Problem Set 2

PPPA 8022
Due in class, on paper, March 7, 2018

Some overall instructions:

- Please use a do-file (or its SAS or SPSS equivalent) for this work – do not program interactively!
- I have provided Stata datasets, but you should feel free to do the analysis in whatever software you prefer. If you need to transfer to another format, use StatTransfer.
- Make formal tables to present your results – don't use statistical software output. Make sure you discuss the answers.
- As an addendum to this problem set, please turn in your .do and .log files, or the equivalent from any other software you use.
- While it is fine (and encouraged) to work with others, your work on this problem set should be your own. This is true for both the write-up and the underlying code.

1. Instrumental Variables

For this problem, we are revisiting a classic: Angrist and Kreuger. We use a random sample (chosen by me) from the 1980 public use micro data file (five percent of long-form respondents; this is the 1980 version of data we used last class). Data are posted on the webpage in full as c1980_ipums_20140220.dta.zip and in a small sample as c1980_ipums_20140220_small.dta.zip. Documentation is at www.ipums.org.

Note that A&K keep only white and black men born between 1930 and 1959. Unfortunately, I didn't include race in my download, so ignore the race restriction.

Some of additional variables are not an exact match. We don't have a continuous education variable like A&K (not sure why not), so make educ into a continuous variable as best you can. We don't have weeks worked, so ignore restrictions relating to that. Use incwage as the dependent variable, rather than weekly earnings.

Get as close to what A&K do as possible, but don't fret over exactly matching all variables (do plan to fret over this, however, when you work on your replication project).

(a) Replicate the first two rows of A&K's Table 1, but don't worry about de-trending the data as A&K do.

(b) Do the A&K first stage, using two sets of instruments: (a) quarter of birth, (b) quarter of birth * birth year. Do the first stage to do the analysis in Table 5, column 8. Make a table to report birth order coefficients from specification (a) and the F for the instruments and the additional R2 from the instruments for both specifications. Interpret whether these instrument seem "good" in a weak instrument sense.

(c) Use your previous specification to make two predicted values for education (one based on specification (a) and the other based on specification (b)). Without ivregress, use these predicted values to estimate a second stage coefficient. Then do a parallel 2SLS analysis using Stata's ivregress (or the equivalent). Compare the coefficients and errors on the variable of interest. What are your findings about education? Why are the coefficients and errors the same or not?

2. Regression Discontinuity

We now turn to data from the 1940 census. Documentation for these data is at www.ipums.org. These data are saved on the course website as ipums1940_20150212.dta.zip. Let's see if the compulsory schooling laws and Angrist and Krueger highlight are amenable to a regression discontinuity analysis.

(a) Using the compulsory school law dates noted on this website (http://www.infoplease.com/ipa/A0112617.html – this is ok for a problem set; for an actual research article, you'd need the real source!) choose a state. I recommend a state with a large population and relative early adoption. Make a regression discontinuity chart where year of birth is the running variable. Please make two charts: one that tells us whether, for the population as a whole, we see a discontinuity in completed education and one that tells us whether we see a parallel discontinuity in income (incwage).

Don't weight observations, and watch out for top codes. I suggest you use the variable "bpl" for the most likely state at time of education.

(b) Can you think of a sub-group where we might be more likely to see a discontinuity? Explain what subgroup that is and why, and replicate your results from (a) using that subgroup.

(c) Write a regression equation that tests whether there is a statistically significant difference in income at the threshold.

(d) Estimate this regression and present the key result in a well-labeled table.

(e) There are many reasons to be suspicious about your results (look at your picture and think about it). Is there a way in which you might want to restrict the sample in the previous regression(s)? Do so, and add to the table in part (d).