Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política
FLP-0468 & FLS-6183
2º semestre / 2018

Problem Set# 3

Please submit a write-up and do file at the beginning of class on October 4, 2018. Please hand in the problem set printed and only send the do-file.

1. For the first exercise, we will return to the first and second cases we analyzed last week. In case 1, we will suppose X and Z are not correlated. In case 2, we will suppose that X and Z are correlated such that corr(X,Z)=0.75 as are the variables that are generated in the simulation in part b of the do file such that:

$$Y \sim (100, 20)$$
$$X \sim (7, 8)$$
$$Z \sim (20, 2)$$
$$Corr(Y, X, Z) = \begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0.75 \\ 0.3 & 0.75 & 1 \end{pmatrix}$$

   Based on the formulas for Omitted Variable Bias (OMB) suggested in Mastering Metrics (MM) calculate the OMB when Z is omitted in Case 1 and Case 2. Please make sure to try to replicate the exercise and use the formulas as outlined in MM.

2. Now, let's build on the examples outlined in MM Chapter 2. For a sub-sample that is smaller than the entire number of cases (n<500), let's assume we have a "within group comparison" such as the case suggested in Table 2.2. Please run the regressions and report them in a separate table (e.g. one table for case 1 and one table for case 2). How do these models compare to the models estimated in 1? (Hint: The first step is to create some type of variable that allows you to select a subsample and within that subsample to have sub-groups).

3. Now, let's continue to build on the example outlined in MM Chapter 2. To do so, let's introduce three group dummies in such a manner that the group dummies are correlated with Z (our treatment effect) in the entire sample (n=500). Please estimate a table similar to the "self-revelation" model reported in Table 2.3 (one table for case 1 and one table for case 2). Can you re-run the calculations above and observe what the OMB is from excluding Z? (Hint: The first step is to create some type of variable to identify the different groups).
4. Table 1.1 in MM compares the health and demographic characteristics of insured and uninsured couples in the NHIS. Execute the Stata code in NHIS2009_hicompare.do through line 35 to make sure that you use the same selection criteria that were used to produce Table 1.1.

a. Panel A compares the health across husbands in this sample with and without health insurance. Use the sum command to calculate average health separately for husbands with and without health insurance. What is the difference in average health by insurance status? Is this difference statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

b. Panel B of Table 1.1 shows that husbands with and without health insurance differ along many demographic dimensions. It is possible that the difference in health between the "Some HI" and "No HI" groups may be smaller if we compare across groups that are more homogeneous. To investigate this, please test if the difference between the health of husbands with Some and No HI significantly different from zero if you restrict to men who:

   i. are employed?

   ii. are employed and have at least 12 years of education?

   iii. are employed, have at least 12 years of education, and earn income of at least $80,000?

c. Use the NHIS data to construct a variable such that a regression of health on this variable reproduces the difference calculated in question (a), above. Compare the difference, t-statistic, and confidence interval for your regression estimate of differences in health with those you computed in (a).

d. In (b) of this Problem Set, we showed that some of the difference in average health between those with and without health insurance in the NHIS can be attributed to the fact that the insured differ from the uninsured along many relevant dimensions. We can also show this using regressions. Starting with your regression from part d above, sequentially add controls for age (age), years of education (yedu), and income (inc). Does any set of controls eliminate the difference in health between insured and uninsured? Explain how the results change as you add controls and what changes in the estimates as you add more controls might mean.

5. Regression application: The effects of class size. The Angrist data archive (http://econ-www.mit.edu/faculty/angrist/data1) contains data from the following article (posted on Stellar): J. Angrist and V. Lavy, "Using Maimonides Rule to estimate the Effect of Class Size on Student Achievement," The Quarterly Journal of Economics, May 1999. This article uses the fact that Israeli class size is capped at 40 to estimate the effects of class size on test scores with an Instrumental Variables / Regression Discontinuity research design. But for now, we'll use the data to explore regression basics.

   (a) Read the article through Section I (at least), download the data, and construct the descriptive stats in Table 1 for 5th graders. From here you should be able to mostly tell

what's what as far as variable names go (note that the unit of observation is the class average). Note that enrollment is called c_size and percent disadvantaged in called tipuach. To exactly reproduce the numbers in Table 1, you must follow footnote 11 and restrict the sample to schools with enrollment of at least 5 and classes of size less than 45. There are also a couple of non-obvious data corrections. There is an average math (avgmath) score and an average verbal (avgverb) score greater than 100 due to a data entry error. The correct values of these scores are 87.606 and 81.246 (not 187.606 and 181.246). Finally, there is a non-missing math score for an observation with mathsize==0 (i.e. no math test takers). This is impossible. Replace avgmath=. if mathsize==0.

(b) Economists and educators have long debated whether it's worth paying the extra labor costs (i.e., teachers' wages) required to reduce class size. What should the sign of the achievement/class-size relationship be if the investment is worthwhile?

(c) Regress average math and verbal scores on class size. What is the sign of this relationship? Is it significantly different from zero? How does it look so far for the class size optimists?