

FLS 6415: Class 2 Homework

August 24, 2017

Remember to answer all the questions in R markdown and produce a PDF. Email your completed homework (R markdown file and PDF) to jonnyphillips@gmail.com by midnight the night before class.

First, some quick review questions about causal inference:

1. In one sentence, describe the fundamental problem of causal inference.
2. We want to investigate the effect of attending university on support for redistribution. How might we define the counterfactual/control?
3. Based on your knowledge of the Brazilian education system and society, describe the actual treatment assignment mechanism that applies to receiving the ‘treatment’ of going to university.
4. The data below were collected from a randomized controlled trial that sent some high-school graduates to university but not others. By magic we have data on *both* potential outcomes for each student. Calculate (a) the average treatment effect, (b) the average treatment effect on the treated, and (c) the average treatment effect on the untreated:

Unit	D_i	Y_{1i}	Y_{0i}
A	0	7	6
B	1	8	4
C	0	6	4
D	1	6	6
E	0	4	5

5. Now imagine we only have observed outcomes and do not see both potential outcomes. Using only **observed** outcomes, calculate the estimated average treatment effect under the assumption that the data was generated from a randomized treatment assignment mechanism.
6. Is it surprising that the estimated average treatment effect in Q5 different from your estimate of the average treatment effect in Q4? Why might they differ?
7. We later find out that unit E ignored the randomized treatment assignment and was actually supposed to have taken the treatment, i.e. they should have gone to university but did not. Re-calculate the average treatment effect using the observed outcomes if Unit E had complied with the randomized treatment assignment and attended university.
8. As an estimate of the true causal effect of university on support for redistribution, do you trust your answer to Q5 or Q7 more? Why?

The second part of the homework is another exercise to increase your fluency in R. It focuses on the political economy causal question of whether institutions cause growth using data from Glaeser et al 2004, “Do Institutions Cause Growth?”. The dataset is an incomplete panel covering 127 countries for four decades from 1960 - 2000. The measure of ‘institutions’ is “Initial executive constraints”, a measure of checks and balances on government.

9. Download the XLS dataset “Do Institutions Cause Growth?” from <http://faculty.tuck.dartmouth.edu/rafael-laporta/research-publications>. Save the “data for table 5” sheet of the XLS file as a CSV. Start a new R markdown file and import the CSV.
10. Create a simple table showing the average global growth rate for each decade (simple average across countries).
11. Create a simple (labelled) chart showing institutions (executive constraints) over time for the following countries: Argentina, Denmark, Dominican Republic, Nigeria and Peru. Create a separate chart showing

growth over time for the same countries.

12. Create a simple (labelled) chart plotting institutions against growth for all the data points in the dataset. Add a linear line of best fit.
13. Conduct a linear regression of the following form: $\text{Growth} \sim \text{Institutions}$. Report the table of variables, coefficients and p-values. Interpret the coefficient and p-value on institutions.
14. We might be worried about an omitted variable that explains both growth and institutions. Add one at a time the following variables to the regression: Log initial GDP per capita, Share of population living in temperate zone, Log initial years of schooling. How does each control variable change the interpretation of the regression results? What happens if you include all the control variables in the same regression?
15. We might be worried that ‘treatment’ here is not independent because we’re repeatedly measuring the same country. Add fixed effects (dummy variables) for each country to the regression (excluding the control variables). How does this change the interpretation of the regression results?
16. (Slightly harder) We might be worried about reverse causation so that institutions explain growth. One rough way of testing this is to see whether growth in, for example, the decade 1960 is related to subsequent institutions in 1970. Adjust your dataset so that each row contains the growth for a particular decade with the institutions at the *end* of that decade, and conduct a regression of $\text{Institutions} \sim \text{Growth}$. (Hint: Use the ‘mutate’ function in dplyr to reorganize the data).
17. Do the regressions in Q13 (and Q16) provide estimates of causal effects? Why or why not?