**Problem Set 2**

PPPA 8022
Due in class, on paper, March 8, 2017

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.

## 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 saved on blackboard 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.*

*(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.*

See columns 1 and 2 of the table at the end. Even without de-trending the data, the results are very similar to A&K's original results. Men born in the first quarter of the year, and to a lesser extent men born in the second quarter, have less education.

*(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 the F for the instruments and the additional R2 from the instruments in each regression; you don't need to report all the coefficients. Interpret whether these instrument seem "good" in a weak instrument sense.*

See columns 3 and 4 of the table at the end. The F-tests for these instruments are in both cases quite low. The F-test value for using three instruments (column 3) is 3.4. This is below levels

that would now be considered acceptable for instrument strength. The F-test value using quarter of birth*birth year is even lower, at 1.4. In both cases, the R2 for the regression increases by 0.001 when I add the instruments. In other words, while the instruments may be individually significant (at least in the first case), they do not explain a substantial amount of the variation in the endogenous variable.

*(c) Use your previous specification to make two predicted value variables for education. Do two A&K second stages, one with each predicted value. 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?*

This regression finds that an additional year of education increases wages by a whopping 17 percent; much larger than the estimates in A&K. This coefficient is significant at the five percent level.

The coefficients using `ivregress` and doing the regression manually are exactly the same – as they should be. Mechanically, the IV coefficient is generated by using the instrumented variable.

However, the standard error for the IV estimation is not correctly calculated using the OLS formula. In addition, the IV standard error should be always larger than the OLS standard error. In my example, the OLS standard error is actually a tad larger (0.081 vs 0.080) than the IV standard error. I expect that this anomaly is driven by rounding errors, since the difference between the values is quite small.

## Table for Question 1

| | Question 2(a) | | Question 2(b): 1st stgs | | Question 2 (c) | | | |
|---|---|---|---|---|---|---|---|---|
| | Table 1, row 1 | Table 1, row 2 | 3 instrumts | bq * birth year | Using predicted value | ivregress | Using predicted value | ivregress |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 1{birth quarter=1} | -0.138*** | -0.070* | -0.102* | | | | | |
| | (0.039) | (0.031) | (0.041) | | | | | |
| 1{birth quarter=2} | -0.098* | -0.047 | -0.103* | | | | | |
| | (0.039) | (0.031) | (0.041) | | | | | |
| 1{birth quarter=3} | 0.018 | -0.005 | -0.012 | | | | | |
| | (0.039) | (0.030) | (0.040) | | | | | |
| Predicted value, | | | | | 0.171* | 0.171* | 0.171* | 0.171* |
| years of education | | | | | (0.081) | (0.080) | (0.081) | (0.080) |
| F test: instruments | 7.629 | 2.453 | 3.407 | 1.356 | | | | |
| p-value of F test | 0.000 | 0.061 | 0.033 | 0.103 | | | | |
| R-squared | 0.000 | 0.000 | 0.034 | 0.034 | 0.056 | 0.070 | 0.056 | 0.070 |
| Observations | 51,162 | 71,816 | 43,163 | 43,163 | 43,163 | 43,163 | 43,163 | 43,163 |

**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 Blackboard 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 two states. I recommend two states with large populations and relative early adoptions. For each state, make a regression discontinuity chart where year of birth is the running variable. Please make two charts per state: one that tells us whether, for the population as a whole, we see a discontinuity in completed education and one that tell us whether we see a parallel discontinuity in income (incwage).*

*You should make four charts in total. 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.*

See pictures at end. Blue dots are for men and pink dots are for women.

*(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.*

I limit the analysis to people with less than a college degree, hypothesizing that those who finish college should not have been much impacted by compulsory schooling laws. This doesn't have much of an effect (at least a visual one) on the estimate.

See pictures at end.

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

See the first equation on page 46 of the Lee and Lemieux article we read for class. You should have written some version of this specification.

*(d) Estimate this regression for your two states (separately or together as you prefer) and present the key result in a well-labeled table.*

See table for part (e)

*(e) You might be suspicious about your results – look at your picture. 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).*

You might be suspicious about your previous results because they show (a) that education declines after the imposition of the compulsory school law, and (b) you have in your sample a lot of people who have yet to complete their education.

