, $i = 1, \ldots, n$. Then, run the second stage regression $\log(wage_i)$ on $\widehat{educ}_i$, $exper_i$, $tenure_i$, and $black_i$, $i = 1, \ldots, n$. Verify that the $\hat{\beta}_j$ are identical to those obtained from part (i), but that the standard errors are somewhat different. The standard errors obtained from the second stage regression when manually carrying out 2SLS are generally inappropriate.

(iii) Now, use the following two-step procedure, which generally yields inconsistent parameter estimates of the β_j, and not just inconsistent standard errors. In step one, regress $educ_i$ on $sibs_i$ only and obtain the fitted values, say $\widetilde{educ}_i$. (Note that this is an incorrect first stage regression.) Then, in the second step, run the regression of $\log(wage_i)$ on $\widetilde{educ}_i$, $exper_i$, $tenure_i$, and $black_i$, $i = 1, \ldots, n$. How does the estimate from this incorrect, two-step procedure compare with the correct 2SLS estimate of the return to education?

C10 Use the data in HTV for this exercise.

(i) Run a simple OLS regression of log(*wage*) on *educ*. Without controlling for other factors, what is the 95% confidence interval for the return to another year of education?

(ii) The variable *ctuit,* in thousands of dollars, is the change in college tuition facing students from age 17 to age 18. Show that *educ* and *ctuit* are essentially uncorrelated. What does this say about *ctuit* as a possible IV for *educ* in a simple regression analysis?

(iii) Now, add to the simple regression model in part (i) a quadratic in experience and a full set of regional dummy variables for current residence and residence at age 18. Also include the urban indicators for current and age 18 residences. What is the estimated return to a year of education?

(iv) Again using *ctuit* as a potential IV for *educ*, estimate the reduced form for *educ*. [Naturally, the reduced form for *educ* now includes the explanatory variables in part (iii).] Show that *ctuit* is now statistically significant in the reduced form for *educ*.

(v) Estimate the model from part (iii) by IV, using *ctuit* as an IV for *educ*. How does the confidence interval for the return to education compare with the OLS CI from part (iii)?

(vi) Do you think the IV procedure from part (v) is convincing?

C11 The data set in VOUCHER, which is a subset of the data used in Rouse (1998), can be used to estimate the effect of school choice on academic achievement. Attendance at a choice school was paid for by a voucher, which was determined by a lottery among those who applied. The data subset was chosen so that any student in the sample has a valid 1994 math test score (the last year available in Rouse's sample). Unfortunately, as pointed out by Rouse, many students have missing test scores, possibly due to attrition (that is, leaving the Milwaukee public school district). These data include students who applied to the voucher program and were accepted, students who applied and were not accepted, and students who did not apply. Therefore, even though the vouchers were chosen by lottery among those who applied, we do not necessarily have a random sample from a population where being selected for a voucher has been randomly determined. (An important consideration is that students who never applied to the program may be systematically different from those who did—and in ways that we cannot know based on the data.)

Rouse (1998) uses panel data methods of the kind we discussed in Chapter 14 to allow student fixed effects; she also uses instrumental variables methods. This problem asks you to do a cross-sectional analysis which winning the lottery for a voucher acts as an instrumental variable for attending a choice school. Actually, because we have multiple years of data on each student, we construct two variables. The first, *choiceyrs*, is the number of years from 1991 to 1994 that a student attended a choice school; this variable ranges from zero to four. The variable *selectyrs* indicates the number of years a student was selected for a voucher. If the student applied for the program in 1990 and received a voucher then $selectyrs = 4$; if he or she applied in 1991 and received a voucher then $selectyrs = 3$; and so on. The outcome of interest is *mnce*, the student's percentile score on a math test administered in 1994.

(i) Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher available for four years? How many students actually attended a choice school for four years?

(ii) Run a simple regression of *choiceyrs* on *selectyrs*. Are these variables related in the direction you expected? How strong is the relationship? Is *selectyrs* a sensible IV candidate for *choiceyrs*?

(iii) Run a simple regression of *mnce* on *choiceyrs*. What do you find? Is this what you expected? What happens if you add the variables *black*, *hispanic*, and *female*?

(iv) Why might *choiceyrs* be endogenous in an equation such as

$$mnce = \beta_0 + \beta_1 choiceyrs + \beta_2 black + \beta_3 hispanic + \beta_4 female + u_1?$$

(v) Estimate the equation in part (iv) by instrumental variables, using *selectyrs* as the IV for *choiceyrs*. Does using IV produce a positive effect of attending a choice school? What do you make of the coefficients on the other explanatory variables?

(vi) To control for the possibility that prior achievement affects participating in the lottery (as well as predicting attrition), add *mnce90*—the math score in 1990—to the equation in part (iv). Estimate the equation by OLS and IV, and compare the results for β_1. For the IV estimate, how much is each year in a choice school worth on the math percentile score? Is this a practically large effect?

(vii) Why is the analysis from part (vi) not entirely convincing? [*Hint*: Compared with part (v), what happens to the number of observations, and why?]

(viii) The variables *choiceyrs*1, *choiceyrs*2, and so on are dummy variables indicating the different number of years a student could have been in a choice school (from 1991 to 1994). The dummy variables *selectyrs*1, *selectyrs*2, and so on have a similar definition, but for being selected from the lottery. Estimate the equation

$$\begin{aligned} mnce = {} & \beta_0 + \beta_1 choiceyrs1 + \beta_2 choiceyrs2 + \beta_3 choiceyrs3 + \beta_4 choiceyrs4 \\ & + \beta_5 black + \beta_6 hispanic + \beta_7 female + \beta_8 mnce90 + u_1 \end{aligned}$$

by IV, using as instruments the four *selectyrs* dummy variables. (As before, the variables *black*, *hispanic*, and *female* act as their own IVs.) Describe your findings. Do they make sense?

C12 Use the data in CATHOLIC to answer this question. The model of interest is

$$math12 = \beta_0 + \beta_1 cathhs + \beta_2 lfaminc + \beta_3 motheduc + \beta_4 fatheduc + u,$$

where *cathhs* is a binary indicator for whether a student attends a Catholic high school.

(i) How many students are in the sample? What percentage of these students attend a Catholic high school?

(ii) Estimate the above equation by OLS. What is the estimate of β_1? What is its 95% confidence interval?

(iii) Using *parcath* as an instrument for *cathhs*, estimate the reduced form *for cathhs*. What is the *t* statistic for *parcath*? Is there evidence of a weak instrument problem?

(iv) Estimate the above equation by IV, using *parcath* as an IV for *cathhs*. How does the estimate and 95% CI compare with the OLS quantities?

(v) Test the null hypothesis that *cathhs* is exogenous. What is the *p*-value of the test?

(vi) Suppose you add the interaction between $cathhs \cdot motheduc$ to the above model. Why is it generally endogenous? Why is $pareduc \cdot motheduc$ a good IV candidate for $cathhs \cdot motheduc$?

(vii) Before you create the interactions in part (vi), first find the sample average of *motheduc* and create $cathhs \cdot (motheduc - \overline{motheduc})$ and $parcath \cdot (motheduc - \overline{motheduc})$. Add the first interaction to the model and use the second as an IV. Of course, *cathhs* is also instrumented. Is the interaction term statistically significant?

(viii) Compare the coefficient on *cathhs* in (vii) to that in part (iv). Is including the interaction important for estimating the average partial effect?

C13 Use the data in LABSUP to answer the following questions. These are data on almost 32,000 black or Hispanic women. Every woman in the sample is married. It is a subset of the data used in Angrist and Evans (1998). Our interest here is in determining how weekly hours worked, *hours*, changes with number of children (*kids*). All women in the sample have at least two children. The two potential instrumental variables for *kids*, which is suspected as being endogenous, work to generate exogenous variation starting with two children. See the original article for further discussion.

(i) Estimate the equation

$$hours = \beta_0 + \beta_1 kids + \beta_2 nonmomi + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 black + \beta_7 hispan + u$$

by OLS and obtain the heteroskedasticity-robust standard errors. Interpret the coefficient on *kids*. Discuss its statistical significance.

(ii) A variable that Angrist and Evans propose as an instrument is *samesex*, a binary variable equal to one if the first two children are the same biological sex. What do you think is the argument for why it is a relevant instrument for *kids*?

(iii) Run the regression

$$kids_i \text{ on } samesex_i, nonmomi_i, educ_i, age_i\; age_i^2, black_i, hispan_i$$

and see if the story from part (ii) holds up. In particular, interepret the coefficient on *samesex*. How statistically significant is *samesex*?

(iv) Can you think of mechanisms by which samesex is correlated with u in the equation in part (i)? (It is fine to assume that biological sex is randomly determined.) [Hint: How might a family's finances be affected based on whether they have two children of the same sex or two children of opposite sex?]

(v) Is it legitimate to check for exogeneity of *samesex* by adding it to the regression in part (i) and testing its significance? Explain.

(vi) Using *samesex* as an IV for *kids*, obtain the IV estimates of the equation in part (i). How does the *kids* coefficient compare with the OLS estimate? Is the IV estimate precise?

(vii) Now add *multi2nd* as an instrument. Obtain the F statistic from the first stage regression and determining whether *samesex* and *multi2nd* are sufficiently strong.

(viii) Using *samesex* and *multi2nd* both as instruments for *kids*, how does the 2SLS estimate compare with the OLS and IV estimates from the previous parts?

(ix) Using the estimation from part (viii), is there strong evidence that *kids* is endogenous in the *hours* equation?

(x) In part (viii), how many overidentification restrictions are there? Does the overidentification test pass?

APPENDIX 15A

15A.1 Assumptions for Two Stage Least Squares

This appendix covers the assumptions under which 2SLS has desirable large sample properties. We first state the assumptions for cross-sectional applications under random sampling. Then, we discuss what needs to be added for them to apply to time series and panel data.

15A.2 Assumption 2SLS.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or random disturbance term. The instrumental variables are denoted as z_j.

It is worth emphasizing that Assumption 2SLS.1 is virtually identical to MLR.1 (with the minor exception that 2SLS.1 mentions the notation for the instrumental variables, z_j). In other words, the model we are interested in is the same as that for OLS estimation of the β_j. Sometimes it is easy to lose sight of the fact that we can apply different estimation methods to the same model. Unfortunately, it is not uncommon to hear researchers say "I estimated an OLS model" or "I used a 2SLS model." Such statements are meaningless. OLS and 2SLS are different *estimation* methods that are applied to the *same* model. It is true that they have desirable statistical properties under different sets of assumptions on the model, but the relationship they are estimating is given by the equation in 2SLS.1 (or MLR.1). The point is similar to that made for the unobserved effects panel data model covered in Chapters 13 and 14: pooled OLS, first differencing, fixed effects, and random effects are different estimation methods for the same model.

15A.3 Assumption 2SLS.2 (Random Sampling)

We have a random sample on y, the x_j, and the z_j.

15A.4 Assumption 2SLS.3 (Rank Condition)

(i) There are no perfect linear relationships among the instrumental variables. (ii) The rank condition for identification holds.

With a single endogenous explanatory variable, as in equation (15.42), the rank condition is easily described. Let $z_1, \ldots, z_m$ denote the exogenous variables, where $z_k, \ldots, z_m$ do not appear in the structural model (15.42). The reduced form of y_2 is

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \cdots + \pi_{k-1} z_{k-1} + \pi_k z_k + \cdots + \pi_m z_m + v_2.$$

Then, we need at least one of $\pi_k, \ldots, \pi_m$ to be nonzero. This requires at least one exogenous variable that does not appear in (15.42) (the order condition). Stating the rank condition with two or more endogenous explanatory variables requires matrix algebra. [See Wooldridge (2010, Chapter 5).]

15A.5 Assumption 2SLS.4 (Exogenous Instrumental Variables)

The error term u has zero mean, and each IV is uncorrelated with u.
Remember that any x_j that is uncorrelated with u also acts as an IV.

15A.6 Theorem 15A.1

Under Assumptions 2SLS.1 through 2SLS.4, the 2SLS estimator is consistent.

