Chapter 14

Using these data, our former student performed two-stage least squares and got the results in Table 14.3.

TABLE 14.3
Using Two-Stage Least Squares to Estimate Our Former's Students Milk Supply Equation

Supply Equation						
	Coefficients	Standard Error	t	P > |t|	Lower 95%	Upper 95%
Intercept	12,057.69	1,909.19	6.32	0.000	8,301.77	15,813.61
All milk price	361.76	126.79	2.85	0.005	112.34	611.19
Milk feed price ratio	−1,282.92	110.23	−11.64	0.000	−1,499.78	− 1,066.07
Prob > F			0.0000			
R-squared			0.4907			
Observations			328			

Note: Dependent variable is all milk production (millions of gallons). Highlighted variables are significant at the 5 percent level.

These results suggested to him that the milk supply curve was indeed upward sloping but that the estimated marginal effect of milk prices on milk production using instrumental variables was lower than the marginal effect estimated through OLS while the estimated marginal effect of the milk–feed price ratio was similar to that estimated through OLS.

Based on this work, from the initial idea through data collection, empirical analysis, interpretation of results, and effective write-up, our former student was judged to have conducted a successful empirical research project.

LOOKING AHEAD TO CHA

There are a number of other situations encountered by introductory econometrics students in which OLS is not the best possible estimator. For instance, we might sometimes wish to (1) learn about the estimated relationship between the dependent variable and the independent variables at points in the distribution other than the mean, (2) estimate models in which the dependent variable takes on non-negative count values, (3) control for sample-selection bias in our dependent variable, or (4) employ quasi-experimental techniques. We discuss the empirical tools required to address these situations in Chapter 15.

Problems

14.1 The following two equations describe property crime and police expenditures across U.S. cities:

$$Crime_i = \alpha_0 + \alpha_1 Police_i + \alpha_2 Income_i + u_{1,i}$$

$$Police_i = \beta_0 + \beta_1 Crime_i + \beta_2 Income_i + \beta_3 Liberal_i + u_{2,1}$$

$Crime_i$ = Number of crimes committed in city i last year

$Police_i$ = Total spending on the police force in city i last year

$Income_i$ = Average income in city i last year

$Liberal_i$ = Share of votes received by the liberal presidential candidate from city i in the last election

a. If you were to estimate these two equations with OLS, would you get consistent estimates? Why or why not?

b. Which of the two equations is fully identified? How do you know this?

c. Describe how you would estimate one or both of the equations using 2SLS. Describe the properties of the estimates in both equations.

d. Come up with another explanatory variable that you could add to one of these equations to make the system fully identified. Justify the inclusion of this variable.

e. Discuss practical and theoretical drawbacks of using 2SLS.

14.2 A public health researcher is trying to estimate the determinants of fertility rates in developing countries. She proposes the following model:

$$FR_i = \beta_0 + \beta_1 Age_i + \beta_2 Education_i + \beta_3 Urban_i + u_i$$

where FR_i is the fertility rate for individual i (measured as the number of children ever born), regressed on their age, their years of education, and whether or not they live in an urban area.

a. Another researcher argues that the Education variable suffers from a simultaneity bias with fertility rates. Identify the consequences of including an endogenous explanatory variable, and briefly outline the steps you would take to consistently estimate this model.

b. This study used two different methods to measure fertility rates: one for rural areas and one for urban areas. Each method generated some measurement error, though the rural methodology was quite a bit less accurate. Given this fact, what are the consequences of measurement error in this sample? What, if anything, can you do to remedy the potential problems associated with measurement error on the fertility variable? Be as specific as possible.

c. Consider your answer to part (b), and assume fertility rates no longer suffer from measurement error, but the education variable does suffer from measurement error. Also assume you have two different measures of education: one self-reported and the other obtained by asking the individual's closest relative. Given this fact, what are the consequences of measurement error in this sample? What, if anything, can you do to remedy the potential problems associated with measurement error on the education variable? Be as specific as possible.

