CHAPTER 10

Endogenous Regressors and Moment-Based Estimation

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Give an intuitive explanation of why correlation between a random x and the error term causes the least squares estimator to be inconsistent.
2. Describe the "errors-in-variables" problem in econometrics and its consequences for the least squares estimator.
3. Describe the properties of a good instrumental variable.
4. Discuss how the method of moments can be used to derive the least squares and instrumental variables estimators, paying particular attention to the assumptions upon which the derivations are based.
5. Explain why it is important for an instrumental variable to be highly correlated with the random explanatory variable for which it is an instrument.
6. Describe how instrumental variables estimation is carried out in the case of surplus instruments.
7. State the approximate large-sample distribution of the instrumental variables estimator for the simple linear regression model, and how it can be used for the construction of interval estimates and hypothesis tests.
8. Describe a test for the existence of contemporaneous correlation between the error term and the contemporaneous explanatory variables in a model, explaining the null and alternative hypotheses, and the consequences of rejecting the null hypothesis.

KEYWORDS

asymptotic properties
conditional expectation
endogenous variables
errors-in-variables
exogenous variables
first-stage regression
Hausman test
instrumental variable
instrumental variable estimator
just-identified
large sample properties
overidentified
population moments
random sampling
reduced-form
sample moments
sampling properties
simultaneous equations bias
surplus moment conditions
two-stage least squares estimation
weak instruments

In this chapter we reconsider the linear regression model. We will initially discuss the simple linear regression model, but our comments apply to the general model as well. The usual assumptions are SR1–SR6, given in Section 2.2.2. In Chapter 8, we relaxed the assumption $\text{var}(e_i|\mathbf{X}) = \sigma^2$ that the error variance is the same for all observations. In Chapter 9 we considered regressions with time-series data in which the assumption of serially uncorrelated errors, $\text{cov}(e_i, e_j|\mathbf{X}) = 0$, for $i \neq j$, cannot be maintained.

In this chapter, we relax the exogeneity assumption. When an explanatory variable is random, the properties of the least squares estimator depend on the characteristics of the independent variable x. The assumption of **strict exogeneity** is SR2 in the simple regression model, $E(e_i|\mathbf{x}) = 0$, and it is MR2 in the multiple regression model, $E(e_i|\mathbf{X}) = 0$. The mathematical form of this assumption is simple but the full meaning is complex. In Section 2.10.2, we gave common simple regression model examples when this assumption might fail. In these cases, with an explanatory variable that is **endogenous**, the usual least squares estimator does not have its desirable properties; it is not an unbiased estimator of the population parameters $\beta_1, \beta_2, \ldots$; it is not a consistent estimator of $\beta_1, \beta_2, \ldots$; tests and interval estimators do not have the anticipated properties, and even having large data samples will not cure the problems.

We review and discuss the properties of the least squares estimator with an endogenous explanatory variable in this chapter, and we suggest a new estimator, the **instrumental variables** estimator, that does have some desirable properties in large samples. The instrumental variables estimator is also called a **method of moments** estimator, and also the **two-stage least squares** estimator. We offer fair warning, however, that this area of econometrics is filled with practical and theoretical difficulties. Our search turns from finding an estimator that is "best" to one that is "adequate," and unfortunately producing convincing research applications requires knowledge, skill, and patience. In order for you to begin properly you should reread (right now!) Section 2.10 on the exogeneity concept and Section 5.7 on the large sample, or asymptotic, properties of the least squares estimator.

10.1 Least Squares Estimation with Endogenous Regressors

As our starting point, let us assume we are working with microeconomic, cross-sectional data obtained by **random sampling**. The standard assumptions for the simple regression model are RS1–RS6, which we repeat here for your convenience.

The Simple Linear Regression Model Under Random Sampling

RS1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

RS2: The data pairs (y_i, x_i) are statistically independent of all other data pairs and have the same joint distribution $f(y_i, x_i)$. They are independent and identically distributed (iid).

RS3: $E(e_i|x_i) = 0$ for $i = 1, \ldots, N$; x is contemporaneously, and strictly, exogenous.

RS4: The random error has constant conditional variance, $\text{var}(e_i|x_i) = \sigma^2$.

RS5: x_i takes at least two different values.

RS6: $e_i \sim N(0, \sigma^2)$

With random sampling, the ith and jth observations are statistically independent, so that the ith error e_i is statistically independent from the jth value of the explanatory variable, x_j. Thus, the

strict exogeneity assumption $E(e_i|x_1, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$ reduces to the simpler contemporaneous exogeneity assumption $E(e_i|x_i) = 0$.

Recall from Chapter 2 that the "gold standard" in research is a randomized controlled experiment. In an ideal (research) world, we would randomly assign x values (the treatment) and examine changes in outcomes y (the effect). If there is a systematic relationship between changes in x and changes in the outcome y, we can claim that changes in x **cause** changes in the outcome y. Any other random factors, "everything else" $= e$, that might affect the outcome are statistically independent of x. We can isolate, or identify, the effects of changes in x alone, and using regression analysis, we can estimate the causal effect $\Delta E(y_i|x_i)/\Delta x_i = \beta_2$.

The importance of the strict exogeneity assumption $E(e_i|x_i) = 0$ is that if it is true then "x is as good as randomly assigned." If $E(e_i|x_i) = 0$, then the best prediction of the random error e_i is simply zero. [See Appendix 4C for the details behind this statement.] There is no information contained in the values of x that helps us predict the random error. We can infer a causal relationship between y_i and x_i when there is covariation between them because variations in the random error e_i are uncorrelated with the variations in the explanatory variable x_i. It is just "as if" we had randomly assigned the treatments, x_i, to experimental subjects. Furthermore under RS1–RS6, the least squares estimators of β_1 and β_2 are the best linear unbiased estimators and the usual interval estimators and hypothesis tests work as they are expected to in samples of all sizes.

10.1.1 Large Sample Properties of the OLS Estimator

In Section 5.7, we introduced "large sample" or "asymptotic" analysis. With large samples of data, strict exogeneity is not required to identify and estimate a causal effect. All that we require is the simpler condition that the x values are uncorrelated with the random errors, e, and that the average of the random errors is zero. Econometricians, statisticians, and mathematicians aim to develop methods that work with as few strong assumptions as possible. We adopt that attitude and replace RS3, strict exogeneity, with

$$\text{RS3}^*: \; E(e_i) = 0 \text{ and } \text{cov}(x_i, e_i) = 0$$

Instead of contemporaneous exogeneity, we simply assume that the random error e_i and the explanatory variable value x_i are **contemporaneously uncorrelated**, which is a weaker condition than $E(e_i|x_i) = 0$. The term **contemporaneous** means "occurring at the same point in time" or, as in this case, occurring for the same cross-sectional observation subscript i. Explanatory variables like this, that are contemporaneously uncorrelated with the regression error, are simply said to be **exogenous**.

If we have obtained a random sample, then the selection of any person is statistically independent of the selection of any other person. Any randomly selected person's characteristics, such as education, income, ability, and race, are statistically independent of the characteristics of any other person selected. Because random sampling automatically implies zero correlation between the ith and jth observations, we only require that the ith value x_i be uncorrelated with e_i. The correlation between the ith error e_i and the jth value of the explanatory variable, x_j, is zero automatically because of random sampling.

Regression assumption RS3* says two things. First, in a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the population average of all unobservable characteristics, or variables omitted from the regression model, is zero, $E(e_i) = 0$. Second, in the population the correlation between the explanatory variable x_i and all the factors combined into the random error e_i is zero, or $\text{cov}(x_i, e_i) = 0$.

We can replace RS3 by RS3* because, if assumption RS3 is true, it follows that RS3* is true, that is, $E(e_i|x_i) = 0 \Rightarrow \text{cov}(x_i, e_i) = 0$ and $E(e_i|x_i) = 0 \Rightarrow E(e_i) = 0$. These relations are proven in Appendix 2G.1. Introducing assumption RS3* is convenient because it is a simpler notion of exogeneity, which is good. However, assumption RS3* is weaker than RS3 and under it we cannot show that the least squares estimator is unbiased, or that any of the other properties hold in small samples. What we can show is that the least squares estimators have

desirable **large sample properties**. Under assumptions RS1, RS2, RS3*, RS4, and RS5 the least squares estimators:

1. are consistent; that is, they converge in probability to the true parameter values as $N \to \infty$;
2. have approximate normal distributions in large samples, whether the random errors are normally distributed or not; and
3. provide interval estimators and test statistics that are valid if the sample is large.

In practice, this means that all the usual interpretations, intervals estimates, hypothesis tests, predictions, and prediction intervals are fine as long as our sample is large and RS1, RS2, RS3*, RS4, and RS5 hold. If samples are large, and if $\text{cov}(x_i, e_i) = 0$ and $E(e_i) = 0$, then it is "almost as good as" randomly assigning treatment values to x_i. We can estimate the population parameters $\beta_1, \beta_2, \ldots$ using the least squares estimator. If there is serial correlation or heteroskedasticity, then the robust standard error methods from Chapters 8 and 9 are fine as long as RS3* holds.

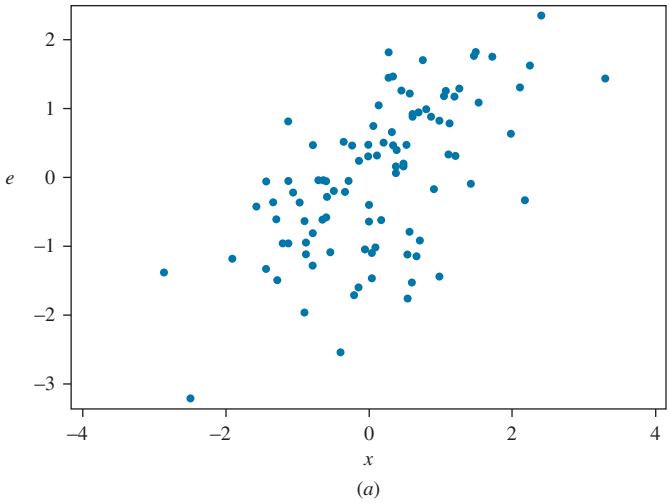
Remark

Do not fall into the trap of thinking "I'll just assume this, or that, if I want this or that result." It is true that access to large samples of data means not having to worry about the complexities of strict exogeneity. But what if you do not have access to large samples? Then statistical inference (estimation, hypothesis testing, and prediction) in small, or finite, samples is important. When the sample size N is not large, the **asymptotic properties** of estimators may be very misleading. Estimators that may be fine in large samples may suffer large biases in small samples. Estimates may appear statistically significant when they are not, and confidence intervals may be too narrow or too wide. If governments, or businesses, make decisions based on faulty inferences then we may suffer large economic or personal losses as a result. It is not just a game.

If assumption RS3* is *not* true, and in particular if $\text{cov}(x_i, e_i) \neq 0$ so that x_i and e_i are contemporaneously correlated, then the least squares estimators are inconsistent. They do not converge to the true parameter values even in very large samples. Furthermore, our usual hypothesis testing or interval estimation procedures are not valid. This means that estimating causal relationships using the least squares estimator when $\text{cov}(x_i, e_i) \neq 0$ may lead to incorrect inferences. When x_i is random, the relationship between x_i and e_i is a crucial factor when deciding whether least squares estimation, either OLS or GLS, is appropriate or not. If the error term e_i is correlated with x_i (or any x_{ik} in the multiple regression model) then the least squares estimator fails. In the next section we explain why correlation between x_i and e_i leads to the failure of the least squares estimator.

10.1.2 Why Least Squares Estimation Fails

In this section, we provide an intuitive explanation why the least squares estimator fails when $\text{cov}(x_i, e_i) \neq 0$. An algebraic proof is in the next section. The regression model **data generation process** adds a random error e_i to the systematic regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ to obtain the observed outcome y_i. In Figure 10.1(a), x_i and e_i values are positively correlated, violating the strict exogeneity assumption. In Figure 10.1(b), the positively sloped regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, which is the object of our analysis, is the solid line. For each value of x_i, the y_i data values, $y_i = \beta_1 + \beta_2 x_i + e_i$, are the sum of the systematic portion $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ and a random error e_i. The data pairs (y_i, x_i) are the dots in Figure 10.1(b). As you see, the true regression function does not pass through the middle of the data in this case and that is because of the correlation between x_i and e_i. The y_i values for larger x_i values tend to have positive errors, $e_i > 0$. The y_i values for smaller x_i values have negative errors, $e_i < 0$. In this case, we can use



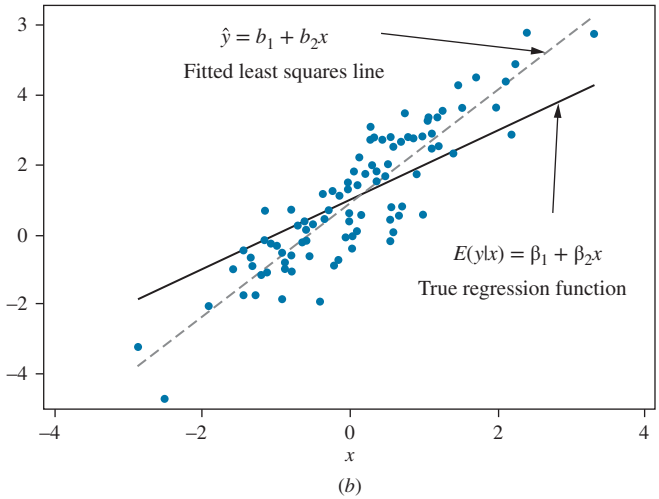


FIGURE 10.1 **(a) Correlated x and e. (b) Plot of data, true and fitted regression functions.**

information provided by the x_i values to provide a better prediction of the random error e_i than simply zero.

Least squares estimation leads to a fitted line passing through the middle of the data, shown as a dashed line in Figure 10.1(b). The slope of the fitted line $\left(\text{the estimate } b_2\right)$ overestimates the true slope of the regression function, $\beta_2 > 0$. The least squares estimator attributes all variation in y_i to variation in x_i. When x_i and e_i are correlated, the variation in y_i comes from two sources: changes in x_i and changes in e_i, and in our example these changes have a positive correlation. If we think about the effect of changes in x_i and e_i on y_i we have

$$\begin{array}{ccccc} \Delta y_i & = & \beta_2 \Delta x_i & + & \Delta e_i \\ (+) & & (+) & & (+) \end{array}$$

If x_i and e_i are positively correlated and $\beta_2 > 0$, increases in x_i and e_i combine to increase y_i. In the least squares estimation process, all the change (increase) in y_i is attributed to the effect of the change (increase) in x_i, and thus the least squares estimator will overestimate β_2.

Throughout this Chapter, we use the relation between wages and years of education as an example. In this case, the omitted variable "intelligence," or ability, is in the regression error, and it is likely to be positively correlated with the years of education a person receives, with more

intelligent individuals usually choosing to obtain more years of education. When regressing wage on years of education, the least squares estimator attributes increases in wages to increases in education. The effect of education is overstated because some of the increase in wages is also due to higher intelligence.

The statistical consequence of a contemporaneous correlation between x_i and e_i is that the least squares estimator is biased, and this bias will not disappear no matter how large the sample is. Consequently, the least squares estimator is **inconsistent** when there is contemporaneous correlation between x_i and e_i.

Remark

If x_i is endogenous the least squares estimator still is a useful **predictive** tool. In Figure 10.1(b) the least squares fitted line fits the data well. Given a value x_0 we can predict y_0 using the fitted line. What we cannot do is interpret the slope of the line as a causal effect.

10.1.3 Proving the Inconsistency of OLS

Let us prove that the least squares estimator is not consistent when $\text{cov}(x_i, e_i) \neq 0$. Our regression model is $y_i = \beta_1 + \beta_2 x_i + e_i$. Continue to assume that $E(e_i) = 0$, so that $E(y_i) = \beta_1 + \beta_2 E(x_i)$. Then,

- Subtract this expectation from the original equation,

$$y_i - E(y_i) = \beta_2\left[x_i - E(x_i)\right] + e_i$$

- Multiply both sides by $x_i - E(x_i)$

$$\left[x_i - E(x_i)\right]\left[y_i - E(y_i)\right] = \beta_2\left[x_i - E(x_i)\right]^2 + \left[x_i - E(x_i)\right]e_i$$

- Take expected values of both sides

$$E\left[x_i - E(x_i)\right]\left[y_i - E(y_i)\right] = \beta_2 E\left[x_i - E(x_i)\right]^2 + E\left\{\left[x_i - E(x_i)\right]e_i\right\},$$

or

$$\text{cov}(x_i, y_i) = \beta_2 \text{var}(x_i) + \text{cov}(x_i, e_i)$$

- Solve for β_2

$$\beta_2 = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)} - \frac{\text{cov}(x_i, e_i)}{\text{var}(x_i)}$$

This equation is the basis for showing when the least squares estimator is consistent, and when it is not.

If we can assume that $\text{cov}(x_i, e_i) = 0$, then

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

We drop the "i" subscript because we are randomly sampling from a population, and the data pairs are not only independently distributed but identically distributed, with the same joint *pdf* $f(x_i, y_i)$, and thus $\text{cov}(x_i, y_i) = \text{cov}(x, y)$ and $\text{var}(x_i) = \text{var}(x)$. The least squares estimator is

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/(N-1)}{\sum(x_i - \bar{x})^2/(N-1)} = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)}$$

This shows that the least squares estimator b_2 is the sample analog of the population relationship, $\beta_2 = \text{cov}(x, y)/\text{var}(x)$. The sample variance and covariance converge to the true variance and covariance as the sample size N increases, using the Law of Large Numbers introduced in Section 10.3.1, so that the least squares estimator converges to β_2. That is, if $\text{cov}(x_i, e_i) = 0$, then

$$b_2 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} \rightarrow \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2$$

showing that the least squares estimator is consistent.

On the other hand, if x_i and e_i are correlated, then

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)} - \frac{\text{cov}(x, e)}{\text{var}(x)}$$

The least squares estimator now converges to

$$b_2 \rightarrow \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2 + \frac{\text{cov}(x, e)}{\text{var}(x)} \neq \beta_2$$

In this case, b_2 is an inconsistent estimator of β_2 and the amount of bias that exists even asymptotically, when samples can be assumed to be large, is $\text{cov}(x, e)/\text{var}(x)$. The direction of the bias depends on the sign of the covariance between x_i and e_i. If factors in the error are positively correlated with the explanatory variable x, then the least squares estimator will overestimate the true parameter. If factors in the error are negatively correlated with the explanatory variable x, then the least squares estimator will underestimate the true parameter.

In the following section, we describe some common situations in which there is correlation between x_i and e_i causing the least squares estimator to fail.

10.2 Cases in Which x and e are Contemporaneously Correlated

There are several common situations in which the least squares estimator fails due to the presence of a contemporaneous correlation between an explanatory variable and the error term. When an explanatory variable and an error term are contemporaneously correlated, the explanatory variable is said to be **endogenous**. This term comes from simultaneous equations models, which we will consider in Chapter 11, and means "determined within the system." When an explanatory variable is contemporaneously correlated with the regression error one is said to have an "endogeneity problem."

10.2.1 Measurement Error

The **errors-in-variables** problem occurs when an explanatory variable is measured with error. If we measure an explanatory variable with error, then it is correlated with the error term, and the least squares estimator is inconsistent. As an illustration, consider the following important example. Let us assume that an individual's personal saving is based on their "permanent" or long-run income. Let y_i = annual savings of the ith person and let x_i^* = the permanent annual income of the ith person. A simple regression model representing this relationship is

$$y_i = \beta_1 + \beta_2 x_i^* + v_i \tag{10.1}$$

We have asterisked (*) the permanent income variable because it is difficult, if not impossible, to observe. For the purposes of a regression, suppose that we attempt to measure permanent income

using x_i = current income. Current income is a measure of permanent income, but it does not measure permanent income exactly. To capture this measurement error specify that

$$x_i = x_i^* + u_i \tag{10.2}$$

where u_i is a random disturbance, with mean 0 and variance σ_u^2. With this statement, we are admitting that observed current income only approximates permanent income, and consequently that we have measured permanent income with error. Furthermore, assume that the measurement error u_i is independent of the regression error v_i. When we use x_i in the regression in place of x_i^*, we do so by replacement, that is, substitute $x_i^* = x_i - u_i$ into (10.1) to obtain

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i^* + v_i = \beta_1 + \beta_2(x_i - u_i) + v_i = \beta_1 + \beta_2 x_i + (v_i - \beta_2 u_i) \\ &= \beta_1 + \beta_2 x_i + e_i \end{aligned} \tag{10.3}$$

In order to estimate (10.3) by OLS, we must determine whether or not x_i is contemporaneously uncorrelated with the random error e_i. The covariance between these two random variables, using the fact that $E(e_i) = 0$ and assuming that x_i^* is exogenous in (10.1), so that $E(x_i^* v_i) = 0$, is

$$\begin{aligned} \operatorname{cov}(x_i, e_i) &= E(x_i e_i) = E\left[(x_i^* + u_i)(v_i - \beta_2 u_i)\right] \\ &= E(-\beta_2 u_i^2) = -\beta_2 \sigma_u^2 \neq 0 \end{aligned} \tag{10.4}$$

The least squares estimator b_2 is an *inconsistent* estimator of β_2 in (10.3) because of the correlation between the explanatory variable x_i and the error term e_i. Consequently, b_2 does not converge to β_2 in large samples. Furthermore, in large or small samples, b_2 is *not* approximately normal with mean β_2 and variance $\operatorname{var}(b_2) = \sigma^2 / \sum(x_i - \bar{x})^2$. When ordinary least squares fails in this way, is there another estimation approach that works? The answer is yes, as we will see in Section 10.3.

