Business 41903

Instructor: Christian Hansen

Problem Set 2

1. Carefully evaluate the following comment. "I'm interested in estimating a model for the quantity of demand for widgets as a function of price, and I have some variables that tell me about the costs of producing widgets that I know are unrelated to demand shocks. I've estimated my model using OLS and 2SLS and have found that OLS produces better within sample forecasts and that the OLS forecasts are also better within a holdout sample I set aside before estimating the model. I also know that if a model doesn't forecast well, it's not a good model; so I'm going to use the OLS estimates to gauge the possible effects on quantity sold of implementing a 10% increase in price."

2. Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns: $\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u$ where $male$ is an indicator which is one if the child is male, $parity$ is the birth order of this child, $faminc$ is family income, and $packs$ is the average number of packs of cigarettes smoked per day during pregnancy.

a. Why might you expect $packs$ to be correlated with u?

b. Suppose that you have data on average cigarette price in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for packs.

c. Use the data in BWGHT.raw to estimate the model. First, use OLS. Then use 2SLS, where *cigprice* is an instrument for packs. Discuss any important differences in the OLS and 2SLS estimates.

d. Estimate the reduced form for *packs*. What do you conclude about identification of the model with *cigprice* as an instrument for *packs*? What bearing does this have on your answer in c.?

3. Use the data in CARD.raw for this problem.

a. Estimate a log(*wage*) equation by OLS with *educ*, *exper*, $exper^2$, *black*, *south*, *smsa*, *reg*661 through *reg*668, and *smsa*66 as explanatory variables. Are there important differences between inference based on heteroskedasticity consistent standard errors and the usual (homoskedastic OLS) standard errors? What assumptions would justify the use of one versus the other?

b. Estimate a reduced form equation for *educ* containing all explanatory variables from part a. and the dummy variable *nearc*4. Do *educ* and *nearc*4 have a practically and statistically significant partial correlation? Is there a difference between inference based on heteroskedasticity consistent standard errors and usual standard errors?

c. Estimate the log(*wage*) equation by IV, using *nearc*4 as an instrument for *educ*. Compare the 95 percent confidence interval for the return to education with that obtained from part a. Use heteroskedasticity consistent standard errors.

d. Now use *nearc*2 along with *nearc*4 as instruments for education. First estimate the reduced form for *educ*, and comment on whether *nearc*2 or *nearc*4 is more strongly related to *educ*. Does it look like there is a stronger relationship available using both instruments? How do the 2SLS estimates compare with the earlier estimates? Continue to use heteroskedasticity consistent standard errors.

e. For a subset of the men in the sample, IQ score is available. Regress *iq* on *nearc*4. What does this suggest?

f. Now regress *iq* on *nearc*4 along with *smsa*66, *reg*661, *reg*662 and *reg*669. How does this compare to the result in part e.? What conclusions would you draw?

4. [Simulation Exercise] Consider data generated from the following model:

$$y_i = x_i\beta + u_i$$

$$x_i = z_i\pi + v_i$$

where $z_i \sim N(0,1)$, $u_i \sim N(0,1)$, $v_i \sim N(0,1)$, $E[u_i v_i] = .9$, and $\beta = 1$. Consider 100 observations. Consider $\pi = 2$, $\pi = 1$, $\pi = .25$, and $\pi = 0$. Use 1000 simulation replications and estimate β using IV.

a. What does the distribution of first stage t-statistics in each case look like? What is the simulation rejection frequency of $H_0 : \pi = 0$ using a 5% level test in each case?

b. What does the distribution of the IV estimator of β look like in each case? Does it appear that the asymptotic normal approximation provides a good approximation to the finite sample distribution? What is the simulation size for a 5% level test of $H_0 : \beta = 1$ using the usual asymptotic approximation in each case? What is the simulation mean of $\widehat{\beta}$? What is the simulation median of $\widehat{\beta}$?

c. Suppose you only ran IV when the first-stage "looked" strong, i.e. when you would reject $H_0 : \pi = 0$ at the 5% level. What is the size of 5% level tests of $H_0 : \beta = 1$ using only the set of replications in which you reject $H_0 : \pi = 0$ at the 5% level for each case? What is the simulation mean of $\widehat{\beta}$ within this set of replications? What is the simulation median of $\widehat{\beta}$ within this set of replications?

d. Another rule of thumb people have advocated is to use 2SLS only when the first-stage F-statistic is greater than 10. Since we are only considering one instrument, this corresponds to when the first-stage t-statistic squared is greater than 10. (With one variable $F = t^2$.) What is the size of 5% level tests of $H_0 : \beta = 1$ using only the set of replications in which $t^2 > 10$? What is the simulation mean of $\widehat{\beta}$ within this set of replications? What is the simulation median of $\widehat{\beta}$ within this set of replications?