15A.7 Assumption 2SLS.5 (Homoskedasticity)

Let $\mathbf{z}$ denote the collection of all instrumental variables. Then, $\mathrm{E}(u^2|\mathbf{z}) = \sigma^2$.

15A.8 Theorem 15A.2

Under Assumptions 2SLS.1 through 2SLS.5, the 2SLS estimators are asymptotically normally distributed. Consistent estimators of the asymptotic variance are given as in equation (15.43), where σ^2 is replaced with $\hat{\sigma}^2 = (n - k - 1)^{-1}\sum_{i=1}^{n}\hat{u}_i^2$, and the $\hat{u}_i$ are the 2SLS residuals.

The 2SLS estimator is also the best IV estimator under the five assumptions given. We state the result here. A proof can be found in Wooldridge (2010, Chapter 5).

15A.9 Theorem 15A.3

Under Assumptions 2SLS.1 through 2SLS.5, the 2SLS estimator is asymptotically efficient in the class of IV estimators that uses linear combinations of the exogenous variables as instruments.

If the homoskedasticity assumption does not hold, the 2SLS estimators are still asymptotically normal, but the standard errors (and t and F statistics) need to be adjusted; many econometrics packages do this routinely. Moreover, the 2SLS estimator is no longer the asymptotically efficient IV estimator, in general. We will not study more efficient estimators here [see Wooldridge (2010, Chapter 8)].

For time series applications, we must add some assumptions. First, as with OLS, we must assume that all series (including the IVs) are weakly dependent: this ensures that the law of large numbers and the central limit theorem hold. For the usual standard errors and test statistics to be valid, as well as for asymptotic efficiency, we must add a no serial correlation assumption.

15A.10 Assumption 2SLS.6 (No Serial Correlation)

Equation (15.54) holds.

A similar no serial correlation assumption is needed in panel data applications. Tests and corrections for serial correlation were discussed in Section 15-7.

CHAPTER 16

Simultaneous Equations Models

In the previous chapter, we showed how the method of instrumental variables can solve two kinds of endogeneity problems: omitted variables and measurement error. Conceptually, these problems are straightforward. In the omitted variables case, there is a variable (or more than one) that we would like to hold fixed when estimating the ceteris paribus effect of one or more of the observed explanatory variables. In the measurement error case, we would like to estimate the effect of certain explanatory variables on y, but we have mismeasured one or more variables. In both cases, we could estimate the parameters of interest by OLS if we could collect better data.

Another important form of endogeneity of explanatory variables is **simultaneity**. This arises when one or more of the explanatory variables is *jointly determined* with the dependent variable, typically through an equilibrium mechanism (as we will see later). In this chapter, we study methods for estimating simple simultaneous equations models (SEMs). Although a complete treatment of SEMs is beyond the scope of this text, we are able to cover models that are widely used.

The leading method for estimating simultaneous equations models is the method of instrumental variables. Therefore, the solution to the simultaneity problem is essentially the same as the IV solutions to the omitted variables and measurement error problems. However, crafting and interpreting SEMs is challenging. Therefore, we begin by discussing the nature and scope of simultaneous equations models in Section 16-1. In Section 16-2, we confirm that OLS applied to an equation in a simultaneous system is generally biased and inconsistent.

Section 16-3 provides a general description of identification and estimation in a two-equation system, while Section 16-4 briefly covers models with more than two equations. Simultaneous equations models are used to model aggregate time series, and in Section 16-5 we include a discussion of some special issues that arise in such models. Section 16-6 touches on simultaneous equations models with panel data.

16-1 The Nature of Simultaneous Equations Models

The most important point to remember in using simultaneous equations models is that each equation in the system should have a ceteris paribus, causal interpretation. Because we only observe the outcomes in equilibrium, we are required to use counterfactual reasoning in constructing the equations of a simultaneous equations model. We must think in terms of potential as well as actual outcomes.

When there are only two states of the world—a worker does or does not participate in a job training program, say—we formally described the potential outcomes setting in Sections 2-7, 3-7, 7-6, and elsewhere. The framework for simultaneous equations models is more complicated because we must represent a continuum of alternative realities. For example, the demand for a product, say, milk, is a function of the price of milk (and other variables). A demand function for milk determines how much milk someone would purchase at each possible price. Rather than formally introduce a notation for a continuum of potential outcomes, for our purposes it suffices to be less formal and to illustrate counterfactual thinking through examples.

The classic example of an SEM is a supply and demand equation for some commodity or input to production (such as labor). For concreteness, let h_s denote the annual labor hours supplied by workers in agriculture, measured at the county level, and let w denote the average hourly wage offered to such workers. A simple labor supply function is

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1, \qquad [16.1]$$

where z_1 is some observed variable affecting labor supply—say, the average manufacturing wage in the county. The error term, u_1, contains other factors that affect labor supply. [Many of these factors are observed and could be included in equation (16.1); to illustrate the basic concepts, we include only one such factor, z_1.] Equation (16.1) is an example of a **structural equation**. This name comes from the fact that the labor supply function is derivable from economic theory and has a causal interpretation. The coefficient α_1 measures how labor supply changes when the wage changes; if h_s and w are in logarithmic form, α_1 is the labor supply elasticity. Typically, we expect α_1 to be positive (although economic theory does not rule out $\alpha_1 \leq 0$). Labor supply elasticities are important for determining how workers will change the number of hours they desire to work when tax rates on wage income change. If z_1 is the manufacturing wage, we expect $\beta_1 \leq 0$: other factors equal, if the manufacturing wage increases, more workers will go into manufacturing than into agriculture.

When we graph labor supply, we sketch hours as a function of wage, with z_1 and u_1 held fixed. A change in z_1 shifts the labor supply function, as does a change in u_1. The difference is that z_1 is observed while u_1 is not. Sometimes, z_1 is called an *observed supply shifter*, and u_1 is called an *unobserved supply shifter*.

How does equation (16.1) differ from those we have studied previously? The difference is subtle. Although equation (16.1) is supposed to hold for all possible values of wage, we cannot generally view wage as varying exogenously for a cross section of counties. If we could run an experiment where we vary the level of agricultural and manufacturing wages across a sample of counties and survey workers to obtain the labor supply h_s for each county, then we could estimate (16.1) by OLS. Unfortunately, this is not a manageable experiment. Instead, we must collect data on average wages in these two sectors along with how many person hours were spent in agricultural production. In deciding how to analyze these data, we must understand that they are best described by the interaction of labor supply *and* demand. Under the assumption that labor markets clear, we actually observe *equilibrium* values of wages and hours worked.

To describe how equilibrium wages and hours are determined, we need to bring in the demand for labor, which we suppose is given by

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2, \qquad [16.2]$$

where h_d is hours demanded. As with the supply function, we graph hours demanded as a function of wage, w, keeping z_2 and u_2 fixed. The variable z_2—say, agricultural land area—is an *observable demand shifter*, while u_2 is an *unobservable demand shifter*.

Just as with the labor supply equation, the labor demand equation is a structural equation: it can be obtained from the profit maximization considerations of farmers. If h_d and w are in logarithmic form, α_2 is the labor demand elasticity. Economic theory tells us that $\alpha_2 < 0$. Because labor and land are complements in production, we expect $\beta_2 > 0$.

Notice how equations (16.1) and (16.2) describe entirely different relationships. Labor supply is a behavioral equation for workers, and labor demand is a behavioral relationship for farmers. Each equation has a ceteris paribus interpretation and stands on its own. They become linked in an econometric analysis only because *observed* wage and hours are determined by the intersection of supply and demand. In other words, for each county i, observed hours h_i and observed wage w_i are determined by the equilibrium condition

$$h_{is} = h_{id}. \qquad [16.3]$$

Because we observe only equilibrium hours for each county i, we denote observed hours by h_i.

When we combine the equilibrium condition in (16.3) with the labor supply and demand equations, we get

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1} \qquad [16.4]$$

and

$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}, \qquad [16.5]$$

where we explicitly include the i subscript to emphasize that h_i and w_i are the equilibrium observed values for county i. These two equations constitute a **simultaneous equations model (SEM)**, which has several important features. First, given z_{i1}, z_{i2}, u_{i1}, and u_{i2}, these two equations determine h_i and w_i. (Actually, we must assume that $\alpha_1 \neq \alpha_2$, which means that the slopes of the supply and demand functions differ; see Problem 1.) For this reason, h_i and w_i are the **endogenous variables** in this SEM. What about z_{i1} and z_{i2}? Because they are determined outside of the model, we view them as **exogenous variables**. From a statistical standpoint, the key assumption concerning z_{i1} and z_{i2} is that they are both uncorrelated with the supply and demand errors, u_{i1} and u_{i2}, respectively. These are examples of **structural errors** because they appear in the structural equations.

A second important point is that, without including z_1 and z_2 in the model, there is no way to tell which equation is the supply function and which is the demand function. When z_1 represents manufacturing wage, economic reasoning tells us that it is a factor in agricultural labor supply because it is a measure of the opportunity cost of working in agriculture; when z_2 stands for agricultural land area, production theory implies that it appears in the labor demand function. Therefore, we know that (16.4) represents labor supply and (16.5) represents labor demand. If z_1 and z_2 are the same—for example, average education level of adults in the county, which can affect both supply and demand—then the equations look identical, and there is no hope of estimating either one. In a nutshell, this illustrates the identification problem in simultaneous equations models, which we will discuss more generally in Section 16-3.

The most convincing examples of SEMs have the same flavor as supply and demand examples. Each equation should have a behavioral, ceteris paribus interpretation on its own. Because we only observe equilibrium outcomes, specifying an SEM requires us to ask such counterfactual questions as: How much labor *would* workers provide if the wage were different from its equilibrium value? Example 16.1 provides another illustration of an SEM in which each equation has a ceteris paribus interpretation.

EXAMPLE 16.1 Murder Rates and Size of the Police Force

Cities often want to determine how much additional law enforcement will decrease their murder rates. A simple cross-sectional model to address this question is

$$murdpc = \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1, \qquad [16.6]$$

where *murdpc* is murders per capita, *polpc* is number of police officers per capita, and *incpc* is income per capita. (Henceforth, we do not include an *i* subscript.) We take income per capita as exogenous in this equation. In practice, we would include other factors, such as age and gender distributions, education levels, perhaps geographic variables, and variables that measure severity of punishment. To fix ideas, we consider equation (16.6).

The question we hope to answer is: If a city exogenously increases its police force, will that increase, on average, lower the murder rate? If we could exogenously choose police force sizes for a random sample of cities, we could estimate (16.6) by OLS. Certainly, we cannot run such an experiment. But can we think of police force size as being exogenously determined, anyway? Probably not. A city's spending on law enforcement is at least partly determined by its expected murder rate. To reflect this, we postulate a second relationship:

$$polpc = \alpha_2 murdpc + \beta_{20} + other\ factors. \qquad [16.7]$$

We expect that $\alpha_2 > 0$: other factors being equal, cities with higher (expected) murder rates will have more police officers per capita. Once we specify the other factors in (16.7), we have a two-equation simultaneous equations model. We are really only interested in equation (16.6), but, as we will see in Section 16-3, we need to know precisely how the second equation is specified in order to estimate the first.

An important point is that (16.7) describes behavior by city officials, while (16.6) describes the actions of potential murderers. This gives each equation a clear ceteris paribus interpretation, which makes equations (16.6) and (16.7) an appropriate simultaneous equations model.

We next give an example of an inappropriate use of SEMs.