Note that in equation (10.4), if $\beta_2 > 0$, there is a negative correlation between x_i and the random error e_i. The least squares estimator will underestimate β_2 and in the literature devoted to measurement error this is called **attenuation bias**. This is a logical result of using $x_i = x_i^* + u_i$. Imagine that the measurement error u_i is very large relative to x_i^*. Then x_i becomes more like a completely random number and there will be little association between y_i and x_i in the data, so that b_2 will be near zero.

10.2.2 Simultaneous Equations Bias

Another situation in which an explanatory variable is correlated with the regression error term arises in simultaneous equations models. While this terminology may not sound familiar, students of economics deal with such models from their earliest introduction to supply and demand. Recall that in a competitive market the prices and quantities of goods are determined jointly by the forces of supply and demand. Thus, if P_i = equilibrium price and Q_i = equilibrium quantity, we can say that P_i and Q_i are endogenous, because they are jointly determined within a simultaneous system of two equations, one equation for the supply curve and one equation for the demand curve. Suppose that we write down the relation

$$Q_i = \beta_1 + \beta_2 P_i + e_i \tag{10.5}$$

We know that changes in price affect the quantities supplied and demanded. But it is also true that changes in quantities supplied and demanded lead to changes in prices. There is a feedback relationship between P_i and Q_i. Because of this feedback, which results because price and quantity are jointly, or simultaneously, determined, we can show that $\operatorname{cov}(P_i, e_i) \neq 0$. The least squares estimation procedure will fail if applied to (10.5) because of the endogeneity problem, and the resulting bias (and inconsistency) is called **simultaneous equations bias**. Supply and demand models permeate economic analysis, and we will treat simultaneous equations models fully in Chapter 11.

10.2.3 Lagged-Dependent Variable Models with Serial Correlation

In Chapter 9, we introduced dynamic models with stationary variables. One way to make models dynamic is to introduce a lagged dependent variable into the right-hand side of an equation, That is, $y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$. The lagged variable y_{t-1} is a random regressor, but as long as it is uncorrelated with the error term e_t then the least squares estimator is consistent. However, it is possible when specifying a dynamic model that the errors will be serially correlated. If the errors e_t follow the AR(1) process $e_t = \rho e_{t-1} + v_t$, then we can see that the lagged dependent variable y_{t-1} must be correlated with the error term e_t, because y_{t-1} depends directly on e_{t-1}, and e_{t-1} directly affects the value of e_t. If $\rho \neq 0$, there will be a correlation between y_{t-1} and e_t. In this case, the OLS estimator applied to the lagged dependent variable model will be biased and inconsistent. Thus, it is very important to test for the presence of serial correlation in models with lagged dependent variables on the right-hand side (see Sections 9.4 and 9.5).

10.2.4 Omitted Variables

When an omitted variable is correlated with an included explanatory variable, then the regression error will be correlated with the explanatory variable. We introduced this idea in Section 6.3.1. A classic example is from labor economics. A person's wage is determined in part by their level of education. Let us specify a log-linear regression model explaining observed hourly wage as

$$\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + e_i \tag{10.6}$$

with $EDUC_i$ = years of education and $EXPER_i$ = years of work experience. What else affects wages? What is omitted from the model? This thought experiment should be carried out each time a regression model is formulated. There are several factors we might think of, such as labor market conditions, region of the country, and union membership. However, labor economists are most concerned about the omission of a variable measuring ability. It is logical that a person's ability, intelligence and industriousness may affect the quality of their work and their wage. These variables are components of the random error e_i, since we usually have no measure for them. The problem is not only that ability might affect wages but more able individuals may also spend more years in school, causing a positive correlation between the error terms e_i and $EDUC_i$, so that $\text{cov}(EDUC_i, e_i) > 0$. If this is true, then we can expect that the least squares estimator of the returns to another year of education will be positively biased, $E(b_2) > \beta_2$, and inconsistent, meaning that the bias will not disappear even in very large samples.

EXAMPLE 10.1 | Least Squares Estimation of a Wage Equation

We will use the data on married women in the data file *mroz* to estimate the wage model in (10.6). Using the $N = 428$ women in the sample who are in the labor force, the least squares estimates and their standard errors are

$$\begin{aligned} \ln(WAGE) &= -0.5220 + 0.1075 \times EDUC \\ \text{(se)} \qquad & \quad (0.1986) \quad (0.0141) \\ &+ 0.0416 \times EXPER - 0.0008 \times EXPER^2 \\ & \quad (0.0132) \qquad\quad (0.0004) \end{aligned}$$

We estimate that an additional year of education increases wages by approximately 10.75%, holding everything else constant. If ability has a positive effect on wages, then this estimate is overstated, as the contribution of ability is attributed to the education variable.

The social and policy importance of the estimate 0.1075 can hardly be exaggerated. Countries invest a large portion of tax revenue to improve education. Why? Spending on education is an investment, and like any other investment investors (taxpaying citizens) expect a rate of return that is competitive with rates of returns for alternative projects. Based on the estimated equation above, additional years of schooling are estimated to increase wages by 10.75%, holding other factors fixed, meaning that individuals are

more likely to be self-sufficient, enjoy a good quality of life, not requiring welfare or public health assistance, and less likely to engage in crime. Suppose, however, that 10.75% overestimates the returns to education for wage income. We might re-evaluate the investment in education and perhaps decide to spend tax dollars on bridges or parks instead of schools. Evaluating the social rate of return to education is a social policy problem. Regression estimates such as those above play heavily into the calculation. Consequently we must do all that we can, as econometricians, to obtain estimates using the best methods. In the next section we begin our examination of alternative estimation methods for models in which regression errors are correlated with regression variables.

10.3 Estimators Based on the Method of Moments

In the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, when x_i is random and $\text{cov}(x_i, e_i) \neq 0$, the least squares estimators are biased and inconsistent, with none of their usual nice properties holding. When faced with such a situation we must consider alternative estimation procedures. In this section we discuss the "method of moments" principle of estimation, which is an alternative to the least squares estimation principle. When all the usual assumptions of the linear model hold, the method of moments leads us to the least squares estimator. If x_i is random and correlated with the error term, the method of moments leads us to an alternative, called instrumental variables estimation or two-stage least squares estimation, that will work in large samples.

10.3.1 Method of Moments Estimation of a Population Mean and Variance

Let us begin with a simple case. The kth moment of a random variable Y is the expected value of the random variable raised to the kth power. That is,

$$E(Y^k) = \mu_k = k\text{th moment of } Y \tag{10.7}$$

The **Law of Large Numbers (LLN)** is a famous theorem. One version says: if $X_1, X_2, \ldots, X_N$ is a random sample from a population, and if $E(X_i) = \mu < \infty$ and $\text{var}(X_i) = \sigma^2 < \infty$, then the sample mean $\overline{X} = \sum X_i/N$ converges (in probability) to the expected value (population mean) μ as the sample size N increases. In this case, $\overline{X}$ is said to be a consistent estimator of μ. It is useful to remember that in most situations **sample moments** are consistent estimators of **population moments**.

We can apply the law of large numbers to obtain a consistent estimator of $E(Y^k) = \mu_k$ by letting $X_i = Y_i^k$ and $E(X_i) = \mu = E(Y_i^k) = \mu_k$. Then, assuming that $\text{var}(Y_i^k) = \sigma_k^2 < \infty$, a consistent estimator of the population moment $E(Y^k) = \mu_k$ is the corresponding sample moment

$$\widehat{E(Y^k)} = \hat{\mu}_k = k\text{th sample moment of } Y = \sum Y_i^k/N \tag{10.8}$$

The **method of moments** estimation procedure equates m population moments to m sample moments to estimate m unknown parameters. As an example, let Y be a random variable with mean $E(Y) = \mu$ and variance, given in the Probability Primer, equation (P.13):

$$\text{var}(Y) = \sigma^2 = E(Y - \mu)^2 = E(Y^2) - \mu^2 \tag{10.9}$$

In order to estimate the two population parameters μ and σ^2, we must equate two population moments to two sample moments. Let $Y_1, Y_2, \ldots, Y_N$ be a random sample from the population.

The first two population and sample moments of Y are

$$\begin{array}{ll} \text{Population moments} & \text{sample moments} \\ E(Y) = \mu_1 = \mu & \hat{\mu} = \sum Y_i / N \\ E(Y^2) = \mu_2 & \hat{\mu}_2 = \sum Y_i^2 / N \end{array} \tag{10.10}$$

Note that for the first population moment μ_1, it is customary to drop the subscript and use μ to denote the population mean of Y. With these two moments, we can solve for the unknown mean and variance parameters. Equate the first sample moment in (10.10) to the first population moment to obtain an estimate of the population mean,

$$\hat{\mu} = \sum Y_i / N = \overline{Y} \tag{10.11}$$

Then use (10.9), replacing the second population moment in (10.10) by its sample value and replacing first moment μ by (10.11)

$$\tilde{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{\sum Y_i^2}{N} - \overline{Y}^2 = \frac{\sum Y_i^2 - N\overline{Y}^2}{N} = \frac{\sum \left(Y_i - \overline{Y}\right)^2}{N} \tag{10.12}$$

The method of moments leads us to the sample mean as an estimator of the population mean. The method of moments estimator of the variance has N in its denominator, rather than the usual $N-1$, so it is not exactly the sample variance we are used to. But in large samples this will not make much difference. In general, method of moments estimators are consistent, and converge to the true parameter values in large samples, but there is no guarantee that they are "best" in any sense.

10.3.2 Method of Moments Estimation in the Simple Regression Model

The definition of a "moment" can be extended to more general situations. Assumption RS3* states that $E(e_i) = 0$ and $\text{cov}(x_i, e_i) = E(x_i e_i) = 0$. Using these two equations, we can derive the OLS estimator by using the method of moments approach. In the linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the two moment conditions $E(e_i) = 0$ and $E(x_i e_i) = 0$ imply

$$E(e_i) = 0 \Rightarrow E(y_i - \beta_1 - \beta_2 x_i) = 0 \tag{10.13}$$

and

$$E(x_i e_i) = 0 \Rightarrow E\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0 \tag{10.14}$$

Equations (10.13) and (10.14) are population moment conditions. The Law of Large Numbers says that under random sampling, sample moments converge to population moments, so

$$\frac{1}{N}\sum(y_i - \beta_1 - \beta_2 x_i) \xrightarrow{p} E(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] \xrightarrow{p} E\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0$$

Setting the two sample moment conditions to zero and replacing the unknown parameters β_1 and β_2 by their estimators b_1 and b_2, we have two equations and two unknowns

$$\frac{1}{N}\sum(y_i - b_1 - b_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[x_i(y_i - b_1 - b_2 x_i)\right] = 0$$

Multiplying these two equations by N we have the two **normal equations** (2A.3) and (2A.4) given in Appendix 2A, and solving them yields the least squares estimators,

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

What we have shown is that under the weaker assumptions, $E(e_i) = 0$ and zero contemporaneous covariance between x_i and e_i, $\text{cov}(x_i, e_i) = E(x_i e_i) = 0$, we can derive the OLS estimators for the simple linear regression model using the method of moments approach. Further, as we have discussed in Section 5.7, the OLS estimators are **consistent estimators** in this case, and have their usual properties in large samples.

10.3.3 Instrumental Variables Estimation in the Simple Regression Model

Problems for least squares estimation arise when x_i is random and contemporaneously correlated with the random error e_i, so that $\text{cov}(x_i, e_i) = E(x_i e_i) \neq 0$. In this case x_i is **endogenous**. As we have discussed in Sections 5.7 and 6.3, and Appendix 6B, the OLS estimator is biased and **inconsistent** when an explanatory variable is endogenous. Also, in the method of moments context, endogeneity makes the moment condition in equation (10.14) invalid.

What are we to do? The method of moments approach gives us an insight into an alternative. Suppose that there is another variable, z_i, with the following properties:

Characteristics of a Good Instrumental Variable

IV1: z_i does not have a direct effect on y_i, and thus it does not belong on the right-hand side of the model $y_i = \beta_1 + \beta_2 x_i + e_i$ as an explanatory variable.

IV2: z_i is not contemporaneously correlated with the regression error term e_i, so that $\text{cov}(z_i, e_i) = E(z_i e_i) = 0$. Variables with the property $\text{cov}(z_i, e_i) = E(z_i e_i) = 0$ are said to be **exogenous**.

IV3: z_i is strongly (or at least not weakly) correlated with x_i, the endogenous explanatory variable.

A variable z_i with these properties is called an **instrumental variable**. This terminology arises because while z does not have a direct effect on y, having it will allow us to estimate the relationship between x and y. It is a *tool*, or **instrument**, that we are using to achieve our objective.

If such a variable z exists, then we can use it to form a moment condition to replace (10.14), that is,

$$E(z_i e_i) = 0 \Rightarrow E\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0 \tag{10.15}$$

Then we can use the two moment equations (10.13) and (10.15) to obtain estimates of β_1 and β_2. Again appealing to the Law of Large numbers, we can assert that sample moments converge to population moments. Therefore,

$$\frac{1}{N}\sum(y_i - \beta_1 - \beta_2 x_i) \xrightarrow{p} E(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] \xrightarrow{p} E\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0$$

Assuming we have a sufficiently large sample, we set the sample moments to zero, yielding the two sample moment conditions

$$\begin{aligned}\frac{1}{N}\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)&=0\\ \frac{1}{N}\sum z_i\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)&=0\end{aligned}\tag{10.16}$$

Solving these equations leads us to method of moments estimators, which in economics are usually called the **instrumental variable (IV) estimators**,

$$\begin{aligned}\hat{\beta}_2&=\frac{N\sum z_iy_i-\sum z_i\sum y_i}{N\sum z_ix_i-\sum z_i\sum x_i}=\frac{\sum(z_i-\bar{z})(y_i-\bar{y})}{\sum(z_i-\bar{z})(x_i-\bar{x})}\\ \hat{\beta}_1&=\bar{y}-\hat{\beta}_2\bar{x}\end{aligned}\tag{10.17}$$

We introduce the notation $\hat{\beta}_1$ and $\hat{\beta}_2$ for the instrumental variables estimators to differentiate them from the OLS estimators b_1 and b_2. If properties IV1, IV2, and IV3 hold, then these new estimators are **consistent**, they converge to the true parameter values as the sample size $N\to\infty$. Also, they have approximate normal distributions in large samples, which we denote by "$\overset{a}{\sim}$". For the simple regression model

$$\hat{\beta}_2\overset{a}{\sim}N\left[\beta_2,\widehat{\text{var}}\left(\hat{\beta}_2\right)\right]$$

where the estimated variance is

$$\widehat{\text{var}}\left(\hat{\beta}_2\right)=\frac{\hat{\sigma}^2_{IV}\sum(z_i-\bar{z})^2}{\left[\sum(z_i-\bar{z})(x_i-\bar{x})\right]^2}\tag{10.18a}$$

The IV estimator of the error variance σ^2 is

$$\hat{\sigma}^2_{IV}=\frac{\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)^2}{N-2}\tag{10.18b}$$

10.3.4 The Importance of Using Strong Instruments

When working with instrumental variables, a constantly repeated question is "How strong are the instruments?" What is a strong instrument? We will develop a full answer to that question in this chapter, but initially, we define a strong instrument z as one that is highly correlated with the endogenous variable x. To show why this definition is useful, apply a bit of algebra to the expression for the variance $\text{var}\left(\hat{\beta}_2\right)$ in equation (10.18a) to obtain an informative equivalent expression.

$$\begin{aligned}\widehat{\text{var}}\left(\hat{\beta}_2\right)&=\frac{\hat{\sigma}^2_{IV}\sum(z_i-\bar{z})^2}{\left[\sum(z_i-\bar{z})(x_i-\bar{x})\right]^2}\\ &=\frac{\hat{\sigma}^2_{IV}}{\left\{\dfrac{\left[\sum(z_i-\bar{z})(x_i-\bar{x})\right]^2/(N-1)}{\sum(z_i-\bar{z})^2\sum(x_i-\bar{x})^2/(N-1)}\right\}\sum(x_i-\bar{x})^2}\\ &=\frac{\hat{\sigma}^2_{IV}}{r^2_{zx}\sum(x_i-\bar{x})^2}\end{aligned}$$

We simply multiplied and divided by $\sum(x_i - \bar{x})^2$ and by $(N-1)$ in the middle equation and did some rearranging. The final expression tells us about the precision of estimation of the coefficient of the endogenous variable. As was the case with the OLS estimator, the variance of $\hat{\beta}_2$ depends on the variation in the explanatory variable about its mean, $\sum(x_i - \bar{x})^2$, and the estimated variance of the error term $\hat{\sigma}^2_{IV}$. Those components are familiar to you. What is new is that the denominator also includes the squared sample correlation r_{zx} between the instrumental variable z and the endogenous variable x. The larger the magnitude of the sample correlation $|r_{zx}|$ the smaller the estimated variance of the IV estimator, and vice versa. When $|r_{zx}|$ is large, the instrumental variable is strong. Stronger instrumental variables lead to smaller estimated variances, smaller standard errors, narrower interval estimates, and generally more precise statistical inference. It is important to choose strong instrumental variables.

To illustrate and make the point about instrument strength dramatic, suppose $\text{cov}(x_i, e_i) = 0$, so that both the OLS and IV estimators are consistent. Comparing the estimated variance of the two estimators, the ratio of the estimated variance of the IV estimator to that of the OLS estimator is

$$\frac{\widehat{\text{var}}(\hat{\beta}_2)}{\widehat{\text{var}}(b_2)} = \frac{\hat{\sigma}^2_{IV}}{\dfrac{r^2_{zx}\sum(x_i - \bar{x})^2}{\hat{\sigma}^2}} = \frac{\hat{\sigma}^2_{IV}/\hat{\sigma}^2}{r^2_{zx}} \simeq \frac{1}{r^2_{zx}}$$

The final approximation uses the fact that if $\text{cov}(x_i, e_i) = 0$, then in large samples the two estimators of σ^2 will converge to the same value so that $\hat{\sigma}^2_{IV}/\hat{\sigma}^2 \simeq 1$. The squared correlation $r^2_{zx} < 1$ and thus we anticipate that the variance estimate for the IV estimator will be larger than the variance estimate for the OLS estimator. The IV estimator is less *efficient* than the OLS estimator, meaning that it makes less efficient use of sample data to estimate the unknown parameters.

We prefer the more efficient consistent estimator because it has a smaller standard error, leading to narrower interval estimates, making statistical inferences more precise. The ratio of standard errors is $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 1/|r_{zx}|$. If the correlation $r_{zx} = 0.5$, then $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 2$, the estimated standard error of the IV estimator is two times as large as the standard error of the OLS estimator. If $r_{zx} = 0.1$, then $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 10$, the estimated standard error of the IV estimator is *10* times as large as the standard error of the OLS estimator.

To put some meat on these bones, recall that in large samples a 95% interval estimate is approximately "estimate $\pm$ 2(standard error)." For the sake of illustration, suppose $b_2 \simeq \hat{\beta}_2 = 5$ and $\text{se}(b_2) = 1$, then the 95% interval estimate using the OLS estimator is $5 \pm 2(1)$ or $[3, 7]$. If $r_{zx} = 0.5$, then the interval estimate based on the IV estimator is $5 \pm 2(2)$ or $[1, 9]$. If $r_{zx} = 0.1$, then the interval estimate based on the IV estimator is $5 \pm 2(10)$ or $[-15, 25]$. This shocking difference will remind you not to use the IV estimator unless you have to. If you do have to use IV estimation, then you must search for a strong instrumental variable, one that is highly correlated with the endogenous x.

10.3.5 Proving the Consistency of the IV Estimator

The demonstration that the instrumental variables estimator is consistent follows the logic used in Section 10.1.3. The IV estimator of β_2 in (10.17) is

$$\hat{\beta}_2 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})/(N-1)}{\sum(z_i - \bar{z})(x_i - \bar{x})/(N-1)} = \frac{\widehat{\text{cov}}(z, y)}{\widehat{\text{cov}}(z, x)}$$

The sample covariance converges to the true covariance in large samples, so we can say

$$\hat{\beta}_2 \to \frac{\text{cov}(z, y)}{\text{cov}(z, x)}$$

If the instrumental variable z is not correlated with x in either the sample data or in the population, then the **instrumental variable estimator** fails. Having z and x uncorrelated in the sample data would mean a zero in the denominator of $\hat{\beta}_2$. Having z and x uncorrelated in the population means $\hat{\beta}_2$ would not converge in large samples. Thus for an instrumental variable to be valid, it must be uncorrelated with the error term e but correlated with the explanatory variable x.

