Chapter Fifteen

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. While 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 variables. In addition, the panel data methods which 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. The next chapter 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; (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 a proxy variable.

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 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

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

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:

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

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.

Sometimes, requirement (15.4) is summarized by saying that "z is exogenous in equation (15.2)." In the context of omitted variables, this means that z should have no partial effect on y and z should not be correlated with other factors that affect y. Equation (15.5) means that z must be related, either positively or negatively, to the endogenous explanatory variable x.

There is a very important difference between the two requirements for an instrumental variable. Because (15.4) is a covariance between z and the unobservable error u, it can never be checked or even tested: we must maintain this assumption by appealing to economic behavior or a gut feeling. 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

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

against the two-sided alternative H_0: $\pi_1 \neq 0$, at a sufficiently small significance level (say, 5% or 1%). 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 unobservable 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, this variable is not correlated with education, so it makes a poor instrumental variable for *educ*.

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* must 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*.

The requirements are less clear-cut for other possible instrumental variable candidates. 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).

Another IV choice for *educ* in (15.1) is number of siblings while growing up (*sibs*). Typically, 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

$$score = \beta_0 + \beta_1 skipped + u, \tag{15.8}$$

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*: better 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. At the same time, the IV must be correlated with *skipped*. One option is to use distance between living quarters and campus. Some students at a large university will commute to campus, which 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.

We now demonstrate that the availability of an instrumental variable can be used to consistently estimate the parameters in equation (15.2). In particular, we show that assumptions (15.4) and (15.5) [equivalently, (15.4) and (15.7)] 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

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

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

$$\beta_1 = \frac{\text{Cov}(z,y)}{\text{Cov}(z,x)}. \tag{15.9}$$

[Notice how this simple algebra fails if z and x are uncorrelated, that is, if $\text{Cov}(z,x) = 0$.] 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:

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

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 identical to the OLS estimator.

A simple application of the law of large numbers shows that the IV estimator is consistent for β_1: $\text{plim}(\hat{\beta}_1) = \beta_1$, 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.

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

$$E(u^2|z) = \sigma^2 = \text{Var}(u). \tag{15.11}$$

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

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}, \tag{15.12}$$

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 a couple of 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,

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

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

$$\frac{\hat{\sigma}^2}{\text{SST}_x \cdot R_{x,z}^2}, \tag{15.13}$$

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 hee. Any modern econometrics package computes the standard error after any IV estimation.

Before we give an example, it is useful 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 $\sigma^2/(\text{SST}_x \cdot R_{x,z}^2)$; they differ only in that $R_{x,z}^2$ appears in the denominator of the IV variance. Since an R-squared is always less than one, the 2SLS variance is always larger than the OLS variance (when OLS is valid). If $R_{x,z}^2$ is small, then the IV variance can be much larger than the OLS variance. Remember, $R_{x,z}^2$ 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.RAW to estimate the return to education in the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u. \tag{15.14}$$

For comparison, we first obtain the OLS estimates:

$$\begin{gathered} \log(\widehat{wage}) = -.185 + .109\ educ \\ (.185)\quad (.014) \\ n = 428,\ R^2 = .118. \end{gathered} \tag{15.15}$$

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):

$$\begin{gathered} \widehat{educ} = 10.24 + .269\ fatheduc \\ (0.28)\quad (.029) \\ n = 428,\ R^2 = .173. \end{gathered} \tag{15.16}$$

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

$$\begin{gathered} \log(\widehat{wage}) = .441 + .059\ educ \\ (.446)\quad (.035) \\ n = 428,\ R^2 = .093. \end{gathered} \tag{15.17}$$

The IV estimate of the return to education is 5.9%, which is about 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, while 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.RAW 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:

$$\widehat{educ} = \underset{(0.11)}{14.14} - \underset{(.030)}{.228}\, sibs$$
$$n = 935,\ R^2 = .057.$$

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

$$\log(\widehat{wage}) = \underset{(0.36)}{5.13} + \underset{(.026)}{.122}\, educ$$
$$n = 935.$$

(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. While 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*) had quantitative meaning. But both types can be 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 pursuasively 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 in 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.