EXAMPLE 16.2 Housing Expenditures and Saving

Suppose that, for a random household in the population, we assume that annual housing expenditures and saving are jointly determined by

$$housing = \alpha_1 saving + \beta_{10} + \beta_{11} inc + \beta_{12} educ + \beta_{13} age + u_1 \qquad [16.8]$$

and

$$saving = \alpha_2 housing + \beta_{20} + \beta_{21} inc + \beta_{22} educ + \beta_{23} age + u_2, \qquad [16.9]$$

where *inc* is annual income and *educ* and *age* are measured in years. Initially, it may seem that these equations are a sensible way to view how housing and saving expenditures are determined. But we have to ask: What value would one of these equations be without the other? Neither has a ceteris paribus interpretation because *housing* and *saving* are chosen by the same household. For example, it makes no sense to ask this question: If annual income increases by \$10,000, how would housing expenditures change, *holding saving fixed*? If family income increases, a household will generally change the optimal mix of housing expenditures and saving. But equation (16.8) makes it seem as if we want to know the effect of changing *inc*, *educ*, or *age* while keeping *saving* fixed. Such a thought experiment is not interesting. Any model based on economic principles, particularly utility maximization, would have households optimally choosing *housing* and *saving* as functions of *inc* and the relative prices of housing and saving. The variables *educ* and *age* would affect preferences for consumption, saving, and risk. Therefore, *housing* and *saving* would each be functions of income, education, age, and other variables that affect the utility maximization problem (such as different rates of return on housing and other saving).

Even if we decided that the SEM in (16.8) and (16.9) made sense, there is no way to estimate the parameters. (We discuss this problem more generally in Section 16-3.) The two equations are indistinguishable, unless we assume that income, education, or age appears in one equation but not the other, which would make no sense.

Though this makes a poor SEM example, we might be interested in testing whether, other factors being fixed, there is a tradeoff between housing expenditures and saving. But then we would just estimate, say, (16.8) by OLS, unless there is an omitted variable or measurement error problem.

Example 16.2 has the characteristics of all too many SEM applications. The problem is that the two endogenous variables are chosen by the same economic agent. Therefore, neither equation can stand on its own. Another example of an inappropriate use of an SEM would be to model weekly hours spent studying and weekly hours working. Each student will choose these variables simultaneously—presumably as a function of the wage that can be earned working, ability as a student, enthusiasm for college, and so on. Just as in Example 16.2, it makes no sense to specify two equations where each is a function of the other. The important lesson is this: just because two variables are determined simultaneously does *not* mean that a simultaneous equations model is suitable. For an SEM to make sense, each equation in the SEM should have a ceteris paribus interpretation in isolation from the other equation. As we discussed earlier, supply and demand examples, and Example 16.1, have this feature. Usually, basic economic reasoning, supported in some cases by simple economic models, can help us use SEMs intelligently (including knowing when not to use an SEM).

GOING FURTHER 16.1

A standard model of advertising for monopolistic firms has firms choosing profit maximizing levels of price and advertising expenditures. Does this mean we should use an SEM to model these variables at the firm level?

16-2 Simultaneity Bias in OLS

It is useful to see, in a simple model, that an explanatory variable that is determined simultaneously with the dependent variable is generally correlated with the error term, which leads to bias and inconsistency in OLS. We consider the two-equation structural model

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \tag{16.10}$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \tag{16.11}$$

and focus on estimating the first equation. The variables z_1 and z_2 are exogenous, so that each is uncorrelated with u_1 and u_2. For simplicity, we suppress the intercept in each equation.

To show that y_2 is generally correlated with u_1, we solve the two equations for y_2 in terms of the exogenous variables and the error term. If we plug the right-hand side of (16.10) in for y_1 in (16.11), we get

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

or

$$(1 - \alpha_2\alpha_1)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2. \tag{16.12}$$

Now, we must make an assumption about the parameters in order to solve for y_2:

$$\alpha_2\alpha_1 \neq 1. \tag{16.13}$$

Whether this assumption is restrictive depends on the application. In Example 16.1, we think that $\alpha_1 \leq 0$ and $\alpha_2 \geq 0$, which implies $\alpha_1\alpha_2 \leq 0$; therefore, (16.13) is very reasonable for Example 16.1.

Provided condition (16.13) holds, we can divide (16.12) by $(1 - \alpha_2\alpha_1)$ and write y_2 as

$$y_2 = \pi_{21}z_1 + \pi_{22}z_2 + v_2, \quad [16.14]$$

where $\pi_{21} = \alpha_2\beta_1/(1 - \alpha_2\alpha_1)$, $\pi_{22} = \beta_2/(1 - \alpha_2\alpha_1)$, and $v_2 = (\alpha_2 u_1 + u_2)/(1 - \alpha_2\alpha_1)$. Equation (16.14), which expresses y_2 in terms of the exogenous variables and the error terms, is the **reduced form equation** for y_2, a concept we introduced in Chapter 15 in the context of instrumental variables estimation. The parameters π_{21} and π_{22} are called **reduced form parameters**; notice how they are nonlinear functions of the **structural parameters**, which appear in the structural equations, (16.10) and (16.11).

The **reduced form error**, v_2, is a linear function of the structural error terms, u_1 and u_2. Because u_1 and u_2 are each uncorrelated with z_1 and z_2, v_2 is also uncorrelated with z_1 and z_2. Therefore, we can consistently estimate π_{21} and π_{22} by OLS, something that is used for two stage least squares estimation (which we return to in the next section). In addition, the reduced form parameters are sometimes of direct interest, although we are focusing here on estimating equation (16.10).

A reduced form also exists for y_1 under assumption (16.13); the algebra is similar to that used to obtain (16.14). It has the same properties as the reduced form equation for y_2.

We can use equation (16.14) to show that, except under special assumptions, OLS estimation of equation (16.10) will produce biased and inconsistent estimators of α_1 and β_1 in equation (16.10). Because z_1 and u_1 are uncorrelated by assumption, the issue is whether y_2 and u_1 are uncorrelated. From the reduced form in (16.14), we see that y_2 and u_1 are correlated if and only if v_2 and u_1 are correlated (because z_1 and z_2 are assumed exogenous). But v_2 is a linear function of u_1 and u_2, so it is generally correlated with u_1. In fact, if we assume that u_1 and u_2 are uncorrelated, then v_2 and u_1 *must* be correlated whenever $\alpha_2 \neq 0$. Even if α_2 equals zero—which means that y_1 does not appear in equation (16.11)— v_2 and u_1 will be correlated if u_1 and u_2 are correlated.

When $\alpha_2 = 0$ and u_1 and u_2 are uncorrelated, y_2 and u_1 are also uncorrelated. These are fairly strong requirements: if $\alpha_2 = 0$, y_2 is not simultaneously determined with y_1. If we add zero correlation between u_1 and u_2, this rules out omitted variables or measurement errors in u_1 that are correlated with y_2. We should not be surprised that OLS estimation of equation (16.10) works in this case.

When y_2 is correlated with u_1 because of simultaneity, we say that OLS suffers from **simultaneity bias**. Obtaining the direction of the bias in the coefficients is generally complicated, as we saw with omitted variables bias in Chapters 3 and 5. But in simple models, we can determine the direction of the bias. For example, suppose that we simplify equation (16.10) by dropping z_1 from the equation, and we assume that u_1 and u_2 are uncorrelated. Then, the covariance between y_2 and u_1 is

$$\begin{aligned}\mathrm{Cov}(y_2,u_1) &= \mathrm{Cov}(v_2,u_1) = [\alpha_2/(1 - \alpha_2\alpha_1)]\mathrm{E}(u_1^2)\\ &= [\alpha_2/(1 - \alpha_2\alpha_1)]\sigma_1^2,\end{aligned}$$

where $\sigma_1^2 = \mathrm{Var}(u_1) > 0$. Therefore, the asymptotic bias (or inconsistency) in the OLS estimator of α_1 has the same sign as $\alpha_2/(1 - \alpha_2\alpha_1)$. If $\alpha_2 > 0$ and $\alpha_2\alpha_1 < 1$, the asymptotic bias is positive. (Unfortunately, just as in our calculation of omitted variables bias from Section 3-3, the conclusions do not carry over to more general models. But they do serve as a useful guide.) For example, in Example 16.1, we think $\alpha_2 > 0$ and $\alpha_2\alpha_1 \leq 0$, which means that the OLS estimator of α_1 would have a positive bias. If $\alpha_1 = 0$, OLS would, on average, estimate a *positive* impact of more police on the murder rate; generally, the estimator of α_1 is biased upward. Because we expect an increase in the size of the police force to reduce murder rates (ceteris paribus), the upward bias means that OLS will underestimate the effectiveness of a larger police force.

16-3 Identifying and Estimating a Structural Equation

As we saw in the previous section, OLS is biased and inconsistent when applied to a structural equation in a simultaneous equations system. In Chapter 15, we learned that the method of two stage least squares can be used to solve the problem of endogenous explanatory variables. We now show how 2SLS can be applied to SEMs.

The mechanics of 2SLS are similar to those in Chapter 15. The difference is that, because we specify a structural equation for each endogenous variable, we can immediately see whether sufficient IVs are available to estimate either equation. We begin by discussing the identification problem.

16-3a Identification in a Two-Equation System

We mentioned the notion of identification in Chapter 15. When we estimate a model by OLS, the key identification condition is that each explanatory variable is uncorrelated with the error term. As we demonstrated in Section 16-2, this fundamental condition no longer holds, in general, for SEMs. However, if we have some instrumental variables, we can still identify (or consistently estimate) the parameters in an SEM equation, just as with omitted variables or measurement error.

Before we consider a general two-equation SEM, it is useful to gain intuition by considering a simple supply and demand example. Write the system in equilibrium form (that is, with $q_s = q_d = q$ imposed) as

$$q = \alpha_1 p + \beta_1 z_1 + u_1 \qquad [16.15]$$

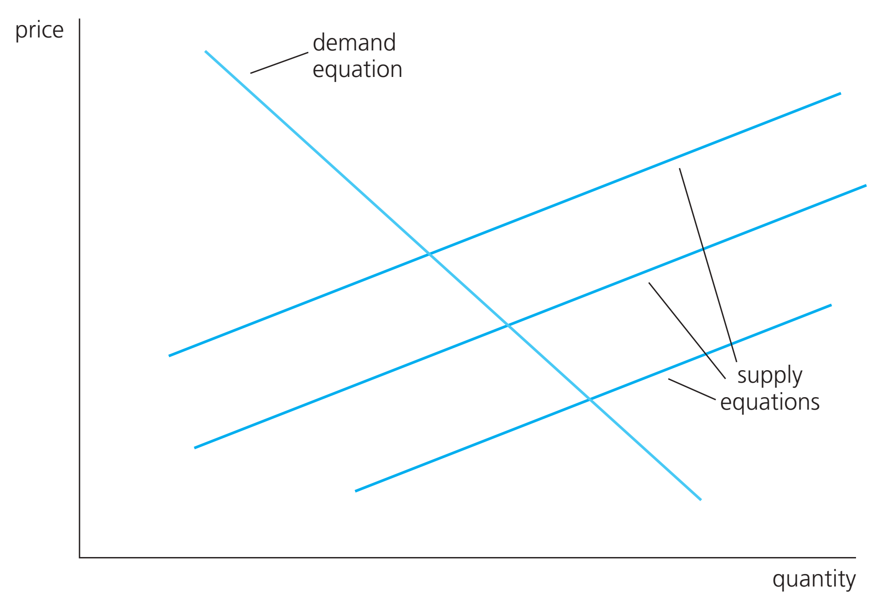
and

$$q = \alpha_2 p + u_2. \qquad [16.16]$$

For concreteness, let q be per capita milk consumption at the county level, let p be the average price per gallon of milk in the county, and let z_1 be the price of cattle feed, which we assume is exogenous to the supply and demand equations for milk. This means that (16.15) must be the supply function, as the price of cattle feed would shift supply $(\beta_1 < 0)$ but not demand. The demand function contains no observed demand shifters.

Given a random sample on (q, p, z_1), which of these equations can be estimated? That is, which is an **identified equation**? It turns out that the *demand* equation, (16.16), is identified, but the supply equation is not. This is easy to see by using our rules for IV estimation from Chapter 15: we can use z_1 as an IV for price in equation (16.16). However, because z_1 appears in equation (16.15), we have no IV for price in the supply equation.

FIGURE 16.1 Shifting supply equations trace out the demand equation. Each supply equation is drawn for a different value of the exogenous variable, z_1.



Intuitively, the fact that the demand equation is identified follows because we have an observed variable, z_1, that shifts the supply equation while not affecting the demand equation. Given variation in z_1 and no errors, we can trace out the demand curve, as shown in Figure 16.1. The presence of the unobserved demand shifter u_2 causes us to estimate the demand equation with error, but the estimators will be consistent, provided z_1 is uncorrelated with u_2.