Now, following the same steps as in Section 10.1.3, we obtain

$$\beta_2 = \frac{\text{cov}(z, y)}{\text{cov}(z, x)} - \frac{\text{cov}(z, e)}{\text{cov}(z, x)}$$

If we can assume that that $\text{cov}(z_i, e_i) = 0$, a condition we imposed on the choice of the instrumental variable z_i, then the instrumental variables estimator $\hat{\beta}_2$ converges in large samples to β_2,

$$\hat{\beta}_2 \to \frac{\text{cov}(z, y)}{\text{cov}(z, x)} = \beta_2$$

Thus, if $\text{cov}(z_i, e_i) = 0$ and $\text{cov}(z_i, x_i) \neq 0$, then the instrumental variable estimator of β_2 is consistent, in a situation in which the OLS estimator is not consistent due to correlation between x_i and e_i.

EXAMPLE 10.2 | IV Estimation of a Simple Wage Equation

To illustrate the instrumental variables estimation method in a simple regression consider a simplified version of the model used in Example 10.1, $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$. Using the data file *mroz* on $N = 428$ married women, the OLS estimates are

$$\widehat{\ln(WAGE)} = -0.1852 + 0.1086 EDUC$$
$$\text{(se)} \qquad (0.1852) \quad (0.0144)$$

The estimated rate of return to education is approximately 10.86%, and $t = 7.55$ indicates that the estimated coefficient is significantly different from zero at even the 1% level of significance. If *EDUC* is endogenous, and correlated with the random error e, then OLS estimation may lead to incorrect inferences. We anticipate that *EDUC* is positively correlated with the omitted variable "ability," meaning that the estimated rate of return 10.86% may overstate the true value.

What might we use as an instrumental variable? One proposal is to use mother's years of education, *MOTHEREDUC*, as an instrument. Does this qualify? In Section 10.3.3, we listed three criteria for an instrumental variable. First, does this variable have a direct effect on the dependent variable? Does it belong in the equation? Mother's education should not play any direct role in the determination of a daughter's wage, so this seems fine. Second, the instrument should not be contemporaneously correlated with the random error, e. Is a mother's education correlated with the omitted variable, her daughter's ability? This is more difficult. Ability includes many attributes, including intelligence, creativity, perseverance, and industriousness to name a few. Some portion of these character traits may be passed into our genetic makeup from our parents. We dodge the scientific debate on this issue and assume that a mother's years of education are uncorrelated with her daughter's ability. Third, the instrument should be highly correlated with the endogenous variable. This we can check! For the 428 women in the sample the correlation between mother's education and daughter's education is 0.3870. This is not very large, but it is not very small either.

The instrumental variables estimates are

$$\widehat{\ln(WAGE)} = 0.7022 + 0.0385 EDUC$$
$$\text{(se)} \qquad (0.4851) \quad (0.0382)$$

The IV estimate of the rate of return to education is 3.85%, one-third of the OLS estimate. The standard error is about 2.65 times larger than the OLS standard error, which is very close to what we reasoned that the ratio might be when both estimators are consistent,

$$\text{se}\left(\hat{\beta}_2\right) \Big/ \text{se}\left(b_2\right) = 0.0382/0.0144 = 2.65 \simeq 1/r_{zx}$$
$$= 1/0.3807 = 2.58$$

10.3.6 IV Estimation Using Two-Stage Least Squares (2SLS)

We can obtain the instrumental variables estimates by another type of calculation, one that will help us extend the IV estimation idea to more general situations. The method called **two-stage**

least squares uses two least squares regressions to calculate the IV estimates. The **first-stage equation** has a dependent variable that is the endogenous regressor x, and the independent variable z, the instrumental variable. That is, the first-stage equation is

$$x = \gamma_1 + \theta_1 z + v$$

where γ_1 is an intercept parameter, θ_1 is a slope parameter, and v is an error term. The steps in 2SLS are as follows:

1. Estimate the first-stage equation by OLS and obtain the fitted value, $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$.
2. In the **second stage**, replace the endogenous variable x in the simple regression $y = \beta_1 + \beta_2 x + e$ with $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$ and then apply OLS estimation to $y = \beta_1 + \beta_2 \hat{x} + e^*$.

The OLS estimates of β_1 and β_2 from the second-stage regression are identically equal to the IV estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Furthermore, the estimated variances and covariances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS formulas with $\hat{\sigma}^2_{IV} = \sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2 \Big/ (N-2)$ replacing the usual estimate of σ^2 and using the fact that $\bar{\hat{x}} = \bar{x}$,

$$\widehat{\text{var}}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2_{IV}}{\sum\left(\hat{x}_i - \bar{x}\right)^2} \tag{10.19}$$

This variance estimate is numerically identical to the previous expression in equation (10.18a). If (10.19) is not used, the second-stage OLS regression computes the variance incorrectly, because OLS software will use

$$\hat{\sigma}^2_{WRONG} = \sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i\right)^2 \Big/ (N-2)$$

putting $\hat{x}_i$ in place of x_i. Always use software designed for IV/2SLS as it will carry out the correct calculation.

EXAMPLE 10.3 | 2SLS Estimation of a Simple Wage Equation

To illustrate the two-stage least squares equivalent of instrumental variables estimation, we estimate the first-stage equation, a regression of the endogenous variable *EDUC* on the instrumental variable *MOTHEREDUC*

$$\widehat{EDUC} = 10.1145 + 0.2674 MOTHEREDUC$$
$$\text{(se)} \qquad (0.3109)\ \ (0.0309)$$

In order for *MOTHEREDUC* to be a strong instrumental variable it must be strongly correlated with *EDUC*. Another way to say this is that *MOTHEREDUC* should be strongly significant in this first-stage equation, and it is. The t-value is 8.66, so the coefficient is significantly different from zero at the 1% level. We will say much more about this approach in Section 10.3.9.

In the second-stage equation, we regress ln(*WAGE*) on the fitted value from the first-stage equation,

$$\widehat{\ln(WAGE)} = 0.7021 + 0.0385 \widehat{EDUC}$$
$$\text{(incorrect se)}\ (0.5021)\ \ (0.0396)$$

The coefficient estimates are the same as in Example 10.2, but note that the standard errors produced by this second OLS estimation are not the same as in Example 10.2. They are **incorrect** because they use $\hat{\sigma}^2_{WRONG}$.

10.3.7 Using Surplus Moment Conditions

The reason for introducing two-stage least squares is that it is an easy way to use extra, additional, instrumental variables. In a simple regression, we need only one instrumental variable, yielding two moment conditions like (10.16), which we solve for the two unknown model parameters.

Sometimes, however, we have more instrumental variables than are necessary. Suppose we have two good instruments, z_1 and z_2 that satisfy conditions IV1–IV3. Compared to (10.16) we have the additional moment condition

$$E(z_2 e) = E\left[z_2(y - \beta_1 - \beta_2 x)\right] = 0$$

There are now three sample moment conditions:

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum z_{i1}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum z_{i2}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

We have three equations with only two unknowns. There are no solutions satisfying all three equations. We could simply throw away one of the conditions (instruments) and use the remaining two to solve for the unknowns. A better solution is to use all the available instruments by combining them. It can be proved that the best way of combining instruments is using the two-stage least squares idea. In the simple regression $y = \beta_1 + \beta_2 x + e$, if x is endogenous, and we have two instruments, z_1 and z_2, the first-stage equation becomes

$$x = \gamma_1 + \theta_1 z_1 + \theta_2 z_2 + v$$

Estimate the first-stage equation by OLS and obtain the fitted value

$$\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z_1 + \hat{\theta}_2 z_2$$

We have combined the two instruments z_1 and z_2 into the single instrument $\hat{x}$. Using $\hat{x}$ as an instrument for x leads to two sample-moment conditions,

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum\hat{x}_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

Solving these conditions, and using $\bar{\hat{x}} = \bar{x}$, we have

$$\hat{\beta}_2 = \frac{\sum\left(\hat{x}_i - \bar{\hat{x}}\right)(y_i - \bar{y})}{\sum\left(\hat{x}_i - \bar{\hat{x}}\right)(x_i - \bar{x})} = \frac{\sum(\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum(\hat{x}_i - \bar{x})(x_i - \bar{x})}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

The estimates obtained using these formulas are identical to the IV/2SLS estimates obtained by applying least squares to $y = \beta_1 + \beta_2 \hat{x} + e^*$. If we have more than two instrumental variables we apply the same strategy of combining several instruments into one.

EXAMPLE 10.4 | Using Surplus Instruments in the Simple Wage Equation

Father's education is also a potential instrument for daughter's education. Using the 428 observations in the data file *mroz*, the correlation between *FATHEREDUC* and *EDUC* is 0.4154. The first-stage equation is

$$EDUC = \gamma_1 + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

The OLS estimated first-stage equation is

$$\widehat{EDUC} = 9.4801 + 0.1564MOTHEREDUC + 0.1881FATHEREDUC$$
$$\text{(se)} \quad (0.3211) \quad (0.0358) \quad (0.0336)$$

The t-statistics for the coefficients of *MOTHEREDUC* and *FATHEREDUC* are 4.37 and 5.59, respectively, and are significant at the 1% level. The test of the joint significance of the two IV is even more important than their individual significance. The F-statistic for the null hypothesis $H_0: \theta_1 = 0, \ \theta_2 = 0$ is 55.83, which is very significant, and we can conclude that at least one of the two IV coefficients is not zero based on this joint test. The importance of the F-test is discussed in Section 10.3.9.

In the second-stage equation, we replace *EDUC* by $\widehat{EDUC}$ and apply least squares to obtain the IV/2SLS estimates

$$\widehat{\ln(WAGE)} = 0.5510 + 0.0505\widehat{EDUC}$$
$$\text{(incorrect se)} \quad (0.4257) \quad (0.0335)$$

The coefficient estimates are the correct IV estimates, but the standard errors reported are incorrect. Using proper IV software yields

$$\widehat{\ln(WAGE)} = 0.5510 + 0.0505\widehat{EDUC}$$
$$\text{(se)} \quad (0.4086) \quad (0.0322)$$

10.3.8 Instrumental Variables Estimation in the Multiple Regression Model

To implement instrumental variables estimation in a multiple regression equation, we need estimation formulas that are more general than equation (10.17). To extend our analysis to a more general setting, consider the multiple regression model $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$. Suppose that among the explanatory variables we know, or suspect, that x_K is an endogenous variable correlated with the error term. The first $K-1$ variables $\left(x_1 = 1, x_2, \ldots, x_{K-1}\right)$ are **exogenous variables** that are uncorrelated with the error term e—they are "included" instruments. Instrumental variables estimation can be carried out using a two-step process, with an OLS regression in each step.

The **first-stage regression** has the endogenous variable x_K on the left-hand side, and **all exogenous and instrumental variables** on the right-hand side. If we have L "external" instrumental variables (we are ***L***ucky to have them) that are from outside the model $z_1, z_2, \ldots, z_L$, then the first-stage regression is

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_L z_L + v_K \tag{10.20}$$

where v_K is a random error term that is uncorrelated with all the right-hand side variables. Estimate the first-stage regression (10.20) by OLS and obtain the fitted value

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \cdots + \hat{\theta}_L z_L \tag{10.21}$$

The fitted value $\hat{x}_K$ is the optimal combination of all the exogenous and instrumental variables.

The **second-stage regression** is based on the original specification with $\hat{x}_K$ replacing x_K,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K \hat{x}_K + e^* \tag{10.22}$$

where e^* is an error term. OLS estimation of (10.22) is justified because in large samples e^* is uncorrelated with the explanatory variables, including $\hat{x}_K$. The OLS estimators from this equation, $\hat{\beta}_1, \ldots, \hat{\beta}_K$, are the **instrumental variables** (IV) estimators, and, because they can be obtained by two least squares regressions, they are also popularly known as the **two-stage least squares** (2SLS) estimators. We will refer to them as IV or 2SLS or IV/2SLS estimators. In the general case with more than one endogenous variable on the right-hand side the steps are similar and are discussed in Section 10.3.10.

We can use the standard formulas for estimator variances and covariances for the least squares estimator of (10.22), which we described in Section 5.3.1, with one modification. While we can use two least squares estimations to obtain proper estimates, least squares software does not produce correct standard errors and t-values. The IV/2SLS estimator of the error variance is based on the residuals from the original model, $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$, so that the proper estimator of the error variance σ^2 is the general version of equation (10.18b)

$$\hat{\sigma}^2_{IV} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right)^2}{N - K}$$

Econometric software will automatically use the proper variance estimator if a two-stage least squares or instrumental variables estimation option is chosen. Using the IV/2SLS estimated standard errors from (10.22), we can carry out t-tests and construct interval estimates of parameters that are valid in large samples. Furthermore, the usual tests of joint hypotheses are valid in large samples **if** the instrumental variables are not weak.

It is informative to recall the discussion in Section 6.4.1. Usually the coefficient of the endogenous variable is most interesting. Thinking about our wage equation example, the coefficient of *EDUC*, years of education, is of critical importance. Let $SSE_{\hat{x}_K}$ be the sum of squared residuals from the regression of $\hat{x}_K$ on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$, then, in large samples,

$$\hat{\beta}_K \overset{a}{\sim} N\left[\beta_K, \operatorname{var}\left(\hat{\beta}_K\right)\right]$$

and the variance estimate is

$$\widehat{\operatorname{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}} \tag{10.23}$$

Equation (10.23) shows that the variance of $\hat{\beta}_K$, the instrumental variables estimator of β_K, depends on, $SSE_{\hat{x}_K}$, the variation in $\hat{x}_K$ that is *not* explained by $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$. See equation (6.33) and the surrounding discussion. Because this is such an important concept we return to it in Section 10.3.9 when analyzing "weak" instrumental variables.

EXAMPLE 10.5 | IV/2SLS Estimation in the Wage Equation

In addition to education a worker's experience is also important in determining their wage. Because additional years of experience have a declining marginal effect on wage use the quadratic model

$$\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 EDUC + e$$

where *EXPER* is years of experience. This is the same specification as in Example 10.1. We assume that *EXPER* is an **exogenous** variable that is uncorrelated with the worker's ability and therefore uncorrelated with the random error e. Two instrumental variables for years of education, *EDUC*, are mother's and father's years of education, *MOTHEREDUC* and *FATHEREDUC*, introduced in the previous examples. The first-stage equation is

$$EDUC = \gamma_1 + \gamma_2 EXPER + \gamma_3 EXPER^2 + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

Using the 428 observations in the data file *mroz* the estimated first-stage equation is reported in Table 10.1. The IV/2SLS estimates, with correctly computed standard errors, are

$$\begin{aligned} \widehat{\ln(WAGE)} &= 0.0481 + 0.0442 EXPER \\ (\text{se}) \quad & (0.4003)\ (0.0134) \\ & - 0.0009 EXPER^2 + 0.0614 EDUC \\ & (0.0004) \qquad (0.0314) \end{aligned}$$

The estimated return to education is approximately 6.1%, and the estimated coefficient is statistically significant with a $t = 1.96$.

TABLE 10.1 First-Stage Equation

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	9.1026	0.4266	21.3396	0.0000
EXPER	0.0452	0.0403	1.1236	0.2618
$EXPER^2$	−0.0010	0.0012	−0.8386	0.4022
MOTHEREDUC	0.1576	0.0359	4.3906	0.0000
FATHEREDUC	0.1895	0.0338	5.6152	0.0000

10.3.9 Assessing Instrument Strength Using the First-Stage Model

In Section 10.3.4, we emphasized the importance of a strong instrument when estimating a simple regression model with an endogenous explanatory variable. There the assessment of the instrument's strength was based on the correlation between the endogenous variable x and the instrument z. In a multiple regression measuring instrument strength is more complicated. The first-stage regression is a key tool in assessing whether an instrument is "strong" or "weak" in the multiple regression setting.

Case 1: Assessing the Strength of One Instrumental Variable Suppose that x_K is endogenous and we have available one external instrumental variable z_1. In terms of the notation above $L = 1$. The first-stage regression equation is

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + v_K \tag{10.24}$$

In a simple regression model, we can look for instrument strength in the correlation between the endogenous variable and the instrument. In the multiple regression model, we must deal with the other exogenous variables $(x_2, \ldots, x_{K-1})$. The key to assessing the strength of the instrumental variable z_1 is the strength of its relationship to x_K **after** controlling for the effects of all the other exogenous variables. This, however, is exactly the purpose of multiple regression analysis. The coefficient θ_1 in the first-stage regression (10.24) measures the effect of z_1 on x_K after accounting for the effects of the other variables.

Not only must there be an effect of z_1 on x_K but also it must be a **statistically significant** effect. How significant? *Very significant*. To reject the hypothesis that the instrument z_1 is weak, a rule of thumb is that the F-test statistic for the null hypothesis $H_0: \theta_1 = 0$ in equation (10.24) should be greater than 10. Using the relationship between the t- and F-tests, $t^2 = F$ described in Section 6.1.3, this translates into the absolute t-statistic for significance being greater than 3.16, which is larger than the usual 5% critical values ±1.96 or the 1% critical values ±2.58. The $F > 10$ rule has been refined by econometric researchers Stock and Yogo, and we discuss their analysis in Appendix 10A. Estimates and tests based on an IV estimator are unreliable when instruments are weak.

Further Analysis of Weak Instruments[1] Another way to illustrate this point is the following. The logic may seem a bit cumbersome, but the final result will be intuitively pleasing.

[1]This section is more advanced.

In Section 10.3.8, we argued that the approximate large sample variance of the IV estimator of β_K is

$$\widehat{\text{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}}$$

where $SSE_{\hat{x}_K}$ is the sum of squared residuals from the regression of $\hat{x}_K$ on $\left(x_2, x_3, \ldots, x_{K-1}\right)$, where $\hat{x}_K$ is the fitted value from the first-stage regression (10.24),

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1$$

By taking one more step, we can obtain an insight into how important the first-stage regression results can be. Let us consider a regression of $\hat{x}_K$ on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$ and z_1. We do not need to do this in practice; we know it will result in a perfect fit, with an $R^2 = 1$. Nevertheless, let us follow the Frisch–Waugh–Lovell approach described in Section 5.2.4.

- First, partial out $\mathbf{x}_{exog}$ from $\hat{x}_K$ and obtain the residuals $\tilde{\hat{x}}_K$.
- Second, partial out $\mathbf{x}_{exog}$ from the instrument z_1 and obtain the residuals $\tilde{z}_1$. The sum of squared residuals is $\sum \tilde{z}^2_{i1}$.
- Regress $\tilde{\hat{x}}_K$ on $\tilde{z}_1$, with no constant. The estimated coefficient is $\hat{\theta}_1$, $R^2 = 1$, and the fitted value $\hat{\theta}_1 \tilde{z}_1$ exactly equals $\tilde{\hat{x}}_K$!
- Because $\tilde{\hat{x}}_K = \hat{\theta}_1 \tilde{z}_1$, we can write $SSE_{\hat{x}_K} = \sum \tilde{\hat{x}}^2_{iK} = \sum \left(\hat{\theta}_1 \tilde{z}_{i1}\right)^2 = \hat{\theta}^2_1 \sum \tilde{z}^2_{i1}$.

The result is an alternative expression for the large sample variance of the IV estimator of β_K given in (10.23),

$$\text{var}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}} = \frac{\hat{\sigma}^2_{IV}}{\hat{\theta}^2_1 \sum \tilde{z}^2_{i1}} \tag{10.25}$$

What factors contribute to the precision of the IV estimator of β_K? The first important factor is the magnitude of the estimate $\hat{\theta}_1$ from the first-stage regression. It is important that this coefficient is **large**! Second, how much variation is there in the external instrument z_1 after removing the linear effects of the included exogenous variables, $\mathbf{x}_{exog}$? What is important is the amount of variation in z_1 **not explained** by the included exogenous variables $\mathbf{x}_{exog}$. Ideally z_1 would be uncorrelated with $\mathbf{x}_{exog}$ and exhibit large variation. If $\hat{\theta}_1$ is numerically small, or if z_1 is highly correlated with $\mathbf{x}_{exog}$, or exhibits little variation, then the precision of the IV estimator $\hat{\beta}_K$ will be worse.