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. Since the numbers given were (eventually) randomly assigned, it seems plausible that draft lottery number is uncorrelated with the error term u. But those

QUESTION 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)?

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 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 15.1 for an example.

Properties of IV with a Poor Instrumental Variable

We have already seen that, while 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. This can be derived in terms of population correlations and standard deviations as

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}, \tag{15.19}$$

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 $\tilde{\beta}_1$—as

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \text{Corr}(x,u) \cdot \frac{\sigma_u}{\sigma_x}. \tag{15.20}$$

Comparing these formulas shows that IV is preferred to OLS on asymptotic bias grounds when $\text{Corr}(z,u)/\text{Corr}(z,x) < \text{Corr}(x,u)$.

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.

E X A M P L E 1 5 . 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.RAW:

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

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

$$\begin{aligned} \log(\widehat{bwght}) &= \underset{(0.91)}{4.45} + \underset{(8.70)}{2.99}\, packs \\ n &= 1{,}388 \end{aligned}$$

(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).

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: $\mathrm{E}(u_1) = 0$. 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 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. Since 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. Since 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:

$$E(u_1) = 0, \text{Cov}(z_1,u_1) = 0, \text{and Cov}(z_2,u_1) = 0. \quad \textbf{(15.24)}$$

Given the zero mean assumption, the latter two assumptions are equivalent to $E(z_1u_1) = E(z_2u_1) = 0$, 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):

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

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:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2, \quad \textbf{(15.26)}$$

where, by definition,

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

and the π_j are unknown parameters. The key identification condition [along with (15.24)] is that

$$\pi_2 \neq 0. \quad \textbf{(15.27)}$$

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; this must be taken on faith.

QUESTION 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 percent 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 the next chapter—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

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1, \tag{15.28}$$

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

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

The reduced form for y_2 is

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots \pi_{k-1} z_{k-1} + \pi_k z_k + v_2, \tag{15.30}$$

and we need some partial correlation between z_k and y_2:

$$\pi_k \neq 0. \tag{15.31}$$

Under (15.29) and (15.31), z_k is a valid IV for y_2. (We do not care about the remaining π_j; some or all of them could be zero.) It makes sense to think that $z_1, \ldots, z_{k-1}$ serve as their own IVs; therefore, the list of exogenous variables is often called the list of instrumental variables. 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.RAW, we obtain, in condensed form,

$$\begin{aligned} \widehat{educ} &= \underset{(0.24)}{16.64} + \underset{(.088)}{.320}\, nearc4 - \underset{(.034)}{.413}\, exper + \dots \\ n &= 3{,}010,\ R^2 = .477. \end{aligned} \tag{15.32}$$

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. 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 from .024 and .239, which is a very wide range. Larger con-

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*		

fidence intervals is a price we must pay to get a consistent estimator of the return to education when we think *educ* is endogenous.

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.

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.

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

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2, \tag{15.33}$$

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^*:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3. \tag{15.34}$$

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:

$$\pi_2 \neq 0 \text{ or } \pi_3 \neq 0. \tag{15.35}$$

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:

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

(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

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

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

With multiple instruments, the IV estimator 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

$$y_1 \text{ on } \hat{y}_2 \text{ and } z_1. \tag{15.38}$$

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 (15.38). This is found to be true by plugging $y_2 = y_2^* + v_2$ into (15.22):

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2. \tag{15.39}$$

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

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1, \tag{15.40}$$

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

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2, \tag{15.41}$$

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.RAW. First, we test H_0: $\pi_3 = 0$, $\pi_4 = 0$ in (15.41) using an F test. The result is $F = 55.40$, 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,

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

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 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),

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1, \tag{15.42}$$

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.

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

$$\frac{\sigma^2}{\mathrm{SST}_2(1 - R_2^2)}, \tag{15.43}$$

where $\sigma^2 = \mathrm{Var}(u_1)$, SST_2 is the total variation in $\hat{y}_2$, and 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 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.) 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, $R^2 = .475$; 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, $R^2 = .995$, 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 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 R_2^2.

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

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1, \tag{15.44}$$

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 the reduced forms of y_2 and y_3. (As before, we can use F statistics to test this.) While 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**.