The supply equation cannot be traced out because there are no exogenous observed factors shifting the demand curve. It does not help that there are unobserved factors shifting the demand function; we need something observed. If, as in the labor demand function (16.2), we have an observed exogenous demand shifter—such as income in the milk demand function—then the supply function would also be identified.

To summarize: *In the system of (16.15) and (16.16), it is the presence of an exogenous variable in the supply equation that allows us to estimate the demand equation.*

Extending the identification discussion to a general two-equation model is not difficult. Write the two equations as

$$y_1 = \beta_{10} + \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\beta}_1 + u_1 \quad [16.17]$$

and

$$y_2 = \beta_{20} + \alpha_2 y_1 + \mathbf{z}_2 \boldsymbol{\beta}_2 + u_2, \quad [16.18]$$

where y_1 and y_2 are the endogenous variables and u_1 and u_2 are the structural error terms. The intercept in the first equation is β_{10}, and the intercept in the second equation is β_{20}. The variable $\mathbf{z}_1$ denotes a set of k_1 exogenous variables appearing in the first equation: $\mathbf{z}_1 = (z_{11}, z_{12}, \ldots, z_{1k_1})$. Similarly, $\mathbf{z}_2$ is the set of k_2 exogenous variables in the second equation: $\mathbf{z}_2 = (z_{21}, z_{22}, \ldots, z_{2k_2})$. In many cases, $\mathbf{z}_1$ and $\mathbf{z}_2$ will overlap. As a shorthand form, we use the notation

$$\mathbf{z}_1 \boldsymbol{\beta}_1 = \beta_{11} z_{11} + \beta_{12} z_{12} + \cdots + \beta_{1k_1} z_{1k_1}$$

and

$$\mathbf{z}_2 \boldsymbol{\beta}_2 = \beta_{21} z_{21} + \beta_{22} z_{22} + \cdots + \beta_{2k_2} z_{2k_2};$$

that is, $\mathbf{z}_1 \boldsymbol{\beta}_1$ stands for all exogenous variables in the first equation, with each multiplied by a coefficient, and similarly for $\mathbf{z}_2 \boldsymbol{\beta}_2$. (Some authors use the notation $\mathbf{z}_1' \boldsymbol{\beta}_1$ and $\mathbf{z}_2' \boldsymbol{\beta}_2$ instead. If you have an interest in the matrix algebra approach to econometrics, see Advanced Treatment E.)

The fact that $\mathbf{z}_1$ and $\mathbf{z}_2$ generally contain different exogenous variables means that we have imposed **exclusion restrictions** on the model. In other words, we *assume* that certain exogenous variables do not appear in the first equation and others are absent from the second equation. As we saw with the previous supply and demand examples, this allows us to distinguish between the two structural equations.

When can we solve equations (16.17) and (16.18) for y_1 and y_2 (as linear functions of all exogenous variables and the structural errors, u_1 and u_2)? The condition is the same as that in (16.13), namely, $\alpha_2 \alpha_1 \neq 1$. The proof is virtually identical to the simple model in Section 16-2. Under this assumption, reduced forms exist for y_1 and y_2.

The key question is: Under what assumptions can we estimate the parameters in, say, (16.17)? This is the identification issue. The **rank condition** for identification of equation (16.17) is easy to state.

Rank Condition for Identification of a Structural Equation. The first equation in a two-equation simultaneous equations model is identified if, and only if, the *second* equation contains at least one exogenous variable (with a nonzero coefficient) that is excluded from the first equation.

This is the necessary and sufficient condition for equation (16.17) to be identified. The **order condition**, which we discussed in Chapter 15, is necessary for the rank condition. The order condition for identifying the first equation states that at least one exogenous variable is excluded from this equation. The order condition is trivial to check once both equations have been specified. The rank condition requires more: at least one of the exogenous variables excluded from the first equation must have a nonzero population coefficient in the second equation. This ensures that at least one of the exogenous variables omitted from the first equation actually appears in the reduced form of y_2, so that we can use these variables as instruments for y_2. We can test this using a t or an F test, as in Chapter 15; some examples follow.

Identification of the second equation is, naturally, just the mirror image of the statement for the first equation. Also, if we write the equations as in the labor supply and demand example in Section 16-1—so that y_1 appears on the left-hand side in *both* equations, with y_2 on the right-hand side—the identification condition is identical.

EXAMPLE 16.3 Labor Supply of Married, Working Women

To illustrate the identification issue, consider labor supply for married women already in the workforce. In place of the demand function, we write the wage offer as a function of hours and the usual productivity variables. With the equilibrium condition imposed, the two structural equations are

$$\begin{aligned} hours = {} & \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 \\ & + \beta_{14} nwifeinc + u_1 \end{aligned} \tag{16.19}$$

and

$$\begin{aligned} \log(wage) = {} & \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper \\ & + \beta_{23} exper^2 + u_2. \end{aligned} \tag{16.20}$$

The variable *age* is the woman's age, in years, *kidslt6* is the number of children less than six years old, *nwifeinc* is the woman's nonwage income (which includes husband's earnings), and *educ* and *exper* are years of education and prior experience, respectively. All variables except *hours* and log(*wage*) are assumed to be exogenous. (This is a tenuous assumption, as *educ* might be correlated with omitted ability in either equation. But for illustration purposes, we ignore the omitted ability problem.) The functional form in this system—where *hours* appears in level form but *wage* is in logarithmic form—is popular in labor economics. We can write this system as in equations (16.17) and (16.18) by defining $y_1 = hours$ and $y_2 = \log(wage)$.

The first equation is the supply function. It satisfies the order condition because two exogenous variables, *exper* and $exper^2$, are omitted from the labor supply equation. These exclusion restrictions are crucial assumptions: we are assuming that, once wage, education, age, number of small children, and other income are controlled for, past experience has no effect on current labor supply. One could certainly question this assumption, but we use it for illustration.

Given equations (16.19) and (16.20), the rank condition for identifying the first equation is that at least one of *exper* and $exper^2$ has a nonzero coefficient in equation (16.20). If $\beta_{22} = 0$ and $\beta_{23} = 0$, there are no exogenous variables appearing in the second equation that do not also appear in the first (*educ* appears in both). We can state the rank condition for identification of (16.19) equivalently in terms of the reduced form for log(*wage*), which is

$$\begin{aligned} \log(wage) = {} & \pi_{20} + \pi_{21} educ + \pi_{22} age + \pi_{23} kidslt6 \\ & + \pi_{24} nwifeinc + \pi_{25} exper + \pi_{26} exper^2 + v_2. \end{aligned} \tag{16.21}$$

For identification, we need $\pi_{25} \neq 0$ or $\pi_{26} \neq 0$, something we can test using a standard F statistic, as we discussed in Chapter 15.

The wage offer equation, (16.20), is identified if at least one of *age*, *kidslt6*, or *nwifeinc* has a nonzero coefficient in (16.19). This is identical to assuming that the reduced form for *hours*—which has the same form as the right-hand side of (16.21)—depends on at least one of *age*, *kidslt6*, or *nwifeinc*. In specifying the wage offer equation, we are *assuming* that *age*, *kidslt6*, and *nwifeinc* have no effect on the offered wage, once hours, education, and experience are accounted for. These would be poor assumptions if these variables somehow have direct effects on productivity, or if women are discriminated against based on their age or number of small children.

In Example 16.3, we take the population of interest to be married women who are in the workforce (so that equilibrium hours are positive). This excludes the group of married women who choose not to work outside the home. Including such women in the model raises some difficult problems. For instance, if a woman does not work, we cannot observe her wage offer. We touch on these issues in Chapter 17; but for now, we must think of equations (16.19) and (16.20) as holding only for women who have $hours > 0$.

EXAMPLE 16.4 Inflation and Openness

Romer (1993) proposes theoretical models of inflation that imply that more "open" countries should have lower inflation rates. His empirical analysis explains average annual inflation rates (since 1973) in terms of the average share of imports in gross domestic (or national) product since 1973—which is his measure of openness. In addition to estimating the key equation by OLS, he uses instrumental variables. While Romer does not specify both equations in a simultaneous system, he has in mind a two-equation system:

$$inf = \beta_{10} + \alpha_1 open + \beta_{11}\log(pcinc) + u_1 \tag{16.22}$$

$$open = \beta_{20} + \alpha_2 inf + \beta_{21}\log(pcinc) + \beta_{22}\log(land) + u_2, \tag{16.23}$$

where *pcinc* is 1980 per capita income, in U.S. dollars (assumed to be exogenous), and *land* is the land area of the country, in square miles (also assumed to be exogenous). Equation (16.22) is the one of interest, with the hypothesis that $\alpha_1 < 0$. (More open economies have lower inflation rates.) The second equation reflects the fact that the degree of openness might depend on the average inflation rate, as well as other factors. The variable log(*pcinc*) appears in both equations, but log(*land*) is *assumed* to appear only in the second equation. The idea is that, ceteris paribus, a smaller country is likely to be more open (so $\beta_{22} < 0$).

Using the identification rule that was stated earlier, equation (16.22) is identified, provided $\beta_{22} \neq 0$. Equation (16.23) is *not* identified because it contains both exogenous variables. But we are interested in (16.22).

> **GOING FURTHER 16.2**
>
> If we have money supply growth since 1973 for each country, which we assume is exogenous, does this help identify equation (16.23)?

16-3b Estimation by 2SLS

Once we have determined that an equation is identified, we can estimate it by two stage least squares. The instrumental variables consist of the exogenous variables appearing in either equation.

EXAMPLE 16.5 Labor Supply of Married, Working Women

We use the data on working, married women in MROZ to estimate the labor supply equation (16.19) by 2SLS. The full set of instruments includes *educ*, *age*, *kidslt6*, *nwifeinc*, *exper*, and *exper*^{2}. The estimated labor supply curve is

$$\begin{aligned}\widehat{hours} = {} & \underset{(574.56)}{2,225.66} + \underset{(470.58)}{1,639.56}\log(wage) - \underset{(59.10)}{183.75}\,educ \\ & - \underset{(9.38)}{7.81}\,age - \underset{(182.93)}{198.15}\,kidslt6 - \underset{(6.61)}{10.17}\,nwifeinc \\ n = {} & 428,\end{aligned} \tag{16.24}$$

where the reported standard errors are computed using a degrees-of-freedom adjustment. This equation shows that the labor supply curve slopes upward. The estimated coefficient on log(*wage*) has the following interpretation: holding other factors fixed, $\Delta\widehat{hours} \approx 16.4(\%\Delta wage)$. We can calculate labor supply elasticities by multiplying both sides of this last equation by 100/*hours*:

$$100 \cdot (\Delta\widehat{hours}/hours) \approx (1{,}640/hours)(\%\Delta wage)$$

or

$$\%\Delta\widehat{hours} \approx (1{,}640/hours)(\%\Delta wage),$$

which implies that the labor supply elasticity (with respect to wage) is simply 1,640/*hours*. [The elasticity is not constant in this model because *hours*, not log(*hours*), is the dependent variable in (16.24).] At the average hours worked, 1,303, the estimated elasticity is 1,640/1,303 ≈ 1.26, which implies a greater than 1% increase in hours worked given a 1% increase in wage. This is a large estimated elasticity. At higher hours, the elasticity will be smaller; at lower hours, such as $hours = 800$, the elasticity is over two.

For comparison, when (16.19) is estimated by OLS, the coefficient on log(*wage*) is −2.05 (se = 54.88), which implies no wage effect on hours worked. To confirm that log(*wage*) is in fact endogenous in (16.19), we can carry out the test from Section 15-5. When we add the reduced form residuals $\hat{v}_2$ to the equation and estimate by OLS, the t statistic on $\hat{v}_2$ is −6.61, which is very significant, and so log(*wage*) appears to be endogenous.