Case 2: Assessing the Strength of More Than One Instrumental Variable

Suppose that x_K is endogenous and we have available L external instrumental variables, z_1, z_2, …, z_L. For a single endogenous variable, we need only a single instrument. Sometimes more instruments are available, and having more strong instruments may improve the instrumental variables estimator. The first-stage regression equation is now

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \overbrace{\theta_1 z_1 + \cdots + \theta_L z_L}^{\text{external } IV} + v_K \tag{10.26}$$

What we require is that **at least one** of the instruments be strong. Given the nature of the requirement, a joint F-test of the null hypothesis $H_0 : \theta_1 = 0, \theta_2 = 0, \ldots, \theta_L = 0$ in (10.26) is relevant, because the alternative is that at least one of the θ_i coefficients is nonzero. If the F-test statistic value is sufficiently large, roughly $F > 10$, we reject the hypothesis that the instruments are "weak" and can proceed with instrumental variables estimation. If the F-value is not sufficiently large, then instrumental variables and **two-stage least squares estimation** is quite possibly worse than "ordinary" least squares.

The fitted value from the first-stage regression (10.26) is

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \cdots + \hat{\theta}_L z_L$$

Applying the Frisch–Waugh–Lovell Theorem, as in the previous section, we find that

$$\widehat{\text{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{\sum\left(\hat{\theta}_1 \tilde{z}_{i1} + \hat{\theta}_2 \tilde{z}_{i2} + \cdots + \hat{\theta}_L \tilde{z}_{iL}\right)^2} \tag{10.27}$$

where $\tilde{z}_{il}$ is the ith residual from a regression of z_l on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$. The precision of the IV estimator of β_K depends on the magnitudes of the first-stage coefficients and the unexplained components of the external instrumental variables.

EXAMPLE 10.6 | Checking Instrument Strength in the Wage Equation

In Example 10.5, there is only one potentially endogenous variable in the wage equation, *EDUC*. The minimum number of instrumental variables is one. Given two instruments, we require that at least one of them be significant in the first-stage equation. The F-test null hypothesis is that both coefficients, θ_1 and θ_2, are zero, and if we reject this null hypothesis we conclude that at least one of them is nonzero. In the first-stage regression in Table 10.1, the estimated coefficient of *MOTHEREDUC* is 0.1576 with a t-value of 4.39, and the estimated coefficient of *FATHEREDUC* is 0.1895 with a t-value of 5.62. The F-statistic value for the null hypothesis that both these coefficients are zero is 55.40, which is significant at the 1% level, but more importantly it is larger than the rule-of-thumb threshold, $F > 10$. In addition to the vitally important F-statistic, the goodness-of-fit measures R^2 and $\overline{R}^2$ are sometimes reported. For the first-stage equation in Table 10.1, these values are $R^2 = 0.1527$ and $\overline{R}^2 = 0.1467$.

Partial Correlation and Partial R^2

In addition to the first-stage F-statistic, R^2 and adjusted-R^2, a partial correlation or partial-R^2 are informative. Applying the partialling-out strategy of the Frisch–Waugh–Lovell Theorem is another way to examine instrument strength. The included exogenous variables in the wage equation are $\mathbf{x}_{exog} = \left(x_1 = 1, EXPER, EXPER^2\right)$. Regress *EDUC* on $\mathbf{x}_{exog}$ and obtain the residuals, *REDUC*.

Suppose that we are using the single instrument *MOTHEREDUC*. Regress *MOTHEREDUC* on $\mathbf{x}_{exog}$ and obtain the residuals, *RMOM*. These residual variables have the included exogenous variables partialled-out. That is, we have removed the linear influences of the included exogenous variables from the endogenous variable *EDUC* and the external IV, *MOTHEREDUC*. The correlation between *REDUC* and *RMOM* is called a **partial correlation**, and in this case it is 0.3854. The R^2 from a regression of *REDUC* on *RMOM* is called the partial-R^2, and in this case it is 0.1485. Because we have one endogenous variable and one external IV, the partial-$R^2 = 0.1485$ is the square of the partial correlation, $0.3854^2 = 0.1485$.

If there are more external instruments, the partial-R^2 is the R^2 of the partialled-out endogenous variable on all the partialled-out external IV. Add *FATHEREDUC* as an IV, regress it on $\mathbf{x}_{exog}$ and obtain the residuals, *RDAD*. The partial-R^2 is then the R^2 from the regression of *REDUC* on *RMOM* and *RDAD*. In this case, partial-$R^2 = 0.2076$ and the adjusted partial-$R^2 = 0.2038$.

10.3.10 Instrumental Variables Estimation in a General Model

To extend our analysis to a more general setting, consider the multiple regression model $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$. Suppose that among the explanatory variables $\left(x_1 = 1, x_2, \ldots, x_K\right)$ we know, or suspect, that several may be correlated with the error term e. Divide the variables into two groups, with the first G variables $\left(x_1 = 1, x_2, \ldots, x_G\right)$ being exogenous variables that are uncorrelated with the error term e. The second group of $B = K - G$ variables $\left(x_{G+1}, x_{G+2}, \ldots, x_K\right)$

is correlated with the regression error, and thus they are endogenous. The multiple regression model, including all K variables, is then

$$y = \overbrace{\beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G}^{G \text{ exogenous variables}} + \overbrace{\beta_{G+1} x_{G+1} + \cdots + \beta_K x_K}^{B \text{ endogenous variables}} + e \tag{10.28}$$

In order to carry out IV estimation we must have at least as many instrumental variables as we have **endogenous variables**. Suppose we have L external instrumental variables, from outside the model, $z_1, z_2, \ldots, z_L$. Such notation is invariably confusing and cumbersome. It may help to keep things straight to think of $G = Good$ explanatory variables and $B = Bad$ explanatory variables and $L = Lucky$ instrumental variables, since we are lucky to have them. Then we have *The Good, the Bad, and the Lucky*.

It is a necessary condition for IV estimation that $L \geq B$. If $L = B$ then there are just enough instrumental variables to carry out IV estimation. The model parameters are said to be **just-identified** or **exactly identified** in this case. The term **identified** is used to indicate that the model parameters can be consistently estimated. If $L > B$ then we have more instruments than are necessary for IV estimation, and the model is said to be **overidentified**.

To implement IV/2SLS, estimate B first-stage equations, one for each explanatory variable that is endogenous. On the left-hand side of the first-stage equations, we have an endogenous variable. On the right-hand side, we have *all* the exogenous variables, including the G explanatory variables that are exogenous, *and* the L instrumental variables, which also must be exogenous. The B first-stage equations are

$$x_{G+j} = \gamma_{1j} + \gamma_{2j} x_2 + \cdots + \gamma_{Gj} x_G + \theta_{1j} z_1 + \cdots + \theta_{Lj} z_L + v_j, \quad j = 1, \ldots, B \tag{10.29}$$

The first-stage parameters (γ's and θ's) take different values in each equation, which is why they have a "j" subscript. We have omitted the observation subscript for simplicity. Since the right-hand side variables are all exogenous, we can estimate (10.29) by OLS. Then obtain the fitted values

$$\hat{x}_{G+j} = \hat{\gamma}_{1j} + \hat{\gamma}_{2j} x_2 + \cdots + \hat{\gamma}_{Gj} x_G + \hat{\theta}_{1j} z_1 + \cdots + \hat{\theta}_{Lj} z_L, \quad j = 1, \ldots, B$$

This comprises the first stage of two-stage OLS estimation.

In the second stage of estimation we apply least squares to

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1} \hat{x}_{G+1} + \cdots + \beta_K \hat{x}_K + e^* \tag{10.30}$$

This two-stage estimation process leads to proper instrumental variables estimates, but it should not be done this way in applied work. Use econometric software designed for two-stage least squares or instrumental variables estimation so that standard errors, t-statistics, and other test statistics will be computed properly.

Assessing Instrument Strength in a General Model

The F-test for **weak instruments** discussed in Section 10.3.9 is not valid for models having more than one endogenous variable on the right side of the equation. Consider the model in (10.28) with $B = 2$,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G + \beta_{G+1} x_{G+1} + \beta_{G+2} x_{G+2} + e \tag{10.31}$$

where $x_2, \ldots, x_G$ are exogenous and uncorrelated with the error term e, while x_{G+1} and x_{G+2} are endogenous. Suppose that we have two external instrumental variables z_1 and z_2, with z_1 being a good instrument for both x_{G+1} and x_{G+2}. The weak instrument F-test may be significant in each first-stage equation even if z_2 is an irrelevant instrument and not at all related to x_{G+1} or x_{G+2}. In such a case, we might conclude that we have two valid instruments when we have only one.

The first-stage equations in this case are

$$x_{G+1} = \gamma_{11} + \gamma_{21}x_2 + \cdots + \gamma_{G1}x_G + \theta_{11}z_1 + \theta_{21}z_2 + v_1$$

$$x_{G+2} = \gamma_{12} + \gamma_{22}x_2 + \cdots + \gamma_{G2}x_G + \theta_{12}z_1 + \theta_{22}z_2 + v_2$$

The weak instrument F-test in the first equation is for the joint significance of θ_{11} and θ_{21}, $H_0: \theta_{11} = 0$, $\theta_{21} = 0$, with the alternative hypothesis that at least *one* of these coefficients is not zero. If θ_{11} is statistically significant, then the joint null hypothesis may be rejected even if $\theta_{21} = 0$. Similarly in the second equation we can obtain a significant F-test outcome even if z_2 is irrelevant as an instrument for x_{G+1} as long as z_1 is statistically significant. In this case we have two individually significant F-tests despite the fact that only one valid instrument z_1 is available, and thus the model in (10.31) is not identified. The more general test required for this case, which builds on the concept of "partial correlation" is discussed in Appendix 10A.

10.3.11 Additional Issues When Using IV Estimation

In this section, we discuss some issues related to IV estimation.

Hypothesis Testing with Instrumental Variables Estimates We may be interested in testing hypotheses about the regression parameters based on the two-stage least squares/instrumental variables estimates. When testing the null hypothesis $H_0: \beta_k = c$, use of the test statistic $t = \left(\hat{\beta}_k - c\right) \Big/ \text{se}\left(\hat{\beta}_k\right)$ is valid in large samples. We know that as $N \to \infty$, the $t_{(N-K)}$ distribution converges to the standard normal distribution $N(0, 1)$. If the degrees of freedom $N - K$ are large, then critical values from the two distributions will be very close. It is common, but not universal, practice to use critical values, and p-values, based on the $t_{(N-K)}$ distribution rather than the more strictly appropriate $N(0, 1)$ distribution. The reason is that tests based on the t-distribution tend to work better in samples of data that are not large.

Another issue is whether to use standard errors that are "robust" to the presence of heteroskedasticity (in cross-section data) or autocorrelation and heteroskedasticity (in time-series data). These options were described in Chapters 8 and 9 for the linear regression model, and they are also available in most software packages for IV estimation. Such corrections to standard errors require large samples in order to work properly.

When using software to test a joint hypothesis, such as $H_0: \beta_2 = c_2, \beta_3 = c_3$, the test may be based on the chi-square distribution with the number of degrees of freedom equal to the number of hypotheses (J) being tested. The test itself may be called a Wald test, or a likelihood ratio (LR) test, or a Lagrange multiplier (LM) test. These testing procedures are all asymptotically equivalent and are discussed in Appendix C.8.4. However, the test statistic reported may also be called an F-statistic with J numerator degrees of freedom and $N - K$ denominator degrees of freedom. This F-value is often calculated by dividing one of the chi-square tests statistics, such as the Wald statistic, by J. The motivation for using the F-test is to achieve better performance in small samples. Asymptotically, the tests will all lead to the same conclusion. See Chapter 6, Appendix 6A, for some related discussion. Once again, joint tests can be made "robust" to potential heteroskedasticity or autocorrelation problems, and this is an option with many software packages.

Generalized Method-of-Moments Estimation If heteroskedasticity or serial correlation is present in a model with one or more endogenous variables, then using instrumental variables estimation with a "robust" covariance matrix ensures that interval estimators, hypothesis tests and prediction intervals use a valid standard error. However, using an instrumental

variables estimator with a robust covariance matrix estimator does not *improve* the efficiency of the estimator, just like using the OLS estimator with a robust covariance matrix estimator does not improve its efficiency. In Chapters 8 and 9 we introduced a **generalized least squares estimator** for linear regression models with error terms that are heteroskedastic and/or serially correlated. In the same way, there is a **generalized method-of-moments (GMM) estimator** that is "asymptotically" more efficient than the instrumental variables estimator. Being "asymptotically more efficient" means that the GMM estimator has smaller variances than the IV estimator in large samples. In order to obtain the gain, we must have at least one surplus instrument. The gain in efficiency is obtained by building into the estimator a heteroskedasticity and/or serial correlation correction. Despite the fact that the GMM estimator improves the large sample precision of estimation its actual performance in samples that are not large might not be good. And like the IV estimator, good instruments are required. Theoretically, the GMM estimator is very attractive because it is a general estimation approach that includes the OLS estimator, the GLS estimator and IV/2SLS as special cases.

The GMM estimation procedure is built into econometric software packages but their proper usage requires an in-depth study of the methodology, which is beyond the scope of this book. It is one of the few topics that is difficult to explain without the tools of matrix algebra. Advanced readers can consult William Greene (2018) *Econometric Analysis, Eighth Edition*, Pearson Prentice-Hall, Chapter 13.

Goodness-of-Fit with Instrumental Variables Estimates We discourage the use of measures like R^2 outside the context of OLS estimation. When there are endogenous variables on the right-hand side of a regression equation, the concept of measuring how well the variation in y is explained by the x variables breaks down, because as we discussed in Section 10.2, these models exhibit feedback. This logical problem is paired with a numerical one. If our model is $y = \beta_1 + \beta_2 x + e$, then the IV residuals are $\hat{e} = y - \hat{\beta}_1 - \hat{\beta}_2 x$. Many software packages will report the goodness-of-fit measure $R^2 = 1 - \sum \hat{e}_i^2 / \sum (y_i - \bar{y})^2$. Unfortunately, this quantity can be negative when based on IV estimates.

10.4 Specification Tests

We have shown that if an explanatory variable is correlated with the regression error term, the OLS estimator fails. If a strong instrumental variable is available, the IV estimator is consistent and approximately normally distributed in large samples. But if we use a weak instrument, or an instrument that is invalid in the sense that it is not uncorrelated with the regression error, then IV estimation can be as bad as, or worse than, using the OLS estimator. We addressed how to detect weak instruments in Section 10.3.9, and go into much greater detail on this problem in Appendix 10A. In this section we ask two other important questions that must be answered in each situation in which instrumental variables estimation is considered:

1. Can we test for whether x is correlated with the error term? This might give us a guide for when to use least squares and when to use IV estimators.
2. Can we test if our instrument is valid, and uncorrelated with the regression error, as required?

10.4.1 The Hausman Test for Endogeneity

In the previous sections, we discussed the fact that the least squares estimator fails if there is correlation between an explanatory variable and the error term. We also provided an estimator, the instrumental variables estimator, that can be used when the least squares estimator fails.

The question we address in this section is how to test for the presence of a correlation between an explanatory variable and the error term, so that we can use the appropriate estimation procedure.

The null hypothesis is $H_0: \text{cov}(x_i, e_i) = 0$ against the alternative that $H_1: \text{cov}(x_i, e_i) \neq 0$. The idea of the test is to compare the performance of the OLS estimator to an instrumental variables estimator. Under the null and alternative hypotheses, we know the following:

- If the null hypothesis is true, both the OLS estimator b and the instrumental variables estimator $\hat{\beta}$ are consistent. Thus, in large samples the difference between them converges to zero. That is, $q = (b - \hat{\beta}) \to 0$. Naturally, if the null hypothesis is true, use the more efficient estimator, which is the least squares estimator.
- If the null hypothesis is false, the OLS estimator is not consistent, and the instrumental variables estimator is consistent. Consequently, the difference between them does not converge to zero in large samples. That is, $q = (b - \hat{\beta}) \to c \neq 0$. If the null hypothesis is not true, use the instrumental variables estimator, which is consistent.

There are several forms of the test, usually called the **Hausman test** in recognition of econometrician Jerry Hausman's pioneering work on this problem, for these null and alternative hypotheses. One form of the test directly examines the differences between the least squares and instrumental variables estimates, as we have described above. Some computer software programs implement this test for the user, which can be computationally difficult to carry out.[2]

An alternative form of the test is very easy to implement, and is the one we recommend. See Section 10.4.2 for an explanation of the test's logic. In the regression $y_i = \beta_1 + \beta_2 x_i + e_i$, we wish to know whether x_i is correlated with e_i. Let z_1 and z_2 be instrumental variables for x. At minimum, one instrument is required for each variable that might be correlated with the error term. Then carry out the following steps:

1. Estimate the first-stage model $x = \gamma_1 + \theta_1 z_1 + \theta_2 z_2 + v$ by OLS, including on the right-hand side all instrumental variables and all exogenous variables not suspected to be endogenous, and obtain the residuals
$$\hat{v} = x - \hat{\gamma}_1 - \hat{\theta}_1 z_1 - \hat{\theta}_2 z_2$$
If more than one explanatory variable is being tested for endogeneity, repeat this estimation for each one.
2. Include the residuals computed in step 1 as an explanatory variable in the original regression, $y = \beta_1 + \beta_2 x + \delta \hat{v} + e$. Estimate this "artificial regression" by OLS, and employ the usual t-test for the hypothesis of significance:
$$H_0: \delta = 0 \quad (\text{no correlation between } x_i \text{ and } e_i)$$
$$H_1: \delta \neq 0 \quad (\text{correlation between } x_i \text{ and } e_i)$$
3. If more than one variable is being tested for endogeneity, the test will be an F-test of joint significance of the coefficients on the included residuals.

The t- and F-tests in steps two and three can be made robust if heteroskedasticity and/or autocorrelation are potential problems.

[2] Some software packages compute Hausman tests with K, or $K - 1$, degrees of freedom, where K is the total number of regression parameters. This is incorrect. Use the correct degrees of freedom B, equal to the number of potentially endogenous right-hand-side variables (see 10.28).

10.4.2 The Logic of the Hausman Test[3]

In Section 10.4.1, we presented the Hausman test for whether or not an explanatory variable is endogenous using an artificial regression. Let us explore how this test works. The simple regression model is

$$y = \beta_1 + \beta_2 x + e \tag{10.32}$$

If x is correlated with the error term e, then x is endogenous and the OLS estimator is biased and inconsistent.

An instrumental variable z must be correlated with x but uncorrelated with e in order to be valid. A correlation between z and x implies that there is a linear association between them. This means that we can describe their relationship as a regression

$$x = \gamma_1 + \theta_1 z + v \tag{10.33}$$

This is the first-stage equation introduced in Section 10.3.6. It is a predictive model with the base assumption $E(x|z) = \gamma_1 + \theta_1 z$. The conditional mean of the endogenous variable x is linearly related to the instrumental variable z. The error term v is simply $v = x - (\gamma_1 + \theta_1 z)$ so that the two sides of (10.33) are equal. There is a correlation between x and z if, and only if, $\theta_1 \neq 0$. We can divide x into two parts, a systematic part and a random part, as

$$x = E(x|z) + v \tag{10.34}$$

where $E(x|z) = \gamma_1 + \theta_1 z$. If we knew γ_1 and θ_1, we could substitute (10.34) into the simple regression model (10.32) to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2 \left[E(x|z) + v\right] + e \\ &= \beta_1 + \beta_2 E(x|z) + \beta_2 v + e \end{aligned} \tag{10.35}$$

Now, suppose for a moment that $E(x|z)$ and v can be observed and are viewed as explanatory variables in the regression $y = \beta_1 + \beta_2 E(x|z) + \beta_2 v + e$. Will least squares work when applied to this equation? The explanatory variable $E(x|z)$ depends only on z and it is not correlated with the error term e if z is a valid instrument. The endogeneity problem, if there is one, comes from a correlation between v (the random part of x) and e. In fact, in the regression (10.32) any correlation between x and e implies correlation between v and e because $v = x - E(x|z)$.

We cannot exactly create the partition in (10.34) because we do not know γ_1 and θ_1. However, we can consistently estimate the first-stage equation (10.33) by OLS. Doing so, we obtain the fitted first-stage equation $\hat{x} = \widehat{E(x|z)} = \hat{\gamma}_1 + \hat{\theta}_1 z$ and the residuals $\hat{v} = x - \hat{x}$. Rearrange these to obtain an estimated analog of (10.34),

$$x = E(x|z) + \hat{v} = \hat{x} + \hat{v} \tag{10.36}$$

Substitute (10.36) into the original equation (10.32) to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2 \left[\hat{x} + \hat{v}\right] + e \\ &= \beta_1 + \beta_2 \hat{x} + \beta_2 \hat{v} + e \end{aligned} \tag{10.37}$$

To reduce confusion, and avoid β_2 appearing twice in same equation, let the coefficient of $\hat{v}$ be denoted as γ, so that (10.37) becomes

$$y = \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e \tag{10.38}$$

[3]Contains advanced material.

