CHAPTER

12 Instrumental Variables Regression

Chapter 9 discussed several problems, including omitted variables, errors in variables, and simultaneous causality, that make the error term correlated with the regressor. Omitted variable bias can be addressed directly by including the omitted variable in a multiple regression, but this is only feasible if you have data on the omitted variable. And sometimes, such as when causality runs *both* from X to Y *and* from Y to X so that there is simultaneous causality bias, multiple regression simply cannot eliminate the bias. If a direct solution to these problems is either infeasible or unavailable, a new method is required.

Instrumental variables (IV) regression is a general way to obtain a consistent estimator of the unknown causal coefficients when the regressor, X, is correlated with the error term, u. To understand how IV regression works, think of the variation in X as having two parts: one part that, for whatever reason, is correlated with u (this is the part that causes the problems) and a second part that is uncorrelated with u. If you had information that allowed you to isolate the second part, you could focus on those variations in X that are uncorrelated with u and disregard the variations in X that bias the OLS estimates. This is, in fact, what IV regression does. The information about the movements in X that are uncorrelated with u is gleaned from one or more additional variables, called **instrumental variables** or simply **instruments**. Instrumental variables regression uses these additional variables as tools or "instruments" to isolate the movements in X that are uncorrelated with u, which in turn permits consistent estimation of the regression coefficients.

The first two sections of this chapter describe the mechanics and assumptions of IV regression: why IV regression works, what is a valid instrument, and how to implement and to interpret the most common IV regression method, two stage least squares. The key to successful empirical analysis using instrumental variables is finding valid instruments, and Section 12.3 takes up the question of how to assess whether a set of instruments is valid. As an illustration, Section 12.4 uses IV regression to estimate the elasticity of demand for cigarettes. Finally, Section 12.5 turns to the difficult question of where valid instruments come from in the first place.

12.1 The IV Estimator with a Single Regressor and a Single Instrument

We start with the case of a single regressor, X, which might be correlated with the error, u. If X and u are correlated, the OLS estimator is inconsistent; that is, it may not be close to the true value of the causal coefficient even when the sample is very large [see Equation (6.1)]. As discussed in Section 9.2, this correlation between X and u can stem from various sources, including omitted variables, errors in variables (measurement errors in the regressors), and simultaneous causality (when causality runs "backward" from Y to X as well as "forward" from X to Y). Whatever the source of the correlation between X and u, if there is a valid instrumental variable, Z, the effect on Y of a unit change in X can be estimated using the instrumental variables estimator.

The IV Model and Assumptions

Let β_1 be the causal effect of X on Y. The model relating the dependent variable Y_i and regressor X_i, without any control variables, is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n, \tag{12.1}$$

where u_i is the error term representing omitted factors that determine Y_i. If X_i and u_i are correlated, the OLS estimator is inconsistent. Instrumental variables estimation uses an additional, "instrumental" variable Z to isolate that part of X that is uncorrelated with u.

Endogeneity and exogeneity. Instrumental variables regression has some specialized terminology to distinguish variables that are correlated with the population error term u from ones that are not. Variables correlated with the error term are called **endogenous variables**, while variables uncorrelated with the error term are called **exogenous variables**. The historical source of these terms traces to models with multiple equations, in which an "endogenous" variable is determined within the model, while an "exogenous" variable is determined outside the model. For example, Section 9.2 considered the possibility that if low test scores produced decreases in the student–teacher ratio because of political intervention and increased funding, causality would run *both* from the student–teacher ratio to test scores *and* from test scores to the student–teacher ratio. This was represented mathematically as a system of two simultaneous equations [Equations (9.3) and (9.4)], one for each causal connection. As discussed in Section 9.2, because both test scores and the student–teacher ratio are determined within the model, both are correlated with the population error term u; that is, in this example, both variables are endogenous. In contrast, an exogenous variable, which is determined outside the model, is uncorrelated with u.

The two conditions for a valid instrument. A valid instrumental variable ("instrument") Z must satisfy two conditions, known as the **instrument relevance condition** and the **instrument exogeneity condition**:

1. Instrument relevance: $\text{corr}(Z_i, X_i) \neq 0$.
2. Instrument exogeneity: $\text{corr}(Z_i, u_i) = 0$.

If an instrument is relevant, then variation in the instrument is related to variation in X_i. If in addition the instrument is exogenous, then that part of the variation of X_i captured by the instrumental variable is exogenous. Thus an instrument that is relevant and exogenous can capture movements in X_i that are exogenous. This exogenous variation can in turn be used to estimate the population coefficient β_1.

The two conditions for a valid instrument are vital for instrumental variables regression, and we return to them (and their extension to multiple regressors and multiple instruments) repeatedly throughout this chapter.

The Two Stage Least Squares Estimator

If the instrument Z satisfies the conditions of instrument relevance and exogeneity, the coefficient β_1 can be estimated using an IV estimator called **two stage least squares (TSLS)**. As the name suggests, the two stage least squares estimator is calculated in two stages. The first stage decomposes X into two components: a problematic component that may be correlated with the regression error and another, problem-free component that is uncorrelated with the error. The second stage uses the problem-free component to estimate β_1.

The first stage begins with a population regression linking X and Z:

$$X_i = \pi_0 + \pi_1 Z_i + v_i, \tag{12.2}$$

where π_0 is the intercept, π_1 is the slope, and v_i is the error term. This regression provides the needed decomposition of X_i. One component is $\pi_0 + \pi_1 Z_i$, the part of X_i that can be predicted by Z_i. Because Z_i is exogenous, this component of X_i is uncorrelated with u_i, the error term in Equation (12.1). The other component of X_i is v_i, which is the problematic component of X_i that is correlated with u_i.

The idea behind TSLS is to use the problem-free component of X_i, $\pi_0 + \pi_1 Z_i$, and to disregard v_i. The only complication is that the values of π_0 and π_1 are unknown, so $\pi_0 + \pi_1 Z_i$ cannot be calculated. Accordingly, the first stage of TSLS applies OLS to Equation (12.2) and uses the predicted value from the OLS regression, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, where $\hat{\pi}_0$ and $\hat{\pi}_1$ are the OLS estimates.

The second stage of TSLS is easy: Regress Y_i on $\hat{X}_i$ using OLS. The resulting estimators from the second-stage regression are the TSLS estimators, $\hat{\beta}_0^{TSLS}$ and $\hat{\beta}_1^{TSLS}$.

Why Does IV Regression Work?

Two examples provide some insight into why IV regression solves the problem of correlation between X_i and u_i.

When Was Instrumental Variables Regression Invented?

Instrumental variables regression was first proposed as a solution to the simultaneous causation problem in econometrics in the appendix to Philip G. Wright's 1928 book, *The Tariff on Animal and Vegetable Oils*. If you want to know how animal and vegetable oils were produced, transported and sold in the early twentieth century, the first 285 pages of the book are for you. Econometricians, however, will be more interested in Appendix B. The appendix provides two derivations of "the method of introducing external factors"—what we now call the instrumental variables estimator—and uses IV regression to estimate the supply and demand elasticities for butter and flaxseed oil. Philip was an obscure economist with a scant intellectual legacy other than this appendix, but his son Sewall went on to become a preeminent population geneticist and statistician. The invention of IV regression has been found to have been a joint intellectual collaboration between father and son. To learn more, see Stock and Trebbi (2003).

The use of principles similar to those employed in IV regression can be traced further back in time to the identification of the origins of an outbreak of cholera in mid-nineteenth century London by John Snow. Snow wanted to investigate whether or not cholera was water-borne, but knew that an association between exposure to impure water and prevalence of cholera would not provide definitive proof of causality as the households exposed to impure water were often also exposed to a wide range of other environmental factors that could also have been behind the outbreak of the disease (omitted variables). Snow identified that two main water companies supplying water to households in London drew their water supply from different parts of the River Thames, and used this information to get around the issue. Lambeth Water Company drew their water from above the sewage discharge, while Southwark and Vauxhall Company drew their water from below the discharge (thereby drawing water with greater impurity). Snow argued that the households served by the two companies were similar except for the purity of the water that they were provided. He used this information to employ an approach analogous to IV regression with exposure to impure water as the endogenous variable and the supplying water company as the instrumental variable. The supplying water company could be considered to be relevant to the exposure to impure water because of where water was drawn from in the river relative to the sewage discharge. The instrument was considered exogenous since there was no plausible direct effect on cholera and it was uncorrelated with other household factors that may have caused cholera. John Snow's work identifying exposure to impure water as one of the causes of the outbreak of cholera led to him being regarded as one of the fathers of epidemiology. This example is explained in greater detail in Deaton (1997), Grootendorst (2007), Greene (2003), and, of course, Snow (1855).

Courtesy of Rosalind W. Harris

Philip G. Wright

Courtesy of Rosalind W. Harris

Sewall Wright

Example 1: Philip Wright's problem. The method of instrumental variables estimation was first published in 1928 in an appendix to a book written by Philip G. Wright (1928), although the key ideas of IV regression were developed collaboratively with his son Sewall Wright (see the box "When Was Instrumental Variables Regression Invented?"). Philip Wright was concerned with an important economic problem of his day: how to set an import tariff (a tax on imported goods) on animal and vegetable oils and fats, such as butter and soy oil. In the 1920s, import tariffs were a major source of tax revenue for the United States. The key to understanding the economic effect of a tariff was having quantitative estimates of the demand and supply curves of the goods. Recall that the supply elasticity is the percentage change in the quantity supplied arising from a 1% increase in the price and that the demand elasticity is the percentage change in the quantity demanded arising from a 1% increase in the price. Philip Wright needed estimates of these elasticities of supply and demand.

To be concrete, consider the problem of estimating the elasticity of demand for butter. Recall from Key Concept 8.2 that the coefficient in a linear equation relating $\ln(Y_i)$ to $\ln(X_i)$ has the interpretation of the elasticity of Y with respect to X. In Wright's problem, this suggests the demand equation

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i, \tag{12.3}$$

where Q_i^{butter} is the i^{th} observation on the quantity of butter consumed, P_i^{butter} is its price, and u_i represents other factors that affect demand, such as income and consumer tastes. In Equation (12.3), a 1% increase in the price of butter yields a β_1 percent change in demand, so β_1 is the demand elasticity.

Philip Wright had data on total annual butter consumption and its average annual price in the United States for 1912 to 1922. It would have been easy to use these data to estimate the demand elasticity by applying OLS to Equation (12.3), but he had a key insight: Because of the interactions between supply and demand, the regressor, $\ln(P_i^{butter})$, was likely to be correlated with the error term.

To see this, look at Figure 12.1a, which shows the market demand and supply curves for butter for three different years. The demand and supply curves for the first period are denoted D_1 and S_1, and the first period's equilibrium price and quantity are determined by their intersection. In year 2, demand increases from D_1 to D_2 (say, because of an increase in income), and supply decreases from S_1 to S_2 (because of an increase in the cost of producing butter); the equilibrium price and quantity are determined by the intersection of the new supply and demand curves. In year 3, the factors affecting demand and supply change again; demand increases again to D_3, supply increases to S_3, and a new equilibrium quantity and price are determined. Figure 12.1b shows the equilibrium quantity and price pairs for these three periods and for eight subsequent years, where in each year the supply and demand curves are subject to shifts associated with factors other than price that affect market supply and demand. This scatterplot is like the one that Wright would have seen when he plotted his data. As he reasoned, fitting a line to these points by OLS will estimate neither a demand curve nor a supply curve because the points have been determined by changes in both demand and supply.

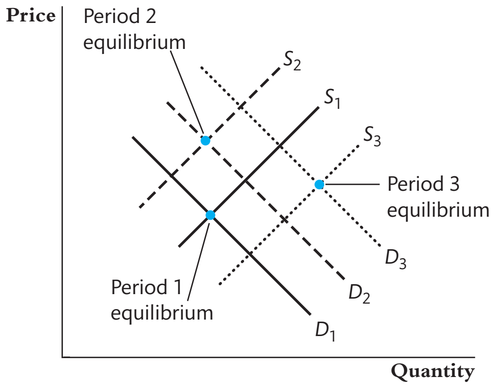
Wright realized that a way to get around this problem was to find some third variable that shifted supply but did not shift demand. Figure 12.1c shows what happens when such a variable shifts the supply curve but demand remains stable. Now all of the equilibrium price and quantity pairs lie on a stable demand curve, and the slope of the demand curve is easily estimated. In the instrumental variable formulation of Wright's problem, this third variable—the instrumental variable—is correlated with price (it shifts the supply curve, which leads to a change in price) but is uncorrelated with u (the demand curve remains stable). Wright considered several potential instrumental variables; one was the weather. For example, below-average rainfall in a dairy region could impair grazing and thus reduce butter production at a given price (it would shift the supply curve to the left and increase the equilibrium price), so dairy-region rainfall satisfies the condition for instrument relevance. But dairy-region rainfall should not have a direct influence on the demand for butter, so the correlation between dairy-region rainfall and u_i would be 0; that is, dairy-region rainfall satisfies the condition for instrument exogeneity.

Example 2: Estimating the effect on test scores of class size. Despite controlling for student and district characteristics, the estimates of the effect on test scores of class size reported in Part II still might have omitted variable bias resulting from unmeasured variables such as learning opportunities outside school or the quality of the teachers. If data on these variables, or on suitable control variables, are unavailable, this omitted variable bias cannot be addressed by including the variables in the multiple regressions.