The wage offer equation (16.20) can also be estimated by 2SLS. The result is

$$\begin{aligned}\widehat{\log(wage)} = &-.656 + .00013\, hours + .110\, educ \\ &\quad(.338)\quad(.00025)\qquad\quad(.016) \\ &+ .035\, exper - .00071\, exper^2 \\ &\quad(.019)\qquad\quad(.00045) \\ n = 428.&\end{aligned} \tag{16.25}$$

This differs from previous wage equations in that *hours* is included as an explanatory variable and 2SLS is used to account for endogeneity of *hours* (and we assume that *educ* and *exper* are exogenous). The coefficient on *hours* is statistically insignificant, which means that there is no evidence that the wage offer increases with hours worked. The other coefficients are similar to what we get by dropping *hours* and estimating the equation by OLS.

Estimating the effect of openness on inflation by instrumental variables is also straightforward.

EXAMPLE 16.6 Inflation and Openness

Before we estimate (16.22) using the data in OPENNESS, we check to see whether *open* has sufficient partial correlation with the proposed IV, log(*land*). The reduced form regression is

$$\begin{aligned}\widehat{open} = &\ 117.08 + .546 \log(pcinc) - 7.57 \log(land) \\ &\ (15.85)\ (1.493)\qquad\qquad\quad (.81) \\ n = 114,&\ R^2 = .449.\end{aligned}$$

The t statistic on log(*land*) is over nine in absolute value, which verifies Romer's assertion that smaller countries are more open. The fact that log(*pcinc*) is so insignificant in this regression is irrelevant.

Estimating (16.22) using log(*land*) as an IV for *open* gives

$$\widehat{inf} = \underset{(15.40)}{26.90} - \underset{(.144)}{.337}\ open + \underset{(2.015)}{.376}\ \log(pcinc) \qquad [16.26]$$
$$n = 114.$$

> **GOING FURTHER 16.3**
>
> How would you test whether the difference between the OLS and IV estimates on *open* are statistically different?

The coefficient on *open* is statistically significant at about the 1% level against a one-sided alternative $(\alpha_1 < 0)$. The effect is economically important as well: for every percentage point increase in the import share of GDP, annual inflation is about one-third of a percentage point lower. For comparison, the OLS estimate is $-.215$ (se $= .095$).

16-4 Systems with More Than Two Equations

Simultaneous equations models can consist of more than two equations. Studying general identification of these models is difficult and requires matrix algebra. Once an equation in a general system has been shown to be identified, it can be estimated by 2SLS.

16-4a Identification in Systems with Three or More Equations

We will use a three-equation system to illustrate the issues that arise in the identification of complicated SEMs. With intercepts suppressed, write the model as

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}z_1 + u_1 \qquad [16.27]$$

$$y_2 = \alpha_{21}y_1 + \beta_{21}z_1 + \beta_{22}z_2 + \beta_{23}z_3 + u_2 \qquad [16.28]$$

$$y_3 = \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3, \qquad [16.29]$$

where the y_g are the endogenous variables and the z_j are exogenous. The first subscript on the parameters indicates the equation number, and the second indicates the variable number; we use α for parameters on endogenous variables and β for parameters on exogenous variables.

Which of these equations can be estimated? It is generally difficult to show that an equation in an SEM with more than two equations is identified, but it is easy to see when certain equations are *not* identified. In system (16.27) through (16.29), we can easily see that (16.29) falls into this category. Because every exogenous variable appears in this equation, we have no IVs for y_2. Therefore, we cannot consistently estimate the parameters of this equation. For the reasons we discussed in Section 16-2, OLS estimation will not usually be consistent.

What about equation (16.27)? Things look promising because z_2, z_3, and z_4 are all excluded from the equation—this is another example of *exclusion restrictions*. Although there are two endogenous variables in this equation, we have three potential IVs for y_2 and y_3. Therefore, equation (16.27) passes the order condition. For completeness, we state the order condition for general SEMs.

Order Condition for Identification. An equation in any SEM satisfies the order condition for identification if the number of *excluded* exogenous variables from the equation is at least as large as the number of right-hand side endogenous variables.

The second equation, (16.28), also passes the order condition because there is one excluded exogenous variable, z_4, and one right-hand side endogenous variable, y_1.

As we discussed in Chapter 15 and in the previous section, the order condition is only necessary, not sufficient, for identification. For example, if $\beta_{34} = 0$, z_4 appears nowhere in the system, which means it is not correlated with y_1, y_2, or y_3. If $\beta_{34} = 0$, then the second equation is not identified, because z_4 is useless as an IV for y_1. This again illustrates that identification of an equation depends on the values of the parameters (which we can never know for sure) in the other equations.

There are many subtle ways that identification can fail in complicated SEMs. To obtain sufficient conditions, we need to extend the rank condition for identification in two-equation systems. This is possible, but it requires matrix algebra [see, for example, Wooldridge (2010, Chapter 9)]. In many applications, one assumes that, unless there is obviously failure of identification, an equation that satisfies the order condition is identified.

The nomenclature on overidentified and just identified equations from Chapter 15 originated with SEMs. In terms of the order condition, (16.27) is an **overidentified equation** because we need only two IVs (for y_2 and y_3) but we have three available (z_2, z_3, and z_4); there is one overidentifying restriction in this equation. In general, the number of overidentifying restrictions equals the total number of exogenous variables in the system minus the total number of explanatory variables in the equation. These can be tested using the overidentification test from Section 15-5. Equation (16.28) is a **just identified equation**, and the third equation is an **unidentified equation**.

16-4b Estimation

Regardless of the number of equations in an SEM, each identified equation can be estimated by 2SLS. The instruments for a particular equation consist of the exogenous variables appearing anywhere in the system. Tests for endogeneity, heteroskedasticity, serial correlation, and overidentifying restrictions can be obtained, just as in Chapter 15.

It turns out that, when any system with two or more equations is correctly specified and certain additional assumptions hold, *system estimation methods* are generally more efficient than estimating each equation by 2SLS. The most common system estimation method in the context of SEMs is *three stage least squares*. These methods, with or without endogenous explanatory variables, are beyond the scope of this text. [See, for example, Wooldridge (2010, Chapters 7 and 8).]

16-5 Simultaneous Equations Models with Time Series

Among the earliest applications of SEMs was estimation of large systems of simultaneous equations that were used to describe a country's economy. A simple Keynesian model of aggregate demand (that ignores exports and imports) is

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1} \tag{16.30}$$

$$I_t = \gamma_0 + \gamma_1 r_t + u_{t2} \tag{16.31}$$

$$Y_t \equiv C_t + I_t + G_t, \tag{16.32}$$

where

$$\begin{aligned} C_t &= \text{consumption}, \\ Y_t &= \text{income}, \\ T_t &= \text{tax receipts}, \\ r_t &= \text{the interest rate}, \\ I_t &= \text{investment, and} \\ G_t &= \text{government spending}. \end{aligned}$$

[See, for example, Mankiw (1994, Chapter 9).] For concreteness, assume t represents year.

The first equation is an aggregate consumption function, where consumption depends on disposable income, the interest rate, and the unobserved structural error u_{t1}. The second equation is a very simple investment function. Equation (16.32) is an *identity* that is a result of national income accounting: it holds by definition, without error. Thus, there is no sense in which we estimate (16.32), but we need this equation to round out the model.

Because there are three equations in the system, there must also be three endogenous variables. Given the first two equations, it is clear that we intend for C_t and I_t to be endogenous. In addition, because of the accounting identity, Y_t is endogenous. We would assume, at least in this model, that T_t, r_t, and G_t are exogenous, so that they are uncorrelated with u_{t1} and u_{t2}. (We will discuss problems with this kind of assumption later.)

If r_t is exogenous, then OLS estimation of equation (16.31) is natural. The consumption function, however, depends on disposable income, which is endogenous because Y_t is. We have two instruments available under the maintained exogeneity assumptions: T_t and G_t. Therefore, if we follow our prescription for estimating cross-sectional equations, we would estimate (16.30) by 2SLS using instruments (T_t, G_t, r_t).

Models such as (16.30) through (16.32) are seldom estimated now, for several good reasons. First, it is very difficult to justify, at an aggregate level, the assumption that taxes, interest rates, and government spending are exogenous. Taxes clearly depend directly on income; for example, with a single marginal income tax rate τ_t in year t, $T_t = \tau_t Y_t$. We can easily allow this by replacing $(Y_t - T_t)$ with $(1 - \tau_t)Y_t$ in (16.30), and we can still estimate the equation by 2SLS if we assume that government spending is exogenous. We could also add the tax rate to the instrument list, if it is exogenous. But are government spending and tax rates really exogenous? They certainly could be in principle, if the government sets spending and tax rates independently of what is happening in the economy. But it is a difficult case to make in reality: government spending generally depends on the level of income, and at high levels of income, the same tax receipts are collected for lower marginal tax rates. In addition, assuming that interest rates are exogenous is extremely questionable. We could specify a more realistic model that includes money demand and supply, and then interest rates could be jointly determined with C_t, I_t, and Y_t. But then finding enough exogenous variables to identify the equations becomes quite difficult (and the following problems with these models still pertain).

Some have argued that certain components of government spending, such as defense spending—see, for example, Hall (1988) and Ramey (1991)—are exogenous in a variety of simultaneous equations applications. But this is not universally agreed upon, and, in any case, defense spending is not always appropriately correlated with the endogenous explanatory variables [see Shea (1993) for discussion and Computer Exercises C6 for an example].

A second problem with a model such as (16.30) through (16.32) is that it is completely static. Especially with monthly or quarterly data, but even with annual data, we often expect adjustment lags. (One argument in favor of static Keynesian-type models is that they are intended to describe the long run without worrying about short-run dynamics.) Allowing dynamics is not very difficult. For example, we could add lagged income to equation (16.31):

$$I_t = \gamma_0 + \gamma_1 r_t + \gamma_2 Y_{t-1} + u_{t2}. \qquad [16.33]$$

In other words, we add a **lagged endogenous variable** (but not I_{t-1}) to the investment equation. Can we treat Y_{t-1} as exogenous in this equation? Under certain assumptions on u_{t2}, the answer is yes. But we typically call a lagged endogenous variable in an SEM a **predetermined variable**. Lags of exogenous variables are also predetermined. If we assume that u_{t2} is uncorrelated with current exogenous variables (which is standard) and all *past* endogenous and exogenous variables, then Y_{t-1} is uncorrelated with u_{t2}. Given exogeneity of r_t, we can estimate (16.33) by OLS.

If we add lagged consumption to (16.30), we can treat C_{t-1} as exogenous in this equation under the same assumptions on u_{t1} that we made for u_{t2} in the previous paragraph. Current disposable income is still endogenous in

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + \beta_3 C_{t-1} + u_{t1}, \qquad [16.34]$$

so we could estimate this equation by 2SLS using instruments (T_t, G_t, r_t, C_{t-l}); if investment is determined by (16.33), Y_{t-1} should be added to the instrument list. [To see why, use (16.32), (16.33), and (16.34) to find the reduced form for Y_t in terms of the exogenous and predetermined variables: T_t, r_t, G_t, C_{t-1}, and Y_{t-1}. Because Y_{t-1} shows up in this reduced form, it should be used as an IV.]

The presence of dynamics in aggregate SEMs is, at least for the purposes of forecasting, a clear improvement over static SEMs. But there are still some important problems with estimating SEMs using aggregate time series data, some of which we discussed in Chapters 11 and 15. Recall that the validity of the usual OLS or 2SLS inference procedures in time series applications hinges on the notion of *weak dependence*. Unfortunately, series such as aggregate consumption, income, investment, and even interest rates seem to violate the weak dependence requirements. (In the terminology of Chapter 11, they have *unit roots*.) These series also tend to have exponential trends, although this can be partly overcome by using the logarithmic transformation and assuming different functional forms. Generally, even the large sample, let alone the small sample, properties of OLS and 2SLS are complicated and dependent on various assumptions when they are applied to equations with I(1) variables. We will briefly touch on these issues in Chapter 18. An advanced, general treatment is given by Hamilton (1994).