14.3 The following two equations describe the interactions between fertility rates and average income of women in a cross-section of countries:

$$Fertility_i = \alpha_0 + \alpha_1 Income_i + \alpha_2 Education_i + \alpha_3 Rural_i + u_1$$

$$Income_i = \beta_0 + \beta_1 Fertility_i + \beta_2 Education_i + u_2$$

where the fertility rate (measured by average number of births per woman in country i) is a function of average female income in that country, the average years of female education, and the fraction of country i's population that lives in rural areas, and female income is a function of the fertility rate and female education.

a. If you were to estimate these two equations separately with OLS, would you get consistent parameter estimates? Explain why or why not.

b. Which of the two equations is fully identified? Explain why.

c. Describe how you would consistently estimate the identified equation.

d. Come up with another explanatory variable, Z_i, that you could add to this system so that both equations can be estimated. What two conditions must Z_i satisfy? Explain why the variable you selected meets these two criteria.

14.4 The following system of equations model the murder rate in a state and the total number of executions in that state:

$$M_i = \beta_0 + \beta_1 P_i + \beta_2 E_i + \beta_3 I_i + u_i$$

$$E_i = \gamma_0 + \gamma_1 M_i + \gamma_2 P_i + e_i$$

where M_i = Murder rate in state i

P_i = Total expenditures on the police force in state i

E_i = Total number of executions in state i

I_i = Average annual income in state i

a. Suppose you estimated each equation with OLS. Would you get consistent estimates? Why or why not?

b. Given what you know about needing both a demand and supply shifter, can both of these equations be estimated? If not, which one cannot be estimated?

c. Describe in detail how you would consistently estimate the equation that can be estimated using two-stage least squares.

14.5 Consider the following regression model:

$$y_i^* = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

where y_i^* is a dependent variable that is measured with error. Specifically, you never observe y_i^*, you actually observe $y_i = y_i^* + \eta$, where η is the measurement error in y_i^*.

a. Under what condition(s) will estimating $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$ yield inconsistent estimates?

b. Assuming that the condition(s) for consistency in part (a) are met, what will be the net effect of measurement error in y?

14.6 Consider a simple time-series model where the independent variable has classical measurement error

$$(1)\ y_t = \beta_0 + \beta_1 x_t^* + \varepsilon_t$$

and

$$(2)\ x_t = x_t^* + u_t$$

where ε_t has zero mean and is uncorrelated with x_t^* and u_t. We observe y_t and x_t only. Assume that u_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).

a. Write $x_t^* = x_t - u_t$, and plug this into equation (1). 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?

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

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

d. What do parts (b) and (c) suggest as a useful strategy for consistently estimating β_0 and β_1?

Exercises

E14.1 Use the data in Education.XLS to run an instrumental variable regression.

a. First estimate the model using OLS. Regress logwage on *Experience*, *Occupation* (equals 1 if blue collar occupation and 0 if white collar), Industry (equals 1 if manufacturing industry and 0 if not), *Married* (equals 1 if married and 0 if not), Union (equals 1 if in a union and 0 if not), *Education* (years of education), and *Black* (equals 1 if black and 0 if not).

b. Now use *South* (equals 1 if live in south and 0 if not) as an instrument for education. Run the first-stage regression by regressing *Education on South* and all other independent variables and get the predicted value for education. Run the second-stage regression by regressing logwage on the predicted value of *Education*, *Experience*, *Occupation*, *Industry*, *Married*, *Union*, and *Black*.

c. Comment on if you think that the dummy variable *South* is a valid instrument; that is, is *South* correlated with years of *Education* and uncorrelated with the error term in the regression in part (a)?

E14.2 Use the data in Demand. XLS to run an instrumental variable regression.

a. First estimate the demand equation model using OLS. Regress *Quantity on Price and Consumer Confidence*.

b. Now use Milk feed price as an instrument for *Price*. Run the first- and second-stage regressions. Does your regression differ from what you found in part (a)?

c. Draw a picture that shows why it is important to use a supply shifter when estimating a demand equation.