Instrumental variables regression provides an alternative approach to this problem. Consider the following hypothetical example: Some California schools

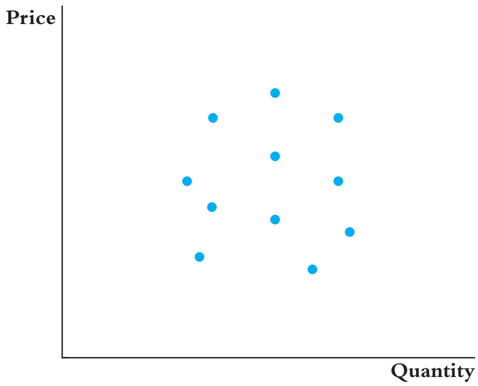
FIGURE 12.1 Equilibrium Price and Quantity Data

(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.



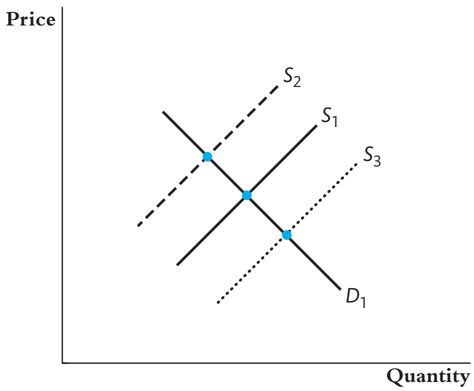
(a) Demand and supply in three time periods

(b) This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



(b) Equilibrium price and quantity for 11 time periods

(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.



(c) Equilibrium price and quantity when only the supply curve shifts are forced to close for repairs because of a summer earthquake. Districts closest to the epicenter are most severely affected. A district with some closed schools needs to "double up" its students, temporarily increasing class size. This means that distance from the epicenter satisfies the condition for instrument relevance because it is correlated with class size. But if distance to the epicenter is unrelated to any of the other factors affecting student performance (such as whether the students are still learning English or disruptive effects of the earthquake on student performance), then it will be exogenous because it is uncorrelated with the error term. Thus the instrumental variable, distance to the epicenter, could be used to circumvent omitted variable bias and to estimate the effect of class size on test scores.

The Sampling Distribution of the TSLS Estimator

The exact distribution of the TSLS estimator in small samples is complicated. However, like the OLS estimator, its distribution in large samples is simple: The TSLS estimator is consistent and is normally distributed.

Formula for the TSLS estimator. Although the two stages of TSLS make the estimator seem complicated, when there is a single X and a single instrument Z, as we assume in this section, there is a simple formula for the TSLS estimator. Let s_{ZY} be the sample covariance between Z and Y, and let s_{ZX} be the sample covariance between Z and X. As shown in Appendix 12.2, the TSLS estimator with a single instrument is

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}. \tag{12.4}$$

That is, the TSLS estimator of β_1 is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X.

Sampling distribution of $\hat{\beta}_1^{TSLS}$ when the sample size is large. The formula in Equation (12.4) can be used to show that $\hat{\beta}_1^{TSLS}$ is consistent and, in large samples, normally distributed. The argument is summarized here, with mathematical details given in Appendix 12.3.

The argument that $\hat{\beta}_1^{TSLS}$ is consistent combines the assumptions that Z_i is relevant and exogenous with the consistency of sample covariances for population covariances. To begin, note that because $Y_i = \beta_0 + \beta_1 X_i + u_i$ in Equation (12.1),

$$\text{cov}(Z_i, Y_i) = \text{cov}(Z_i, \beta_0 + \beta X_i + u_i) = \beta_1 \text{cov}(Z_i, X_i) + \text{cov}(Z_i, u_i), \tag{12.5}$$

where the second equality follows from the properties of covariances [Equation (2.34)]. By the instrument exogeneity assumption, $\text{cov}(Z_i, u_i) = 0$, and by the instrument relevance assumption, $\text{cov}(Z_i, X_i) \neq 0$. Thus, if the instrument is valid, Equation (12.5) implies that

$$\beta_1 = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)}. \tag{12.6}$$

That is, the population coefficient β_1 is the ratio of the population covariance between Z and Y to the population covariance between Z and X.

As discussed in Section 3.7, the sample covariance is a consistent estimator of the population covariance; that is, $s_{ZY} \xrightarrow{p} \text{cov}(Z_i, Y_i)$ and $s_{ZX} \xrightarrow{p} \text{cov}(Z_i, X_i)$. It follows from Equations (12.4) and (12.6) that the TSLS estimator is consistent:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)} = \beta_1. \tag{12.7}$$

The formula in Equation (12.4) also can be used to show that the sampling distribution of $\hat{\beta}_1^{TSLS}$ is normal in large samples. The reason is the same as for every other least squares estimator we have considered: The TSLS estimator is an average of random variables, and when the sample size is large, the central limit theorem tells us that averages of random variables are normally distributed. Specifically, the numerator of the expression for $\hat{\beta}_1^{TSLS}$ in Equation (12.4) is $s_{ZY} = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})$, an average of $(Z_i - \overline{Z})(Y_i - \overline{Y})$. A bit of algebra, sketched out in Appendix 12.3, shows that because of this averaging, the central limit theorem implies that, in large samples, $\hat{\beta}_1^{TSLS}$ has a sampling distribution that is approximately $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$, where

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}. \tag{12.8}$$

Statistical inference using the large-sample distribution. The variance $\sigma^2_{\hat{\beta}_1^{TSLS}}$ can be estimated by estimating the variance and covariance terms appearing in Equation (12.8), and the square root of the estimate of $\sigma^2_{\hat{\beta}_1^{TSLS}}$ is the standard error of the IV estimator. This is done automatically in TSLS regression commands in econometric software packages. Because $\hat{\beta}_1^{TSLS}$ is normally distributed in large samples, hypothesis tests about β_1 can be performed by computing the t-statistic, and a 95% large-sample confidence interval is given by $\hat{\beta}_1^{TSLS} \pm 1.96\, SE(\hat{\beta}_1^{TSLS})$.

Application to the Demand for Cigarettes

Philip Wright was interested in the demand elasticity of butter, but Wright's thinking could be explored with a view to estimating other important quantities. One example is the spending elasticity for mortality, the percentage change in avoidable mortality resulting from a 1% increase in healthcare expenditure, where researchers have also used an IV estimator to overcome simultaneous equation bias to inform health policy debates. Other examples concern other commodities, besides butter, such as cigarettes, which today figure more prominently in public policy debates.

The answer to this question depends on the elasticity of demand for cigarettes. If the elasticity is -1, then the 20% target in consumption can be achieved by a 20% increase in price. If the elasticity is -0.5, then the price must rise 40% to decrease consumption by 20%. Of course, we do not know the demand elasticity of cigarettes: We must estimate it from data on prices and sales. But, as with butter, because of the interactions between supply and demand, the elasticity of demand for cigarettes cannot be estimated consistently by an OLS regression of log quantity on log price.

We therefore use TSLS to estimate the elasticity of demand for cigarettes using annual data for the 48 contiguous U.S. states for 1985 through 1995 (the data are described in Appendix 12.1). For now, all the results are for the cross section of states in 1995; results using data for earlier years (panel data) are presented in Section 12.4.

The instrumental variable, $SalesTax_i$, is the portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack (in real dollars, deflated by the Consumer Price Index). Cigarette consumption, $Q_i^{cigarettes}$, is the number of packs of cigarettes sold per capita in the state, and the price, $P_i^{cigarettes}$, is the average real price per pack of cigarettes including all taxes.

Before using TSLS, it is essential to ask whether the two conditions for instrument validity hold. We return to this topic in detail in Section 12.3, where we provide some statistical tools that help in this assessment. Even with those statistical tools, judgment plays an important role, so it is useful to think about whether the sales tax on cigarettes plausibly satisfies the two conditions.

First consider instrument relevance. Because a high sales tax increases the after-tax sales price $P_i^{cigarettes}$, the sales tax per pack plausibly satisfies the condition for instrument relevance.

Next consider instrument exogeneity. For the sales tax to be exogenous, it must be uncorrelated with the error in the demand equation; that is, the sales tax must affect the demand for cigarettes only indirectly through the price. This seems plausible: General sales tax rates vary from state to state, but they do so mainly because different states choose different mixes of sales, income, property, and other taxes to finance public undertakings. Those choices about public finance are driven by political considerations, not by factors related to the demand for cigarettes. We discuss the credibility of this assumption more in Section 12.4, but for now we keep it as a working hypothesis.

In modern statistical software, the first stage of TSLS is estimated automatically, so you do not need to run this regression yourself to compute the TSLS estimator. Even so, it is a good idea to look at the first-stage regression. Using data for the 48 states in 1995, it is

$$\widehat{\ln(P_i^{cigarettes})} = \underset{(0.03)}{4.62} + \underset{(0.005)}{0.031}SalesTax_i. \tag{12.9}$$

As expected, higher sales taxes mean higher after-tax prices. The R^2 of this regression is 47%, so the variation in sales tax on cigarettes explains 47% of the variance of cigarette prices across states.

In the second stage of TSLS, $\ln(Q_i^{cigarettes})$ is regressed on $\widehat{\ln(P_i^{cigarettes})}$ using OLS. The resulting estimated regression function is

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08\widehat{\ln(P_i^{cigarettes})}. \tag{12.10}$$

This estimated regression function is written using the regressor in the second stage, the predicted value $\widehat{\ln(P_i^{cigarettes})}$. It is, however, conventional and less cumbersome simply to report the estimated regression function with $\ln(P_i^{cigarettes})$ rather than $\widehat{\ln(P_i^{cigarettes})}$. Reported in this notation, the TSLS estimates and heteroskedasticity-robust standard errors are

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.53)}{9.72} - \underset{(0.32)}{1.08}\ln(P_i^{cigarettes}). \tag{12.11}$$

The TSLS estimate suggests that the demand for cigarettes is surprisingly elastic in light of their addictive nature: An increase in the price of 1% reduces consumption by 1.08%. But, recalling our discussion of instrument exogeneity, perhaps this estimate should not yet be taken too seriously. Even though the elasticity was estimated using an instrumental variable, there might still be omitted variables that are correlated with the sales tax per pack. A leading candidate is income: States with higher incomes might depend relatively less on a sales tax and more on an income tax to finance state government. Moreover, the demand for cigarettes presumably depends on income. Thus we would like to reestimate our demand equation including income as a control variable. To do so, however, we must first extend the IV regression model to include additional regressors.

12.2 The General IV Regression Model

The general IV regression model has four types of variables: the dependent variable, Y; problematic endogenous regressors, like the price of cigarettes, which are correlated with the error term and which we will label X; additional regressors W, which are either control variables or **included exogenous variables**; and instrumental variables, Z. In general, there can be multiple endogenous regressors (X's), multiple additional regressors (W's), and multiple instrumental variables (Z's).

For IV regression to be possible, there must be at least as many instrumental variables (Z's) as endogenous regressors (X's). In Section 12.1, there was a single endogenous regressor and a single instrument. Having (at least) one instrument for this single endogenous regressor was essential. Without the instrument, we could not have computed the instrumental variables estimator: there would be no first-stage regression in TSLS.

The relationship between the number of instruments and the number of endogenous regressors has its own terminology. The regression coefficients are said to be

KEY CONCEPT 12.1

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad (12.12)$$

$i = 1, \ldots, n$, where

- Y_i is the dependent variable;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown coefficients;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i or are control variables;
- u_i is the error term, which represents measurement error and/or omitted factors; and
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors $(m > k)$, they are underidentified if $m < k$, and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

exactly identified if the number of instruments (m) equals the number of endogenous regressors (k); that is, $m = k$. The coefficients are **overidentified** if the number of instruments exceeds the number of endogenous regressors; that is, $m > k$. They are **underidentified** if the number of instruments is less than the number of endogenous regressors; that is, $m < k$. The coefficients must be either exactly identified or overidentified if they are to be estimated by IV regression.

The general IV regression model and its terminology are summarized in Key Concept 12.1.

Included exogenous variables and control variables in IV regression. The W variables in Equation (12.12) can be either exogenous variables, in which case $E(u_i|W_i) = 0$, or they can be control variables that need not have a causal interpretation but are included to ensure that the instrument is uncorrelated with the error term. For example, Section 12.1 raised the possibility that the sales tax might be correlated with income, which economic theory tells us is a determinant of cigarette demand. If so, the sales tax would be correlated with the error term in the cigarette demand equation, $\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$, and thus would not be an exogenous instrument. Including income in the IV regression, or including variables that control for income, would remove this source of potential correlation between the instrument and the error term. In general, if W is an effective control variable in IV regression, then including W makes the instrument uncorrelated with u, so the TSLS estimator of the coefficient on X is consistent; if W is correlated with u, however, then the TSLS coefficient on W is subject to omitted variable bias and does not have a causal interpretation. The logic of control variables in IV regression therefore parallels the logic of control variables in OLS, discussed in Section 7.5.

The mathematical condition for W to be an effective control variable in IV regression is similar to the condition on control variables in OLS discussed in Section 7.5. Specifically, including W must ensure that the conditional mean of u does not depend on Z, so conditional mean independence holds; that is, $E(u_i|Z_i, W_i) = E(u_i|W_i)$. For clarity, in the body of this chapter we focus on the case that W variables are exogenous, so that $E(u_i|W_i) = 0$. Appendix 12.6 explains how the results of this chapter extend to the case that W is a control variable, in which case the conditional mean 0 condition, $E(u_i|W_i) = 0$, is replaced by the conditional mean independence condition, $E(u_i|Z_i, W_i) = E(u_i|W_i)$.

TSLS in the General IV Model

TSLS with a single endogenous regressor. When there is a single endogenous regressor X and some additional included exogenous variables, the equation of interest is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \tag{12.13}$$

where, as before, X_i might be correlated with the error term, but $W_{1i}, \ldots, W_{ri}$ are not.

The population first-stage regression of TSLS relates X to the exogenous variables—that is, the W's and the instruments (Z's):

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i, \tag{12.14}$$

where $\pi_0, \pi_1, \ldots, \pi_{m+r}$ are unknown regression coefficients and v_i is an error term.

