exogenous, that is, uncorrelated with the error term of the structural equation; (2) it must be partially correlated with the endogenous explanatory variable. Finding a variable with these two properties is usually challenging.

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

When we have valid instrumental variables, we can test whether an explanatory variable is endogenous, using the test in Section 15.5. In addition, though we can never test whether all IVs are exogenous, we can test that at least some of them are—assuming that we have more instruments than we need for consistent estimation (that is, the model is overidentified). Heteroskedasticity and serial correlation can be tested for and dealt with using methods similar to the case of models with exogenous explanatory variables.

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

Key Terms

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

Problems

1 Consider a simple model to measure the effects of taking a preparatory course (a binary variable, *course*) on eventual score on a college admissions exam:

$$score = \beta_0 + \beta_1 course + u.$$

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

(ii) Is *course* likely to be related to parents' income? If so, does this mean parental income is a good IV for *course*? Explain.

(iii) Suppose that 20% of students at every school were randomly given tuition waivers for the course. Carefully explain how you would use this information to construct an IV for *course*.

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

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

where the variables are defined as in Chapter 6.

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

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

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

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

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

3 Consider the simple regression model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(iii) How would you consistently estimate the a_j?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Computer Exercises

C1 Use the data in WAGE2.RAW for this exercise.

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

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

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

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

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

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

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

State and test the identification assumption.

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

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

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

(i) Estimate the model

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

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

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

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

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

C3 Use the data in CARD.RAW for this exercise.

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

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

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

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

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

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

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

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

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

(ii) Some economists feel that the Consumer Price Index mismeasures the true rate of inflation, so that the OLS from part (i) suffers from measurement error bias.