If we omit $\hat{v}$ from (10.38) the regression becomes

$$y = \beta_1 + \beta_2 \hat{x} + e \tag{10.39}$$

The least squares estimates of β_1 and β_2 in (10.39) *are* the IV/2SLS estimates discussed in Section 10.3.6. Then, recall from Section 6.6.1, equation (6.23), that if we omit a variable from a regression that is uncorrelated with the included variable(s) there is no omitted variables bias, and in fact the least squares estimates are unchanged! This holds true in (10.39) because the least squares residuals $\hat{v}$ are uncorrelated with $\hat{x}$ and the intercept variable. Thus, the least squares estimates of β_1 and β_2 in (10.38) and (10.39) are identical and are equal to the IV/2SLS estimates. Consequently, the least squares estimators of β_1 and β_2 in (10.38) are consistent whether or not x is exogenous, because they are the IV estimators.

What about γ? If x is exogenous, and hence v and e are uncorrelated, then the least squares estimator of γ in (10.38) will also converge in large samples to β_2. However, if x is endogenous then the least squares estimator of γ in (10.38) will *not* converge to β_2 in large samples because $\hat{v}$, like v, is correlated with the error term e. This observation makes it possible to test for whether x is exogenous by testing the equality of the estimates of β_2 and γ in (10.38). If we reject the null hypothesis $H_0: \beta_2 = \gamma$ then we reject the exogeneity of x, and conclude that it is endogenous.

Carrying out the test is made simpler by playing a trick on (10.38). Add and subtract $\beta_2 \hat{v}$ to the right-hand side to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e + \beta_2 \hat{v} - \beta_2 \hat{v} \\ &= \beta_1 + \beta_2 (\hat{x} + \hat{v}) + (\gamma - \beta_2) \hat{v} + e \\ &= \beta_1 + \beta_2 x + \delta \hat{v} + e \end{aligned} \tag{10.40}$$

Thus, instead of testing $H_0: \beta_2 = \gamma$ we can simply use an ordinary t-test of the null hypothesis $H_0: \delta = 0$ in (10.40), which is exactly the test we described in Section 10.4.1. This is much nicer because software automatically prints out the t-statistic for this hypothesis test. This test can be made robust to heteroskedasticity and/or autocorrelation if desired.

10.4.3 Testing Instrument Validity

A valid instrument z must be contemporaneously uncorrelated with the regression error term, so that $\text{cov}(z_i, e_i) = 0$. If this condition fails then the resulting moment condition, like (10.16), is invalid and the IV estimator will not be consistent. Unfortunately, not every instrument can be tested for validity. In order to compute the IV estimator for an equation with B possibly endogenous variables, we must have at least B instruments. The validity of this minimum number of required instruments cannot be tested. In the case in which we have $L > B$ instruments available, we can test the validity of the $L - B$ extra, or surplus, moment conditions.[4]

An intuitive approach is the following. From the set of L instruments, form groups of B instruments and compute the IV estimates using each different group. If all the instruments are valid, then we would expect all the IV estimates to be similar. Rather than do this, there is a test of the validity of the **surplus moment conditions** that is easier to compute. The steps are

1. Compute the IV estimates $\hat{\beta}_k$ using all available instruments, including the G variables $x_1 = 1, x_2, \ldots, x_G$ that are presumed to be exogenous, and the L instruments $z_1, \ldots, z_L$.
2. Obtain the residuals $\hat{e}_{IV} = y - \hat{\beta}_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_K x_K$.

[4]Econometric jargon for surplus moment conditions is "overidentifying restrictions." A surplus of moment conditions means we have more than enough for identification, hence "overidentifying." Moment conditions like (10.16) can be thought of as restrictions on parameters.

3. Regress $\hat{e}_{IV}$ on all the available instruments described in step one.
4. Compute NR^2 from this regression, where N is the sample size and R^2 is the usual goodness-of-fit measure.
5. If all of the surplus moment conditions are valid, then $NR^2 \sim \chi^2_{(L-B)}$.[5] If the value of the test statistic exceeds the $100(1-\alpha)$th percentile (i.e., the critical value) from the $\chi^2_{(L-B)}$ distribution, then we conclude that at least one of the surplus moment conditions is not valid.

If we reject the null hypothesis that all the surplus moment conditions are valid, then we are faced with trying to determine which instrument(s) are invalid, and how to weed them out.

EXAMPLE 10.7 | Specification Tests for the Wage Equation

In Section 10.3.6, we examined a ln(*WAGE*) equation for married women, using the two instruments "mother's education" and "father's education" for the potentially endogenous explanatory variable education (*EDUC*).

To implement the Hausman test we first obtain the first-stage regression estimates, which are shown in Table 10.1. Using these estimates we calculate the least squares residuals $\hat{v} = EDUC - \widehat{EDUC}$. Insert the residuals in the ln(*WAGE*) equation as an extra variable, and estimate the resulting augmented regression using OLS. The resulting estimates are shown in Table 10.2.

The Hausman test of the endogeneity is based on the t-test of significance of the first-stage regression residuals, $\hat{v}$. If we reject the null hypothesis that the coefficient is zero, we conclude that education is endogenous. Note that the coefficient of the first-stage regression residuals (*VHAT*) is significant at the 10% level of significance using a two-tail test. While this is not strong evidence of the endogeneity of education, it is sufficient cause for concern to consider using instrumental variables estimation. Second, note that the coefficient estimates of the remaining variables, but not their standard errors, are identical to their instrumental variables estimates. This feature of the regression-based Hausman test is explained in Section 10.4.2.

In order to be valid, the instruments *MOTHEREDUC* and *FATHEREDUC* should be uncorrelated with the regression error term. As discussed in Section 10.4.3, we cannot test the validity of both instruments, only the "overidentifying" or surplus instrument. Since we have two instruments and only one potentially endogenous variable, we have $L - B = 1$ extra instrument. The test is carried out by regressing the residuals from the ln(*WAGE*) equation, calculated using the instrumental variables estimates, on all available exogenous and instrumental variables. The test statistic is NR^2 from this artificial regression, and R^2 is the usual goodness-of-fit measure. If the surplus instruments are valid, then the test statistic has an asymptotic $\chi^2_{(1)}$ distribution, where the degrees of freedom are the number of surplus instruments. If the test statistic value is greater than the critical value from this distribution, then we reject the null hypothesis that the surplus instrument is valid. For the artificial regression $R^2 = 0.000883$, and the test statistic value is $NR^2 = 428 \times 0.000883 = 0.3779$. The 0.05 critical value for the chi-square distribution with one degree of freedom is 3.84, so we fail to reject the surplus instrument as valid. With this result we are reassured that our instrumental variables estimator for the wage equation is consistent.

TABLE 10.2 Hausman Test Auxiliary Regression

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.0481	0.3946	0.1219	0.9030
EDUC	0.0614	0.0310	1.9815	0.0482
EXPER	0.0442	0.0132	3.3363	0.0009
$EXPER^2$	−0.0009	0.0004	−2.2706	0.0237
VHAT	0.0582	0.0348	1.6711	0.0954

[5] This test is valid if errors are homoskedastic and is sometimes called the Sargan test. If the errors are heteroskedastic, there is a more general test called Hansen's J-test that is provided by some software. A very advanced reference is Hayashi, *Econometrics*, Princeton, 2000, pp. 227–228.

10.5 Exercises

10.5.1 Problems

10.1 Using state level data, a researcher wishes to examine the relationship between the median rent paid (*RENT*) as a function of median house values (*MDHOUSE* in $1000). The percentage of the state population living in an urban area (*PCTURBAN*) is used as an additional control. Use the results in Table 10.3 to answer the following questions.

TABLE 10.3 Estimates for Exercise 10.1

	(1) *RENT*	(2) *MDHOUSE*	(3) *MDHOUSE*	(4) *RENT*	(5) *RENT*	(6) *EHAT*
C	125.9	−19.78	7.225	121.1	121.1	−53.50
	(14.19)	(10.23)	(8.936)	(12.87)	(15.51)	(22.66)
PCTURBAN	0.525	0.205	0.616	0.116	0.116	−0.257
	(0.249)	(0.113)	(0.131)	(0.254)	(0.306)	(0.251)
MDHOUSE	1.521			2.184	2.184	
	(0.228)			(0.282)	(0.340)	
FAMINC		2.584				3.851
		(0.628)				(1.393)
REG4		15.89				−16.87
		(3.157)				(6.998)
VHAT				−1.414		
				(0.411)		
N	50	50	50	50	50	50
R^2	0.669	0.679	0.317	0.737	0.609	0.198
SSE	20259.6	3907.4	8322.2	16117.6	23925.6	19195.8

Standard errors in parentheses.

a. The OLS estimates of the model are in column (1). Why might we be concerned that *MDHOUSE*, the median price of houses, is endogenous in this regression?

b. Two instruments are considered: median family income (*FAMINC* in $1000) and a regional dummy variable *REG4*. Using the models in columns (2) and (3), test if the instruments are weak.

c. In column (4), the least squares residuals (*VHAT*) from the regression in column (2) are added as a regressor to the basic regression. The estimates are obtained using OLS. What is the usefulness of this regression? What does it indicate about the results in (1)?

d. In column (5) are IV/2SLS estimates using the instruments listed in part (b). What differences do you observe between these results and the OLS results in column (1)? Note that the estimates (though not the standard errors) are the same in columns (4) and (5). Is this a mistake? Explain.

e. In column (6) the residuals from the estimation in column (5) are regressed upon the variables shown. What information is contained in these results?

10.2 The labor supply of married women has been a subject of a great deal of economic research. Consider the following supply equation specification

$$HOURS = \beta_1 + \beta_2 WAGE + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 NWIFEINC + e$$

where *HOURS* is the supply of labor, *WAGE* is hourly wage, *EDUC* is years of education, *KIDSL6* is the number of children in the household who are less than 6 years old, and *NWIFEINC* is household income from sources other than the wife's employment.

a. Discuss the signs you expect for each of the coefficients.
b. Explain why this supply equation cannot be consistently estimated by OLS regression.
c. Suppose we consider the woman's labor market experience *EXPER* and its square, $EXPER^2$, to be instruments for *WAGE*. Explain how these variables satisfy the logic of instrumental variables.
d. Is the supply equation identified? Explain.
e. Describe the steps [not a computer command] you would take to obtain IV/2SLS estimates.

10.3 In the regression model $y = \beta_1 + \beta_2 x + e$, assume x is endogenous and that z is a valid instrument. In Section 10.3.5, we saw that $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$.

a. Divide the denominator of $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$ by $\text{var}(z)$. Show that $\text{cov}(z, x)/\text{var}(z)$ is the coefficient of the simple regression with dependent variable x and explanatory variable z, $x = \gamma_1 + \theta_1 z + v$. [*Hint*: See Section 10.2.1.] Note that this is the first-stage equation in two-stage least squares.
b. Divide the numerator of $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$ by $\text{var}(z)$. Show that $\text{cov}(z, y)/\text{var}(z)$ is the coefficient of a simple regression with dependent variable y and explanatory variable z, $y = \pi_0 + \pi_1 z + u$. [*Hint*: See Section 10.2.1.]
c. In the model $y = \beta_1 + \beta_2 x + e$, substitute for x using $x = \gamma_1 + \theta_1 z + v$ and simplify to obtain $y = \pi_0 + \pi_1 z + u$. What are π_0, π_1, and u in terms of the regression model parameters and error and the first-stage parameters and error? The regression you have obtained is a **reduced-form** equation.
d. Show that $\beta_2 = \pi_1/\theta_1$.
e. If $\hat{\pi}_1$ and $\hat{\theta}_1$ are the OLS estimators of π_1 and θ_1, show that $\hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1$ is a consistent estimator of $\beta_2 = \pi_1/\theta_1$. The estimator $\hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1$ is an **indirect least squares** estimator.

10.4 Suppose that x is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero; it is an indicator variable. Make the assumption $E(e_i|z_i) = 0$.

a. Show that $E(y_i|z_i) = \beta_1 + \beta_2 E(x_i|z_i)$.
b. Assume $E(x_i|z_i) \neq 0$. Does z_i satisfy conditions IV1–IV3? Explain.
c. Write out the **conditional expectation** in (a) for the two cases with $z_i = 1$ and $z_i = 0$. Solve the two resulting equations for β_2.
d. Suppose we have a random sample (y_i, x_i, z_i), $i = 1, \ldots, N$. Give an intuitive argument that a consistent estimator of $E(y_i|z_i = 1)$ is the sample average of the y_i values for the subset of observations for which $z_i = 1$, which we might call $\bar{y}_1$.
e. Following the strategy in part (d) form $\bar{y}_1$, $\bar{y}_0$, $\bar{x}_1$, and $\bar{x}_0$. Show that the empirical implementation of the expression in (c) is $\hat{\beta}_{WALD} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$, which is the **Wald Estimator**, in honor of Abraham Wald.
f. Explain how $E(x_i|z_i = 1) - E(x_i|z_i = 0)$ might be viewed as a measure of the strength of the instrumental variable z_i.

10.5 Suppose that x_i is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero with probabilities p and $1 - p$, respectively, that is, $\Pr(z_i = 1) = p$ and $\Pr(z_i = 0) = 1 - p$.

a. Show that $E(z_i) = p$.
b. Show that $E(y_i z_i) = pE(y_i|z_i = 1)$.
c. Use the law of iterated expectations to show that $E(y_i) = pE(y_i|z_i = 1) + (1 - p)\,E(y_i|z_i = 0)$.
d. Substitute (a), (b), and (c) results into $E(y_i z_i) - E(y_i)E(z_i)$ to show that $\text{cov}(y_i, z_i) = p(1 - p)\,E(y_i|z_i = 1) - p(1 - p)\,E(y_i|z_i = 0)$.
e. Use the arguments in (a)–(d) to show that $\text{cov}(x_i, z_i) = p(1 - p)\left[E(x_i|z_i = 1) - E(x_i|z_i = 0)\right]$.
f. Assuming $E(e_i) = 0$ show $\left[y_i - E(y_i)\right] = \beta_2\left[x_i - E(x_i)\right] + e_i$.
g. Multiply both sides of the expression in (f) by $z_i - E(z_i)$ and take expectations to show $\text{cov}(y_i, z_i) = \beta_2 \text{cov}(x_i, z_i)$ if $\text{cov}(e_i, z_i) = 0$.
h. Using (d), (f), and (g) show that $\beta_2 = \dfrac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$
i. Show that the empirical implementation of (h) leads to $\hat{\beta}_{WALD} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$.

10.6 Suppose that x_i is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero.

a. Let $N_1 = \sum z_i$ be the number of z_i values such that $z_i = 1$. Show that $\sum z_i x_i = N_1 \bar{x}_1$ where $\bar{x}_1$ is the sample average of the x_i values corresponding to $z_i = 1$.

b. Let $N_0 = N - \sum z_i = N - N_1$ be the number of z_i values such that $z_i = 0$. Show that $\sum x_i = N_1 \bar{x}_1 + N_0 \bar{x}_0$ where $\bar{x}_0$ is the sample average of the x_i values corresponding to $z_i = 0$.

c. Show that $N\sum x_i z_i - \sum z_i \sum x_i = N_1 N_0 (\bar{x}_1 - \bar{x}_0)$

d. Show that $N\sum y_i z_i - \sum z_i \sum y_i = N_1 N_0 (\bar{y}_1 - \bar{y}_0)$

e. Use the results in (c) and (d) to show that the IV estimator of β_2 in (10.17) reduces to $\hat{\beta}_2 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$.

f. The estimated variance of the IV estimator is given in (10.18a). Show that $\sum (z_i - \bar{z})(x_i - \bar{x}) = \sum z_i x_i - N\bar{z}\bar{x} = N_1 N_0 (\bar{x}_1 - \bar{x}_0)$.

g. Using the result in part (f), suppose $(\bar{x}_1 - \bar{x}_0) \simeq 0$. How does this indicate that the IV z_i is weak?

h. $\sum (z_i - \bar{z})(x_i - \bar{x}) / \sum (z_i - \bar{z})^2$ is the OLS estimate of the slope coefficient from a regression of x_i on z_i. True or False? How does this value relate to the weak instrument discussion in part (g)? If this coefficient is small, with a low t-value, does it imply that z_i is a weak IV? Explain.

10.7 Angrist and Krueger (1991) use quarter of birth as an instrumental variable to estimate the returns to schooling, using a sample of 327,509 from the 1980 census. The model of interest is $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$.

a. Let $\overline{\ln(WAGE)}$ denote the average of the natural log of weekly wage. For men born in the first quarter of the year the average is 5.8916, and for men born in the fourth quarter of the year the average is 5.9027. What is the approximate percentage difference in wages for the two groups of men?

b. The standard error of the difference in means from part (a) is 0.00274. Is the difference in $\overline{\ln(WAGE)}$ statistically significant? What is the two-tail p-value?

c. Let $\overline{EDUC}$ denote the average years of schooling. For men born in the first quarter of the year the average is 12.6881, and for men born in the fourth quarter of the year the average is 12.7969. What is the approximate percentage difference in years of schooling for the two groups of men? Is there a reason why men born in the fourth quarter have higher average schooling than men born in the first quarter?

d. The standard error of the difference in means from part (c) is 0.0132. Is the difference in $\overline{EDUC}$ statistically significant? What is the two-tail p-value.

e. Compute the Wald estimate of the return to schooling, $\hat{\beta}_{2,WALD}$ using the results above. What is the instrumental variable z being used in this case? The Wald estimator is introduced in Exercise 10.4.

f. Explain why the result in (d) is important to the success of the Wald estimator.

10.8 Knowledge is Power Program (KIPP) Schools are charter schools with largely minority students. These schools differ in a number of ways from public schools, but emphasize longer days and more time spent in school. The question is: "How much benefit is there to attending a KIPP school?"[6]

a. Let $y_i = MATH_i$ be the outcome of a math achievement test. This outcome is standardized by subtracting the average and dividing by the standard deviation, so that $y = 0$ is the average score, and $y = 1$ is a score that is one standard deviation above average, and so on. Let $x_i = ATTENDED_i$ be an indicator variable with the value one if a student attended a KIPP school and zero otherwise. In the regression $y_i = \beta_1 + \beta_2 x_i + e_i$, suppose that the OLS estimate of β_2 is $b_2 = 0.467$, with a standard error of 0.103. Based on this regression result, does attending a KIPP school seem to improve math test score? Is the estimate of the amount of improvement a meaningful amount? If the average math score of those attending the KIPP school is 0.095, what is the average score of those who do not attend the KIPP school?

b. Explain why we might worry that *ATTENDED* is an endogenous variable.

c. Offers of admission are randomly assigned to the pool of KIPP applicants. Some of those offered admission wind up attending and some do not. Let *WINNER* be an indicator variable taking the value one if a student receives an offer to attend, and zero otherwise. Suppose that 78.7% of offers to attend are accepted. Does *WINNER* satisfy the conditions for an instrumental variable?

[6]This exercise is adapted from Angrist and Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.

d. Suppose that $z_i = WINNER_i$. In the terms of this example, explain the components of

$$\beta_2 = \frac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$$

See Exercises 10.4 and 10.5 for background discussion of the expression.

e. The average math score of those receiving an offer to attend the KIPP school was −0.003, which is very close to average. The average score of those not offered a seat was −0.358, which is about one-third of a standard deviation below average. Interestingly, some students wind up attending the KIPP school despite not being randomly selected from the applicants. Assume that the proportion of students attending the KIPP school who were not "winners" be 4.6%. Obtain the Wald estimator of β_2 by replacing the population averages in part (d) with sample averages. How does this estimate compare to the OLS estimate in part (a)? Does attending a KIPP school appear to have a meaningful positive effect on scores of those attending?

10.9 Consider the wage equation used in Example 10.5. Suppose we have a variable designed to measure *ABILITY*. This variable is an index created using 10 different tests of cognitive ability. Using data on 2,178 white males in 1980, the ability variable has a sample mean of 0.04 and a standard deviation of 0.96.

a. The estimated relationship between years of education and the ability measure is $\widehat{EDUC} = 12.30 + 0.977ABILITY$ with a t-value of 25.81. Is this result consistent with the usual "omitted variables bias" explanation of the endogeneity of education? Explain.

b. Using these data and the model in Example 10.5, the estimated coefficient on *EDUC* is 0.0609 with standard error 0.005. Adding *ABILITY* to the equation reduces the estimated coefficient on *EDUC* to 0.054 with standard error 0.006. Is this the effect that you anticipate? Explain.

c. Assuming that *ABILITY* and *EXPER* are exogenous, along with instrumental variables *MOTHEREDUC* and *FATHEREDUC*, what is the specification of the first-stage equation? That is, what variables are on the right-hand side?

d. Estimating the first-stage equation in (c), we find that the t-values on *MOTHEREDUC* and *FATHEREDUC* are 2.55 and 4.72, respectively. The F-test of their joint significance is 33.82. Are these instruments adequately strong for their use in IV/2SLS? Explain.

e. Let $\hat{v}$ denote the OLS residuals from part (d). If we estimate the model in Example 10.5, and include the variables *ABILITY* and $\hat{v}$, the t-statistic for $\hat{v}$ is −0.94. What does this result tell us about the endogeneity of *EDUC* after controlling for ability?

