Problem Set 6 (Introduction to Instrumental Variables)

ECON 441, 2019 Fall

Please solve all problems below, and do not forget to submit the Stata outputs somehow (copy here, print it separately, or upload it as an attachment to the submission on Canvas). This is NOT a group exercise, so everyone needs to write up their own solutions.

1 Wooldridge, Chapter 15, Exercise 1 (20 pts)

Consider a simple model to estimate the effect of personal computer (also called 'PC') ownership on college grade average for graduating student at a large public university (in the mid-1990s):

$$GPA = \beta_0 + \beta_1 PC + U. \tag{1}$$

(Here the 'PC' is a dummy variable indicating if the student owned a PC. GPA is the grade point average.)

1. Why might PC ownership be correlated with U? (That is, give a factor that would cause endogeneity.)

2. Explain why PC is likely related to parents' annual income. Does this mean parental income would be a good (=valid) instrument for PC? Why or why not?

3. Suppose that four years ago the university gave grants to buy computers to roughly one-half of the students, and the students who got the grant were randomly chosen (by a lottery, for example). Explain how you would use this information to create a valid instrumental variable for PC. (Give a definition of a variable and explain why it would be a valid IV in your opinion.)

2 Wooldridge, Chapter 15, Exercise C2 (25 pts)

Use the FERTIL2 data set for this question. Estimate the model

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + U \tag{2}$$

by OLS.

1. Interpret the coefficient on *educ*. (You have endogeneity.) If 100 women receive 1 year more education, how many fewer children are they expected to have (as a group)?

2. The variable *frsthalf* is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that this variable is uncorrelated with the unobservable term (U; the exclusion restriction), show that *frsthalf* is a valid instrument. (Give at least code, statistic that you are looking at.)

3. Estimate the model from the first part now with using *frsthalf* as an instrument for *educ*. (Code, output/estimates with White se-s + n and R^2) How does the effect of education compare from this estimator compared to what we found in the first part? (Give a sentence of interpretation.)

4. Add the variables $electric, tv$ and *bycicle* the regression model, and assume that these variables are exogenous. Estimate the model with OLS and 2SLS. Compare the estimated coefficients on *educ*. Interpret the coefficient on tv, and explain why television ownership has a negative effect on fertility.

3 Wooldridge, Chapter 15, Exercise C10 (30 pts)

We use the data set HTV.dta for this question.

1. Run a simple OLS regression of log(*wage*) on *educ*. Without controlling for other factors, what is the 95% confidence interval for the (average) return to another year of education in the sample?

2. The variable *ctuit*, measured in thousands of dollars, is the change in college tuition facing students from age 17 to age 18. Show that *educ* and *ctuit* are essentially uncorrelated. What does this say about *ctuit* as a possible instrument for *educ* in a simple linear model? (Hint: What 2 conditions do we need for the validity for *ctuit* as an IV?)

3. Now add to the simple regression model in the first part a linear and a quadratic term in experience, a full set of regional dummy variables for current residence and residence at age 18. Also include the urban indicators for current and age 18 residences. What is the estimated return for an additional year of education now?

4. Again using *ctuit* as potential IV for *educ*, estimate the reduced form of *educ*, by which we mean the first stage regression. Do not forget to include all exogenous variables from the RHS of the second stage along with the potential instrument. Show that now *ctuit* is statistically significant in the first stage for *educ*. (Give the estimates with standard errors as usual, and reference the test you conducted.)

5. Estimate the model from part c) by using *ctuit* as an instrumental variable for *educ*. How does the 95% confidence interval for the coefficient of *educ* change (Report the results again; say if the CI is wider and where the center shifted, if any direction.)

6. Do you think the IV procedure from the previous part is convincing?

4 Wooldridge, Chapter 15, Exercise C12 (25 pts)

Use the data in CATHOLIC to answer this question. The model of interest is

$$math12 = \beta_0 + \beta_1 cathhs + \beta_2 faminc + \beta_3 motheduc + \beta_4 fatheduc + U, \qquad (3)$$

where *cathhs* is a binary indicator for whether a student is attends a Catholic high school.

1. How many students are there in the sample? What is the percentage of kids going to Catholic high school?

2. Estimate the above equation by OLS. (Report your results.) What is the estimate of β_1? What is its 95% confidence interval?

3. Using *parcath* as an instrument for *cathhs*, estimate the reduced form (= first stage) regression for *cathhs*. What is the t or $F-$ statistic for *parcath* in this regression? Do you think we have a weak-instrument problem?

4. Estimate the above equation using the *parcath* as an instrument for *cathhs*. (What

is the name of the estimator?) How does the estimate and the 95% CI changes compared to the OLS estimate? (Interpret the coefficient, don't just say 'lower'/'higher')

5. Test the null hypothesis that *cathhs* was exogenous at the first place. What is the p-value of the test. (You may assume homoskedasticity.)