Equation (12.14) is sometimes called the **reduced form** equation for X. It relates the endogenous variable X to all the available exogenous variables, both those included in the regression of interest (W) and the instruments (Z).

In the first stage of TSLS, the unknown coefficients in Equation (12.14) are estimated by OLS, and the predicted values from this regression are $\hat{X}_1, \ldots, \hat{X}_n$.

In the second stage of TSLS, Equation (12.13) is estimated by OLS except that X_i is replaced by its predicted value from the first stage. That is, Y_i is regressed on $\hat{X}_i, W_{1i}, \ldots, W_{ri}$ using OLS. The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{1+r}$ is the TSLS estimator.

KEY CONCEPT 12.2

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (12.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s)**: Regress X_{1i} on the instrumental variables $(Z_{1i}, \ldots, Z_{mi})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. Compute the predicted values from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.
2. **Second-stage regression**: Regress Y_i on the predicted values of the endogenous variables $(\hat{X}_{1i}, \ldots, \hat{X}_{ki})$ and the included exogenous variables and/or control variables $(W_{1i}, \ldots, W_{ri})$ using OLS, including an intercept. The TSLS estimators $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in econometric software.

Extension to multiple endogenous regressors. When there are multiple endogenous regressors $X_{1i}, \ldots, X_{ki}$, the TSLS algorithm is similar except that each endogenous regressor requires its own first-stage regression. Each of these first-stage regressions has the same form as Equation (12.14); that is, the dependent variable is one of the X's, and the regressors are all the instruments (Z's) and all the included exogenous variables (W's). Together, these first-stage regressions produce predicted values of each of the endogenous regressors.

In the second stage of TSLS, Equation (12.12) is estimated by OLS except that the endogenous regressors (X's) are replaced by their respective predicted values ($\hat{X}$'s). The resulting estimator of $\beta_0, \beta_1, \ldots, \beta_{k+r}$ is the TSLS estimator.

In practice, the two stages of TSLS are done automatically within TSLS estimation commands in econometric software. The general TSLS estimator is summarized in Key Concept 12.2.

Instrument Relevance and Exogeneity in the General IV Model

The conditions of instrument relevance and exogeneity need to be modified for the general IV regression model.

When there is one included endogenous variable but multiple instruments, the condition for instrument relevance is that at least one Z is useful for predicting X given W. When there are multiple included endogenous variables, this condition is more complicated because we must rule out perfect multicollinearity in the second-stage population regression. Intuitively, when there are multiple included

The Two Conditions for Valid Instruments

KEY CONCEPT 12.3

A set of m instruments $Z_{1i}, \ldots, Z_{mi}$ must satisfy the following two conditions to be valid:

1. **Instrument Relevance**

   - *In general*, let $\hat{X}^*_{1i}$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and the included exogenous regressors (W's), and let "1" denote the constant regressor that takes on the value 1 for all observations. Then $(\hat{X}^*_{1i}, \ldots, \hat{X}^*_{ki}, W_{1i}, \ldots, W_{ri}, 1)$ are not perfectly multicollinear.
   - *If there is only one* X, then for the previous condition to hold, at least one Z must have a nonzero coefficient in the population regression of X on the Z's and the W's.

2. **Instrument Exogeneity**

   The instruments are uncorrelated with the error term; that is, $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$.

endogenous variables, the instruments must provide enough information about the exogenous movements in these variables to sort out their separate effects on Y.

The general statement of the instrument exogeneity condition is that each instrument must be uncorrelated with the error term u_i. The general conditions for valid instruments are given in Key Concept 12.3.

The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator

Under the IV regression assumptions, the TSLS estimator is consistent and has a sampling distribution that, in large samples, is approximately normal.

The IV regression assumptions. The IV regression assumptions are modifications of the least squares assumptions for causal inference in the multiple regression model in Key Concept 6.4.

The first IV regression assumption modifies the conditional mean assumption in Key Concept 6.4 to apply only to the included exogenous variables. Just like the second least squares assumption for the multiple regression model, the second IV regression assumption is that the draws are i.i.d., as they are if the data are collected by simple random sampling. Similarly, the third IV assumption is that large outliers are unlikely.

KEY CONCEPT 12.4

The IV Regression Assumptions

The variables and errors in the IV regression model in Key Concept 12.1 satisfy the following:

1. $E(u_i \mid W_{1i}, \ldots, W_{ri}) = 0$;
2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;
3. Large outliers are unlikely: The X's, W's, Z's, and Y have nonzero finite fourth moments; and
4. The two conditions for a valid instrument in Key Concept 12.3 hold.

The fourth IV regression assumption is that the two conditions for instrument validity in Key Concept 12.3 hold. The instrument relevance condition in Key Concept 12.3 subsumes the fourth least squares assumption in Key Concepts 6.4 and 6.6 (no perfect multicollinearity) by assuming that the regressors in the second-stage regression are not perfectly multicollinear. The IV regression assumptions are summarized in Key Concept 12.4.

Sampling distribution of the TSLS estimator. Under the IV regression assumptions, the TSLS estimator is consistent and normally distributed in large samples. This is shown in Section 12.1 (and Appendix 12.3) for the special case of a single endogenous regressor, a single instrument, and no included exogenous variables. Conceptually, the reasoning in Section 12.1 carries over to the general case of multiple instruments and multiple included endogenous variables. The expressions in the general case are complicated, however, and are deferred to Chapter 19.

Inference Using the TSLS Estimator

Because the sampling distribution of the TSLS estimator is normal in large samples, the general procedures for statistical inference (hypothesis tests and confidence intervals) in regression models extend to TSLS regression. For example, 95% confidence intervals are constructed as the TSLS estimator ± 1.96 standard errors. Similarly, joint hypotheses about the population values of the coefficients can be tested using the F-statistic, as described in Section 7.2.

Calculation of TSLS standard errors. There are two points to bear in mind about TSLS standard errors. First, the standard errors reported by OLS estimation of the second-stage regression are incorrect because they do not recognize that it is the second stage of a two-stage process. Specifically, the second-stage OLS standard errors fail to adjust for the second-stage regression using the predicted values of the included endogenous variables. Formulas for standard errors that make the necessary adjustment are incorporated into (and automatically used by) TSLS regression commands in econometric software. Therefore, this issue is not a concern in practice if you use a specialized TSLS regression command.

Second, as always the error u might be heteroskedastic. It is therefore important to use heteroskedasticity-robust versions of the standard errors for precisely the same reason that it is important to use heteroskedasticity-robust standard errors for the OLS estimators of the multiple regression model.

Application to the Demand for Cigarettes

In Section 12.1, we estimated the elasticity of demand for cigarettes using data on annual consumption in 48 U.S. states in 1995 using TSLS with a single regressor (the logarithm of the real price per pack) and a single instrument (the real sales tax per pack). Income also affects demand, however, so it is part of the error term of the population regression. As discussed in Section 12.1, if the state sales tax is related to state income, it is correlated with a variable in the error term of the cigarette demand equation, which violates the instrument exogeneity condition. If so, the IV estimator in Section 12.1 is inconsistent. That is, the IV regression suffers from a version of omitted variable bias. We can solve this problem by including income in the regression.

We therefore consider an alternative specification in which the logarithm of income is included in the demand equation. In the terminology of Key Concept 12.1, the dependent variable Y is the logarithm of consumption, $\ln(Q_i^{cigarettes})$; the endogenous regressor X is the logarithm of the real after-tax price, $\ln(P_i^{cigarettes})$; the included exogenous variable W is the logarithm of the real per capita state income, $\ln(Inc_i)$; and the instrument Z is the real sales tax per pack, $SalesTax_i$. The TSLS estimates and (heteroskedasticity-robust) standard errors are

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\ln(P_i^{cigarettes}) + \underset{(0.31)}{0.21}\ln(Inc_i). \tag{12.15}$$

This regression uses a single instrument, $SalesTax_i$, but, in fact, another candidate instrument is available. In addition to general sales taxes, states levy special taxes that apply only to cigarettes and other tobacco products. These cigarette-specific taxes ($CigTax_i$) constitute a possible second instrumental variable. The cigarette-specific tax increases the price of cigarettes paid by the consumer, so it arguably meets the condition for instrument relevance. If it is uncorrelated with the error term in the state cigarette demand equation, it is an exogenous instrument.

With this additional instrument in hand, we now have two instrumental variables, the real sales tax per pack and the real state cigarette-specific tax per pack. With two instruments and a single endogenous regressor, the demand elasticity is overidentified; that is, the number of instruments ($SalesTax_i$ and $CigTax_i$, so $m = 2$) exceeds the number of included endogenous variables ($P_i^{cigarettes}$, so $k = 1$). We can estimate the demand elasticity using TSLS, where the regressors in the first-stage regression are the included exogenous variable, $\ln(Inc_i)$, and both instruments.

The resulting TSLS estimate of the regression function using the two instruments $SalesTax_i$ and $CigTax_i$ is

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\ln(P_i^{cigarettes}) + \underset{(0.25)}{0.28}\ln(Inc_i). \tag{12.16}$$

Compare Equations (12.15) and (12.16): The standard error of the estimated price elasticity is smaller by one-third in Equation (12.16) [0.25 in Equation (12.16) versus 0.37 in Equation (12.15)]. The reason the standard error is smaller in Equation (12.16) is that this estimate uses more information than Equation (12.15): In Equation (12.15), only one instrument (the sales tax) is used, but in Equation (12.16), two instruments (the sales tax and the cigarette-specific tax) are used. Using two instruments explains more of the variation in cigarette prices than using just one, and this is reflected in smaller standard errors on the estimated demand elasticity.

Are these estimates credible? Ultimately, credibility depends on whether the set of instrumental variables—here, the two taxes—plausibly satisfies the two conditions for valid instruments. It is therefore vital that we assess whether these instruments are valid, and it is to this topic that we now turn.

12.3 Checking Instrument Validity

Whether instrumental variables regression is useful in a given application hinges on whether the instruments are valid: Invalid instruments produce meaningless results. It therefore is essential to assess whether a given set of instruments is valid in a particular application.

Assumption 1: Instrument Relevance

The role of the instrument relevance condition in IV regression is subtle. One way to think of instrument relevance is that it plays a role akin to the sample size: The more relevant are the instruments—that is, the more the variation in X is explained by the instruments—the more information is available for use in IV regression. A more relevant instrument produces a more accurate estimator, just as a larger sample size produces a more accurate estimator. Moreover, statistical inference using TSLS is predicated on the TSLS estimator having a normal sampling distribution, but according to the central limit theorem, the normal distribution is a good approximation in large—but not necessarily small—samples. If having a more relevant instrument is like having a larger sample size, this suggests, correctly, that the more relevant is the instrument, the better is the normal approximation to the sampling distribution of the TSLS estimator and its t-statistic.

Instruments that explain little of the variation in X are called **weak instruments**. In the cigarette example, the distance of the state from cigarette manufacturing plants arguably would be a weak instrument: Although a greater distance increases shipping costs (thus shifting the supply curve in and raising the equilibrium price), cigarettes are lightweight, so shipping costs are a small component of the price of cigarettes. Thus the amount of price variation explained by shipping costs, and thus distance to manufacturing plants, probably is quite small.

This section discusses why weak instruments are a problem, how to check for weak instruments, and what to do if you have weak instruments. It is assumed throughout that the instruments are exogenous.

Why weak instruments are a problem. If the instruments are weak, then the normal distribution provides a poor approximation to the sampling distribution of the TSLS estimator, even if the sample size is large. Thus there is no theoretical justification for the usual methods for performing statistical inference, even in large samples. In fact, if instruments are weak, then the TSLS estimator can be badly biased in the direction of the OLS estimator. In addition, 95% confidence intervals constructed as the TSLS estimator ± 1.96 standard errors can contain the true value of the coefficient far less than 95% of the time. In short, if instruments are weak, TSLS is no longer reliable.

To see that there is a problem with the large-sample normal approximation to the sampling distribution of the TSLS estimator, consider the special case, introduced in Section 12.1, of a single included endogenous variable, a single instrument, and no included exogenous regressor. If the instrument is valid, then $\hat{\beta}_1^{TSLS}$ is consistent because the sample covariances s_{ZY} and s_{ZX} are consistent; that is, $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX} \xrightarrow{p} \mathrm{cov}(Z_i, Y_i)/\mathrm{cov}(Z_i, X_i) = \beta_1$ [Equation (12.7)]. But now suppose that the instrument is not just weak but in fact is irrelevant, so that $\mathrm{cov}(Z_i, X_i) = 0$. Then $s_{ZX} \xrightarrow{p} \mathrm{cov}(Z_i, X_i) = 0$, so, taken literally, the denominator on the right-hand side of the limit $\mathrm{cov}(Z_i, Y_i)/\mathrm{cov}(Z_i, X_i)$ is 0! Clearly, the argument that $\hat{\beta}_1^{TSLS}$ is consistent breaks down when the instrument relevance condition fails. As shown in Appendix 12.4, this breakdown results in the TSLS estimator having a nonnormal sampling distribution, even if the sample size is very large. In fact, when the instrument is irrelevant, the large-sample distribution of $\hat{\beta}_1^{TSLS}$ is not the distribution of a normal random variable but rather the distribution of a *ratio* of two normal random variables! As discussed in Appendix 12.4, this ratio-of-normals distribution is centered at the large-sample value of the OLS estimator.

While this circumstance of totally irrelevant instruments might not be encountered in practice, it raises a question: How relevant must the instruments be for the normal distribution to provide a good approximation in practice? The answer to this question in the general IV model is complicated. Fortunately, however, there is a simple rule of thumb available for the most common situation in practice, the case of a single endogenous regressor.

