14.771 Problem Set 1

Nils Enevoldsen

Question 1

1. Suppose a colleague suggests simply running a regression of years of education obtained on the density of schools. Why might this be problematic? List some factors you think may bias this estimate and discuss the direction of bias you would expect from each factor.
2. Suppose a colleague suggests simply running a regression of adult outcomes (say, wages) on years of education. Why might this be problematic? List some factors you think may bias this estimate and discuss the direction of bias you would expect from each factor.

1 Question 2

1. We will consider differences in outcomes of individuals based on where they were born, rather than where they were educated. Explain why this is necessary. In what situation might this strategy be problematic? Discuss how this could constitute a source a bias.
2. We will consider differences in outcomes of individuals based on where they were born, rather than where they were educated. We will also focus on the sample of individuals for whom wage data are available (where applicable, the paper typically presents results for both the entire sample and this subset; this is clearly not possible for all of the results related to wages). Discuss why each of these is necessary, and how each could constitute a source of bias.

2 Question 3

Now, we will try to replicate some of the results from the paper. First, define the "young" as those born between 1968 and 1972, who were exposed to the INPRES school construction program the entire time they were in primary school (they were 2 to 6 years old in 1974). Define the "old" as those born between 1957 and 1962, who should have had little or no exposure to the program (they were 12 to 17 in 1974). For the moment, ignore the "very old" who were born before 1957. Next, define "high" program areas as those in which the residual of a regression of the number of schools on the number of children is positive, and all other regions as "low" program areas. [Note: This program intensity variable is already generated for you in the dataset, and is labeled "recp".]

1. Generate a new variable - yeduc_diff - which is the difference in the average years of education obtained between those born in high and low program areas.

   (Hint: Collapse your data to do this; make sure to limit your sample and weight appropriately).

   Plot this difference on the y axis with cohorts (i.e., birth year) on the x axis. Add separate fitted regression lines for the old and young cohorts. Comment on the plot. How does this pattern fit with

the author's argument?

(Hint: Use Stata's "twoway lfit" command to add regression lines to your plot).

1. Calculate the average difference in years of education obtained between the high and low program areas for the old and the young. Subtract your measure for the old from your measure for the young. Provide standard errors for these differences. Repeat the procedure for the log of hourly wages. What do we call this estimator?

   (Hint: Your results should match those found in Panel A of Table 3. Don't forget to limit your sample and weight appropriately).

2. What do these results tell us? Briefly explain what identification assumptions are necessary for these estimates to be valid. What is your assessment of the validity of the identification assumptions in light of the plots you produced in question 3.1?

3. Now define one more group - the "very old" - born before 1957 (in practice, because we only have data beginning in 1950, this encompasses individuals born between 1950 and 1956, who were thus 18 to 24 years old in 1974). Repeat the above analysis from 3.2 and 3.3, comparing the old and very old. Why is this exercise useful? What should you see if the identifying assumptions hold? Comment on your plot.

   (Hint: Don't forget to limit your sample and weight appropriately).

3 Question 4

The results obtained in Question 3 are suggestive, but not statistically significant. The author then uses a broader set of data and a regression framework to try to obtain more precise results.

1. Discuss the results in Table 4.

2. The author includes sets of specifications with additional controls. Why? What do you think about the results? Do you see any potential problems with any of the controls?

3. Discuss (but don't replicate) equations (2) and (3), as well as the corresponding empirical results (Figure 1 and Table 5, respectively). Who is the control group in each case? Does the restriction placed in equation (3) seem valid? Why does the author impose this restriction?

4 Question 5

1. After attempting to establish that the school construction program had an impact on both education and wages, Duflo makes the case that it also constitutes an exogenous shock to educational attainment (i.e., it can be used as an instrument for schooling), and thus can be used to estimate the returns to education. Explain the logic behind her argument. What exactly is the instrument? What are some possible concerns with this instrument? How does she try to address these concerns?

2. Aggregate the data by cohort (young and old) and region of birth. Run a regression of (a) education on school construction program intensity and (b) wages on school construction program intensity. Calculate the ratio of (b) to (a). What is this called? Discuss how the results compare to Table 7.

3. Does the IV make a large difference in the results? What does this imply?

4. What is the interpretation of an IV such as this, where the treatment takes on more than one value? (What exactly is the treatment in this case?) How does that impact our interpretation of estimates for the returns to education?

5. What is the author's conclusion and what channels of causation does she discuss? Can you think of any other explanations for the results - both causal or due to bias? How would you test your theories if you had access to unlimited data?

5 Question 6

1. What do you think about the magnitude of the author's estimates - are they small, large? Do they seem reasonable?
2. As documented in this paper, the INPRES school construction program led to an increase in educational attainment in Indonesia. Discuss some potential medium-run consequences this might have (for instance, in the labor market).