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 (1999, Chapter 5).]

QUESTION 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:

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

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?

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 $\text{SSR}_r \geq \text{SSR}_{ur}$; 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 (1999, 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

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u, \tag{15.45}$$

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1). \tag{15.46}$$

If the classical errors-in-variables (CEV) assumptions hold, the bias in the OLS estimator of β_1 is towards 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 must 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. Since 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

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \tag{15.47}$$

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

$$\begin{aligned}\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 \\ + \alpha_1 test_1 + (u - \alpha_1 e_1),\end{aligned} \tag{15.48}$$

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 $test_1$.

E X A M P L E 1 5 . 6

(Using Two Test Scores as Indicators of Ability)

We use the data in WAGE2.RAW 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.

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,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1, \tag{15.49}$$

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

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2. \tag{15.50}$$

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.

E X A M P L E 1 5 . 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%).

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.

An interesting feature of the regression from part (ii) is that the estimates on all of the variables (except $\hat{v}_2$) are identical to the 2SLS estimates. For example, estimating (15.51) by OLS gives $\hat{\beta}_j$ that are identical to the 2SLS estimates from equation (15.49). This is a simple way to see if you have done the proper regression in testing for endogeneity. It also gives another interpretation of 2SLS: including $\hat{v}_2$ in the OLS regression (15.51) clears up the endogeneity of y_2.

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.

Testing Overidentification Restrictions

When we introduced the simple instrumental variables estimator in Section 15.1, we emphasized that an IV must satisfy two requirements: it must be uncorrelated with the error and correlated with the endogenous explanatory variable. We have seen in fairly complicated models how to decide whether the second requirement can be tested using a t or an F test in the reduced form regression. We claimed that the first requirement cannot be tested because it involves a correlation between the IV and an unobserved error. However, if we have more than one instrumental variable, we can effectively test whether some of them are uncorrelated with the structural error.

As an example, again consider equation (15.49) with two additional instrumental variables, z_3 and z_4. We know we can estimate (15.49) using only z_3 as an IV for y_2. Given the IV estimates, we can compute the residuals, $\hat{u}_1 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 y_2 - \hat{\beta}_2 z_1 - \hat{\beta}_3 z_2$. Because z_4 is not used at all in the estimation, we can check whether z_4 and $\hat{u}_1$ are correlated in the sample. If they are, z_4 is not a valid IV for y_2. Of course, this tells us nothing about whether z_3 and u_1 are correlated; in fact, for this to be a useful test, we must *assume* that z_3 and u_1 are uncorrelated. Nevertheless, if z_3 and z_4 are chosen using the same logic—such as mother's education and father's education—finding that z_4 is correlated with u_1 casts doubt on using z_3 as an IV.

Because the roles of z_3 and z_4 can be reversed, we can also test whether z_3 is correlated with u_1, provided z_4 and u_1 are assumed to be uncorrelated. Which test should we use? It turns out that our test choice does not matter. We must assume that at least one IV is exogenous. Then, we can test the **overidentifying restrictions** that are used in 2SLS. For our purposes, the number of overidentifying restrictions is simply the number of extra instrumental variables. Suppose we have only one endogenous explanatory variable. If we have only a single IV for y_2, we have *no* overidentifying restrictions, and there is nothing that can be tested. If we have two IVs for y_2, as in the previous example, we have one overidentifying restriction. If we have three IVs, we have two overidentifying restrictions, and so on.

Testing overidentifying restrictions is rather simple. We must obtain the 2SLS residuals and then run an auxiliary regression.

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.

E X A M P L E 1 5 . 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). Therefore, 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).

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 (1999, 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. 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 $z_1, z_2, \dots, z_m$ denote all the exogenous variables (incuding those used as IVs for the endogenous explanatory variables). Then, under reasonable assumptions [spelled out, for example, in Wooldridge (1999, Chapter 5)], an asymptotically valid statistic is the usual F statistic for joint significance in a regression of $\hat{u}^2$ on $z_1, z_2, \dots, z_m$. The null hypothesis of homoskedasticity is rejected if the z_j are jointly significant.