KEY CONCEPT 12.5

A Rule of Thumb for Checking for Weak Instruments

The first-stage F-statistic is the F-statistic testing the hypothesis that the coefficients on the instruments $Z_{1i}, \ldots, Z_{mi}$ equal 0 in the first stage of two stage least squares. When there is a single endogenous regressor, a first-stage F-statistic less than 10 indicates that the instruments are weak, in which case the TSLS estimator is biased (even in large samples) and TSLS t-statistics and confidence intervals are unreliable.

Checking for weak instruments when there is a single endogenous regressor. One way to check for weak instruments when there is a single endogenous regressor is to compute the F-statistic testing the hypothesis that the coefficients on the instruments are all 0 in the first-stage regression of TSLS. This **first-stage *F*-statistic** provides a measure of the information content contained in the instruments: The more information content, the larger the expected value of the F-statistic. One simple rule of thumb is that you do not need to worry about weak instruments if the first-stage F-statistic exceeds 10. (Why 10? See Appendix 12.5.) This is summarized in Key Concept 12.5.

What do I do if I have weak instruments? If you have many instruments, some of those instruments are probably weaker than others. If you have a small number of strong instruments and many weak ones, you will be better off discarding the weakest instruments and using the most relevant subset for your TSLS analysis. Your TSLS standard errors might increase when you drop weak instruments, but keep in mind that your original standard errors were not meaningful anyway!

If, however, the coefficients are exactly identified, you cannot discard the weak instruments. Even if the coefficients are overidentified, you might not have enough strong instruments to achieve identification, so discarding some weak instruments will not help. In this case, you have two options. The first option is to find additional, stronger instruments. This is easier said than done: It requires an intimate knowledge of the problem at hand and can entail redesigning the data set and the nature of the empirical study. The second option is to proceed with your empirical analysis using the weak instruments, but employing methods other than TSLS. Although this chapter has focused on TSLS, some other methods for instrumental variable analysis are less sensitive to weak instruments than TSLS, and some of these methods are discussed in Appendix 12.5.

Assumption 2: Instrument Exogeneity

If the instruments are not exogenous, then TSLS is inconsistent: The TSLS estimator converges in probability to something other than the causal coefficient. After all, the idea of instrumental variables regression is that the instrument contains information

The First IV Regression

After he and his son Sewall derived the IV estimator (see the box "Who Invented Instrumental Variables Regression?"), Philip Wright set out to see how it worked in practice. In a letter to Sewall of March 15, 1926, Philip wrote out a table (reproduced here in part) of annual data on variables relating to U.S. production of flaxseed from 1903 through 1925. Flaxseed was grown for its oil, also called linseed oil, which was used in oil-based paint for buildings. Philip wanted to estimate the elasticity of supply. To get a percent–percent relationship, he first transformed the data to be percentage deviations from a long-term trend.

Philip then needed to make a key decision: What instrument should he use? He chose building permits on the East Coast. He reasoned that if there were more new buildings, there would be more demand for oil-based paint and thus for flaxseed, so the instrument would be relevant. He further reasoned that fluctuations in building permits on the East Coast were largely driven by broader economic conditions that had nothing to do with disturbances to flaxseed supply in a given year, so that building permits would be exogenous. Said differently, fluctuations in building permits on the East Coast were a determinant of demand but not of supply.

After laborious computations—by hand, of course—Philip obtained the IV estimate of the supply elasticity, −0.88. This elasticity has the wrong sign: It suggests that the supply curve slopes *down*. In the March 15 letter, Philip called this result "obviously absurd."

So what went wrong? Although Philip did not know it, his IV regression had a first-stage *F*-statistic of 1.75, far less than the rule-of-thumb cutoff of 10. As explained in the text and in Appendix 12.4, when the instrument is irrelevant, its distribution centers on the OLS estimate, which in Wright's data is −0.66. This first IV regression had a very weak instrument, and the result was biased toward OLS.

But Philip persevered. For estimating the demand elasticity, he had as an instrument rainfall in the Upper Midwest, where flaxseed was grown. More rain makes for a better harvest, so rainfall is plausibly relevant; because rainfall in the Midwest does not affect the demand for oil paint, it is plausibly exogenous. Rainfall, it turns out, has a first-stage *F* of 12.8 and yields an IV estimate of the demand elasticity of −0.48. This estimate indicates that the demand curve slopes down (as it should) and that demand is inelastic, which is consistent with there being no good substitute for linseed oil for paints during this period.

The First Five Observations of the First IV Regression Data Set, from Philip Wright's Letter to Sewall Wright of March 15, 1926.

			B		B	B	A
Supply	Real prices [Money price ÷ index]	Output (million bu.)	Acreage (millions)	Yield	Rainfall (inches)	Ratio [illegible]	Building permits [illegible]
1903	.126	27.3	3.23	8.4	3.40	93	128
4	.153	23.4	2.26	10.3	2.19	75	140
5	.123	28.5	2.53	11.2	4.27	95	186
6	.126	25.6	2.51	10.2	3.30	93	18
7	.133	25.9	2.86	9.0	2.66	119	187

The first two data columns are the real price and quantity ("output") of flaxseed. The "B" variables— acreage planted, yield, rainfall in the Upper Midwest, and the ratio of flaxseed yield that year to spring wheat yield the previous year—shift supply but not demand, so they are potential instruments for the demand elasticity. The "A" variable—building permits on the East Coast—shifts demand but not supply, so it is a potential instrument for the supply elasticity.

about variation in X_i that is unrelated to the error term u_i. If, in fact, the instrument is not exogenous, it cannot pinpoint this exogenous variation in X_i, and it stands to reason that IV regression fails to provide a consistent estimator. The math behind this argument is summarized in Appendix 12.4.

Can you statistically test the assumption that the instruments are exogenous? Yes and no. On the one hand, it is not possible to test the hypothesis that the instruments are exogenous when the coefficients are exactly identified. On the other hand, if the coefficients are overidentified, it is possible to test the overidentifying restrictions—that is, to test the hypothesis that the "extra" instruments are exogenous under the maintained assumption that there are enough valid instruments to identify the coefficients of interest.

First consider the case that the coefficients are exactly identified, so you have as many instruments as endogenous regressors. Then it is impossible to develop a statistical test of the hypothesis that the instruments are, in fact, exogenous. That is, empirical evidence cannot be brought to bear on the question of whether these instruments satisfy the exogeneity restriction. In this case, the only way to assess whether the instruments are exogenous is to draw on expert opinion and your personal knowledge of the empirical problem at hand. For example, Philip Wright's knowledge of agricultural supply and demand led him to suggest that below-average rainfall would plausibly shift the supply curve for fats and oils but would not directly shift the demand curve.

Assessing whether the instruments are exogenous *necessarily* requires making an expert judgment based on personal knowledge of the application. If, however, there are more instruments than endogenous regressors, then there is a statistical tool that can be helpful in this process: the so-called test of overidentifying restrictions.

The overidentifying restrictions test. Suppose you have a single endogenous regressor and two instruments. Then you could compute two different TSLS estimators: one using the first instrument and the other using the second. These two estimators will not be the same because of sampling variation, but if both instruments are exogenous, then they will tend to be close to each other. But what if these two instruments produce very different estimates? You might sensibly conclude that there is something wrong with one or the other of the instruments or with both. That is, it would be reasonable to conclude that one or the other or both of the instruments are not exogenous.

The **test of overidentifying restrictions** implicitly makes this comparison. We say implicitly because the test is carried out without actually computing all of the different possible IV estimates. Here is the idea. Exogeneity of the instruments means that they are uncorrelated with u_i. This suggests that the instruments should be approximately uncorrelated with $\hat{u}_i^{TSLS}$, where $\hat{u}_i^{TSLS} = Y_i - (\hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X_{1i} + \cdots + \hat{\beta}_{k+r}^{TSLS} W_{ri})$

The Overidentifying Restrictions Test (The *J*-Statistic)

KEY CONCEPT 12.6

Let $\hat{u}_i^{TSLS}$ be the residuals from TSLS estimation of Equation (12.12). Use OLS to estimate the regression coefficients in

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i, \quad (12.17)$$

where e_i is the regression error term. Let F denote the homoskedasticity-only F-statistic testing the hypothesis that $\delta_1 = \cdots = \delta_m = 0$. The overidentifying restrictions test statistic is $J = mF$. Under the null hypothesis that all the instruments are exogenous, if e_i is homoskedastic, in large samples J is distributed χ^2_{m-k}, where $m - k$ is the *degree of overidentification*—that is, the number of instruments minus the number of endogenous regressors.

is the residual from the estimated TSLS regression using all the instruments (approximately rather than exactly because of sampling variation). (Note that these residuals are constructed using the true X's rather than their first-stage predicted values.) Accordingly, if the instruments are, in fact, exogenous, then the coefficients on the instruments in a regression of $\hat{u}_i^{TSLS}$ on the instruments and the included exogenous variables should all be 0, and this hypothesis can be tested.

This method for computing the overidentifying restrictions test is summarized in Key Concept 12.6. This statistic is computed using the homoskedasticity-only F-statistic. The test statistic is commonly called the J-statistic and is computed as $J = mF$.

In large samples, if the instruments are not weak and the errors are homoskedastic, then, under the null hypothesis that the instruments are exogenous, the J-statistic has a chi-squared distribution with $m - k$ degrees of freedom (χ^2_{m-k}). It is important to remember that even though the number of restrictions being tested is m, the degrees of freedom of the asymptotic distribution of the J-statistic is $m - k$. The reason is that it is possible to test only the *over*identifying restrictions, of which there are $m - k$. The modification of the J-statistic for heteroskedastic errors is given in Section 19.7.

The easiest way to see that you cannot test the exogeneity of the regressors when the coefficients are exactly identified $(m = k)$ is to consider the case of a single included endogenous variable $(k = 1)$. If there are two instruments, then you can compute two TSLS estimators, one for each instrument, and you can compare them to see if they are close. But if you have only one instrument, then you can compute only one TSLS estimator, and you have nothing to which to compare it. In fact, if the coefficients are exactly identified, so that $m = k$, then the overidentifying test statistic J is exactly 0.

12.4 Application to the Demand for Cigarettes[1]

Our attempt to estimate the elasticity of demand for cigarettes left off with the TSLS estimates summarized in Equation (12.16), in which income was an included exogenous variable and there were two instruments, the general sales tax and the cigarette-specific tax. We can now undertake a more thorough evaluation of these instruments.

As in Section 12.1, it makes sense that the two instruments are relevant because taxes are a big part of the after-tax price of cigarettes, and shortly we will look at this empirically. First, however, we focus on the difficult question of whether the two tax variables are plausibly exogenous.

The first step in assessing whether an instrument is exogenous is to think through the arguments for why it may or may not be. This requires thinking about which factors account for the error term in the cigarette demand equation and whether these factors are plausibly related to the instruments.

Why do some states have higher per capita cigarette consumption than others? One reason might be variation in incomes across states, but state income is included in Equation (12.16), so this is not part of the error term. Another reason is that there are historical factors influencing demand. For example, states that grow tobacco have higher rates of smoking than most other states. Could this factor be related to taxes? Quite possibly: If tobacco farming and cigarette production are important industries in a state, then these industries could exert influence to keep cigarette-specific taxes low. This suggests that an omitted factor in cigarette demand—whether the state grows tobacco and produces cigarettes—could be correlated with cigarette-specific taxes.

One solution to this possible correlation between the error term and the instrument would be to include information on the size of the tobacco and cigarette industry in the state; this is the approach we took when we included income as a regressor in the demand equation. But because we have panel data on cigarette consumption, a different approach is available that does not require this information. As discussed in Chapter 10, panel data make it possible to eliminate the influence of variables that vary across entities (states) but do not change over time, such as the historical circumstances that lead to a large tobacco and cigarette industry in a state. Two methods for doing this were given in Chapter 10: constructing data on *changes* in the variables between two different time periods and using fixed effects regression. To keep the analysis here as simple as possible, we adopt the former approach and perform regressions of the type described in Section 10.2, based on the changes in the variables between two different years.

The time span between the two different years influences how the estimated elasticities are to be interpreted. Because cigarettes are addictive, changes in price will take some time to alter behavior. At first, an increase in the price of cigarettes might have little effect on demand. Over time, however, the price increase might contribute

[1]This section assumes knowledge of the material in Sections 10.1 and 10.2 on panel data with $T = 2$ time periods.

The Externalities of Smoking

It is often said that smoking creates negative externalities or costs, such as those of healthcare and cleaning, which are imposed on third parties by the act of smoking. Outright bans on smoking in various locations—at the workplace or in public locations—have been suggested or imposed in Western Europe in recent years, such as in France and England in 2007, and in the Netherlands in 2008. Economists, however, often object to this, and suggest imposing taxes to correct for these.

It is usually suggested that these taxes should be imposed at such a level that the external cost is reduced to zero, by its burden being shifted onto the smoker in this way. We could use econometric techniques to estimate this external cost, and subsequently the required tax.

Such estimation is no simple matter, however. The U.K. Government estimates that the smoking-related cost to the National Health Service (NHS) in 2015 was £2.6 billion, but this does not adjust for costs that would have been imposed anyway.[1] How much would it have cost to treat these people for other illnesses had they not smoked? Are there potentially other benefits and costs that this misses? If smokers die young, how do we value the foregone benefit of their lost life years? What about the value of the employment that smoking generates?

One recent academic review of available evidence points to various different such potential costs and benefits, but ultimately concludes that the external costs of smoking "far outweigh any benefits."[2] This suggests that if tax is the lever we wish to use to change smoking behavior, taxes on tobacco should rise. We must recognize, however, that the exact value of these benefits and costs is dependent on what we actually consider to be benefits and costs and can only be estimated. While econometricians can advise on policy questions such as these, they will still remain questions of political contention.