Does the previous discussion mean that SEMs are not usefully applied to time series data? Not at all. The problems with trends and high persistence can be avoided by specifying systems in first differences or growth rates. But one should recognize that this is a different SEM than one specified in levels. [For example, if we specify consumption growth as a function of disposable income growth and interest rate changes, this is different from (16.30).] Also, as we discussed earlier, incorporating dynamics is not especially difficult. Finally, the problem of finding truly exogenous variables to include in SEMs is often easier with disaggregated data. For example, for manufacturing industries, Shea (1993) describes how output (or, more precisely, growth in output) in other industries can be used as an instrument in estimating supply functions. Ramey (1991) also has a convincing analysis of estimating industry cost functions by instrumental variables using time series data.

The next example shows how aggregate data can be used to test an important economic theory, the permanent income theory of consumption, usually called the *permanent income hypothesis* (PIH). The approach used in this example is not, strictly speaking, based on a simultaneous equations model, but we can think of consumption and income growth (as well as interest rates) as being jointly determined.

EXAMPLE 16.7 Testing the Permanent Income Hypothesis

Campbell and Mankiw (1990) used instrumental variables methods to test various versions of the PIH. We will use the annual data from 1959 through 1995 in CONSUMP to mimic one of their analyses. Campbell and Mankiw used quarterly data running through 1985.

One equation estimated by Campbell and Mankiw (using our notation) is

$$gc_t = \beta_0 + \beta_1 gy_t + \beta_2 r3_t + u_t, \qquad [16.35]$$

where

$gc_t = \Delta\log(c_t) =$ annual growth in real per capita consumption (excluding durables),

$gy_t =$ growth in real disposable income, and

$r3_t =$ the (ex post) real interest rate as measured by the return on three-month T-bill rates: $r3_t = i3_t - inf_t$, where the inflation rate is based on the Consumer Price Index.

The growth rates of consumption and disposable income are not trending, and they are weakly dependent; we will assume this is the case for $r3_t$ as well, so that we can apply standard asymptotic theory.

The key feature of equation (16.35) is that the PIH implies that the error term u_t has a zero mean conditional on all information observed at time $t-1$ or earlier: $\mathrm{E}(u_t|I_{t-1}) = 0$. However, u_t is *not* necessarily uncorrelated with gy_t or $r3_t$; a traditional way to think about this is that these variables are jointly determined, but we are not writing down a full three-equation system.

Because u_t is uncorrelated with all variables dated $t - 1$ or earlier, valid instruments for estimating (16.35) are lagged values of gc, gy, and $r3$ (and lags of other observable variables, but we will not use those here). What are the hypotheses of interest? The pure form of the PIH has $\beta_1 = \beta_2 = 0$. Campbell and Mankiw argue that β_1 is positive if some fraction of the population consumes current income, rather than permanent income. The PIH with a nonconstant real interest rate implies that $\beta_2 > 0$.

When we estimate (16.35) by 2SLS, using instruments gc_{-1}, gy_{-1}, and $r3_{-1}$ for the endogenous variables gy_t and $r3_t$, we obtain

$$\begin{aligned} \widehat{gc_t} &= \underset{(.0032)}{.0081} + \underset{(.135)}{.586}\, gy_t - \underset{(.00076)}{.00027} r3_t \\ n &= 35,\ R^2 = .678. \end{aligned} \qquad [16.36]$$

Therefore, the pure form of the PIH is strongly rejected because the coefficient on gy is economically large (a 1% increase in disposable income increases consumption by over .5%) and statistically significant $(t = 4.34)$. By contrast, the real interest rate coefficient is very small and statistically insignificant. These findings are qualitatively the same as Campbell and Mankiw's.

The PIH also implies that the errors $\{u_t\}$ are serially uncorrelated. After 2SLS estimation, we obtain the residuals, $\hat{u}_t$, and include $\hat{u}_{t-1}$ as an additional explanatory variable in (16.36); we still use instruments gc_{t-1}, gy_{t-1}, $r3_{t-1}$, *and* $\hat{u}_{t-1}$ acts as its own instrument (see Section 15-7). The coefficient on $\hat{u}_{t-1}$ is $\hat{\rho} = .187$ $(\text{se} = .133)$, so there is some evidence of positive serial correlation, although not at the 5% significance level. Campbell and Mankiw discuss why, with the available quarterly data, positive serial correlation might be found in the errors even if the PIH holds; some of those concerns carry over to annual data.

GOING FURTHER 16.4

Suppose that for a particular city you have monthly data on per capita consumption of fish, per capita income, the price of fish, and the prices of chicken and beef; income and chicken and beef prices are exogenous. Assume that there is no seasonality in the demand function for fish, but there is in the supply of fish. How can you use this information to estimate a constant elasticity demand-for-fish equation? Specify an equation and discuss identification. (*Hint*: You should have 11 instrumental variables for the price of fish.)

Using growth rates of trending or I(1) variables in SEMs is fairly common in time series applications. For example, Shea (1993) estimates industry supply curves specified in terms of growth rates.

If a structural model contains a time trend—which may capture exogenous, trending factors that are not directly modeled—then the trend acts as its own IV.

16-6 Simultaneous Equations Models with Panel Data

Simultaneous equations models also arise in panel data contexts. For example, we can imagine estimating labor supply and wage offer equations, as in Example 16.3, for a group of people working over a given period of time. In addition to allowing for simultaneous determination of variables within each time period, we can allow for unobserved effects in each equation. In a labor supply function, it would be useful to allow an unobserved taste for leisure that does not change over time.

The basic approach to estimating SEMs with panel data involves two steps: (1) eliminate the unobserved effects from the equations of interest using the fixed effects transformation or first differencing and (2) find instrumental variables for the endogenous variables in the transformed equation. This can be very challenging because, for a convincing analysis, we need to find instruments that change over time. To see why, write an SEM for panel data as

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1} \tag{16.37}$$

$$y_{it2} = \alpha_2 y_{it1} + \mathbf{z}_{it2}\boldsymbol{\beta}_2 + a_{i2} + u_{it2}, \tag{16.38}$$

where i denotes cross section, t denotes time period, and $\mathbf{z}_{it1}\boldsymbol{\beta}_1$ or $\mathbf{z}_{it2}\boldsymbol{\beta}_2$ denotes linear functions of a set of exogenous explanatory variables in each equation. The most general analysis allows the unobserved effects, a_{i1} and a_{i2}, to be correlated with *all* explanatory variables, even the elements in z. However, we assume that the idiosyncratic structural errors, u_{it1} and u_{it2}, are uncorrelated with the $\mathbf{z}$ in both equations and across all time periods; this is the sense in which the $\mathbf{z}$ are exogenous. Except under special circumstances, y_{it2} is correlated with u_{it1} and y_{it1} is correlated with u_{it2}.

Suppose we are interested in equation (16.37). We cannot estimate it by OLS, as the composite error $a_{i1} + u_{it1}$ is potentially correlated with all explanatory variables. Suppose we difference over time to remove the unobserved effect, a_{i1}:

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \Delta\mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \tag{16.39}$$

(As usual with differencing or time-demeaning, we can only estimate the effects of variables that change over time for at least some cross-sectional units.) Now, the error term in this equation is uncorrelated with $\Delta\mathbf{z}_{it1}$ by assumption. But Δy_{it2} and Δu_{it1} are possibly correlated. Therefore, we need an IV for Δy_{it2}.

As with the case of pure cross-sectional or pure time series data, possible IVs come from the *other* equation: elements in $\mathbf{z}_{it2}$ that are not also in $\mathbf{z}_{it1}$. In practice, we need *time-varying* elements in $\mathbf{z}_{it2}$ that are not also in $\mathbf{z}_{it1}$. This is because we need an instrument for Δy_{it2}, and a change in a variable from one period to the next is unlikely to be highly correlated with the *level* of exogenous variables. In fact, if we difference (16.38), we see that the natural IVs for Δy_{it2} are those elements in $\Delta\mathbf{z}_{it2}$ that are not also in $\Delta\mathbf{z}_{it1}$.

As an example of the problems that can arise, consider a panel data version of the labor supply function in Example 16.3. After differencing, suppose we have the equation

$$\Delta hours_{it} = \beta_0 + \alpha_1 \Delta\log(wage_{it}) + \Delta(other\ factors_{it}),$$

and we wish to use $\Delta exper_{it}$ as an instrument for $\Delta\log(wage_{it})$. The problem is that, because we are looking at people who work in every time period, $\Delta exper_{it} = 1$ for all i and t. (Each person gets another year of experience after a year passes.) We cannot use an IV that is the same value for all i and t, and so we must look elsewhere. One possibility as an instrument for $\Delta\log(wage_{it})$ is the change in the minimum wage at the state or local level. (As of January 2018, more than 40 localities in the United States have minimum wages above the state minimum wage.) Naturally, in the labor supply function, and, therefore, in the reduced form for $\Delta\log(wage_{it})$, one should include a full set of dummy variables for the different time periods in order to render changes in the minimum wage exogenous to the individual labor supply equation.

Often, participation in an experimental program can be used to obtain IVs in panel data contexts. In Example 15.10, we used receipt of job training grants as an IV for the change in hours of training in determining the effects of job training on worker productivity. In fact, we could view that in an SEM context: job training and worker productivity are jointly determined, but receiving a job training grant is exogenous in equation (15.57).

One can sometimes come up with clever, convincing instrumental variables in panel data applications, as the following example illustrates.

EXAMPLE 16.8 Effect of Prison Population on Violent Crime Rates

In order to estimate the causal effect of prison population increases on crime rates at the state level, Levitt (1996) used instances of prison overcrowding litigation as instruments for the growth in prison population. The equation Levitt estimated is in first differences; we can write an underlying fixed effects model as

$$\log(crime_{it}) = \theta_t + \alpha_l \log(prison_{it}) + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1}, \quad [16.40]$$

where θ_t denotes different time intercepts, and *crime* and *prison* are measured per 100,000 people. (The prison population variable is measured on the last day of the previous year.) The vector $\mathbf{z}_{it1}$ contains log of police per capita, log of income per capita, the unemployment rate, proportions of black and those living in metropolitan areas, and age distribution proportions.

Differencing (16.40) gives the equation estimated by Levitt:

$$\Delta\log(crime_{it}) = \xi_t + \alpha_1 \Delta\log(prison_{it}) + \Delta\mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \quad [16.41]$$

Simultaneity between crime rates and prison population, or more precisely in the growth rates, makes OLS estimation of (16.41) generally inconsistent. Using the violent crime rate and a subset of the data from Levitt (in PRISON, for the years 1980 through 1993, for $51 \cdot 14 = 714$ total observations), we obtain the pooled OLS estimate of α_1, which is $-.181$ $(\text{se} = .048)$. We also estimate (16.41) by pooled 2SLS, where the instruments for $\Delta\log(prison)$ are two binary variables, one each for whether a final decision was reached on overcrowding litigation in the current year or in the previous two years. The pooled 2SLS estimate of α_1 is -1.032 $(\text{se} = .370)$. Therefore, the 2SLS estimated effect is much larger; not surprisingly, it is much less precise, too. Levitt found similar results when using a longer time period (but with early observations missing for some states) and more instruments.

Testing for AR(1) serial correlation in $r_{it1} = \Delta u_{it1}$ is easy. After the pooled 2SLS estimation, obtain the residuals, $\hat{r}_{it1}$. Then, include one lag of these residuals in the original equation, and estimate the equation by 2SLS, where $\hat{r}_{it1}$ acts as its own instrument. The first year is lost because of the lagging. Then, the usual 2SLS *t* statistic on the lagged residual is a valid test for serial correlation. In Example 16.8, the coefficient on $\hat{r}_{it1}$ is only about .076 with $t = 1.67$. With such a small coefficient and modest *t* statistic, we can safely assume serial independence.