E14.3 Use the data in Measurement Error. XLS to correct for measurement error.

a. First estimate the model using OLS. Regress logwage on *Experience*, *Occupation* (equals 1 if blue collar occupation and 0 if white collar), *Industry* (equals 1 if manufacturing industry and 0 if not), *Married* (equals 1 if married and 0 if not), Union (equals 1 if in a union and 0 if not), *Education* (years of education), and *Black* (equals 1 if black and 0 if not).

b. Now use spouse experience as an instrument for experience to correct for the fact that experience may be measured with error. Run the first-stage regression by regressing *Education on Spouse Experience* and all other independent variables and get the predicted value for *Experience*. Run the second-stage regression by regressing logwage on the predicted value of *Experience*, *Education*, *Occupation*, *Industry*, *Married*, *Union*, and *Black*.

Chapter 15

Chapter Fifteen *Quantile Regression, Count Data, Sample Selection Bias, and Quasi Experimental Methods*

LOOKING AHEAD TO CHA

At this point, we have filled our empirical toolkit with an extensive number of tools that apply to a wide variety of situations. Beyond being able to correctly estimate marginal effects, being able to coherently present the chosen estimation methods and empirical results is a skill for which the benefit cannot be underestimated. With this fact in mind, Chapter 16 provides tips for performing and writing up an empirical research project.

Problems

15.1 How do quantile regression estimates differ from OLS estimates?

15.2 Explain how the following situations suffer from sample selection bias.

a. Data on the salaries of people in the labor force.

b. Telephone survey respondents.

c. Number of years of education.

d. Drug trials for a lethal disease.

15.3 Compare and contrast difference-in-differences estimation to panel data techniques.

15.4 Given how difficult it is to estimate true marginal effects in economics, why aren't quasi-experimental techniques used more often?

Exercises

E15.1 Use the file California Data.xls for the following questions.

a. Regress Loan Amount on income, black, other race, and male. Comment on the results using OLS.

b. Estimate the same model using quantile regression at the 50th percentile. How do your results in parts (a) and (b) differ?

c. Estimate the model at the 10th, 25th, 75th, and 90th percentiles. Do these results differ from the results you found in parts (a) and (b)?

d. From a policy perspective, why do you think it is important to not just focus on average values but also consider the marginal values at different points in the distribution?

E15.2 Use the file Hunting Data.xls, which is collected from the 2006 National Survey of Fishing, Hunting, and Wildlife Associated Recreation, for the following questions.

a. Regress Hunt Days on male married, divorced, somecol, hsonly, ba, somepost, postgrad, black, other, income, and south central. Comment on the results. Why is OLS not the appropriate estimation strategy in this circumstance?

b. Estimate the same model as in part (a) with a Poisson model. Comment on the results.

c. Estimate the model with a negative binomial model. Test for over-dispersion. Do your results from parts (b) and (c) differ?

d. Which model do you think is the most appropriate to use, and why?

E15.3 Use the file , which is collected from the 2003 National Health Interview Survey, to answer the following questions.

a. Estimate a Heckman selection model in order to identify factors that affect the number of cigarettes smoked per day. The identification variable to be used in the first stage is dadattack (if the person's father had a heart attack) while the dependent variable is cigperday (number of cigarettes smoked per day), and the explanatory variables are black, Hispanic, male, age, family size, married, and divorced. Comment on the results.

b. Estimate the same model with OLS and compare and contrast the results with those you found in part (a).

E15.4 Using the file Suicide Data.xls, which is collected from the National Center for Health Statistic's 1998 mortality data, estimate a difference-in-differences model to estimate the effects of a policy change in the assisted suicide law in Oregon in 1997 in two ways.

a. First, estimate the sample mean of the number of suicides in Oregon in 1996 and the number of suicides in Washington in 1996, and then subtract the Washington mean from the Oregon mean. Second, estimate the sample mean of the number of suicides in Oregon in 1998 and the number of suicides in Washington in 1998, and then subtract the Washington mean from the Oregon mean. Finally, subtract the value from the first step from the value in the second step. This is the difference-in-differences estimator. Comment on the results.

b. Estimate the difference-in-differences estimator using a regression. Regress the number of suicides in Oregon, the year 1998 and Oregon $\times$ 1998. Verify that the difference-in-differences estimator is the same as what you found in part (a). Comment on the results.

c. From a policy perspective, why is the method employed in part (b) preferable to the method employed in part (a)?

d. Explain how this method does a better job of estimating the true effects of the policy change in Oregon than if only the 1998 data had been used.