[1] The data on the cost of smoking to the NHS in England in 2015 is an ad hoc statistical publication from July 2017. The analysis was undertaken by Public Health England (PHE) to support the development of the new Tobacco Control Plan for England. For more information, see "Cost of smoking to the NHS in England: 2015," on https://www.gov.uk/

[2] Read the article "The Economic Impact of Smoking and of Reducing Smoking Prevalence: Review of Evidence," by Victor U. Ekpu and Abraham K. Brown, published by the U.S. National Library of Medicine National Institutes of Health, https://www.ncbi.nlm.nih.gov, July 14, 2015.

to some smokers' desire to quit, and, importantly, it could discourage nonsmokers from taking up the habit. Thus the response of demand to a price increase could be small in the short run but large in the long run. Said differently, for an addictive product like cigarettes, demand might be inelastic in the short run—that is, it might have a short-run elasticity near 0—but it might be more elastic in the long run.

In this analysis, we focus on estimating the long-run price elasticity. We do this by considering quantity and price changes that occur over 10-year periods. Specifically, in the regressions considered here, the 10-year change in log quantity, $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$, is regressed against the 10-year change in log price, $\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$, and the 10-year change in log income, $\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$. Two instruments are used: the change in the sales tax over 10 years, $SalesTax_{i,1995} - SalesTax_{i,1985}$, and the change in the cigarette-specific tax over 10 years, $CigTax_{i,1995} - CigTax_{i,1985}$.

TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q_{i,1995}^{cigarettes}) - \ln(Q_{i,1985}^{cigarettes})$

Regressor	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	−0.94 (0.21) [−1.36, −0.52]	−1.34 (0.23) [−1.80, −0.88]	−1.20 (0.20) [−1.60, −0.81]
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34) [−0.16, 1.21]	0.43 (0.30) [−0.16, 1.02]	0.46 (0.31) [−0.16, 1.09]
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage F-statistic	33.7	107.2	88.6
Overidentifying restrictions J-test and p-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The J-test of overidentifying restrictions is described in Key Concept 12.6 (its p-value is given in parentheses), and the first-stage F-statistic is described in Key Concept 12.5. Heteroskedasticity-robust standard errors are given in parentheses beneath coefficients, and 95% confidence intervals are given in brackets.

The results are presented in Table 12.1. As usual, each column in the table presents the results of a different regression. All regressions have the same regressors, and all coefficients are estimated using TSLS; the only difference among the three regressions is the set of instruments used. In column (1), the only instrument is the sales tax; in column (2), the only instrument is the cigarette-specific tax; and in column (3), both taxes are used as instruments.

In IV regression, the reliability of the coefficient estimates hinges on the validity of the instruments, so the first things to look at in Table 12.1 are the diagnostic statistics assessing the validity of the instruments.

First, are the instruments relevant? We need to look at the first-stage F-statistics. The first-stage regression in column (1) is

$$\begin{aligned}\widehat{\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})} &= \underset{(0.03)}{0.53} - \underset{(0.22)}{0.22}[\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})] \\ &\quad + \underset{(0.0044)}{0.0255}(SalesTax_{i,1995} - SalesTax_{i,1985}). \quad (12.18)\end{aligned}$$

Because there is only one instrument in this regression, the first-stage F-statistic is the square of the t-statistic testing that the coefficient on the instrumental variable, $SalesTax_{i,1995} - SalesTax_{i,1985}$, is 0; this is $F = t^2 = (0.0255/0.0044)^2 = 33.7$. For the regressions in columns (2) and (3), the first-stage F-statistics are 107.2 and 88.6, so in all three cases the first-stage F-statistics exceed 10. We conclude that the instruments are not weak, so we can rely on the standard methods for statistical inference (hypothesis tests and confidence intervals) using the TSLS coefficients and standard errors.

Second, are the instruments exogenous? Because the regressions in columns (1) and (2) each have a single instrument and a single included endogenous regressor, the coefficients in those regressions are exactly identified. Thus we cannot deploy the J-test in either of those regressions. The regression in column (3), however, is overidentified because there are two instruments and a single included endogenous regressor, so there is one $(m - k = 2 - 1 = 1)$ overidentifying restriction. The J-statistic is 4.93; this has a χ^2_1 distribution, so the 5% critical value is 3.84 (Appendix Table 3) and the null hypothesis that both the instruments are exogenous is rejected at the 5% significance level (this deduction also can be made directly from the p-value of 0.026, reported in the table).

The reason the J-statistic rejects the null hypothesis that both instruments are exogenous is that the two instruments produce rather different estimated coefficients. When the only instrument is the sales tax [column (1)], the estimated price elasticity is -0.94, but when the only instrument is the cigarette-specific tax, the estimated price elasticity is -1.34. Recall the basic idea of the J-statistic: If both instruments are exogenous, then the two TSLS estimators using the individual instruments are consistent and differ from each other only because of random sampling variation. If, however, one of the instruments is exogenous and one is not, then the estimator based on the endogenous instrument is inconsistent, which is detected by the J-statistic. In this application, the difference between the two estimated price elasticities is sufficiently large that it is unlikely to be the result of pure sampling variation, so the J-statistic rejects the null hypothesis that both the instruments are exogenous.

The J-statistic rejection means that the regression in column (3) is based on invalid instruments (the instrument exogeneity condition fails). What does this imply about the estimates in columns (1) and (2)? The J-statistic rejection says that at least one of the instruments is endogenous, so there are three logical possibilities: The sales tax is exogenous but the cigarette-specific tax is not, in which case the column (1) regression is reliable; the cigarette-specific tax is exogenous but the sales tax is not, so the column (2) regression is reliable; or neither tax is exogenous, so neither regression is reliable. The statistical evidence cannot tell us which possibility is correct, so we must use our judgment.

We think that the case for the exogeneity of the general sales tax is stronger than that for the cigarette-specific tax because the political process can link changes in the cigarette-specific tax to changes in the cigarette market and smoking policy. For example, if smoking decreases in a state because it falls out of fashion, there will be fewer smokers and a weakened lobby against cigarette-specific tax increases, which in turn could lead to higher cigarette-specific taxes. Thus changes in tastes (which are part of u) could be correlated with changes in cigarette-specific taxes (the instrument). This suggests discounting the IV estimates that use the cigarette-only tax as an instrument and adopting the price elasticity estimated using the general sales tax as an instrument, -0.94.

The estimate of -0.94 indicates that cigarette consumption is somewhat elastic: An increase in price of 1% leads to a decrease in consumption of 0.94%. This may seem surprising for an addictive product like cigarettes. But remember that this elasticity is computed using changes over a 10-year period, so it is a long-run elasticity. This estimate suggests that increased taxes can make a substantial dent in cigarette consumption, at least in the long run.

When the elasticity is estimated using 5-year changes from 1985 to 1990 rather than the 10-year changes reported in Table 12.1, the elasticity (estimated with the general sales tax as the instrument) is -0.79; for changes from 1990 to 1995, the elasticity is -0.68. These estimates suggest that demand is less elastic over horizons of 5 years than over 10 years. This finding of greater price elasticity at longer horizons is consistent with the large body of research on cigarette demand. Demand elasticity estimates in that literature typically fall in the range -0.3 to -0.5, but these are mainly short-run elasticities; some studies suggest that the long-run elasticity could be perhaps twice the short-run elasticity.[2]

12.5 Where Do Valid Instruments Come From?

In practice, the most difficult aspect of IV estimation is finding instruments that are both relevant and exogenous. There are two main approaches, which reflect two different perspectives on econometric and statistical modeling.

The first approach is to use economic theory to suggest instruments. For example, Philip Wright's understanding of the economics of agricultural markets led him to look for an instrument that shifted the supply curve but not the demand curve; this in turn led him to consider weather conditions in agricultural regions. One area where this approach has been particularly successful is the field of financial economics. Some economic models of investor behavior involve statements about how investors forecast, which then imply sets of variables that are uncorrelated with the error term. Those models sometimes are nonlinear in the data and in the parameters, in which case the IV estimators discussed in this chapter cannot be used. An extension of IV methods to nonlinear models, called generalized method of moments estimation, is used instead. Economic theories are, however, abstractions that often do not take into account the nuances and details necessary for analyzing a particular data set. Thus this approach does not always work.

The second approach to constructing instruments is to look for some exogenous source of variation in X arising from what is, in effect, a random phenomenon that

[2]A sobering economic study by Adda and Cornaglia (2006) suggests that smokers compensate for higher taxes by smoking more intensively, thus extracting more nicotine per cigarette. If you are interested in learning more about the economics of smoking, see Chaloupka and Warner (2000), Gruber (2001), and Carpenter and Cook (2008).

induces shifts in the endogenous regressor. For example, in our hypothetical example in Section 12.1, earthquake damage increased average class size in some school districts, and this variation in class size was unrelated to potential omitted variables that affect student achievement. This approach typically requires knowledge of the problem being studied and careful attention to the details of the data, and it is best explained through examples.

Three Examples

We now turn to three empirical applications of IV regression that illustrate how different researchers used their expert knowledge of their empirical problem to find instrumental variables.

Do economic institutions affect economic development? The single question that has troubled economists since Adam Smith the most is why some nations are rich while others remain poor. Unpicking the various mechanisms that lead to economic growth and evaluating the contribution of each mechanism requires a combination of theory and empirical analysis. However, such empirical analysis is not as straightforward as it seems. For example, the role played by institutions, such as legal institutions that facilitate the ownership of property. It is quite plausible that strong institutions that foster property rights could lead to higher economic growth if they incentivized a more efficient use of scarce resources.

Disentangling this particular issue is challenging, precisely because economic institutions and economic growth are so interconnected. This means that a simple regression of some measure of economic development (GDP per capita) against a measure of institutions, such as protection against expropriation (the strength of the property rights in a country), will yield a biased estimate of the causal effect of institutions on economic development even if the analyst controls for a number of other factors affecting economic development, such as whether a country is landlocked or not. This results from the serious potential for simultaneous causality bias in this analysis: Stronger institutions can lead to greater economic development. Conversely, however, economic growth could enable the creation of these kinds of institutions and institutional arrangements. As a result there is a "chicken and egg" situation where it is not clear which comes first. As in the butter example in Figure 12.1, because of this simultaneous causality, an OLS regression of economic development on a measure of institutions will estimate some complicated combination of these two effects. This problem cannot be solved by finding better control variables.

This simultaneous causality bias, however, can be eliminated by finding a suitable instrumental variable and using TSLS. The instrument must be correlated with the measure of institutions (it must be relevant), but it must also be uncorrelated with the error term in the economic development equation of interest (it must be exogenous). That is it must affect the measure of institutions but be unrelated to any of the unobserved factors that determine economic development.

Things that might affect the ability to have strong economic institutions are very likely to be related to the economic performance of a country. So where does one find something that affects institutions but has no direct effect on economic development? Because it takes a long time for institutions to become established, one idea is to consider the history of how economic institutions were first developed. Plausibly there may be factors from hundreds of years ago that were relevant in the initial founding of institutions, but are not related to the level of economic development today except for through their impact on institutions. Specifically, Acemoglu et al. (2001) consider the colonial origins of economic institutions. They argue that the potential mortality rate among settlers was influential in determining whether European countries established "Neo-Europes" involving setting up European-style institutions that protected private property rights or instead set up "extractive states." They further argue that these differences in institutions persist to the present day.

Are measures of potential settler mortality valid instruments? Although Acemoglu et al. did not report first-stage F-statistics, settler mortality alone was found to explain 27% of the levels of current institutions, suggesting that this instrument is relevant.[3] The argument that the instruments are exogenous requires that settler mortality only affects economic development through the effect on institutions. As a robustness check, to investigate whether settler mortality may have been caused by diseases that still exist and that may hamper economic performance today, Acemoglu et al. include prevalence of malaria in their regression. They find that the inclusion of this regressor makes little difference to the resulting regression coefficients. In addition, because Acemoglu et al. break down the causal pathway through which settler mortality affects current institutions into three parts, there are three instruments and, therefore, overidentifying restrictions can be tested. The failure to reject the null hypotheses of these tests bolsters the case that the instruments are valid.

Using these instruments and TSLS, Acemoglu et al. estimated the effect on economic development of institutions to be substantial. This estimated effect was twice as large as the effect estimated using OLS, suggesting that OLS suffered from large simultaneous causality bias. In addition, they find that in the TSLS model neither the coefficient on the dummy for Africa nor a country's distance from the equator are statistically significant suggesting that "Africa is poorer than the rest of the world not because of pure geographic or cultural factors, but because of worse institutions."[4]

Does cutting class sizes increase test scores? As we saw in the empirical analysis of Part II, schools with small classes tend to be wealthier, and their students have access to enhanced learning opportunities both in and out of the classroom. In Part II, we

---

[3]For further reading see Daron Acemoglu, Simon Johnson, James A. Robinson's *The Colonial Origins of Comparative Development: An Empirical Investigation*, The American Economic Review, Vol. 91, No. 5, December, 2001.

[4]If you are interested in learning more about this empirical analysis and the response to it by other economists, see the original paper Acemoglu et al. (2001), the comment on it by Albouy (2012), and the reply to the comment Acemoglu et al. (2012).

used multiple regression to tackle the threat of omitted variables bias by controlling for various measures of student affluence, ability to speak English, and so forth. Still, a skeptic could wonder whether we did enough: If we left out something important, our estimates of the class size effect would still be biased.

This potential omitted variables bias could be addressed by including the right control variables, but if these data are unavailable (some, like outside learning opportunities, are hard to measure), then an alternative approach is to use IV regression. This regression requires an instrumental variable correlated with class size (relevance) but uncorrelated with the omitted determinants of test performance that make up the error term, such as parental interest in learning, learning opportunities outside the classroom, quality of the teachers and school facilities, and so forth (exogeneity).