10.10 Consider the model in Example 10.5. Suppose we have the idea that the effect of education may differ for individuals who have siblings. Suppose *SIBS* = number of siblings, which we assume is exogenous. We add to the model the variable $EDUC \times SIBS$.

a. Assuming we treat *EDUC* as endogenous, what type of variable is $EDUC \times SIBS$? Is it exogenous or endogenous? Explain your reasoning.

b. In addition to *MOTHEREDUC* and *FATHEREDUC*, are $MOTHEREDUC \times SIBS$ and $FATHEREDUC \times SIBS$ potentially useful IV? Explain how they satisfy, or might satisfy, the three conditions IV1–IV3.

c. Using OLS with a large sample of individuals, we find the estimated coefficient of *EDUC* to be 0.0903 ($t = 46.74$) and the estimated coefficient of $EDUC \times SIBS$ to be −0.0001265 ($t = -0.91$). Explain why we should not simply omit the variable $EDUC \times SIBS$ in the wage equation based on this result.

d. The first-stage equations for *EDUC* and $EDUC \times SIBS$ include *EXPER*, $EXPER^2$, and the four variables listed in (b). The F-tests for the joint significance of the IV have p-values of 0.0000. Can we safely conclude that our IV are strong for both *EDUC* and $EDUC \times SIBS$?

e. We calculate the residuals from the two first-stage equations. Let the residuals from the *EDUC* equation be $\hat{v}_1$ and the residuals from the $EDUC \times SIBS$ equation be $\hat{v}_2$. We estimate the structural model by OLS including both $\hat{v}_1$ and $\hat{v}_2$ as explanatory variables. Their t-values are −10.29 and −1.63, respectively, and the joint F-test of their significance is 55.87. Can we safely conclude that both *EDUC* and $EDUC \times SIBS$ are endogenous?

f. Using IV/2SLS, we find that the estimated coefficient of *EDUC* is 0.1462 with a t-value of 25.25, and the estimated coefficient of $EDUC \times SIBS$ is 0.0007942 with a t-value of 4.53. The estimated covariance between these two coefficients is 4.83×10^{-7}. Estimate the marginal effect of another

year of education on wages for a person with no siblings. What is the estimated marginal effect of education if a person has five siblings?

10.11 Consider the wage equation in Example 10.5.

a. Two possible instruments for *EDUC* are *NEARC4* and *NEARC2*, where these are dummy variables indicating whether the individual lived near a 4-year college or a 2-year college at age 10. Speculate as to why these might be potentially valid IV.

b. Explain the steps (not the computer command) required to carry out the regression-based Hausman test, assuming we use both IV.

c. Using a large data set, the p-value for the regression-based Hausman test for the model in Example 10.5, using only *NEARC4* as an IV is 0.28; using only *NEARC2* the p-value is 0.0736, and using both IV the p-value is 0.0873 [with robust standard errors it is 0.0854]. What should we conclude about the endogeneity of *EDUC* in this model?

d. We compute the IV/2SLS residuals, using both *NEARC4* and *NEARC2* as IV. In the regression of these 2SLS residuals on all exogenous variables and the IV, with $N = 3010$ observations, all regression p-values are greater than 0.30 and the $R^2 = 0.000415$. What can you conclude based on these results?

e. The main reason we seldom use OLS to estimate the coefficients of equations with endogenous variables is that other estimation methods are available that yield better fitting equations. Is this statement true or false, or are you uncertain? Explain the reasoning of your answer.

f. The F-test of the joint significance of *NEARC4* and *NEARC2* in the first-stage regression is 7.89. The 95% interval estimates for the coefficient of education using OLS is 0.0678 to 0.082, and using 2SLS it is 0.054 to 0.260. Explain why the width of the interval estimates is so different.

10.12 Estimating cost and production functions for industrial plants is important. Decisions are based on estimated average and marginal cost, and average and marginal products. Suppose a manufacturing plant for a particular firm has output modeled as $Q = \beta_1 + \beta_2 MGT\_EFF + \beta_3 CAP + \beta_4 LAB + e$, where Q is the output in a particular manufacturing plant, *MGT_EFF* is a managerial efficiency index, *CAP* is capital stock input index and *LAB* is labor input index.

a. What is the interpretation of β_2? What sign should it have?

b. Measuring *MGT_EFF* is difficult. Suppose we propose to estimate the model

$$Q = \beta_1 + \beta_2 XPER + \beta_3 CAP + \beta_4 LAB + e$$

where *XPER* is the plant manager's experience, measured in years. What should the sign of β_2 be now? Why might we worry that *XPER* is endogenous? [*Hint*: Think carefully about this one.]

c. We use data from 75 plants to estimate the model in (b). The least squares estimates are

$$\hat{Q} = 1.7623 + 0.1468\,XPER + 0.4380\,CAP + 0.2392\,LAB$$
$$\text{(se)}\ (1.0550)\ (0.0634)\qquad (0.1176)\qquad (0.0998)$$

Are the signs of the coefficients and their significance consistent with your expectations? Explain.

d. If *XPER* is endogenous, what is the direction of the bias of the OLS estimator? Explain. [*Hint*: Remember your answer to part (b).]

e. Suppose we consider *AGE*, the age of the plant manager, as an instrument. Does it satisfy the criteria for an IV based on your economic reasoning? Why or why not?

f. In the OLS regression of *XPER* on *CAP*, *LAB,* and *AGE,* the t-value for the coefficient of *AGE* is 3.13. What information does this provide us about the feasibility of carrying out IV/2SLS?

g. We add the residuals from part (f) to the model in (b) to obtain

$$Q = \beta_1 + \beta_2 XPER + \beta_3 CAP + \beta_4 LAB + \beta_5 RESID + e$$

The t-statistic for the null hypothesis $H_0: \beta_5 = 0$ from this regression is -2.2. What should we infer from it?

h. The two-stage least squares estimates are

$$\hat{Q} = -2.4867 + 0.5121\,XPER + 0.3321\,CAP + 0.2400\,LAB$$
$$\text{(se)}\ (2.7230)\ (0.2205)\qquad (0.1545)\qquad (0.1209)$$

What are the differences in these estimates versus the OLS estimates? Are the differences consistent with your expectations, relative to the OLS estimates? Explain.

i. Reasoning that *AGE* is an adequate IV, a staff economist decides to add $AGE \times LAB$ and $AGE \times CAP$ as IV also. Are these likely to be valid IV and uncorrelated with the regression error term? To test this, the two-stage least squares residuals are regressed on *CAP*, *LAB*, *AGE*, $AGE \times LAB$, and $AGE \times CAP$. The resulting R^2 is 0.0045. What do you think about the validity of the IV now?

j. The economist regresses *XPER* on *CAP*, *LAB*, *AGE*, $AGE \times LAB$, and $AGE \times CAP$. The *F*-test of the joint significance of *AGE*, $AGE \times LAB$, and $AGE \times CAP$ is 3.3. Do you think it is advisable to use the interaction variables as IV in the estimation? Justify your answer.

10.13 Households plan consumption expenditures and saving with consideration of their long-run income. We wish to estimate $SAVING = \beta_1 + \beta_2 LRINCOME + e$, where *LRINCOME* is long-run income.

a. Long-run income is difficult to define and measure. Using data on 50 households' annual savings (*SAVINGS*, \$1000 units) and annual income (*INCOME*, \$1000 units), we estimate a savings equation by OLS to obtain

$$\widehat{SAVINGS} = 4.3428 - 0.0052 INCOME$$
$$\text{(se)} \qquad (0.8561) \quad (0.0112)$$

Why might we expect the OLS estimator of the marginal propensity to save to be biased and inconsistent? What is the likely direction of the bias?

b. Suppose that in addition to current income we know average household income over the past 10 years (*AVGINC*, \$1000 units). Why might this be a suitable instrumental variable?

c. The estimated first-stage regression is

$$\widehat{INCOME} = -35.0220 + 1.6417 AVGINC$$
$$(t) \qquad (-1.83) \quad (5.80)$$

Does *AVGINC* qualify as a strong instrument? Explain.

d. Let the residuals from part (c) be $\hat{v}$. Adding this variable to the savings equation and estimating the result by OLS gives

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME - 0.0755 \hat{v}$$
$$\text{(se)} \qquad (1.1720) \quad (0.0154) \qquad\qquad (0.0201)$$

Based on this result should we rely on the OLS estimates of the savings equation?

e. Using the fitted values from part (c) in place of *INCOME* and applying OLS, we obtain

$$\widehat{SAVINGS} = 0.9883 + 0.0392 \widehat{INCOME}$$
$$\text{(se)} \qquad (1.2530) \quad (0.0165)$$

Compare these coefficient estimates to those in part (a). Are these estimates more in line with your prior expectations than those in (a), or not?

f. Are the OLS standard errors in part (e) correct or not? Explain.

g. Using IV/2SLS software, with instrument *AVGINC*, we obtain the estimates

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME$$
$$\text{(se)} \qquad (1.5240) \quad (0.0200)$$

Construct a 95% interval estimate of the effect of *INCOME* on *SAVINGS*. Compare and contrast it to the 95% interval estimate based on the results in part (a).

h. In parts (d), (e), and (g), the estimated coefficient of *INCOME* is 0.0392. Is this an accident? Explain.

i. Explain how to test whether *AVGINC* is a valid instrument, and uncorrelated with the regression error.

10.14 The Capital Asset Pricing Model (CAPM) [see Exercise 2.16] says that the risk premium on security *j* is related to the risk premium on the market portfolio, that is

$$r_j - r_f = \alpha_j + \beta_j (r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. We measure the market portfolio using the Standard & Poors value weighted index, and the risk-free rate by the 30-day LIBOR monthly rate of return.

a. Using 180 monthly observations from January 1988, the OLS estimate of IBM's beta is 0.9769 with a standard error of 0.0978. If our constructed values of the market return and the risk-free rate are measured with error is the OLS estimator unbiased and consistent? If it is biased, what is the direction of the bias?

b. It has been suggested that it is possible to construct an IV by ranking the values of the explanatory variable and using the rank as the IV. That is, we sort $(r_m - r_f)$ from smallest to largest, and assign the values $RANK = 1, 2, \ldots, 180$. Does this variable potentially satisfy the conditions IV1–IV3?

c. The estimated first-stage regression of $(r_{IBM} - r_f)$ on *RANK* yields an overall F-test of model significance of 93.77. What can we conclude about the strength of the IV *RANK*?

d. If we compute the first-stage residuals and add them to the CAPM model, the resulting coefficient has a t-value of 60.60. What does this result suggest to us about the OLS estimator in the CAPM model?

e. Using *RANK* as an IV and estimating the CAPM model by IV/2SLS yield an estimate of IBM's beta of 1.0025 with a standard error of 0.1019. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

10.5.2 Computer Exercises

10.15 Consider the simple wage model in Example 10.2. Use the 428 observations on married women who participate in the labor force.

a. Using the instrumental variables estimator in equation (10.17), divide the numerator and denominator by $(N-1)$ and show that the IV estimator is the ratio of sample covariances, $\hat{\beta}_2 = \widehat{\text{cov}}(z_i, y_i)/\widehat{\text{cov}}(z_i, x_i)$.

b. Using your computer software, calculate $\widehat{\text{cov}}(MOTHEREDUC_i, \ln(WAGE_i))$ and $\widehat{\text{cov}}(MOTHEREDUC_i, EDUC_i)$. Compare their ratio to the IV estimate in Example 10.2.

c. In Example 10.5, we added experience and its square to the model specification. To implement the ratio of covariances estimator in part (a), we first remove (partial-out) the influence of experience and its square from *MOTHEREDUC*, *EDUC*, and ln(*WAGE*). Regress each of variables on *EXPER* and $EXPER^2$ and save the residuals, calling them *RMOTHEREDUC*, *REDUC*, and *RLWAGE*. Calculate $\widehat{\text{cov}}(RMOTHEREDUC_i, RLWAGE_i)$ and $\widehat{\text{cov}}(RMOTHEREDUC_i, REDUC_i)$. Compare their ratio to the IV estimate in Example 10.5.

d. Using your IV/2SLS software, estimate the model $RLWAGE = \beta_2 REDUC + error$, omitting the constant term, using *RMOTHEREDUC* as an instrumental variable. Compare the resulting estimate to that in part (c).

10.16 Consider the wage model in Example 10.5 and the 428 observations on married women who participate in the labor force. Use only *MOTHEREDUC* as an instrument in this exercise.

a. Estimate the first-stage equation by OLS and obtain the fitted values

$$\widehat{EDUC} = \hat{\gamma}_1 + \hat{\gamma}_2 EXPER + \hat{\gamma}_3 EXPER^2 + \hat{\theta}_1 MOTHEREDUC$$

b. Use OLS to estimate the second-stage equation

$$\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 \widehat{EDUC} + error$$

c. Obtain the least squares residuals, $\hat{e}$, from the estimation in part (b). Calculate $\sum \hat{e}_i$. Explain why the value you obtain is theoretically correct.

d. Using the coefficient estimates from part (b), calculate the residuals

$$\hat{e}_{IV} = \ln(WAGE) - \hat{\beta}_1 - \hat{\beta}_2 EXPER - \hat{\beta}_3 EXPER^2 - \hat{\beta}_4 EDUC$$

Calculate $\sum \hat{e}_{IV}$. Explain why the value you obtain is theoretically correct.

e. Calculate $\sum \hat{e}_i^2/(N-4)$ and $\sum \hat{e}_{IV}^2/(N-4)$. Which of these is the correct estimator of the error variance, σ^2?

f. Estimate the regression $\widehat{EDUC} = a_1 + a_2 EXPER + a_3 EXPER^2 + error$ and obtain the sum of squared residuals. Use equation (10.25) and one of the values from part (e) to obtain $\widehat{\text{var}}\left(\hat{\beta}_4\right)$.

g. Using software for IV/2SLS estimate the wage model $\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 EDUC + e$ using the instrumental variable *MOTHEREDUC*. How do the estimates compare to those in part (b)? Does the reported standard error $\text{se}\left(\hat{\beta}_4\right)$ agree with the calculated variance in part (f)?

10.17 Consider the wage model in Example 10.5 and the 428 observations on married women who participate in the labor force. Use only *MOTHEREDUC* as an instrument in this exercise.

a. Estimate the first-stage equation by OLS and obtain the fitted values

$$\widehat{EDUC} = \hat{\gamma}_1 + \hat{\gamma}_2 EXPER + \hat{\gamma}_3 EXPER^2 + \hat{\theta}_1 MOTHEREDUC$$

Save the least squares residuals. Call them *REDUCHAT*. Calculate the sum of squared residuals, $\sum REDUCHAT_i^2$.

b. Estimate the regression $\widehat{EDUC} = a_1 + a_2 EXPER + a_3 EXPER^2 + error$ and save the OLS residuals. Call them *REDUC*. Calculate the sum of squared residuals, $\sum REDUC_i^2$.

c. Estimate the regression $MOTHEREDUC = c_1 + c_2 EXPER + c_3 EXPER^2 + error$ and save the OLS residuals. Call them *RMOM*. Calculate the sum of squared residuals, $\sum RMOM_i^2$.

d. Estimate the regression $REDUC = \theta_1 RMOM + error$. Compare the estimated value of θ_1 from this regression to the estimated θ_1 from the first-stage equation. What R^2 value did you obtain from this regression? What is the sum of squared residuals?

e. Show that $\sum RMOM_i^2 = \hat{\theta}_1^2 \sum REDUC_i^2$.

f. Refer to equation (10.25) and discuss the importance of the quantities in (e) for the precision of the IV/2SLS estimator.

10.18 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of a parent's college education as an instrumental variable.

a. Create two new variables. *MOTHERCOLL* is a dummy variable equaling one if *MOTHEREDUC* > 12, zero otherwise. Similarly, *FATHERCOLL* equals one if *FATHEREDUC* > 12 and zero otherwise. What percentage of parents have some college education in this sample?

b. Find the correlations between *EDUC*, *MOTHERCOLL*, and *FATHERCOLL*. Are the magnitudes of these correlations important? Can you make a logical argument why *MOTHERCOLL* and *FATHERCOLL* might be better instruments than *MOTHEREDUC* and *FATHEREDUC*?

c. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* as the instrumental variable. What is the 95% interval estimate for the coefficient of *EDUC*?

d. For the problem in part (c), estimate the first-stage equation. What is the value of the F-test statistic for the hypothesis that *MOTHERCOLL* has no effect on *EDUC*? Is *MOTHERCOLL* a strong instrument?

e. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* and *FATHERCOLL* as the instrumental variables. What is the 95% interval estimate for the coefficient of *EDUC*? Is it narrower or wider than the one in part (c)?

f. For the problem in part (e), estimate the first-stage equation. Test the joint significance of *MOTHERCOLL* and *FATHERCOLL*. Do these instruments seem adequately strong?

g. For the IV estimation in part (e), test the validity of the surplus instrument. What do you conclude?

10.19 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of a parent's college education as an instrumental variable.

a. Create two new variables. *MOTHERCOLL* is a dummy variable equaling one if *MOTHEREDUC* > 12, zero otherwise. Similarly *FATHERCOLL* equals one if *FATHEREDUC* > 12, and zero otherwise. Also, create *COLLSUM* = *MOTHERCOLL* + *FATHERCOLL* and *COLLBOTH* = *MOTHERCOLL* × *FATHERCOLL*. What values do *COLLSUM* and *COLLBOTH* take? What percentage of women in the sample have both a mother and a father with some college education.

b. Find the correlations between *EDUC*, *COLLSUM*, and *COLLBOTH*. Are the magnitudes of these correlations important? Can you make a logical argument why *COLLSUM* and *COLLBOTH* might be better instruments than *MOTHEREDUC* and *FATHEREDUC*?

c. Estimate the wage equation in Example 10.5 using 2SLS with *COLLSUM* as the instrumental variable. What is the 95% interval estimate for the coefficient of *EDUC*?

d. For the problem in part (c), estimate the first-stage equation. What is the value of the F-test statistic for the hypothesis that *COLLSUM* has no effect on *EDUC*? Is *COLLSUM* a strong instrument?

e. Using OLS estimate the regression model with *EDUC* as dependent variable, and include as explanatory variables experience, and its square, along with *MOTHERCOLL* and *FATHERCOLL*, and a constant term. Test the null hypothesis that the coefficients of *MOTHERCOLL* and *FATHERCOLL* are equal at the 5% level.

f. Based on the results in part (e), are we justified in using *COLLSUM* = *MOTHERCOLL* + *FATHERCOLL* as an IV? Are we better off using *COLLSUM* only or using *MOTHERCOLL* and *FATHERCOLL*?

10.20 The CAPM [see Exercises 10.14 and 2.16] says that the risk premium on security j is related to the risk premium on the market portfolio. That is

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. We measure the market portfolio using the Standard & Poor's value weighted index, and the risk-free rate by the 30-day LIBOR monthly rate of return. As noted in Exercise 10.14, if the market return is measured with error, then we face an errors-in-variables, or measurement error, problem.

a. Use the observations on Microsoft in the data file *capm5* to estimate the CAPM model using OLS. How would you classify the Microsoft stock over this period? Risky or relatively safe, relative to the market portfolio?

b. It has been suggested that it is possible to construct an IV by ranking the values of the explanatory variable and using the rank as the IV, that is, we sort $(r_m - r_f)$ from smallest to largest, and assign the values $RANK = 1, 2, \ldots, 180$. Does this variable potentially satisfy the conditions IV1–IV3? Create *RANK* and obtain the first-stage regression results. Is the coefficient of *RANK* very significant? What is the R^2 of the first-stage regression? Can *RANK* be regarded as a strong IV?

c. Compute the first-stage residuals, $\hat{v}$, and add them to the CAPM model. Estimate the resulting augmented equation by OLS and test the significance of $\hat{v}$ at the 1% level of significance. Can we conclude that the market return is exogenous?

d. Use *RANK* as an IV and estimate the CAPM model by IV/2SLS. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

e. Create a new variable $POS = 1$ if the market return $(r_m - r_f)$ is positive, and zero otherwise. Obtain the first-stage regression results using both *RANK* and *POS* as instrumental variables. Test the joint significance of the IV. Can we conclude that we have adequately strong IV? What is the R^2 of the first-stage regression?

f. Carry out the Hausman test for endogeneity using the residuals from the first-stage equation in (e). Can we conclude that the market return is exogenous at the 1% level of significance?

g. Obtain the IV/2SLS estimates of the CAPM model using *RANK* and *POS* as instrumental variables. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

h. Obtain the IV/2SLS residuals from part (g) and use them (not an automatic command) to carry out a Sargan test for the validity of the surplus IV at the 5% level of significance.

