Problem Set 3
Due Lecture 8, March 4

PPPA 8022
Spring 2026

Some overall instructions

- Please use a do-file (or its R, SAS or SPSS equivalent) for this work. Do not program interactively. While interactive programming may seem faster at first, inevitably you find mistakes and lose track of edits to the data – and it is slower.
- Turn in a typed up set of answers that answers the questions below. Also turn in a Stata .do file and its associated .log file or the equivalent in whatever software you are using.
- Use this link to submit the problem set. You need to be logged into GW email to submit. If it still gives you trouble, open an incognito window, log in and try again.
- Make formal tables to present your results. Do not present statistical software output.
- I have provided Stata datasets, but you should feel free to do the analysis in whatever software you prefer. If you prefer R, use the `haven` package and the `read_dta()` function.
- 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
- If the question is insufficiently clear, explain the assumptions you made to reach your final estimates.
- Data are
  - Question 1
    - big
    - small
  - Question 2: [here]

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. The full sample, from which I choose the smaller random sample, is five percent of long-form respondents (who are about 1 in 6 residents); these data are the 1980 version of data we used last class. Data are linked

at the top of the problem set. 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), so make `educ` into a continuous variable as best you can. I did the following

```
** make education into a continuous variable **;
gen yrs_educ = 0;
replace yrs_educ = 0 if educ == 0;
replace yrs_educ = 4 if educ == 1;
replace yrs_educ = 8 if educ == 2;
```

etc... 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. Specifically, you want to replicate the set of rows where the outcome is total years of education, with birth cohorts are as noted. Omit the final column with the F test. Explain whether your results are qualitatively similar or not.

(b) Do two versions of the A&K first stage estimations for the analysis in Table 5, column 8. For the first first stage estimation use quarter of birth as the instrument. For the second first stage use quarter of birth * birth year. Make a table with these estimates, and also include the value of the F test for the instruments and the additional R^2 that comes from the instruments in each specification. Interpret whether these instruments seem "good" in a weak instrument sense.

Specifically, I used covariates

- 1 if in a metro (`in_smsa = 1 if metro == 2| metro == 3| metro == 4`)
- 1 if married (`married = 1 if marst == 1| marst == 2`)
- census region fixed effects
- birth year fixed effects
- age in quarters and age in quarters squared

(c) Now do the second stage estimation associated with each of these first stages by hand (that is, using OLS and not a built-in IV command). For both of the first stage estimations

in part (b), create a predicted value for education ($\hat{X}_1$ and $\hat{X}_2$; see Stata's `predict`). For each first stage, regress the resulting predicted values and the other covariates from Table 5, column 8, on wages. Report the results in a well-labeled table.

(d) Do this same two stage least squares analysis using a built-in IV regression command (this could be Stata's `ivreg2`, which you'll need to install, or Stata's built-in `ivregress`, or the equivalent of your choice). Report the results in a well-labeled table. Compare the coefficients and standard errors on the variable of interest in each specification from (c) and (d). The coefficients should be the same, and your standard errors should differ. If the coefficients are not the same, you have done something wrong! (Hint: check the number of observations in each regression.)

2. Regression Discontinuity

We now turn to data from the 1940 census, linked at the top of this problem set. Documentation for these data is at `www.ipums.org`. 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 (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 relatively early adoption.

For the state you've chosen, and the time of treatment determined by the compulsory schooling law adoption, make a regression discontinuity chart where year of birth is the running variable. Make two charts: one that tells us whether, for the population as a whole, we see a discontinuity in completed education at the time of the law's implementation 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` to measure state at time of education.

(b) 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 charts from (a) using that subgroup.

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

(d) Estimate this regression and present the key result (we don't need all the coefficients) in a well-labeled table.