Where does one look for an instrument that induces random, exogenous variation in class size, but is unrelated to the other determinants of test performance? Hoxby (2000) suggested biology. Because of random fluctuations in timings of births, the size of the incoming kindergarten class varies from one year to the next. Although the actual number of children entering kindergarten might be endogenous (recent news about the school might influence whether parents send a child to a private school), she argued that the *potential* number of children entering kindergarten—the number of 4-year-olds in the district—is mainly a matter of random fluctuations in the birth dates of children.

Is potential enrollment a valid instrument? Whether it is exogenous depends on whether it is correlated with unobserved determinants of test performance. Surely biological fluctuations in potential enrollment are exogenous, but potential enrollment also fluctuates because parents with young children choose to move into an improving school district and out of one in trouble. If so, an increase in potential enrollment could be correlated with unobserved factors such as the quality of school management, rendering this instrument invalid. Hoxby addressed this problem by reasoning that growth or decline in the potential student pool for this reason would occur smoothly over several years, whereas random fluctuations in birth dates would produce short-term "spikes" in potential enrollment. Thus she used as her instrument not potential enrollment, but the deviation of potential enrollment from its long-term trend. These deviations satisfy the criterion for instrument relevance (the first-stage F-statistics all exceed 100). She makes a good case that this instrument is exogenous, but, as in all IV analysis, the credibility of this assumption is ultimately a matter of judgment.

Hoxby implemented this strategy using detailed panel data on elementary schools in Connecticut in the 1980s and 1990s. The panel data set permitted her to include school fixed effects, which, in addition to the instrumental variables strategy, attack the problem of omitted variables bias at the school level. Her TSLS estimates suggested that the effect on test scores of class size is small; most of her estimates were statistically insignificantly different from 0.

Does aggressive treatment of heart attacks prolong lives? Aggressive treatments for victims of heart attacks (technically, acute myocardial infarctions, or AMIs) hold the potential for saving lives. Before a new medical procedure—in this example, cardiac catheterization[5]—is approved for general use, it goes through clinical trials, a series of randomized controlled experiments designed to measure its effects and side effects. But strong performance in a clinical trial is one thing; actual performance in the real world is another.

A natural starting point for estimating the real-world effect of cardiac catheterization is to compare patients who received the treatment to those who did not. This leads to regressing the length of survival of the patient against the binary treatment variable (whether the patient received cardiac catheterization) and other control variables that affect mortality (age, weight, other measured health conditions, and so forth). The population coefficient on the indicator variable is the increment to the patient's life expectancy provided by the treatment. Unfortunately, the OLS estimator is subject to bias: Cardiac catheterization does not "just happen" to a patient randomly; rather, it is performed because the doctor and patient decide that it might be effective. If their decision is based in part on unobserved factors relevant to health outcomes not in the data set, the treatment decision will be correlated with the regression error term. If the healthiest patients are the ones who receive the treatment, the OLS estimator will be biased (treatment is correlated with an omitted variable), and the treatment will appear more effective than it really is.

This potential bias can be eliminated by IV regression using a valid instrumental variable. The instrument must be correlated with treatment (must be relevant) but must be uncorrelated with the omitted health factors that affect survival (must be exogenous).

Where does one look for something that affects treatment but does not affect the health outcome other than through its effect on treatment? McClellan, McNeil, and Newhouse (1994) suggested geography. Most hospitals in their data set did not offer cardiac catheterization, so many patients were closer to "regular" hospitals that did not offer this treatment than to cardiac catheterization hospitals. McClellan, McNeil, and Newhouse therefore used as an instrumental variable the difference between the distance from the AMI patient's home to the nearest cardiac catheterization hospital and the distance to the nearest hospital of any sort; this distance is 0 if the nearest hospital is a cardiac catheterization hospital, and otherwise it is positive. If this relative distance affects the probability of receiving this treatment, then it is relevant. If it is distributed randomly across AMI victims, then it is exogenous.

Is relative distance to the nearest cardiac catheterization hospital a valid instrument? McClellan, McNeil, and Newhouse do not report first-stage F-statistics, but they do provide other empirical evidence that it is not weak. Is this distance measure exogenous? They make two arguments. First, they draw on their medical expertise and

[5]Cardiac catheterization is a procedure in which a catheter, or tube, is inserted into a blood vessel and guided all the way to the heart to obtain information about the heart and coronary arteries.

knowledge of the health care system to argue that distance to a hospital is plausibly uncorrelated with any of the unobservable variables that determine AMI outcomes. Second, they have data on some of the additional variables that affect AMI outcomes, such as the weight of the patient, and in their sample, distance is uncorrelated with these *observable* determinants of survival; this, they argue, makes it more credible that distance is uncorrelated with the *unobservable* determinants in the error term as well.

Using 205,021 observations on Americans aged at least 64 who had an AMI in 1987, McClellan, McNeil, and Newhouse reached a striking conclusion: Their TSLS estimates suggest that cardiac catheterization has a small, possibly 0, effect on health outcomes; that is, cardiac catheterization does not substantially prolong life. In contrast, the OLS estimates suggest a large positive effect. They interpret this difference as evidence of bias in the OLS estimates.

McClellan, McNeil, and Newhouse's IV method has an interesting interpretation. The OLS analysis used actual treatment as the regressor, but because actual treatment is itself the outcome of a decision by patient and doctor, they argue that the actual treatment is correlated with the error term. Instead, TSLS uses *predicted* treatment, where the variation in predicted treatment arises because of variation in the instrumental variable: Patients closer to a cardiac catheterization hospital are more likely to receive this treatment.

This interpretation has two implications. First, the IV regression actually estimates the effect of the treatment not on a "typical" randomly selected patient but rather on patients for whom distance is an important consideration in the treatment decision. The effect on those patients might differ from the effect on a typical patient, which provides one explanation of the greater estimated effectiveness of the treatment in clinical trials than in McClellan, McNeil, and Newhouse's IV study. Second, it suggests a general strategy for finding instruments in this type of setting: Find an instrument that affects the probability of treatment, but does so for reasons that are unrelated to the outcome except through their effect on the likelihood of treatment. Both these implications have applicability to experimental and "quasi-experimental" studies, the topic of Chapter 13.

12.6 Conclusion

From the humble start of estimating how much less butter people will buy if its price rises, IV methods have evolved into a general approach for estimating regressions when one or more variables are correlated with the error term. Instrumental variables regression uses the instruments to isolate variation in the endogenous regressors that is uncorrelated with the error in the regression of interest; this is the first stage of two stage least squares. This in turn permits estimation of the effect of interest in the second stage of two stage least squares.

Successful IV regression requires valid instruments—that is, instruments that are both relevant (not weak) and exogenous. If the instruments are weak, then the TSLS estimator can be biased, even in large samples, and statistical inferences based on TSLS t-statistics and confidence intervals can be misleading. Fortunately, when there is a single endogenous regressor, it is possible to check for weak instruments simply by checking the first-stage F-statistic.

If the instruments are not exogenous—that is, if one or more instruments are correlated with the error term—the TSLS estimator is inconsistent. If there are more instruments than endogenous regressors, instrument exogeneity can be examined by using the J-statistic to test the overidentifying restrictions. However, the core assumption—that there are at least as many exogenous instruments as there are endogenous regressors—cannot be tested. It is therefore incumbent on both the empirical analyst and the critical reader to use their own understanding of the empirical application to evaluate whether this assumption is reasonable.

The interpretation of IV regression as a way to exploit known exogenous variation in the endogenous regressor can be used to guide the search for potential instrumental variables in a particular application. This interpretation underlies much of the empirical analysis in the area that goes under the broad heading of program evaluation, in which experiments or quasi-experiments are used to estimate the effect of programs, policies, or other interventions on some outcome measure. A variety of additional issues arises in those applications—for example, the interpretation of IV results when, as in the cardiac catheterization example, different "patients" might have different responses to the same "treatment." These and other aspects of empirical program evaluation are taken up in Chapter 13.

Summary

1. Instrumental variables regression is a way to estimate causal coefficients when one or more regressors are correlated with the error term.
2. Endogenous variables are correlated with the error term in the equation of interest; exogenous variables are uncorrelated with this error term.
3. For an instrument to be valid, it must be (1) correlated with the included endogenous variable and (2) exogenous.
4. IV regression requires at least as many instruments as included endogenous variables.
5. The TSLS estimator has two stages. First, the included endogenous variables are regressed against the included exogenous variables and the instruments. Second, the dependent variable is regressed against the included exogenous variables and the predicted values of the included endogenous variables from the first-stage regression(s).
6. Weak instruments (instruments that are nearly uncorrelated with the included endogenous variables) make the TSLS estimator biased and TSLS confidence intervals and hypothesis tests unreliable.
7. If an instrument is not exogenous, the TSLS estimator is inconsistent.

Key Terms

instrumental variables (IV) regression (427)
instrumental variable (instrument) (427)
endogenous variable (428)
exogenous variable (428)
instrument relevance condition (429)
instrument exogeneity condition (429)
two stage least squares (429)
included exogenous variables (437)
exactly identified (438)
overidentified (438)
underidentified (438)
reduced form (439)
first-stage regression (440)
second-stage regression (440)
weak instruments (445)
first-stage F-statistic (446)
test of overidentifying restrictions (448)

Review the Concepts

12.1 In the demand curve model of Equation (12.3), is $\ln(P_i^{butter})$ positively or negatively correlated with the error, u_i? If β_1 is estimated by OLS, would you expect the estimated value to be larger or smaller than the true value of β_1? Explain.

12.2 Describe the key characteristics of a valid instrument. If you were a researcher, how would you determine if the variable you have selected for an endogenous regressor is a valid instrument or not?

12.3 In their study of the effect of institutions on economic development, suppose Acemoglu et al. had used the prevalence of malaria as an instrument. Would this instrument be relevant? Would it be exogenous? Would it be a valid instrument?

12.4 In their study of the effectiveness of cardiac catheterization, McClellan, McNeil, and Newhouse (1994) used as an instrument the difference in distances to cardiac catheterization and regular hospitals. How could you determine whether this instrument is relevant? How could you determine whether this instrument is exogenous?

Exercises

12.1 This question refers to the panel data IV regressions summarized in Table 12.1.

a. Suppose the federal government is considering a new tax on cigarettes that is estimated to increase the retail price by \$0.25 per pack. If the current price per pack is \$6.75, use the IV regression in column (1) to predict the change in demand. Construct a 95% confidence interval for the change in demand.

b. Suppose the United States enters a recession, and income falls by 5%. Use the regression in column (1) to predict the change in demand.

c. Suppose you have additional data on the prices and quantities of cigarettes in 1993, 1994, 1996, and 1997. How do you think the estimated coefficients would change with an eight-year horizon? With a twelve-year horizon?

d. Suppose that the F-statistic in column (1) were 63.7 instead of 33.7. Would the regression provide a reliable answer to the question posed in (a)? Why or why not?

12.2 Consider the regression model with a single regressor: $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose the least squares assumptions in Key Concept 4.3 are satisfied.

a. Show that X_i is a valid instrument. That is, show that Key Concept 12.3 is satisfied with $Z_i = X_i$.

b. Show that the IV regression assumptions in Key Concept 12.4 are satisfied with this choice of Z_i.

c. Show that the IV estimator constructed using $Z_i = X_i$ is identical to the OLS estimator.

12.3 A classmate is interested in estimating the variance of the error term in Equation (12.1).

a. Suppose she uses the estimator from the second-stage regression of TSLS: $\hat{\sigma}_a^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}\hat{X}_i)^2$, where $\hat{X}_i$ is the fitted value from the first-stage regression. Is this estimator consistent? (For the purposes of this question, suppose that the sample is very large and the TSLS estimators are essentially identical to β_0 and β_1.)

b. Is $\hat{\sigma}_b^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}X_i)^2$ consistent?

12.4 Consider TSLS estimation of the effect of a single included endogenous variable, X_i, on Y_i using one binary instrument, Z_i, which takes values of either 0 or 1. Noting that $\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z}) = \sum_{i=1}^{n}Z_i(Y_i - \overline{Y})$, show that the Wald estimator can be derived from the TSLS estimator in this circumstance to estimate the effect of X_i on Y_i: $\hat{\beta}_{Wald} = (\overline{Y_{|Z=1}} - \overline{Y_{|Z=0}})/(\overline{X_{|Z=1}} - \overline{X_{|Z=0}})$ where $\overline{Y_{|Z=1}}$ equals the mean of values of Y_i for which $Z_i = 1$.

12.5 Consider the IV regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i,$$

where X_i is correlated with u_i and Z_i is an instrument. Suppose that the first three assumptions in Key Concept 12.4 are satisfied. Which IV assumption is not satisfied when

a. Z_i is independent of (Y_i, X_i, W_i)?

b. $Z_i = W_i$?

c. $W_i = 1$ for all i?

d. $Z_i = X_i$?

12.6 Suppose a researcher is considering developing an IV regression model with one regressor, X_i, and one instrument, Z_i. If she has a sample of $n = 113$, what range must the correlation coefficient be between X_i and Z_i in order for Z_i to be considered a strong instrument? [*Hint:* See Equation (7.14).]

12.7 A classmate has developed an IV regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}. She has a strong theoretical basis as to why $corr(Z_{1i}, u_i) = 0$, namely that Z_{1i} is the result of a random lottery. Preliminary work, however, showed that the first-stage F-statistic from this exactly identified model was insufficiently large for Z_{1i} to be considered a relevant instrument set by itself. As a result she includes an additional instrument, Z_{2i}, which is strongly relevant but is less likely to satisfy the condition of instrument exogeneity. In the instrumental variable regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}, the value of the J-statistic is $J = 7.5$.

a. Does this suggest that $E(u_i \mid Z_{1i}, Z_{2i}) \neq 0$? Explain.

b. Does this suggest that $E(u_i \mid Z_{2i}) \neq 0$? Explain.