10.21 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of alternative constructed instrumental variables. Estimate the model in Example 10.5 using IV/2SLS using both *MOTHEREDUC* and *FATHEREDUC* as IV. These will serve as our baseline results.

a. Write down the first-stage equation using econometric notation, as in equation (10.26), with $\gamma_1, \gamma_2, \gamma_3$ as the unknown coefficients of the intercept, *EXPER* and its square, and θ_1, θ_2 as the coefficients of *MOTHEREDUC* and *FATHEREDUC*, respectively. Test the null hypothesis that $\theta_1 = \theta_2$ at the 5% level. What do you conclude?

b. Assume that $\theta_1 = \theta_2 = \theta$. Substitute into the first-stage equation to obtain a "restricted" model. What variable involving *MOTHEREDUC* and *FATHEREDUC* now appears on the right-hand side?

c. Create a new variable $PARENTSUM = MOTHEREDUC + FATHEREDUC$. Obtain IV/2SLS estimates using this as the IV. How do the estimates compare to the baseline results? Is this IV strong?

d. Create two new variables $MOMED2 = MOTHEREDUC^2$ and $DADED2 = FATHEREDUC^2$. Use these new variables and both *MOTHEREDUC* and *FATHEREDUC* as IV. Estimate the first-stage equation using these four IV. Test their joint significance using an F-test. Are these instruments adequately strong? Do any seem irrelevant based on t-tests of significance? Find the simple correlations among the four IV. Are any large?

e. Obtain IV/2SLS estimates of the model in Example 10.5 using the four IV in part (d). How do these estimates compare to the baseline results and to those in part (c)?

f. Based on the results in this question, which set of IV/2SLS estimates would you prefer to report? The baseline estimates, the results in part (c), or the results in part (e). Explain your choice.

10.22 Consider the data file *mroz* on working wives and the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$. Use the 428 observations on married women who participate in the labor force.

a. Write down in algebraic form the three moment conditions, like (10.13) and (10.14), that would lead to the OLS estimates of the model above.

b. Calculate the OLS estimates and residuals, $\hat{e}_i$. What is the sum of the least squares residuals? What is the sum of squared least squares residuals? What is $\sum EDUC_i \times \hat{e}_i$? What is $\sum EXPER_i \times \hat{e}_i$? Relate these results to the moment conditions in (a).

c. Calculate the fitted values $\widehat{\ln(WAGE)} = b_1 + b_2 EDUC + b_3 EXPER$. What is the sample average of the fitted values? What is the sample average of $\ln(WAGE)$, $\overline{\ln(WAGE)}$?

d. Find each of the following:

$$SST = \sum\left[\ln(WAGE_i) - \overline{\ln(WAGE)}\right]^2, \quad SSE = \sum \hat{e}_i^2, \quad SSR = \sum\left[\widehat{\ln(WAGE_i)} - \overline{\ln(WAGE)}\right]^2$$

Compute $SSR + SSE$, $R^2 = SSR/SST$ and $R^2 = 1 - SSE/SST$. Explain what these calculations show about measuring goodness-of-fit.

10.23 This question is an extension of Exercise 10.22. Consider the data file *mroz* on working wives and the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$. Use the 428 observations on married women who participate in the labor force. Let the instrumental variable be *MOTHEREDUC*.

a. Write down in algebraic form the three moment conditions, like (10.16) , that would lead to the IV/2SLS estimates of the model above.

b. Calculate the IV/2SLS estimates and residuals, $\hat{e}_{IV}$. What is the sum of the IV residuals? What is $\sum MOTHEREDUC_i \times \hat{e}_{IV,i}$? What is $\sum EXPER_i \times \hat{e}_{IV,i}$? Relate these results to the moment conditions in (a).

c. What is $\sum EDUC_i \times \hat{e}_{IV,i}$? What is the sum of squared IV residuals? How do these two results compare with the corresponding OLS results in Exercise 10.22(b)?

d. Calculate the IV/2SLS fitted values $FLWAGE = \hat{\beta}_1 + \hat{\beta}_2 EDUC + \hat{\beta}_3 EXPER$. What is the sample average of the fitted values? What is the sample average of $\ln(WAGE)$, $\overline{\ln(WAGE)}$?

e. Find each of the following:

$$SST = \sum\left[\ln(WAGE_i) - \overline{\ln(WAGE)}\right]^2, \quad SSE\_IV = \sum \hat{e}_{IV}^2,$$
$$SSR\_IV = \sum\left[FLWAGE - \overline{\ln(WAGE)}\right]^2$$

Compute $SSR\_IV + SSE\_IV$, $R^2_{IV,1} = SSR\_IV/SST$, and $R^2_{IV,2} = 1 - SSE\_IV/SST$. How do these values compare to those in Exercise 10.22(d)?

f. Does your IV/2SLS software report an R^2 value. Is it either of the ones in (e)? Explain why the usual concept of R^2 fails to hold for IV/2SLS estimation.

10.24 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of alternative standard errors for the IV estimator. Estimate the model in Example 10.5 using IV/2SLS using both *MOTHEREDUC* and *FATHEREDUC* as IV. These will serve as our baseline results.

a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity?

b. Regress $\hat{e}_{IV}^2$ against a constant and *EXPER*. Apply the NR^2 test from Chapter 8 to test for the presence of heteroskedasticity.

c. Obtain the IV/2SLS estimates with the software option for Heteroskedasticity Robust Standard Errors. Are the robust standard errors larger or smaller than those for the baseline model? Compute the 95% interval estimate for the coefficient of *EDUC* using the robust standard error.

d. Obtain the IV/2SLS estimates with the software option for Bootstrap standard errors, using $B = 200$ bootstrap replications. Are the bootstrap standard errors larger or smaller than those for the baseline model? How do they compare to the heteroskedasticity robust standard errors in (c)? Compute the 95% interval estimate for the coefficient of *EDUC* using the bootstrap standard error.

10.25 To examine the quantity theory of money, Brumm (2005) ["Money Growth, Output Growth, and Inflation: A Reexamination of the Modern Quantity Theory's Linchpin Prediction," *Southern Economic Journal*, 71(3), 661–667] specifies the equation

$$INFLATION = \beta_1 + \beta_2 MONEY\ \ GROWTH + \beta_3 OUTPUT\ \ GROWTH + e$$

where *INFLATION* is the growth rate of the general price level, *MONEY GROWTH* is the growth rate of the money supply, and *OUTPUT GROWTH* is the growth rate of national output. According to theory we should observe that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$. Use the data file *brumm*. It consists of 1995 data on 76 countries. We wish to test

i. the *strong* joint hypothesis that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$.

ii. the *weak* joint hypothesis $\beta_2 = 1$ and $\beta_3 = -1$

a. It is argued that *OUTPUT GROWTH* may be endogenous. Four instrumental variables are proposed, *INITIAL* = initial level of real GDP, *SCHOOL* = a measure of the population's educational attainment, *INV* = average investment share of GDP, and *POPRATE* = average population growth rate. Using these instruments, obtain instrumental variables (2SLS) estimates of the inflation equation.

b. Test the strong and weak hypotheses using the IV estimates.

c. Compute the IV/2SLS residuals, $\hat{e}_{IV}$. Identify the observation with the largest absolute residual, $|\hat{e}_{IV}|$. How does it compare to the next smallest residual?

d. Let us examine the effect of the observation with the largest residual. Drop the corresponding observation from the data, reestimate the model using IV/2SLS, and carry out the tests of the strong and weak hypotheses. How much do things change, if any?

e. Obtain the IV/2SLS residuals from part (d), $\tilde{e}_{IV}$. Regress $\tilde{e}_{IV}^2$ on *MONEY*. Calculate the heteroskedasticity test statistic NR^2. Compare it to the 95th percentile of the $\chi^2_{(1)}$ distribution. Is there evidence of heteroskedasticity?

f. Using the 75 remaining observations from (d) obtain the IV/2SLS estimates with heteroskedasticity robust standard errors. Carry out the tests of the strong and weak hypotheses. How to the test results compare to those in (d)?

g. Using the remaining 75 observations from (d), estimate the first-stage equation and test the joint significance of the IV. Repeat the tests robust to heteroskedasticity. Is there evidence that the instruments are strong?

h. Regress $\tilde{e}_{IV}$ against the four IV and *MONEY*. Are any of the coefficients significant? If the IV are valid, do we expect any significant coefficients in this regression? Explain.

Appendix 10A Testing for Weak Instruments

The F-test for weak instruments discussed in Section 10.3.9 is not valid for models with more than one endogenous variable on the right side of the equation.[7] Using **canonical correlations** there is a solution to the problem of identifying weak instruments when an equation has more than one endogenous variable. Canonical correlations are a generalization of the usual concept of

[7] The $F > 10$ rule of thumb comes from D. Staiger and J.H. Stock (1997) "Instrumental Variables with Weak Instruments," *Econometrica* 65, pp. 557–586.

a correlation between two variables and attempt to describe the association between two **sets** of variables. The association in which we are interested is the association between the pair of endogenous variables (x_{G+1}, x_{G+2}) and the pair of additional, external, instrumental variables (z_1, z_2) **after** controlling for the effect of the other G exogenous variables $\mathbf{x}_1 \equiv (1, x_2, \ldots, x_G)$. The effects of the G exogenous variables are "removed" by first regressing (x_{G+1}, x_{G+2}) and (z_1, z_2) on $\mathbf{x}_1$ and then computing the residuals $(\tilde{x}_{G+1}, \tilde{x}_{G+2})$ and $(\tilde{z}_1, \tilde{z}_2)$. This process is often called **partialing out** the effect of $\mathbf{x}_1$.

Suppose that $x_1^* = h_{11}\tilde{x}_{G+1} + h_{21}\tilde{x}_{G+2}$ is a linear combination of the "partialed out" endogenous variables $(\tilde{x}_{G+1}, \tilde{x}_{G+2})$ and $z_1^* = k_{11}\tilde{z}_1 + k_{21}\tilde{z}_2$ is a linear combination of the "partialed out" instrumental variables $(\tilde{z}_1, \tilde{z}_2)$. Using **canonical correlation analysis**, we can determine values h_{11}, h_{21}, k_{11}, and k_{21}, resulting in the largest correlation between x_1^* and z_1^*.[8] It is called the **first** canonical correlation, r_1. Similarly, we can determine values h_{12}, h_{22}, k_{12}, and k_{22}, resulting in the second largest correlation between $x_2^* = h_{12}\tilde{x}_{G+1} + h_{22}\tilde{x}_{G+2}$ and $z_2^* = k_{12}\tilde{z}_1 + k_{22}\tilde{z}_2$, which is called the **second** canonical correlation, r_2—and so on.

If we have two variables in the first set of variables and two variables in the second set, then there are two canonical correlations, r_1 and r_2. If we have B variables in the first group (the endogenous variables with the effects of $\mathbf{x}_1$ removed) and $L \geq B$ variables in the second group (the group of instruments with the effects of $\mathbf{x}_1$ removed), then there are B possible canonical correlations, $r_1 \geq r_2 \geq \cdots \geq r_B$. If the **smallest** canonical correlation $r_B = 0$, then we do not have enough relationships between the instruments and the endogenous variables, and **the equation is not identified**.

10A.1 A Test for Weak Identification

Using the smallest canonical correlation, we are able to test whether any relationship between the instruments and the endogenous variables is sufficiently strong for reliable econometric inferences.[9] Let N denote the sample size, B the number of right-hand side endogenous variables, G the number of exogenous variables included in the equation (including the intercept), L the number of "external" instruments that are not included in the model, and r_B the minimum canonical correlation. A test for weak identification, the situation that arises when the instruments are correlated with the endogenous regressors but only weakly, is based on the **Cragg–Donald *F*-test statistic**[10]

$$\text{Cragg–Donald } F = \left[(N-L)/L\right] \times \left[r_B^2/\left(1-r_B^2\right)\right] \tag{10A.1}$$

The Cragg–Donald statistic reduces to the usual weak instruments F-test when the number of endogenous variables is $B = 1$. Critical values for this test statistic have been tabulated by James Stock and Motohiro Yogo (2005),[11] so that we can test the null hypothesis that the instruments

[8] Certain normalizations on h and k constants are necessary to make the solutions unique. The algebra and calculations are beyond the scope of this book. An online search will reveal many sources but virtually all use matrix algebra and multidimensional calculus. Harold Hotelling did research in mathematical statistics and economic theory and introduced the concept of canonical correlation in a 1935 publication."The most predictable criterion," in the *Journal of Educational Psychology*.

[9] The tests based on canonical correlations are neatly summarized in "Enhanced Routines for Instrumental Variables/ Generalized Method of Moments Estimation and Testing," by Christopher F. Baum, Mark E. Schaffer, and Steven Stillman, *The Stata Journal* (2007), 7, pp. 465–506. Further discussion is provided by Alastair R. Hall, Glenn D. Rudebusch and David W. Wilcox (1996) "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37(2), pp. 283–298.

[10] Cragg, J. G. and S. G. Donald (1993) "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory,* 9, 222–240. D. Poskitt and C. Skeels (2009), "Assessing the Magnitude of the Concentration Parameter in a Simultaneous Equations Model." *The Econometrics Journal,* 12, pp. 26–44, showed that the Cragg–Donald statistic could be conveniently written in terms of the smallest canonical correlation.

[11] "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds, Donald W. K. Andrews and James H. Stock. Cambridge University Press, Chapter 5.

are weak against the alternative that they are not, for two particular consequences of weak instruments.

- **Relative Bias:** In the presence of weak instruments, the amount of bias in the IV estimator can become large. Stock and Yogo consider the bias when estimating the coefficients of the endogenous variables. They examine the maximum IV estimator bias relative to the bias of the least squares estimator. Stock and Yogo give the illustration of estimating the return to education. If a researcher believes that the least squares estimator suffers a maximum bias of 10%, and if the relative bias is 0.1, then the maximum bias of the IV estimator is 1%.
- **Rejection Rate (Test Size):** When estimating a model with endogenous regressors, testing hypotheses about the coefficients of the endogenous variables is frequently of interest. If we choose the $\alpha = 0.05$ level of significance, we expect that a true null hypothesis is rejected 5% of the time in repeated samples. If instruments are weak, then the actual rejection rate of the null hypothesis, also known as the **test size**, may be larger. Stock and Yogo's second criterion is the maximum rejection rate of a true null hypothesis if we choose $\alpha = 0.05$. For example, we may be willing to accept a maximum rejection rate of 10% for a test at the 5% level, but we may not be willing to accept a rejection rate of 20% for a 5% level test.

To test the null hypothesis that instruments are weak against the alternative that they are not, we compare the Cragg–Donald F-test statistic to a critical value chosen from Table 10A.1 or Table 10A.2.

TABLE 10A.1 **Critical Values for the Weak Instrument Test Based on IV Test Size (5% level of significance)[12]**

	$B = 1$ Maximum Test Size				$B = 2$ Maximum Test Size			
L	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	19.93	11.59	8.75	7.25	7.03	4.58	3.95	3.63
3	22.30	12.83	9.54	7.80	13.43	8.18	6.40	5.45
4	24.58	13.96	10.26	8.31	16.87	9.93	7.54	6.28

TABLE 10A.2 **Critical Values for the Weak Instrument Test Based on IV Relative Bias (5% level of significance)[13]**

	$B = 1$ Maximum Relative Bias				$B = 2$ Maximum Relative Bias			
L	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
3	13.91	9.08	6.46	5.39				
4	16.85	10.27	6.71	5.34	11.04	7.56	5.57	4.73

[12]These values are from Table 5.2, page 101, in Stock and Yogo (2005), *op cit*. The authors thank James Stock and Motohiro Yogo for permission to use these results. (Their tables are more extensive than the ones we provide.)

[13]These values are from Table 5.1, page 100, in James H. Stock and Motohiro Yogo (2005), *op cit*. In their paper Stock and Yogo explain that the $F > 10$ rule introduced by Staiger and Stock (1997), op cit., is for $B = 1$ approximately the critical value for a maximum relative bias of 0.10 for all values of L. Their critical values can be considered refinements of the Staiger–Stock rule of thumb.

1. **First** choose either the maximum relative bias or maximum test size criterion. You must also choose the maximum relative bias or maximum test size you are willing to accept.

2a. If you choose the maximum test size criterion, select from Table 10A.1 the critical value associated with a maximum test size of 0.10, 0.15, 0.20, or 0.25 for $B = 1$ or $B = 2$ endogenous variables using $L = 1$ to $L = 4$ instrumental variables.

2b. If you choose the maximum relative bias criterion, select from Table 10A.2 the critical value associated with a maximum relative bias of 0.05, 0.10, 0.20, or 0.30 for $B = 1$ or $B = 2$ endogenous variables using $L = 3$ or $L = 4$ instrumental variables. There are no critical values using this criterion if $L < 3$.

3. Reject the null hypothesis that the instruments are weak if the Cragg–Donald F-test statistic is larger than the tabled critical value. If the F-test statistic is not larger than the critical value, then do not reject the null hypothesis that the instruments are weak.

EXAMPLE 10.8 | Testing for Weak Instruments

In Section 10.2.4 we introduced an example of a wage equation for married working women using Thomas Mroz's data. Consider the following *HOURS* supply equation specification:

$$HOURS = \beta_1 + \beta_2 MTR + \beta_3 EDUC + \beta_4 KIDSL6 + \beta_5 NWIFEINC + e \quad (10A.4)$$

The variable $NWIFEINC = (FAMINC - WAGE \times HOURS)/1000$ is household income attributable to sources other than the wife's income. The variable *MTR* is the marginal tax rate facing the wife, including Social Security taxes. In this equation we expect the signs of coefficients on *MTR*, *KIDSL6*, and *NWIFEINC* to be negative, and the coefficient on *EDUC* is of uncertain sign. In this example, we treat the marginal tax rate as endogenous.[14] Initially we treat *EDUC* as exogenous and use the wife's previous years of work experience, *EXPER*, as an instrumental variable for *MTR*.

Weak IV Example 1: Endogenous: *MTR*; Instrument: *EXPER*

Suppose that we choose the maximum test size criterion and are willing to accept a maximum test size of 0.15 for a 5% test. In Table 10A.1, we see that for $B = 1$ (one right-hand side endogenous variable) and $L = 1$ (one instrument) that the Stock-Yogo critical value is 8.96. The estimated first-stage equation for *MTR* is Model (1) of Table 10A.3. The F-statistic for the hypothesis that the coefficient of experience is zero is 30.61. The Cragg–Donald F-statistic is also 30.61 in this case. Since the Cragg–Donald F-test statistic is larger than the Stock-Yogo critical value 8.96, we reject the null hypothesis that the instruments are weak and accept the alternative that they are not weak. This conclusion is conditional upon the test criterion we have chosen and the maximum size selected. The relative bias criterion cannot be used in this case because it requires at least three instruments. The estimated coefficient of *MTR* in the estimated *HOURS* supply equation in Model (1) of Table 10A.4 is negative and significant at the 5% level.

Weak IV Example 2: *Endogenous: MTR; Instruments: EXPER, EXPER*^{2}*, LARGECITY*

For the sake of illustration, consider using the $L = 3$ instruments *EXPER*, *EXPER*^{2}, and the indicator variable *LARGECITY*, which = 1 if the city is large. Suppose we choose the maximum relative bias criterion and are willing to tolerate a maximum relative bias of 0.10. From Table 10A.2 the Stock–Yogo critical value is 9.08. If the Cragg–Donald F-test statistic is greater than this value, we reject the null hypothesis that the instruments are weak. The first-stage equation estimates are reported in Model (2) of Table 10A.3. The Cragg–Donald F-statistic is 13.22. We conclude that using this test the instruments are not weak. If, however, we are only willing to accept a 0.05 relative bias, then the Stock–Yogo critical value is 13.91. Since the Cragg–Donald F-statistic is less than this value, we cannot reject the null hypothesis that the instruments are weak. The estimated coefficient of *MTR* in the estimated *HOURS* supply equation in Model (2) of Table 10A.4 is negative and significant at the 5% level, although the magnitudes of all the coefficients are smaller in absolute value for this estimation than for the model in Model (1). Qualitatively the estimates of Model (1) and Model (2), using $L = 1$ instrument and $L = 3$ instruments are much the same, with likely thanks to the strong instrument *EXPER*. This example illustrates the point that having more instrumental variables is not necessarily beneficial from the standpoint of weak instrument diagnostics.

[14] This idea is explored by Mroz (1987, p. 786).