An alternative approach to estimating SEMs with panel data is to use the fixed effects transformation and then to apply an IV technique such as pooled 2SLS. A simple procedure is to estimate the time-demeaned equation by pooled 2SLS, which would look like

$$\ddot{y}_{it1} = \alpha_1 \ddot{y}_{it2} + \ddot{\mathbf{z}}_{it1}\boldsymbol{\beta}_1 + \ddot{u}_{it1}, \quad t = 1, 2, \ldots, T, \quad [16.42]$$

where $\ddot{\mathbf{z}}_{it1}$ and $\ddot{\mathbf{z}}_{it2}$ are IVs. This is equivalent to using 2SLS in the dummy variable formulation, where the unit-specific dummy variables act as their own instruments. Ayres and Levitt (1998) applied 2SLS to a time-demeaned equation to estimate the effect of LoJack electronic theft prevention devices on car theft rates in cities. If (16.42) is estimated directly, then the *df* needs to be corrected to $N(T - 1) - k_1$, where k_1 is the total number of elements in α_1 and $\boldsymbol{\beta}_1$. Including unit-specific dummy variables and applying pooled 2SLS to the original data produces the correct *df*. A detailed treatment of 2SLS with panel data is given in Wooldridge (2010, Chapter 11).

Summary

Simultaneous equations models are appropriate when firmly grounded in counterfactual reasoning. In particular, each equation in the system should have a ceteris paribus interpretation. Good examples are when separate equations describe different sides of a market or the behavioral relationships of different economic agents. Supply and demand examples are leading cases, but there are many other applications of SEMs in economics and the social sciences.

An important feature of SEMs is that, by fully specifying the system, it is clear which variables are assumed to be exogenous and which ones appear in each equation. Given a full system, we are able to determine which equations can be identified (that is, can be estimated). In the important case of a two-equation system, identification of (say) the first equation is easy to state: at least one exogenous variable must be excluded from the first equation that appears with a nonzero coefficient in the second equation.

As we know from previous chapters, OLS estimation of an equation that contains an endogenous explanatory variable generally produces biased and inconsistent estimators. Instead, 2SLS can be used to estimate any identified equation in a system. More advanced system methods are available, but they are beyond the scope of our treatment.

The distinction between omitted variables and simultaneity in applications is not always sharp. Both problems, not to mention measurement error, can appear in the same equation. A good example is the labor supply of married women. Years of education (*educ*) appears in both the labor supply and the wage offer functions [see equations (16.19) and (16.20)]. If omitted ability is in the error term of the labor supply function, then wage and education are both endogenous. The important thing is that an equation estimated by 2SLS can stand on its own.

SEMs can be applied to time series data as well. As with OLS estimation, we must be aware of trending, integrated processes in applying 2SLS. Problems such as serial correlation can be handled as in Section 15-7. We also gave an example of how to estimate an SEM using panel data, where the equation is first differenced to remove the unobserved effect. Then, we can estimate the differenced equation by pooled 2SLS, just as in Chapter 15. Alternatively, in some cases, we can use time-demeaning of all variables, including the IVs, and then apply pooled 2SLS; this is identical to putting in dummies for each cross-sectional observation and using 2SLS, where the dummies act as their own instruments. SEM applications with panel data are very powerful, as they allow us to control for unobserved heterogeneity while dealing with simultaneity. They are becoming more and more common and are not especially difficult to estimate.

Key Terms

Endogenous Variables
Exclusion Restrictions
Exogenous Variables
Identified Equation
Just Identified Equation
Lagged Endogenous Variable
Order Condition
Overidentified Equation
Predetermined Variable
Rank Condition
Reduced Form Equation
Reduced Form Error
Reduced Form Parameters
Simultaneity
Simultaneity Bias
Simultaneous Equations Model (SEM)
Structural Equation
Structural Errors
Structural Parameters
Unidentified Equation

Problems

1 Write a two-equation system in "supply and demand form," that is, with the same variable y_t (typically, "quantity") appearing on the left-hand side:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_1 = \alpha_2 y_2 + \beta_2 z_2 + u_2.$$

(i) If $\alpha_1 = 0$ or $\alpha_2 = 0$, explain why a reduced form exists for y_1. (Remember, a reduced form expresses y_1 as a linear function of the exogenous variables and the structural errors.) If $\alpha_1 \neq 0$ and $\alpha_2 = 0$, find the reduced form for y_2.

(ii) If $\alpha_1 \neq 0$, $\alpha_2 \neq 0$, and $\alpha_1 \neq \alpha_2$, find the reduced form for y_1. Does y_2 have a reduced form in this case?

(iii) Is the condition $\alpha_1 \neq \alpha_2$ likely to be met in supply and demand examples? Explain.

2 Let *corn* denote per capita consumption of corn in bushels at the county level, let *price* be the price per bushel of corn, let *income* denote per capita county income, and let *rainfall* be inches of rainfall during the last corn-growing season. The following simultaneous equations model imposes the equilibrium condition that supply equals demand:

$$corn = \alpha_1 price + \beta_1 income + u_1$$
$$corn = \alpha_2 price + \beta_2 rainfall + \gamma_2 rainfall^2 + u_2.$$

Which is the supply equation, and which is the demand equation? Explain.

3 In Problem 3 of Chapter 3, we estimated an equation to test for a tradeoff between minutes per week spent sleeping (*sleep*) and minutes per week spent working (*totwrk*) for a random sample of individuals. We also included education and age in the equation. Because *sleep* and *totwrk* are jointly chosen by each individual, is the estimated tradeoff between sleeping and working subject to a "simultaneity bias" criticism? Explain.

4 Suppose that annual earnings and alcohol consumption are determined by the SEM

$$\log(earnings) = \beta_0 + \beta_1 alcohol + \beta_2 educ + u_1$$
$$alcohol = \gamma_0 + \gamma_1 \log(earnings) + \gamma_2 educ + \gamma_3 \log(price) + u_2,$$

where *price* is a local price index for alcohol, which includes state and local taxes. Assume that *educ* and *price* are exogenous. If β_1, β_2, γ_1, γ_2, and γ_3 are all different from zero, which equation is identified? How would you estimate that equation?

5 A simple model to determine the effectiveness of condom usage on reducing sexually transmitted diseases among sexually active high school students is

$$infrate = \beta_0 + \beta_1 conuse + \beta_2 percmale + \beta_3 avginc + \beta_4 city + u_1,$$

where

infrate = the percentage of sexually active students who have contracted venereal disease.

conuse = the percentage of boys who claim to use condoms regularly.

avginc = average family income.

city = a dummy variable indicating whether a school is in a city.

The model is at the school level.

(i) Interpreting the preceding equation in a causal, ceteris paribus fashion, what should be the sign of β_1?

(ii) Why might *infrate* and *conuse* be jointly determined?

(iii) If condom usage increases with the rate of venereal disease, so that $\gamma_1 > 0$ in the equation

$$conuse = \gamma_0 + \gamma_1 infrate + other\ factors,$$

what is the likely bias in estimating β_1 by OLS?

(iv) Let *condis* be a binary variable equal to unity if a school has a program to distribute condoms. Explain how this can be used to estimate β_1 (and the other betas) by IV. What do we have to assume about *condis* in each equation?

6 Consider a linear probability model for whether employers offer a pension plan based on the percentage of workers belonging to a union, as well as other factors:

$$pension = \beta_0 + \beta_1 percunion + \beta_2 avgage + \beta_3 avgeduc$$
$$+ \beta_4 percmale + \beta_5 percmarr + u_1.$$

(i) Why might *percunion* be jointly determined with *pension*?

(ii) Suppose that you can survey workers at firms and collect information on workers' families. Can you think of information that can be used to construct an IV for *percunion*?

(iii) How would you test whether your variable is at least a reasonable IV candidate for *percunion*?

7 For a large university, you are asked to estimate the demand for tickets to women's basketball games. You can collect time series data over 10 seasons, for a total of about 150 observations. One possible model is

$$lATTEND_t = \beta_0 + \beta_1 lPRICE_t + \beta_2 WINPERC_t + \beta_3 RIVAL_t + \beta_4 WEEKEND_t + \beta_5 t + u_t,$$

where

$PRICE_t$ = the price of admission, probably measured in real terms—say, deflating by a regional consumer price index.

$WINPERC_t$ = the team's current winning percentage.

$RIVAL_t$ = a dummy variable indicating a game against a rival.

$WEEKEND_t$ = a dummy variable indicating whether the game is on a weekend.

The l denotes natural logarithm, so that the demand function has a constant price elasticity.

(i) Why is it a good idea to have a time trend in the equation?

(ii) The supply of tickets is fixed by the stadium capacity; assume this has not changed over the 10 years. This means that quantity supplied does not vary with price. Does this mean that price is necessarily exogenous in the demand equation? (*Hint*: The answer is no.)

(iii) Suppose that the nominal price of admission changes slowly—say, at the beginning of each season. The athletic office chooses price based partly on last season's average attendance, as well as last season's team success. Under what assumptions is last season's winning percentage $(SEASPERC_{t-1})$ a valid instrumental variable for $lPRICE_t$?

(iv) Does it seem reasonable to include the (log of the) real price of men's basketball games in the equation? Explain. What sign does economic theory predict for its coefficient? Can you think of another variable related to men's basketball that might belong in the women's attendance equation?

(v) If you are worried that some of the series, particularly *lATTEND* and *lPRICE*, have unit roots, how might you change the estimated equation?

(vi) If some games are sold out, what problems does this cause for estimating the demand function? (*Hint*: If a game is sold out, do you necessarily observe the true demand?)

8 How big is the effect of per-student school expenditures on local housing values? Let *HPRICE* be the median housing price in a school district and let *EXPEND* be per-student expenditures. Using panel data for the years 1992, 1994, and 1996, we postulate the model

$$lHPRICE_{it} = \theta_t + \beta_1 lEXPEND_{it} + \beta_2 lPOLICE_{it} + \beta_3 lMEDINC_{it} + \beta_4 PROPTAX_{it} + a_{i1} + u_{it1},$$

where $POLICE_{it}$ is per capita police expenditures, $MEDINC_{it}$ is median income, and $PROPTAX_{it}$ is the property tax rate; l denotes natural logarithm. Expenditures and housing price are simultaneously determined because the value of homes directly affects the revenues available for funding schools.

Suppose that, in 1994, the way schools were funded was drastically changed: rather than being raised by local property taxes, school funding was largely determined at the state level. Let $lSTATEALL_{it}$ denote the log of the state allocation for district i in year t, which is exogenous in the preceding equation, once we control for expenditures and a district fixed effect. How would you estimate the β_j?

Computer Exercises

C1 Use SMOKE for this exercise.

(i) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness, or productivity effects) is

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1,$$

where *cigs* is number of cigarettes smoked per day, on average. How do you interpret β_1?

(ii) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is

$$\begin{aligned} cigs = {} & \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 \\ & + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2, \end{aligned}$$

where *cigpric* is the price of a pack of cigarettes (in cents) and *restaurn* is a binary variable equal to unity if the person lives in a state with restaurant smoking restrictions. Assuming these are exogenous to the individual, what signs would you expect for γ_5 and γ_6?

(iii) Under what assumption is the income equation from part (i) identified?

(iv) Estimate the income equation by OLS and discuss the estimate of β_1.

(v) Estimate the reduced form for *cigs*. (Recall that this entails regressing *cigs* on all exogenous variables.) Are log(*cigpric*) and *restaurn* significant in the reduced form?

(vi) Now, estimate the income equation by 2SLS. Discuss how the estimate of β_1 compares with the OLS estimate.

(vii) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?

C2 Use MROZ for this exercise.

(i) Reestimate the labor supply function in Example 16.5, using log(*hours*) as the dependent variable. Compare the estimated elasticity (which is now constant) to the estimate obtained from equation (16.24) at the average hours worked.

(ii) In the labor supply equation from part (i), allow *educ* to be endogenous because of omitted ability. Use *motheduc* and *fatheduc* as IVs for *educ*. Remember, you now have two endogenous variables in the equation.

(iii) Test the overidentifying restrictions in the 2SLS estimation from part (ii). Do the IVs pass the test?

C3 Use the data in OPENNESS for this exercise.

(i) Because log(*pcinc*) is insignificant in both (16.22) and the reduced form for *open*, drop it from the analysis. Estimate (16.22) by OLS and IV without log(*pcinc*). Do any important conclusions change?

(ii) Still leaving log(*pcinc*) out of the analysis, is *land* or log(*land*) a better instrument for *open*? (*Hint*: Regress *open* on each of these separately and jointly.)