12.8 Consider a product market with a supply function $Q_i^s = \beta_0 + \beta_1 P_i + u_i^s$, a demand function $Q_i^d = \gamma_0 + u_i^d$, and a market equilibrium condition $Q_i^s = Q_i^d$, where u_i^s and u_i^s are mutually independent i.i.d. random variables, both with a mean of 0.

a. Show that P_i and u_i^s are correlated.

b. Show that the OLS estimator of β_1 is inconsistent.

c. How would you estimate β_0, β_1, and γ_0?

12.9 A researcher is interested in the effect of more secure property rights on income across countries. He collects recent data from 60 countries and runs the OLS regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i is a country's GDP per capita and X_i is an index taking values between 0 and 10 reflecting the protection against expropriation where a higher value indicates greater protection against expropriation, that is, more secure property rights.

a. Explain why the OLS estimates are likely to be unreliable and indicate in which direction they might be biased. (*Hint:* In which direction does causality run in this example?)

b. All of the countries in the researcher's sample were former colonies. Institutions securing property rights could originate from early institutions established alongside European settlements. The decision for Europeans to settle or otherwise could reflect concerns for mortality among settlers. Explain how settler mortality might be used as an instrument to estimate the effect of more secure property rights on income across countries.

12.10 Two classmates are comparing their answers to an assignment. One classmate has specified an instrumental variable regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$, using Z_i as an instrument. The other student has specified the same model, but has omitted W_i.

a. The first student says that if Z_i and W_i are correlated, then the second student's IV estimator will not be consistent. Is the first student right about this?

b. The second student argues that if in the true model $\beta_2 = 0$, then their IV estimator will be consistent. Is the second student correct in saying this?

Empirical Exercises

E12.1 How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise, you will estimate this effect using data for married women from the 1980 U.S. Census.[6] The data are available on the text website, **http://www.pearsonglobaleditions.com**, in the file **Fertility** and described in the file **Fertility_Description**. The data set contains information on married women aged 21–35 with two or more children.

a. Regress *weeksworked* on the indicator variable *morekids*, using OLS. On average, do women with more than two children work less than women with two children? How much less?

b. Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labor supply (*weeksworked*).

c. The data set contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

d. Explain why *samesex* is a valid instrument for the IV regression of *weeksworked* on *morekids*.

[6]These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Joshua Angrist, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 1998, 88(3): 450–477.

e. Is *samesex* a weak instrument?

f. Estimate the IV regression of *weeksworked* on *morekids*, using *samesex* as an instrument. How large is the fertility effect on labor supply?

g. Do the results change when you include the variables *agem1*, *black*, *hispan*, and *othrace* in the labor supply regression (treating these variable as exogenous)? Explain why or why not.

E12.2 Does viewing a violent movie lead to violent behavior? If so, the incidence of violent crimes, such as assaults, should rise following the release of a violent movie that attracts many viewers. Alternatively, movie viewing may substitute for other activities (such as alcohol consumption) that lead to violent behavior, so that assaults should fall when more viewers are attracted to the cinema. On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Movies**, which contains data on the number of assaults and movie attendance for 516 weekends from 1995 through 2004.[7] A detailed description is given in **Movies_Description**, available on the website. The data set includes weekend U.S. attendance for strongly violent movies (such as *Hannibal*), mildly violent movies (such as *Spider-Man*), and nonviolent movies (such as *Finding Nemo*). The data set also includes a count of the number of assaults for the same weekend in a subset of counties in the United States. Finally, the data set includes indicators for year, month, whether the weekend is a holiday, and various measures of the weather.

a. i. Regress the logarithm of the number of assaults [$ln\_assaults = \ln(assaults)$] on the year and month indicators. Is there evidence of seasonality in assaults? That is, do there tend to be more assaults in some months than others? Explain.

ii. Regress total movie attendance ($attend = attend\_v + attend\_m + attend\_n$) on the year and month indicators. Is there evidence of seasonality in movie attendance? Explain.

b. Regress *ln_assaults* on *attend_v*, *attend_m*, *attend_n*, the year and month indicators, and the weather and holiday control variables available in the data set.

i. Based on the regression, does viewing a strongly violent movie increase or decrease assaults? By how much? Is the estimated effect statistically significant?

ii. Does attendance at strongly violent movies affect assaults differently than attendance at moderately violent movies? Differently than attendance at nonviolent movies?

---

[7]These are aggregated versions of data provided by Gordon Dahl of University of California–San Diego and Stefano DellaVigna of University of California–Berkeley and were used in their paper "Does Movie Violence Increase Violent Crime?" *Quarterly Journal of Economics*, 2009, 124(2): 677–734.

iii. A strongly violent blockbuster movie is released, and the weekend's attendance at strongly violent movies increases by 6 million; meanwhile, attendance falls by 2 million for moderately violent movies and by 1 million for nonviolent movies. What is the predicted effect on assaults? Construct a 95% confidence interval for the change in assaults. [*Hint:* Review Section 7.3 and material surrounding Equations (8.7) and (8.8).]

c. It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments—*pr_attend_v*, *pr_attend_m*, and *pr_attend_n*—that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of the construction of these instruments are available in the Dahl and DellaVigna paper referenced in footnote 5.) Run the regression from (b) (including year, month, holiday, and weather controls) but now using *pr_attend_v*, *pr_attend_m*, and *pr_attend_n* as instruments for *attend_v*, *attend_m*, and *attend_n*. Use this IV regression to answer (b)(i)–(b)(iii).

d. The intuition underlying the instruments in (c) is that attendance in a given week is correlated with attendance in surrounding weeks. For each movie category, the data set includes attendance in surrounding weeks. Run the regression using the instruments *attend_v_f*, *attend_m_f*, *attend_n_f*, *attend_v_b*, *attend_m_b*, and *attend_n_b* instead of the instruments used in (c). Use this IV regression to answer (b)(i)–(b)(iii).

e. There are nine instruments listed in (c) and (d), but only three are needed for identification. Carry out the test for overidentification summarized in Key Concept 12.6. What do you conclude about the validity of the instruments?

f. Based on your analysis, what do you conclude about the effect of violent movies on (short-run) violent behavior?

E12.3 (This requires Appendix 12.5) On the text website, **http://www.pearsonglobaleditions.com**, you will find the data set **WeakInstrument**, which contains 200 observations on (Y_i, X_i, Z_i) for the instrumental regression $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Construct $\hat{\beta}_1^{TSLS}$, its standard error, and the usual 95% confidence interval for β_1.

b. Compute the F-statistic for the regression of X_i on Z_i. Is there evidence of a "weak instrument" problem?

c. Compute a 95% confidence interval for β_1, using the Anderson–Rubin procedure. (To implement the procedure, assume that $-5 \leq \beta_1 \leq 5$.)

d. Comment on the differences in the confidence intervals in (a) and (c). Which is more reliable?

APPENDIX

12.1 The Cigarette Consumption Panel Data Set

The data set consists of annual data for the 48 contiguous U.S. states from 1985 to 1995. Quantity consumed is measured by annual per capita cigarette sales in packs per fiscal year, as derived from state tax collection data. The price is the real (that is, inflation-adjusted) average retail cigarette price per pack during the fiscal year, including taxes. Income is real per capita income. The general sales tax is the average tax, in cents per pack, due to the broad-based state sales tax applied to all consumption goods. The cigarette-specific tax is the tax applied to cigarettes only. All prices, income, and taxes used in the regressions in this chapter are deflated by the Consumer Price Index and thus are in constant (real) dollars. We are grateful to Professor Jonathan Gruber of MIT for providing us with these data.

APPENDIX

12.2 Derivation of the Formula for the TSLS Estimator in Equation (12.4)

The first stage of TSLS is to regress X_i on the instrument Z_i by OLS and then compute the OLS predicted value $\hat{X}_i$; the second stage is to regress Y_i on $\hat{X}_i$ by OLS. Accordingly, the formula for the TSLS estimator, expressed in terms of the predicted value $\hat{X}_i$, is the formula for the OLS estimator in Key Concept 4.2, with $\hat{X}_i$ replacing X_i. That is, $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y}/s_{\hat{X}}^2$, where $s_{\hat{X}}^2$ is the sample variance of $\hat{X}_i$ and $s_{\hat{X}Y}$ is the sample covariance between Y_i and $\hat{X}_i$.

Because $\hat{X}_i$ is the predicted value of X_i from the first-stage regression, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, the definitions of sample variances and covariances imply that $s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY}$ and $s_{\hat{X}}^2 = \hat{\pi}_1^2 s_Z^2$ (Exercise 12.4). Thus, the TSLS estimator can be written as $\hat{\beta}_1^{TSLS} = s_{\hat{X}Y}/s_{\hat{X}}^2 = s_{ZY}/(\hat{\pi}_1 s_Z^2)$. Finally, $\hat{\pi}_1$ is the OLS slope coefficient from the first stage of TSLS, so $\hat{\pi}_1 = s_{ZX}/s_Z^2$. Substitution of this formula for $\hat{\pi}_1$ into the formula $\hat{\beta}_1^{TSLS} = s_{ZY}/(\hat{\pi}_1 s_Z^2)$ yields the formula for the TSLS estimator in Equation (12.4).

APPENDIX

12.3 Large-Sample Distribution of the TSLS Estimator

This appendix studies the large-sample distribution of the TSLS estimator in the case considered in Section 12.1—that is, with a single instrument, a single included endogenous variable, and no included exogenous variables.

To start, we derive a formula for the TSLS estimator in terms of the errors; this formula forms the basis for the remaining discussion, similar to the expression for the OLS estimator in Equation (4.28) in Appendix 4.3.

From Equation (12.1), $Y_i - \overline{Y} = \beta_1(X_i - \overline{X}) + (u_i - \overline{u})$. Accordingly, the sample covariance between Z and Y can be expressed as

$$\begin{aligned} s_{ZY} &= \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})[\beta_1(X_i - \overline{X}) + (u_i - \overline{u})] \\ &= \beta_1 s_{ZX} + \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})(u_i - \overline{u}) \\ &= \beta_1 s_{ZX} + \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})u_i, \end{aligned} \tag{12.19}$$

where $s_{ZX} = [1/(n-1)]\Sigma_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})$ and where the final equality follows because $\Sigma_{i=1}^{n}(Z_i - \overline{Z}) = 0$. Substituting the definition of s_{ZX} and the final expression in Equation (12.19) into the definition of $\hat{\beta}_1^{TSLS}$ and multiplying the numerator and denominator by $(n-1)/n$ yields

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}. \tag{12.20}$$

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the IV Regression Assumptions in Key Concept 12.4 Hold

Equation (12.20) for the TSLS estimator is similar to Equation (4.28) in Appendix 4.3 for the OLS estimator except that Z rather than X appears in the numerator and that the denominator is the covariance between Z and X rather than the variance of X. Because of these similarities and because Z is exogenous, the argument in Appendix 4.3 that the OLS estimator is normally distributed in large samples extends to $\hat{\beta}_1^{TSLS}$.

Specifically, when the sample is large, $\overline{Z} \cong \mu_Z$, so the numerator is approximately $\overline{q} = (\frac{1}{n})\Sigma_{i=1}^{n} q_i$, where $q_i = (Z_i - \mu_Z)u_i$. Because the instrument is exogenous, $E(q_i) = 0$. By the IV regression assumptions in Key Concept 12.4, q_i is i.i.d. with variance $\sigma_q^2 = \text{var}[(Z_i - \mu_Z)u_i]$. It follows that var $(\overline{q}) = \sigma_{\overline{q}}^2 = \sigma_q^2/n$, and, by the central limit theorem, $\overline{q}/\sigma_{\overline{q}}$ is, in large samples, distributed $N(0, 1)$.

Because the sample covariance is consistent for the population covariance, $s_{ZX} \xrightarrow{p} \text{cov}(Z_i, X_i)$, which, because the instrument is relevant, is nonzero. Thus, by Equation (12.20), $\hat{\beta}_1^{TSLS} \cong \beta_1 + \bar{q}/\text{cov}(Z_i, X_i)$, so in large samples $\hat{\beta}_1^{TSLS}$ is approximately distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$, where $\sigma^2_{\hat{\beta}_1^{TSLS}} = \sigma^2_{\bar{q}}/[\text{cov}(Z_i, X_i)]^2 = (1/n)\text{var}[(Z_i - \mu_Z)u_i]/[\text{cov}(Z_i, X_i)]^2$, which is the expression given in Equation (12.8).

APPENDIX

12.4 Large-Sample Distribution of the TSLS Estimator When the Instrument Is Not Valid

This appendix considers the large-sample distribution of the TSLS estimator in the setup of Section 12.1 (one X, one Z) when one or the other of the conditions for instrument validity fails. If the instrument relevance condition fails, the large-sample distribution of the TSLS estimator is not normal; in fact, its distribution is that of a ratio of two normal random variables. If the instrument exogeneity condition fails, the TSLS estimator is inconsistent.

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the Instrument Is Weak

First consider the case that the instrument is irrelevant, so that $\text{cov}(Z_i, X_i) = 0$. Then the argument in Appendix 12.3 entails division by 0. To avoid this problem, we need to take a closer look at the behavior of the term in the denominator of Equation (12.20) when the population covariance is 0.