If we apply this 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 $\text{Var}(u|z_1, z_2, \dots, z_m)$, 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

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

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

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

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.

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

QUESTION 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 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 $gGOV_t$?

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

and the no serial correlation assumption is stated as

$$\mathrm{E}(u_t u_s|z_t, z_s) = 0, \text{ for all } t \neq s, \tag{15.54}$$

where 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 Problem 15.15.

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

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

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

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

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):

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

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, $\tilde{z}_{tj} \equiv z_{tj} - \rho z_{t-1,j}$. 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.

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.RAW 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 and father's education levels (*meduc*, *feduc*). The 2SLS estimate of β_{educ} is $-.153$ (se = .039), compared with the OLS estimate $-.128$ (se = .018). The 2SLS estimate shows a somewhat larger effect of education on fertility, but the 2SLS standard 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 is due to sampling error.

Instrumental variables estimation can be combined with panel data methods, particularly first differencing, to consistently estimate parameters 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:

$$\Delta\log(scrap_i) = \delta_0 + \beta_1 \Delta hrsemp_i + \Delta u_i. \tag{15.57}$$

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.RAW differenced between 1987 and 1988, the first stage regression is

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

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

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

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 $\hat{\beta}_1 = -.0076$ (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.

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 (1999, 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 consistently estimate the parameters in a linear model 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 often 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, while 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
Identification
Instrumental Variables
Instrumental Variables (IV) Estimator
Natural Experiment
Omitted Variables
Order Condition
Overidentifying Restrictions
Rank Condition
Reduced Form Equation
Structural Equation
Two Stage Least Squares (2SLS) Estimator

PROBLEMS

15.1 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*.

15.2 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 in order 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*.]

15.3 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).

15.4 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 $gMIN_t$.

15.5 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?

15.6 (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 α_j?

15.7 The following is a simple model to measure the effect of a school choice program on standardized test performance [see Rouse (1998) for motivation]:

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

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*?)

15.8 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 twenty-mile radius of a girl's home to be a valid IV for *girlhs*.

15.9 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?

15.10 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 these variables 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 effect on attending a Catholic high school. Do you think *CathRel* is a convincing instrument for *CathHS*?

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

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t^* + u_t \\ x_t &= x_t^* + e_t, \end{aligned} \tag{15.58}$$

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

15.12 Use the data in WAGE2.RAW for this exercise.

(i) In Example 15.2, using *sibs* 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).

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

(i) Estimate this model by OLS

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

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) *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.

15.14 Use the data in CARD.RAW for this exercise.

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

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

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?

15.15 Use the data in INTDEF.RAW 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?)

15.16 Use the data in CARD.RAW for this exercise.

(i) In Table 15.1, the difference between the IV and OLS estimates of the return to education are economically important. Obtain the reduced form residuals, $\hat{v}_2$, from (15.32). (See Table 15.1 for the other variables to include in the regression.) 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).

15.17 Use the data in MURDER.RAW 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 occured 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)?

15.18 Use the data in PHILLIPS.RAW 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.

APPENDIX 15A

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.

ASSUMPTION 2SLS.1 (LINEAR IN PARAMETERS)

The model in the population can be written as

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

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

ASSUMPTION 2SLS.2 (RANDOM SAMPLING)

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

ASSUMPTION 2SLS.3 (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.

ASSUMPTION 2SLS.4 (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 $z_1, \dots, z_m$ denote the exogenous variables, where $z_k, \dots, z_m$ do not appear in the structural model (15.42). The reduced form of y_2 is

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

Then, we need at least one of $\pi_k, \dots, \pi_m$ 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 (1999, Chapter 5).]

THEOREM 15A.1
Under Assumptions 2SLS.1 through 2SLS.4, the 2SLS estimator is consistent.

ASSUMPTION 2SLS.5 (HOMOSKEDASTICITY)
Let $\boldsymbol{z}$ denote the collection of all instrumental variables. Then $E(u^2|\mathbf{z}) = \sigma^2$.

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 $\hat{\sigma}^2 = (n - k - 1)^2 \sum_{i=1}^{n} \hat{u}_i^2$, 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 (1999).

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 (1999, 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.

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.