Instrumental Variables Simulation Homework

Jennifer Hill, Ray Lu & Zarni Htet

Objective

The goal of this exercise is to simulate data consistent with the assumptions of the IV estimaator we discussed in class (and described in the Angrist, Imbens, Rubin article posted on the Classes site). We will also evaluate the properties of different approaches to estimating the Complier Average Causal Effect.

Setting

To help conceptualize data that might be consistent with the IV assumptions, we will generate data from a hypothetical randomized encouragment design. In particular, imagine a study in which 1000 students entering an undergraduate degree program in the sciences in a major university were randomly assigned to one of two conditions. One group was encouraged via an email from the chair of their department to participate in a one week math boot camp just before the start of their first semester. Students in the other (not encouraged) group were also allowed to participate but received no special encouragement. In fact they would have had to discover on their own the existence of the program on the university website. The outcome variable is derived from the student test scores on the final exam for required math course for the sciences. In particular the Y variable that you will simulate below represents the *difference* between that score and the threshold for passing. Thus a negative value for a student reflects that the student did not pass.

Part A: Generate and explore the data for the population

In this section you will simulate data consistent with the assumptions. We will generate data for a sample of 1000 individuals.

Question 1: Simulate the data as god/goddess/supreme being of your choice.

(a) Simulate compliance status. Assume that 25% of individuals are compliers, 60% are never takers, and 15% are always takers. Generate D(0) and D(1) vectors to reflect this. You can also generate a vector indicating compliance type, C, if that is helpful to you.

(b) Which compliance group has been omitted from consideration? What assumption does that imply?

(c) Simulate the potential outcomes in a way that meets the following criteria:

(d) The exclusion restriction is satisfied.

(ii) The average effect of Z on Y for the compliers is 4.

(iii) The average Y(Z=0) for never takers is 0; The average Y(0) for compliers is 3; The average Y(Z=0) for always takers is 6.

(iv) The residual standard deviation is 1 for everyone in the sample (generated independently for each potential outcome).

(v) Calculate the SATE (average effect of Z on Y) for each of the compliance groups.

(e) What is another name for the SATE for the compliers?

(f) Calculate the ITT using your simulated data.

(g) Put D(0), D(1), Y(0), Y(1) into one dataset called dat.full. (You can also include a variable, C, indicating compliance group if you created one.)

Question 2: Playing the role of the researcher to randomize

Now switch to the role of the researcher. Pretend that you are running the experiment that we are examining for this assignment. Generate a binary indicator for the ignorable treatment *assignment* (as distinct from treatment receipt.... so this is Z, not D). Probability of receiving the treatment should be .5.

Question 3: Back to playing god

Use dat.full to create a dataset that the researcher would actually get to see given the Z generated in Question 2. It should only have D, Z, and Y in it. Call it dat.obs.

Question 4: Researcher again

(a) *Estimate* the percent of compliers, never takers and always takers assuming that there are no defiers. Use only information in dat.obs.

(b) Estimate the naive regression estimate of the effect of the treatment on the outcome. Which estimand that we discussed in class is this equivalent to?

(c) Estimate the intention-to-treat effect.

(d) Calculate an estimate of the CACE by dividing the ITT estimate by the percent of compliers in the sample.

(e) Estimate the CACE by performing two stage least squares on your own (that is without using an IV function in the R package AER).

(f) Provide an estimate of CACE and its standard error using the ivreg command.

(g) Simulate a sampling distribution for the estimator used in (f). Is the estimator unbiased? Also report the standard deviation of the sampling distribution and compare to the standard error in (f). [We are back to God mode here – apologies for any confusion!]

Question 3: Exploring the connection between the DGP and the assumptions.

Now we're back to the role of god of Statistics.

(a) Describe the assumptions required to obtain and unbiased estimate of the treatment effect, as described in AIR. We have generated data that satisfy these assumptions. Suppose instead you were handed data from the study described above. Comment on the plausibility of each of the required assumptions in that setting.

(b) Suppose that the data generating process above included a covariate that predicted both Z and Y. Which of the assumptions described in (a) would that change and how?

(c) Suppose that the directions for Q1.c.iii was amended as follows " (iii) The average Y(0) for never takers is 0; The average Y(0) for compliers is 3; The average Y(0) for always takers is 6. The average Y(1) for never takers is 2." Which of the assumptions described in (a) would that violate?

(d) Redo one of the commands from Question 1 (just provide the code – you don't have to run it) such that the monotonicy assumption is violated.

(e) How could we alter the study design to preclude the existence of always takers? Would this be ethical?