We start by rewriting Equation (12.20). Because of the consistency of the sample average, in large samples $\bar{Z}$ is close to μ_Z, and $\bar{X}$ is close to μ_X. Thus the term in the denominator of Equation (12.20) is approximately $(\frac{1}{n})\sum_{i=1}^n (Z_i - \mu_Z)(X_i - \mu_X) = (\frac{1}{n})\sum_{i=1}^n r_i = \bar{r}$, where $r_i = (Z_i - \mu_Z)(X_i - \mu_X)$. Let $\sigma_r^2 = \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$, let $\sigma^2_{\bar{r}} = \sigma_r^2/n$, and let $\bar{q}$, σ_q^2, and $\sigma^2_{\bar{q}}$ be as defined in Appendix 12.3. Then Equation (12.20) implies that, in large samples,

$$\hat{\beta}_1^{TSLS} \cong \beta_1 + \frac{\bar{q}}{\bar{r}} = \beta_1 + \left(\frac{\sigma_{\bar{q}}}{\sigma_{\bar{r}}}\right)\left(\frac{\bar{q}/\sigma_{\bar{q}}}{\bar{r}/\sigma_{\bar{r}}}\right) = \beta_1 + \left(\frac{\sigma_q}{\sigma_r}\right)\left(\frac{\bar{q}/\sigma_{\bar{q}}}{\bar{r}/\sigma_{\bar{r}}}\right). \tag{12.21}$$

If the instrument is irrelevant, then $E(r_i) = \text{cov}(Z_i, X_i) = 0$. Thus $\bar{r}$ is the sample average of the random variables r_i, $i = 1, \ldots, n$, which are i.i.d. (by the second least squares assumption), have variance $\sigma_r^2 = \text{var}[(Z_i - \mu_Z)(X_i - \mu_X)]$ (which is finite by the third IV regression assumption), and have a mean of 0 (because the instruments are irrelevant). It follows that the central limit theorem applies to $\bar{r}$; specifically, $\bar{r}/\sigma_{\bar{r}}$ is approximately distributed $N(0, 1)$. Therefore, the final expression of Equation (12.21) implies that, in large samples, the distribution of $\hat{\beta}_1^{TSLS} - \beta_1$ is the distribution of aS, where $a = \sigma_q/\sigma_r$ and S is the ratio of two random variables, each of which has a standard normal distribution (these two standard normal random variables are correlated).

In other words, when the instrument is irrelevant, the central limit theorem applies to the denominator as well as the numerator of the TSLS estimator, so in large samples the distribution of the TSLS estimator is the distribution of the ratio of two normal random variables. Because X_i and u_i are correlated, these normal random variables are correlated, and the large-sample distribution of the TSLS estimator when the instrument is irrelevant is complicated. In fact, the large-sample distribution of the TSLS estimator with irrelevant instruments is centered on the probability limit of the OLS estimator. Thus when the instrument is irrelevant, TSLS does not eliminate the bias in OLS and, moreover, has a nonnormal distribution, even in large samples.

A weak instrument represents an intermediate case between an irrelevant instrument and the normal distribution derived in Appendix 12.3. When the instrument is weak but not irrelevant, the distribution of the TSLS estimator continues to be nonnormal, so the general lesson here about the extreme case of an irrelevant instrument carries over to weak instruments.

Large-Sample Distribution of $\hat{\beta}_1^{TSLS}$ When the Instrument Is Endogenous

The numerator in the final expression in Equation (12.20) converges in probability to $\text{cov}(Z_i, u_i)$. If the instrument is exogenous, this is 0, and the TSLS estimator is consistent (assuming that the instrument is not weak). If, however, the instrument is not exogenous, then, if the instrument is not weak, $\hat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1 + \text{cov}(Z_i, u_i)/\text{cov}(Z_i, X_i) \neq \beta_1$. That is, if the instrument is not exogenous, the TSLS estimator is inconsistent.

APPENDIX

12.5 Instrumental Variables Analysis with Weak Instruments

This appendix discusses some methods for instrumental variables analysis in the presence of potentially weak instruments. The appendix focuses on the case of a single included endogenous regressor [Equations (12.13) and (12.14)].

Testing for Weak Instruments

The rule of thumb in Key Concept 12.5 is that a first-stage F-statistic less than 10 indicates that the instruments are weak. One motivation for this rule of thumb arises from an approximate expression for the bias of the TSLS estimator. Let β_1^{OLS} denote the probability limit of the OLS estimator β_1, and let $\beta_1^{OLS} - \beta_1$ denote the asymptotic bias of the OLS estimator (if the regressor is endogenous, then $\hat{\beta}_1 \xrightarrow{p} \beta_1^{OLS} \neq \beta_1$). It is possible to show that, when there are many instruments, the bias of the TSLS estimator is approximately $E(\hat{\beta}_1^{TSLS}) - \beta_1 \approx (\beta_1^{OLS} - \beta_1)/[E(F) - 1]$, where $E(F)$ is the expectation of the first-stage F-statistic. If $E(F) = 10$, then the bias of TSLS relative to the bias of OLS is approximately 1/9, or just over 10%, which is small enough to be acceptable in many applications. Replacing $E(F) > 10$ with $F > 10$ yields the rule of thumb in Key Concept 12.5.

The motivation in the previous paragraph involved an approximate formula for the bias of the TSLS estimator when there are many instruments. In most applications, however, the number of instruments, m, is small. Stock and Yogo (2005) provide a formal test for weak instruments that avoids the approximation that m is large. In the Stock–Yogo test, the null hypothesis is that the instruments are weak, and the alternative hypothesis is that the instruments are strong, where strong instruments are defined to be instruments for which the bias of the TSLS estimator is at most 10% of the bias of the OLS estimator. The test entails comparing the first-stage F-statistic (for technical reasons, the homoskedasticity-only version) to a critical value that depends on the number of instruments. As it happens, for a test with a 5% significance level, this critical value ranges between 9.08 and 11.52, so the rule of thumb of comparing F to 10 is a good approximation to the Stock–Yogo test.

Hypothesis Tests and Confidence Sets for β

If the instruments are weak, the TSLS estimator is biased and has a nonnormal distribution. Thus the TSLS t-test of $\beta_1 = \beta_{1,0}$ is unreliable, as is the TSLS confidence interval for β_1. There are, however, other tests of $\beta_1 = \beta_{1,0}$, along with confidence intervals based on those tests, that are valid whether instruments are strong, weak, or even irrelevant. When there is a single endogenous regressor, the preferred test is Moreira's (2003) conditional likelihood ratio (CLR) test. An older test, which works for any number of endogenous regressors, is based on the Anderson–Rubin (1949) statistic. Because the Anderson–Rubin statistic is conceptually less complicated, we describe it first.

The Anderson–Rubin test of $\beta_1 = \beta_{1,0}$ proceeds in two steps. In the first step, compute a new variable, $Y_i^* = Y_i - \beta_{1,0}X_i$. In the second step, regress Y_i^* against the included exogenous regressors (W's) and the instruments (Z's). The Anderson–Rubin statistic is the F-statistic testing the hypothesis that the coefficients on the Z's are all 0. Under the null hypothesis that $\beta_1 = \beta_{1,0}$, if the instruments satisfy the exogeneity condition (condition 2 in Key Concept 12.3), they will be uncorrelated with the error term in this regression, and the null hypothesis will be rejected in 5% of all samples.

As discussed in Sections 3.3 and 7.4, a confidence set can be constructed as the set of values of the parameters that are not rejected by a hypothesis test. Accordingly, the set of values of β_1 that are not rejected by a 5% Anderson–Rubin test constitutes a 95% confidence set for β_1. When the Anderson–Rubin F-statistic is computed using the homoskedasticity-only formula, the Anderson–Rubin confidence set can be constructed by solving a quadratic equation (see Empirical Exercise 12.3). The logic behind the Anderson–Rubin statistic never assumes instrument relevance, and the Anderson–Rubin confidence set will have a coverage probability of 95% in large samples, whether the instruments are strong, weak, or even irrelevant.

The CLR statistic also tests the hypothesis that $\beta_1 = \beta_{1,0}$. Likelihood ratio statistics compare the value of the likelihood (see Appendix 11.2) under the null hypothesis to its value under the alternative and reject it if the likelihood under the alternative is sufficiently greater than under the null. Familiar test statistics in this text, such as the homoskedasticity-only F-statistic in multiple regression, can be derived as likelihood ratio statistics under the assumption of homoskedastic normally distributed errors. Unlike any of the other tests discussed in this text, however, the critical value of the CLR test depends on the data—specifically, on a statistic that measures the strength of the instruments. By using the right critical value, the CLR test is valid whether instruments are strong, weak, or irrelevant. CLR confidence intervals can be computed as the set of values of β_1 that are not rejected by the CLR test.

The CLR test is equivalent to the TSLS t-test when instruments are strong and has very good power when instruments are weak. With suitable software, the CLR test is easy to use. The disadvantage of the CLR test is that it does not generalize readily to more than one endogenous regressor. In that case, the Anderson–Rubin test (and confidence set) is recommended; however, when instruments are strong (so TSLS is valid) and the coefficients are overidentified, the Anderson–Rubin test is inefficient in the sense that it is less powerful than the TSLS t-test.

Estimation of β

If the instruments are irrelevant, then without further restrictions it is not possible to obtain an unbiased estimator of β_1, even in large samples. With weak instruments, CLR or Anderson–Rubin confidence intervals for the coefficients are preferable to point estimation.

The problems of estimation, testing, and confidence intervals in IV regression with weak instruments constitute an area of ongoing research. To learn more about this topic, visit the website for this text.

APPENDIX

12.6 TSLS with Control Variables

In Key Concept 12.4, the W variables are assumed to be exogenous. This appendix considers the case in which W is not exogenous but instead is a control variable included to make Z exogenous. The logic of control variables in TSLS parallels the logic in OLS: If a control variable effectively controls for an omitted factor, then the instrument is uncorrelated with the error term. Because the control variable is correlated with the error term, the coefficient on a control variable does not have a causal interpretation. The mathematics of control variables in TSLS also parallels the mathematics of control variables in OLS and entails relaxing the assumption that the error has conditional mean 0 given Z and W to be that the conditional mean of the error does not depend on Z. This appendix draws on Appendix 6.5 (OLS with control variables), which should be reviewed first.

Consider the IV regression model in Equation (12.12) with a single X and a single W:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i. \tag{12.22}$$

We replace IV regression assumption 1 in Key Concept 12.4 [which states that $E(u_i \mid W_i) = 0$] with the assumption that, conditional on W_i, the mean of u_i does not depend on Z_i:

$$E(u_i \mid W_i, Z_i) = E(u_i \mid W_i). \tag{12.23}$$

The next steps in the argument parallel those for regression with control variables in Equations (6.23)–(6.25) in Appendix 6.5. Assume that $E(u_i \mid W_i)$ is linear in W_i, so $E(u_i \mid W_i) = \gamma_0 + \gamma_1 W_i$, where γ_0 and γ_1 are coefficients. Then

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i - E(u_i \mid W_i, Z_i) + E(u_i \mid W_i, Z_i) \\ &= \beta_0 + \beta_1 X_i + \beta_2 W_i + \varepsilon_i + \gamma_0 + \gamma_1 W_i \end{aligned} \tag{12.24}$$

where the first line adds and subtracts $E(u_i \mid W_i, Z_i)$ to the right hand side of Equation (12.22), and the second line and defines $\varepsilon_i = u_i - E(u_i \mid W_i, Z_i)$ and uses the conditional mean independence assumption plus linearity to write $E(u_i \mid W_i, Z_i) = E(u_i \mid W_i) = \gamma_0 + \gamma_1 W_i$. We thus have that,

$$Y_i = \delta_0 + \beta_1 X_i + \delta_1 W_i + \varepsilon_i, \tag{12.25}$$

where $\delta_0 = \beta_0 + \gamma_0$ and $\delta_1 = \beta_2 + \gamma_1$. Now $E(\varepsilon_i \mid W_i, Z_i) = E[u_i - E(u_i \mid W_i, Z_i) \mid W_i, Z_i] = E(u_i \mid W_i, Z_i) - E(u_i \mid W_i, Z_i) = 0$, which in turn implies $\mathrm{corr}(Z_i, \varepsilon_i) = 0$. Thus IV regression assumption 1 and the instrument exogeneity requirement (condition 2 in Key Concept 12.3) both hold for Equation (12.24) with error term ε_i, Thus, if IV regression assumption 1 is replaced by conditional mean independence in Equation (12.23), the original IV regression assumptions in Key Concept 12.4 apply to the modified regression in Equation (12.25).

Because the IV regression assumptions of Key Concept 12.4 hold for Equation (12.25), all the methods of inference (for both weak and strong instruments) discussed in this chapter apply to Equation (12.25). In particular, if the instruments are strong, the coefficients in Equation (12.25) will be estimated consistently by TSLS and TSLS tests, and confidence intervals will be valid.

Just as in OLS with control variables, in general the TSLS coefficient on the control variable W does not have a causal interpretation. TSLS consistently estimates δ_1 in Equation (12.25), but δ_1 is the sum of β_2, the direct causal effect of W, and γ_1, which reflects the correlation between W and the omitted factors in u_i for which W controls.

In the cigarette consumption regressions in Table 12.1, it is tempting to interpret the coefficient on the 10-year change in log income as the income elasticity of demand. If, however, income growth is correlated with increases in education and if more education reduces smoking, income growth would have its own causal effect (β_2, the income elasticity) plus an effect arising from its correlation with education (γ_1). If the latter effect is negative ($\gamma_1 < 0$), the income coefficients in Table 12.1 (which estimate $\delta_1 = \beta_2 + \gamma_1$) would underestimate the income elasticity. As long as the conditional mean independence assumption in Equation (12.23) holds, however, the TSLS estimator of the price elasticity is consistent, even if the estimate of the income elasticity is not.