TABLE 10A.3 First-stage Equations

MODEL Dependent/ independent	(1) *MTR*	(2) *MTR*	(3) *MTR*	(4) *EDUC*	(5) *MTR*	(6) *EDUC*
C	0.87930 (74.33)	0.88470 (71.93)	0.79907 (103.22)	8.71459 (25.83)	0.82960 (93.34)	8.17622 (20.34)
EXPER	−0.00142 (−5.53)	−0.00217 (−2.65)			−0.00168 (−6.23)	0.02957 (2.43)
EDUC	−0.00718 (−7.76)	−0.00689 (−7.45)				
KIDSL6	0.02037 (3.86)	0.02039 (3.89)	0.02189 (3.92)	0.61812 (2.54)	0.01559 (2.87)	0.72921 (2.96)
NWIFEINC	−0.00551 (−27.40)	−0.00539 (−26.35)	−0.00565 (−27.15)	0.04961 (5.46)	−0.00585 (−28.96)	0.05304 (5.81)
$EXPER^2$		0.00002 (1.01)				
LARGECITY		−0.01163 (−2.70)				
MOTHEREDUC			−0.00111 (−1.40)	0.15202 (4.40)	−0.00134 (−1.76)	0.15601 (4.54)
FATHEREDUC			−0.00180 (−2.40)	0.16371 (5.01)	−0.00202 (−2.81)	0.16754 (5.15)
N	428	428	428	428	428	428
Weak IV *F*	30.61	13.22	8.14	49.02	18.86	35.03
Number IV *L*	1	3	2	2	3	3
Number Endog *B*	1	1	2	2	2	2

t-statistics in parentheses.

TABLE 10A.4 IV Estimation of Hours Equation

MODEL	(1)	(2)	(3)	(4)
C	17423.7211 (5.56)	14394.1144 (5.68)	−24491.5995 (−0.31)	18067.8425 (5.11)
MTR	−18456.5896 (−5.08)	−14934.3696 (−5.09)	29709.4677 (0.33)	−18633.9223 (−4.85)
EDUC	−145.2928 (−4.40)	−118.8846 (−4.28)	258.5590 (0.32)	−189.8611 (−3.04)
KIDSL6	151.0229 (1.07)	58.7879 (0.48)	−1144.4779 (−0.46)	190.2755 (1.20)
NWIFEINC	−103.8983 (−5.27)	−85.1934 (−5.32)	149.2325 (0.31)	−102.1516 (−5.11)
N	428	428	428	428
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60

t-statistics in parentheses.

Weak IV Example 3 *Endogenous: MTR, EDUC; Instruments: MOTHEREDUC, FATHEREDUC*

Now treat both marginal tax rate *MTR* and education *EDUC* as endogenous, so that $B = 2$. Following Section 10.3.6 we use mother's and father's education, *MOTHEREDUC* and *FATHEREDUC*, as instruments, so that $L = 2$. Suppose that we are willing to accept a maximum test size of 15% for a 5% test. From Table 10A.1 the critical value for the weak instrument test is 4.58. The first-stage equations for *MTR* and *EDUC* are Model (3) and Model (4) of Table 10A.3. These instruments are strong for *EDUC* as we have earlier seen, with the first-stage weak instrument F-test statistic 49.02. For *MTR* [Model (3)] these two instruments are less strong. *FATHEREDUC* is significant at the 5% level, and the first-stage weak instrument F-test statistic is 8.14, which has a p-value of 0.0003. While this does not satisfy the $F \geq 10$ rule of thumb, it is "close," and we may have concluded that these two instruments were adequately strong. The Cragg–Donald F-test statistic value is only 0.101, which is far below the critical value 4.58 for 15% maximum test size (for a 5% test on *MTR* and *EDUC*). We cannot reject the null hypothesis that the instruments are *weak*, despite the favorable first-stage F-test values. The estimates of the *HOURS* supply equation, Model (3) of Table 10A.3, shows parameter estimates that are wildly different from those in Model (1) and Model (2), and the very small t-statistic values imply very large standard errors, another consequence of instrumental variables estimation in the presence of weak instruments.

Weak IV Example 4 *Endogenous: MTR, EDUC; Instruments: MOTHEREDUC, FATHEREDUC, EXPER*

If we include the additional instrument *EXPER*, so that $L = 3$, we obtain the first-stage estimates in Model (5) and Model (6) of Table 10A.3. Once again the first-stage weak instrument F-test statistic values appear strong, with values for *MTR* of 18.86 and for *EDUC* of 35.03. Using the $F > 10$ rule of thumb, we would be comfortable that our instruments are strong. The Cragg–Donald F-test statistic value is 8.60, which tells a slightly different story. Our instruments are not quite as strong as the first-stage weak instrument F-test statistics imply. If we choose a maximum test size of 0.15, we can reject the null hypothesis of weak instruments. If, however, we are prepared to accept only a maximum 10% rejection rate for a 5% test, the critical value is 13.43, and we do not reject the null hypothesis that the instruments are weak. The instrumental variables estimates of the *HOURS* supply equation are Model (4) of Table 10A.4 and we see that they are more in line with Model (1) and Model (2) than those in Model (3).

10A.2 Testing for Weak Identification: Conclusions

If instrumental variables are "weak," then the instrumental variables, or two-stage least squares, estimator is unreliable. When there is a single endogenous variable, the first-stage F-test of the joint significance of the external instruments is an indicator of instrument strength. The $F > 10$ rule of thumb has been refined by Stock and Yogo, who provide tables of critical values for the null hypothesis "the instruments are weak" using two criteria: the bias of the IV estimator relative to the bias of the least squares estimator, and the maximum size of a 5% test of the coefficients of the endogenous variables. If there is more than one endogenous variable on the right-hand side of an equation, then the F-test statistics from the first-stage equations do not provide reliable information about instrument strength. In this case the Cragg–Donald F-test statistic should be used to test for weak instruments, along with the Stock-Yogo tables of critical values.

Econometric research continues for alternatives to the IV/2SLS estimator in the weak instrument case. Some progress has been made; these results are summarized in Appendix 11B. The discussion is deferred until the next chapter, as the advances have their genesis in discussions of estimation of simultaneous equations models.

Appendix 10B Monte Carlo Simulation

In this appendix we do two sorts of simulations. First, we generate a sample of artificial data and give numerical illustrations of the estimators and tests discussed in the chapter. In the chapter the illustrations used real data. The advantage gained here is that we can see how the estimators and tests perform using data we know comes from a particular data generation process. Secondly, we carry out a Monte Carlo simulation to illustrate the repeated **sampling properties** of the least squares and IV/2SLS estimators under various conditions.

10B.1 Illustrations Using Simulated Data

In this section, we demonstrate, using a simulated sample of data, that the OLS estimator fails when $\text{cov}(x_i, e_i) \neq 0$, and that instrumental variables estimators "work" when conditions listed in Section 10.3.3 are satisfied. For the simulated data, we specify a simple regression model in which the parameter values are $\beta_1 = 1$ and $\beta_2 = 1$. Thus, the systematic part of the regression model is $E(y|x) = \beta_1 + \beta_2 x = 1 + 1 \times x$. By adding to $E(y|x)$ an error term value, which will be a random number we create, we can create a sample value of y.

We want to explore the properties of the OLS estimator when x and e are correlated. Using random number generators, we create $N = 100$ pairs of x and e values, such that each has a normal distribution with mean zero and variance one. The population correlation between the x and e values is ρ_{xe}. We then create an artificial sample of y values by adding e to the systematic portion of the regression,

$$y = E(y|x) + e = \beta_1 + \beta_2 x + e = 1 + 1 \times x + e$$

The data values are contained in the data file *ch10*. The OLS estimates are

$$\begin{aligned} \hat{y}_{\text{OLS}} &= 0.9789 + 1.7034x \\ (\text{se}) \quad & (0.088) \quad (0.090) \end{aligned}$$

When x and e are positively correlated, the estimated slope tends to be too large—here, $b_2 = 1.7034$ compared to the true $\beta_2 = 1$. Furthermore, the systematic overestimation of the slope will not go away in larger samples, so the least squares estimators are not correct on average even in large samples. The least squares estimators are inconsistent.

In the process of creating the artificial data (data file *ch10*) we also created two instrumental variables, both uncorrelated with the error term. The correlation between the first instrument z_1 and x is $\rho_{xz_1} = 0.5$, and the correlation between the second instrument z_2 and x is $\rho_{xz_2} = 0.3$. The IV estimates using z_1 are

$$\begin{aligned} \hat{y}_{IV\_z_1} &= 1.1011 + 1.1924x \\ (\text{se}) \quad & (0.109) \quad (0.195) \end{aligned}$$

and the IV estimates using z_2 are

$$\begin{aligned} \hat{y}_{IV\_z_2} &= 1.3451 + 0.1724x \\ (\text{se}) \quad & (0.256) \quad (0.797) \end{aligned}$$

Using z_1, the stronger instrument, yields an estimate of the slope of 1.1924 with a standard error of 0.195, about twice the standard error of the OLS estimate. Using the weaker instrument z_2 produces a slope estimate of 0.1724, which is far from the true value, and a standard error of 0.797, about eight times as large as the least squares standard error. The results with the weaker instrument are far less satisfactory than the estimates based on the stronger instrument z_1.

Another problem that an instrument can have is that it is not uncorrelated with the error term as it is supposed to be. The variable z_3 is correlated with x, with correlation $\rho_{xz_3} = 0.5$, but it is correlated with the error term e, with correlation $\rho_{ez_3} = 0.3$. Thus, z_3 is not a valid instrument. What happens if we use instrumental variables estimation with the invalid instrument? The results are

$$\begin{aligned} \hat{y}_{IV-z_3} &= 0.9640 + 1.7657x \\ (\text{se}) \quad & (0.095) \quad (0.172) \end{aligned}$$

As you can see, using the invalid instrument produces a slope estimate even further from the true value than the least squares estimate. Using an invalid instrumental variable means that the instrumental variables estimator will be inconsistent, just like the least squares estimator.

What is the outcome of two-stage least squares estimation using the two instruments z_1 and z_2? Obtain the first-stage regression of x on the two instruments z_1 and z_2,

$$\begin{aligned} \hat{x} &= 0.1947 + 0.5700z_1 + 0.2068z_2 \\ (\text{se}) \ (0.095) \quad &(0.089) \quad (0.077) \end{aligned} \tag{10B.1}$$

Using the predicted value $\hat{x}$ to replace x, then applying least squares to the modified equation, as in (10.22), we obtain the instrumental variables estimates

$$\begin{aligned} \hat{y}_{IV\_z_1,z_2} &= 1.1376 + 1.0399x \\ (\text{se}) \quad &(0.116) \quad (0.194) \end{aligned} \tag{10B.2}$$

The standard errors are based on an estimated error variance as in (10.18b). Using the two valid instruments yields an estimate of the slope of 1.0399, which, in this example, is close to the true value of $\beta_2 = 1$.

10B.1.1 The Hausman Test

To implement the Hausman test we estimate the first-stage equation, which is shown in (10A.1) using the instruments z_1 and z_2. Compute the residuals

$$\hat{v} = x - \hat{x} = x - 0.1947 - 0.5700z_1 - 0.2068z_2$$

Include the residuals as an extra variable in the regression equation and apply least squares,

$$\begin{aligned} \hat{y} &= 1.1376 + 1.0399x + 0.9957\hat{v} \\ (\text{se}) \ (0.080) \quad &(0.133) \quad (0.163) \end{aligned}$$

The t-statistic for the null hypothesis that the coefficient of $\hat{v}$ is zero is 6.11. The critical value comes from the t-distribution with 97 degrees of freedom and is 1.985, so we reject the null hypothesis that x is uncorrelated with the error term and correctly conclude that it is endogenous.

10B.1.2 Test for Weak Instruments

The test for weak instruments again begins with estimation of the first-stage regression. If we consider using just z_1 as an instrument, the estimated first-stage equation is

$$\begin{aligned} \hat{x} &= 0.2196 + 0.5711z_1 \\ (t) \quad &(6.24) \end{aligned}$$

The t-statistic 6.24 corresponds to an F-value of 38.92, which is well above the guideline value of 10. If we use just z_2 as an instrument, the estimated first-stage equation is

$$\begin{aligned} \hat{x} &= 0.2140 + 0.2090z_2 \\ (t) \quad &(2.28) \end{aligned}$$

While the t-statistic 2.28 indicates statistical significance at the 0.05 level, the corresponding F-value is $5.21 < 10$, indicating that z_2 is a weak instrument. The first-stage equation using both instruments is shown in (10B.1), and the F-test for their joint significance is 24.28, indicating that we have at least one strong instrument.

10B.1.3 Testing the Validity of Surplus Instruments

If we use z_1 and z_2 as instruments, there is one extra. The number of instruments is $L = 2$, and the number of endogenous regressors is $B = 1$. The IV estimates are shown in (10B.2). Calculate the residuals from this equation and then regress them on intercept, z_1 and z_2, to obtain $\hat{e} = 0.0189 + 0.0881z_1 - 0.1818z_2$. The R^2 from this regression is 0.03628, and $NR^2 = 3.628$. The 0.05 critical value for the chi-square distribution with one degree of freedom is 3.84, so we fail to reject the validity of the surplus moment condition.

If we use z_1, z_2, and z_3 as instruments, there are two surplus moment conditions. The IV estimates using these three instruments are $\hat{y}_{IV\_z_1,z_2,z_3} = 1.0626 + 1.3535x$. Obtaining the residuals and regressing them on the instruments yields

$$\hat{e} = 0.0207 - 0.1033z_1 - 0.2355z_2 + 0.1798z_3$$

The R^2 from this regression is 0.1311, and $NR^2 = 13.11$. The 0.05 critical value for the chi-square distribution with two degrees of freedom is 5.99, so we reject the validity of the two surplus moment conditions. This test does not identify the problem instrument, but since we first tested the validity of z_1 and z_2 and failed to reject their validity, and then found that adding z_3 led us to reject the validity of the surplus moment conditions, the instrument z_3 seems to be the culprit.

10B.2 The Sampling Properties of IV/2SLS

To illustrate the repeated sampling properties of the OLS and IV/2SLS estimators, we use an experimental design based on the discussion in Section 10.4.2. In the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, if x_i is correlated with the error term e_i then x_i is endogenous, and the least squares estimator is biased and inconsistent. An instrumental variable z_i must be correlated with x_i but uncorrelated with e_i in order to be valid. A correlation between z_i and x_i implies that there is a linear association between them. This means that we can describe their relationship as a regression $x_i = \gamma_1 + \theta_1 z_i + v_i$. There is a correlation between x_i and z_i if, and only if, $\theta_1 \neq 0$. If we knew γ_1 and θ_1, we could substitute $E(x_i|z_i) = \gamma_1 + \theta_1 z_i$ into the simple regression model to obtain $y_i = \beta_1 + \beta_2 E(x_i|z_i) + \beta_2 v_i + e_i$. Suppose for a moment that $E(x_i|z_i)$ and v_i can be observed and are viewed as explanatory variables in the regression $y_i = \beta_1 + \beta_2 E(x_i|z_i) + \beta_2 v_i + e_i$. The explanatory variable $E(x_i|z_i)$ is not correlated with the error term e_i because it depends only on z_i. Any correlation between x_i and e_i implies correlation between v_i and e_i because $v_i = x_i - E(x_i|z_i)$.

In the simulation,[15] we use the data generation process $y_i = x_i + e_i$, so that the intercept parameter is 0 and the slope parameter is 1. The first-stage regression is $x_i = \theta z_{i1} + \theta z_{i2} + \theta z_{i3} + v_i$. Note that we have $L = 3$ instruments, each of which has an independent standard normal $N(0,1)$ distribution. The parameter θ controls the instrument strength. If $\theta = 0$, the instruments are not correlated with x_i and instrumental variables estimation will fail. The larger θ becomes the stronger the instruments become. Finally, we create the random errors e_i and v_i to have standard normal distributions with correlation ρ, which controls the endogeneity of x. If $\rho = 0$, then x is not endogenous. The larger ρ becomes the stronger the endogeneity. We create 10,000 samples of size $N = 100$ and then try out OLS and IV/2SLS under several scenarios. We let $\theta = 0.1$ (weak instruments) and $\theta = 0.5$ (strong instruments). We let $\rho = 0$ (x exogenous) and $\rho = 0.8$ (x highly endogenous).

In Table 10B.1, the reported values are

- $\overline{F}$ is the average first-stage F: compare these values to 10. Note that the average value of F is about 2 when $\theta = 0.1$ indicating weak instruments. The average value of F is about 21 when $\theta = 0.5$ indicating strong instruments.

[15] This design is similar to that used by Jinyong Hahn and Jerry Hausman (2003) "Weak Instruments: Diagnosis and Cures in Empirical Economics," *American Economic Review*, 93(2), pp. 118–125.

TABLE 10B.1 **Monte Carlo Simulation Results**

ρ	θ	$\overline{F}$	$\overline{b}_2$	$s.d.(b_2)$	$t(b_2)$	$t(H)$	$\overline{\hat{\beta}}_2$	$s.d.(\hat{\beta}_2)$	$t(\hat{\beta}_2)$
0.0	0.1	1.98	1.0000	0.1000	0.0499	0.0510	0.9941	0.6378	0.0049
0.0	0.5	21.17	0.9999	0.0765	0.0484	0.0518	0.9998	0.1184	0.0441
0.8	0.1	2.00	1.7762	0.0610	1.0000	0.3077	1.3311	0.9483	0.2886
0.8	0.5	21.18	1.4568	0.0610	1.0000	0.9989	1.0111	0.1174	0.0636

- $\overline{b}_2$ is the average of the OLS estimates of $\beta_2 = 1$. The least squares estimator is unbiased when $\rho = 0$, but when $\rho = 0.8$, the least squares estimator shows severe bias.
- $s.d.(b_2)$ is the sample standard deviation of the 10,000 Monte Carlo values of b_2. It tells us how much variation the OLS estimates exhibit in repeated sampling.
- $t(b_2)$ is the percentage of rejections of the true null hypothesis $\beta_2 = 1$ using the 0.05 level of significance t-test based on the OLS estimator. If there is no endogeneity, the percent rejections is very close to the 0.05 value, but if there is strong endogeneity, the OLS estimator rejects the true null hypothesis 100% of the time. That is not good.
- $t(H)$ is the percentage rejections of the regression-based Hausman test for endogeneity using the 0.05 level of significance. If there is no endogeneity, the test rejects 5% of the time, which is what we expect. If there is strong endogeneity but weak instruments, $\theta = 0.1$, the test rejects only 31% of the time, failing to indicate the endogeneity problem. If instruments are not strong, nothing is going to work well. If the instruments are strong, then the test for endogeneity is very successful in detecting strong endogeneity.
- $\overline{\hat{\beta}}_2$ is the average of the instrumental variables estimates of $\beta_2 = 1$. The IV estimator is unbiased when $\rho = 0$. When endogeneity is strong, with weak instruments the IV estimator has a 33% bias, but when instruments are strong it has an average very close to the true value.
- $s.d.(\hat{\beta}_2)$ is the sample standard deviation of the IV estimates in the 10,000 Monte Carlo samples. If there is no endogeneity, note how large its standard deviation is relative to the least squares estimator. With weak instruments its standard deviation is six times that of the least squares estimator. Even with strong instruments, it is substantially larger. The IV estimator is **inefficient** relative to the least squares estimator when endogeneity is absent. When endogeneity is present, the effect of weak instruments shows up in the large standard deviation of the estimates. When instruments are stronger, the standard deviation of the IV estimates falls from 0.95 to 0.12, a substantial improvement.
- Finally, we see the rate of rejections of the true null hypothesis $\beta_2 = 1$ under the scenarios. When x is endogenous and the instruments are weak, the t-test rejects far too often, but it is better than the t-test based on the least squares estimator. Otherwise, the rejection rate is close to the 5% that we expect.

These results are based on a sample size of $N = 100$, which is neither large nor small. What results do you anticipate with larger or smaller samples?

Advice about what to do when there is uncertainty as to whether a regressor is endogenous or not is somewhat mixed. In Table 10.2, the Hausman test statistic p-value is 0.0954. The prevailing attitude is probably summarized by Jeffrey Wooldridge,[16] who says, "We find evidence of endogeneity of *EDUC* at the 10% significance level against a two-sided alternative, and so 2SLS is probably a good idea (assuming that we trust the instruments.)" On the other hand, Patrik

[16] *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition, The MIT Press, 2010, p. 132.

Guggenberger[17] advises, that if testing the coefficient of the endogenous regressor is the objective, then we should avoid considering the Hausman test result and use 2SLS. On the other hand, if we consider how close the estimates are to the true value on average, the "mean square error," Chmelarova and Hill[18] advise that perhaps IV/2SLS should be used only if a Hausman pretest has a much smaller p-value. This result is revealed somewhat in the Monte Carlo simulation. In the case in which $\rho = 0.8$ and $\theta = 0.1$, the mean square error for the least squares estimator is

$$\sum_{m=1}^{10000} \left(b_{2m} - \beta_2\right)^2 / 10000 = 0.6062$$

while for the IV estimator it is

$$\sum_{m=1}^{1000} \left(\hat{\beta}_{2m} - \beta_2\right)^2 / 10000 = 1.0088$$

In other words, in this experimental setting with strong endogeneity and weak instruments, the least squares estimator is, on average, closer to the true parameter value than the IV estimator.

[17] "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test," *Econometric Theory*, 2010, 26(2), pp. 369–382.

[18] "The Hausman Pretest Estimator," *Economics Letters*, 2010, 108, 96–99.