(iii) Now, return to (16.22). Add the dummy variable *oil* to the equation and treat it as exogenous. Estimate the equation by IV. Does being an oil producer have a ceteris paribus effect on inflation?

C4 Use the data in CONSUMP for this exercise.

(i) In Example 16.7, use the method from Section 15-5 to test the single overidentifying restriction in estimating (16.35). What do you conclude?

(ii) Campbell and Mankiw (1990) use *second* lags of all variables as IVs because of potential data measurement problems and informational lags. Reestimate (16.35), using only gc_{t-2}, gy_{t-2}, and $r3_{t-2}$ as IVs. How do the estimates compare with those in (16.36)?

(iii) Regress gy_t on the IVs from part (ii) and test whether gy_t is sufficiently correlated with them. Why is this important?

C5 Use the *Economic Report of the President* (2005 or later) to update the data in CONSUMP, at least through 2003. Reestimate equation (16.35). Do any important conclusions change?

C6 Use the data in CEMENT for this exercise.

(i) A static (inverse) supply function for the monthly growth in cement price (*gprc*) as a function of growth in quantity (*gcem*) is

$$gprc_t = \alpha_1 gcem_t + \beta_0 + \beta_1 gprcpet + \beta_2 feb_t + \cdots + \beta_{12} dec_t + u_t^s,$$

where *gprcpet* (growth in the price of petroleum) is assumed to be exogenous and *feb*, . . . , *dec* are monthly dummy variables. What signs do you expect for α_1 and β_1? Estimate the equation by OLS. Does the supply function slope upward?

(ii) The variable *gdefs* is the monthly growth in real defense spending in the United States. What do you need to assume about *gdefs* for it to be a good IV for *gcem*? Test whether *gcem* is partially correlated with *gdefs*. (Do not worry about possible serial correlation in the reduced form.) Can you use *gdefs* as an IV in estimating the supply function?

(iii) Shea (1993) argues that the growth in output of residential (*gres*) and nonresidential (*gnon*) construction are valid instruments for *gcem*. The idea is that these are demand shifters that should be roughly uncorrelated with the supply error u_t^s. Test whether *gcem* is partially correlated with *gres* and *gnon*; again, do not worry about serial correlation in the reduced form.

(iv) Estimate the supply function, using *gres* and *gnon* as IVs for *gcem*. What do you conclude about the static supply function for cement? [The dynamic supply function is, apparently, upward sloping; see Shea (1993).]

C7 Refer to Example 13.9 and the data in CRIME4,

(i) Suppose that, after differencing to remove the unobserved effect, you think $\Delta\log(polpc)$ is simultaneously determined with $\Delta\log(crmrte)$; in particular, increases in crime are associated with increases in police officers. How does this help to explain the positive coefficient on $\Delta\log(polpc)$ in equation (13.33)?

(ii) The variable *taxpc* is the taxes collected per person in the county. Does it seem reasonable to exclude this from the crime equation?

(iii) Estimate the reduced form for $\Delta\log(polpc)$ using pooled OLS, including the potential IV, $\Delta\log(taxpc)$. Does it look like $\Delta\log(taxpc)$ is a good IV candidate? Explain.

(iv) Suppose that, in several of the years, the state of North Carolina awarded grants to some counties to increase the size of their county police force. How could you use this information to estimate the effect of additional police officers on the crime rate?

C8 Use the data set in FISH, which comes from Graddy (1995), to do this exercise. The data set is also used in Computer Exercise C9 in Chapter 12. Now, we will use it to estimate a demand function for fish.

(i) Assume that the demand equation can be written, in equilibrium for each time period, as

$$\log(totqty_t) = \alpha_1 \log(avgprc_t) + \beta_{10} + \beta_{11} mon_t + \beta_{12} tues_t + \beta_{13} wed_t + \beta_{14} thurs_t + u_{t1},$$

so that demand is allowed to differ across days of the week. Treating the price variable as endogenous, what additional information do we need to estimate the demand-equation parameters consistently?

(ii) The variables $wave2_t$ and $wave3_t$ are measures of ocean wave heights over the past several days. What two assumptions do we need to make in order to use $wave2_t$ and $wave3_t$ as IVs for $\log(avgprc_t)$ in estimating the demand equation?

(iii) Regress $\log(avgprc_t)$ on the day-of-the-week dummies and the two wave measures. Are $wave2_t$ and $wave3_t$ jointly significant? What is the *p*-value of the test?

(iv) Now, estimate the demand equation by 2SLS. What is the 95% confidence interval for the price elasticity of demand? Is the estimated elasticity reasonable?

(v) Obtain the 2SLS residuals, $\hat{u}_{t1}$. Add a single lag, $\hat{u}_{t-1,1}$ in estimating the demand equation by 2SLS. Remember, use $\hat{u}_{t-1,1}$ as its own instrument. Is there evidence of AR(1) serial correlation in the demand equation errors?

(vi) Given that the supply equation evidently depends on the wave variables, what two assumptions would we need to make in order to estimate the price elasticity of supply?

(vii) In the reduced form equation for $\log(avgprc_t)$, are the day-of-the-week dummies jointly significant? What do you conclude about being able to estimate the supply elasticity?

C9 For this exercise, use the data in AIRFARE, but only for the year 1997.

(i) A simple demand function for airline seats on routes in the United States is

$$\log(passen) = \beta_{10} + \alpha_1 \log(fare) + \beta_{11}\log(dist) + \beta_{12}[\log(dist)]^2 + u_1,$$

where

$$passen = \text{average passengers per day,}$$
$$fare = \text{average airfare, and}$$
$$dist = \text{the route distance (in miles).}$$

If this is truly a demand function, what should be the sign of α_1?

(ii) Estimate the equation from part (i) by OLS. What is the estimated price elasticity?

(iii) Consider the variable *concen*, which is a measure of market concentration. (Specifically, it is the share of business accounted for by the largest carrier.) Explain in words what we must assume to treat *concen* as exogenous in the demand equation.

(iv) Now assume *concen* is exogenous to the demand equation. Estimate the reduced form for log(*fare*) and confirm that *concen* has a positive (partial) effect on log(*fare*).

(v) Estimate the demand function using IV. Now what is the estimated price elasticity of demand? How does it compare with the OLS estimate?

(vi) Using the IV estimates, describe how demand for seats depends on route distance.

C10 Use the entire panel data set in AIRFARE for this exercise. The demand equation in a simultaneous equations unobserved effects model is

$$\log(passen_{it}) = \theta_{t1} + \alpha_1 \log(fare_{it}) + a_{i1} + u_{it1},$$

where we absorb the distance variables into a_{i1}.

(i) Estimate the demand function using fixed effects, being sure to include year dummies to account for the different intercepts. What is the estimated elasticity?

(ii) Use fixed effects to estimate the reduced form

$$\log(fare_{it}) = \theta_{t2} + \pi_{21} concen_{it} + a_{i2} + v_{it2}.$$

Perform the appropriate test to ensure that $concen_{it}$ can be used as an IV for $\log(fare_{it})$.

(iii) Now estimate the demand function using the fixed effects transformation along with IV, as in equation (16.42). What is the estimated elasticity? Is it statistically significant?

C11 A common method for estimating *Engel curves* is to model expenditure shares as a function of total expenditure, and possibly demographic variables. A common specification has the form

$$sgood = \beta_0 + \beta_1 ltotexpend + demographics + u,$$

where *sgood* is the fraction of spending on a particular good out of total expenditure and *ltotexpend* is the log of total expenditure. The sign and magnitude of β_1 are of interest across various expenditure categories.

To account for the potential endogeneity of *ltotexpend*—which can be viewed as an omitted variables or simultaneous equations problem, or both—the log of family income is often used as an instrumental variable. Let *lincome* denote the log of family income. For the remainder of this question, use the data in EXPENDSHARES, which comes from Blundell, Duncan, and Pendakur (1998).

(i) Use *sfood*, the share of spending on food, as the dependent variable. What is the range of values of *sfood*? Are you surprised there are no zeros?

(ii) Estimate the equation

$$sfood = \beta_0 + \beta_1 ltotexpend + \beta_2 age + \beta_3 kids + u \quad [16.43]$$

by OLS and report the coefficient on *ltotexpend*, $\hat{\beta}_{OLS,1}$, along with its heteroskedasticity-robust standard error. Interpret the result.

(iii) Using *lincome* as an IV for *ltotexpend*, estimate the reduced form equation for *ltotexpend*; be sure to include *age* and *kids*. Assuming *lincome* is exogenous in (16.43), is *lincome* a valid IV for *ltotexpend*?

(iv) Now estimate (16.43) by instrumental variables. How does $\hat{\beta}_{IV,1}$ compare with $\hat{\beta}_{OLS,1}$? What about the robust 95% confidence intervals?

(v) Use the test in Section 15-5 to test the null hypothesis that *ltotexpend* is exogenous in (16.43). Be sure to report and interpret the *p*-value. Are there any overidentifying restrictions to test?

(vi) Substitute *salcohol* for *sfood* in (16.43) and estimate the equation by OLS and 2SLS. Now what do you find for the coefficients on *ltotexpend*?

C12 Use the data in PRISON.DTA to answer the following questions. Refer to Example 16.8. In the data set, variables beginning with "g" are growth rates from one year to the next, obtained as the changes in the natural log. For example, $gcriv_{it} = \log(criv_{it}) - \log(criv_{i,t-1})$. Variables beginning with "c" are changes in levels from one year to the next, for example, $cunem_{it} = unem_{it} - unem_{i,t-1}$.

(i) Estimate the equation

$$\begin{aligned} gcriv_{it} = {} & \xi_t + \alpha_1 gpris_{it} + \beta_1 gincpc_{it} + \beta_2 gpolpc_{it} + \beta_3 cag0\_14_{it} + \beta_4 cag15\_17_{it} \\ & + \beta_5 cag18\_24_{it} + \beta_6 cag25\_34_{it} + \beta_7 cunem_{it} + \beta_8 cblack_{it} + \beta_9 cmetro_{it} + \Delta u_{it} \end{aligned}$$

by OLS and verify that you obtain $\hat{\alpha}_1 = -0.181$ (se $= .048$). The parameters ξ_t are to remind you to include year dummies for 1981 through 1993.

(ii) Estimate the reduced form equation for $gpris_{it}$, where $final1_{it}$ and $final2_{it}$ are the instruments:

$$\begin{aligned} gpris_{it} = {} & \eta_t + \gamma_1 final1_{it} + \gamma_2 final2_{it} + \pi_1 gincpc_{it} + \pi_2 gpolpc_{it} + \pi_3 cag0\_14_{it} \\ & + \pi_4 cag15\_17_{it} + \pi_5 cag18\_24_{it} + \pi_6 cag25\_34_{it} + \pi_7 cunem_{it} + \pi_8 cblack_{it} \\ & + \pi_9 cmetro_{it} + \Delta u_{it} \end{aligned}$$

Verify that γ_1 and γ_2 are both negative. Are they each statistically significant? What is the F statistic for $H_0: \gamma_1 = \gamma_2 = 0$? (Remember again to put in a full set of year dummies.)

(iii) Obain the 2SLS estimates of the equation in part (i), using $final1_{it}$ and $final2_{it}$ as instruments for $gpris_{it}$. Verify that you obtain $\hat{\alpha}_1 = -1.032$ (se $= .370$).

(iv) If you have access to econometrics software that computes standard errors robust to heteroskedasticity and serial correlation, obtain them for the 2SLS estimate in part (iii). What happens to the standard error of $\hat{\alpha}_1$?

(v) Reestimate the reduced form in part (ii) using the differences, $\Delta final1_{it}$ and $\Delta final2_{it}$, as the instruments. (You will lose 1980 in differencing the instruments.) Do $\Delta final1_{it}$ and $\Delta final2_{it}$ seem like sufficiently strong instruments for $gpris_{it}$? In particular, do you prefer using the differences or levels as the IVS? Estimate the reduced form in part (ii) dropping 1980 to be sure that you reach your conclusion on instrument strength using the same set of data.