ECO 348K, Problem Set #5, due 4/8/22. Please submit answers + Stata code + .log files on Canvas.

Q1. Open the data LotteryExample.dta that I posted on Canvas (note: LotteryExample_s13.dta can be opened in Stata 13 and is posted under the [data] folder). For this dataset, I created 100,000 fake student records that include the following variables: parent_income, ability, lottery, attends, and test_score. They are defined as follows:

parent_income: total parent income earned in the year before charter school admission.

ability: a measure of the student's intellectual abilities.

lottery: equal to one if the student won the admissions lottery, and it equals zero if not.

attends: equal to one if the student attends the charter school, and it equals zero if not.

test_score: student's standardized test score at the end of 12^{th} grade.

Finally, when I created this data, I explicitly coded the variable **test_score** as follows:

*g test_score = rand_test + 5*attends + 10*(parent_income/80000) + 20*ability*

where **rand_test** is a normally distributed random number with mean 50 and standard deviation 10.

You are going to use this data to estimate the effect of attending a charter high school on the student's standardized test scores at the end of high school. Students can be admitted to the charter high school in one of three ways: (1) their parents pay for admission, (2) they score sufficiently high on an entry exam, or (3) they win an admissions lottery. This means that **parent_income**, **ability**, and **lottery** all have a positive direct effect on **attends**. Assume that these three variables (**parent_income**, **ability**, and **lottery**) do not directly affect each other. Finally, assume that **attends**, **parent_income**, and **ability** all have a positive direct effect on **test_score**.

a. Assume that **ability** and **parent_income** are **not** observed in the data (even though they are in this dataset for instructional reasons). Draw a DAG that describes the relationship between **attends**, **parent_income**, **ability** and **test_score**. If you regressed **test_score** on **attends**, would your estimate equal the causal effect of charter school attendance on test scores? Why or why not? If there is bias, what direction is the bias?

b. Draw a DAG that describes the relationship between **attends**, **lottery**, **ability**, and **parent_income**.

c. Run a regression of **test_score** on **attends** and report it here. Normally, you don't observe the data generating process and the true coefficient for each variable, but since I created this data, you can observe those things. Compare the estimated coefficient to the true coefficient from the code used to generate the variable test_score (i.e., the data generating process). Is the omitted variable bias positive or negative in this regression?

d. Now, for this question only, assume that you can observe **ability** and **parent_income**. Run a regression of **test_score** on **attends**, **ability**, and **parent_income**. Report that regression here. Discuss how these coefficients compare to the true coefficients (note: the true coefficient on parent_income is 10/80,000).

e. Use the information from (a) and (b) to draw a DAG that illustrates the instrumental variables strategy you could use to estimate the causal effect of charter school attendance on test scores. State the instrument that you are going to use for the variable **attends**. State the assumptions necessary for that variable to be a valid instrument if treatment effects heterogeneous.

f. Run an IV regression using the instrument you described in (e) and report the results here. Compare the estimated coefficient to the true coefficient.

g. Estimate the reduced form and first stage regressions and show how these coefficients relate to your coefficient from part (f).

h. In the context of this example, define the following groups: compliers, defiers, never-takers, always-takers. In the context of this example, describe the LATE. Assume that there are no defiers.

Q2.

a We read a paper in this course called, "The Returns to College Admission for Academically Marginal Students." In that paper, the author uses an IV regression to estimate the local average treatment effect of receiving a bachelor's degree on earnings. First, describe the instrument that he uses. Second, in the context of the paper, describe the assumptions necessary for that variable to be a valid instrument if treatment effects are homogenous. Finally, discuss whether it is likely to satisfy those underlying assumptions for a valid instrument.

b. Earlier in the course, we discussed Project STAR, an experiment that randomly assigned elementary school students to large or small classes. One issue in this experiment was non-compliance. A small fraction of students assigned to small classes actually ended up in large classes and vice versa. Assume that this is the only issue with the experiment. How could you use IV to estimate the local average treatment effect of attending a small classroom? Describe the regression(s) that you would run, the necessary assumptions, and whether your instrument is likely to satisfy those assumptions. Assume homogenous treatment effects.

Stata guide:

These are the Stata commands & functions that I used to get the answers for all of the questions above. Although this is how I did it, you can use any commands you see fit unless stated otherwise in the problem.

reg y x, r – used to run a regression of y on x, using heteroskedasticity-robust standard errors.

ivregress 2sls y (x=z), r – used to run an IV regression of y on x, instrumenting for x with z, and using heteroskedasticity-robust standard errors.