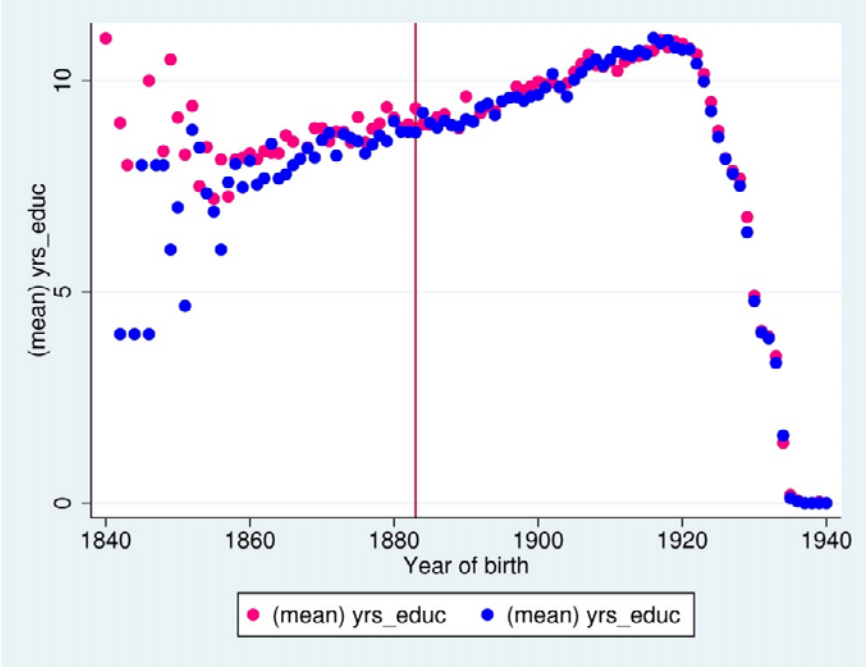
The table below shows results for Illinois and Louisiana, for the equation as in (d), with and without people who are under age 20. The dependent variable is income. We don't really expect to see anything, because we didn't see a discontinuity in education – this part is just to make sure you know how to set up a regression discontinuity equation. And, happily, as this cohort is no more educated, they have no higher incomes (except for one anomaly with Maryland).

Specifically, we are interested in the coefficient on the variable "D" below. It is almost always not significantly different from zero.

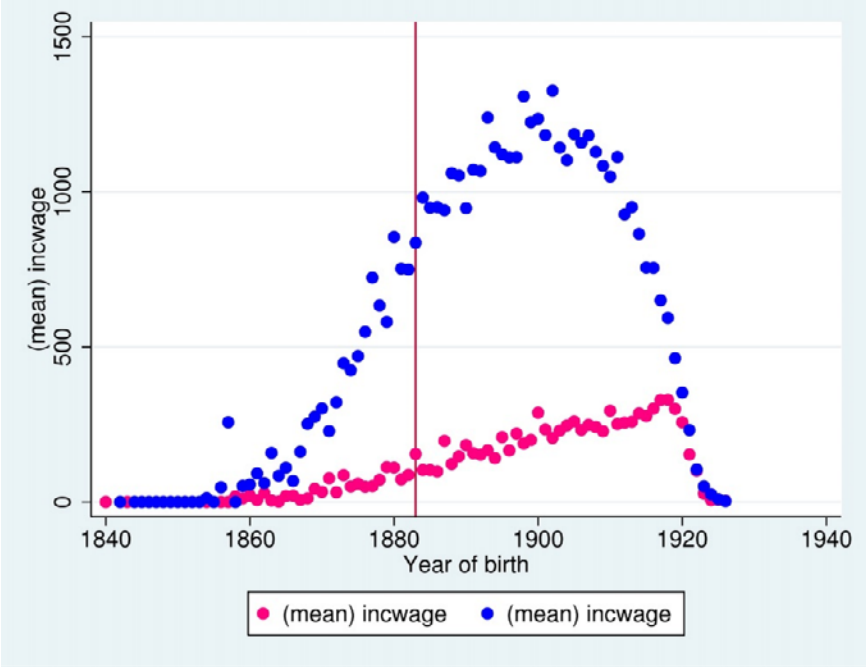
| | Louisiana | | Maryland | | Illinois | |
|---|---|---|---|---|---|---|
| | All ages | Only > age 20 in 1940 | All ages | Only > age 20 in 1940 | All ages | Only > age 20 in 1940 |
| D | -33.054 | -7.602 | 108.911* | 80.488 | 167.721 | 85.563 |
| | (34.93) | (41.30) | (52.29) | (61.40) | (322.11) | (374.68) |
| (X-c) | 15.650* | 15.650* | 9.749 | 9.749 | 0 | 0 |
| | (7.19) | (7.96) | (10.25) | (11.23) | (143.95) | (164.21) |
| $(X-c)^2$ | 0.171 | 0.171 | -0.959 | -0.959 | 0 | 0 |
| | (0.53) | (0.59) | (0.74) | (0.81) | (14.69) | (16.76) |
| $(X-c)^3$ | -0.003 | -0.003 | -0.023 | -0.023 | 0 | 0 |
| | (0.01) | (0.01) | (0.02) | (0.02) | (0.39) | (0.44) |
| (X-c)*D | -1.353 | -8.908 | -12.339 | -1.439 | 47.989 | 65.538 |
| | (8.44) | (10.43) | (12.59) | (15.79) | (144.87) | (165.89) |
| $(X-c)^2$*D | -0.32 | 0.183 | 1.667* | 0.758 | -0.415 | -1.275 |
| | (0.58) | (0.72) | (0.84) | (1.09) | (14.70) | (16.78) |
| (X-c)3*D | -0.006 | -0.016 | -0.004 | 0.016 | -0.008 | 0.003 |
| | (0.01) | (0.01) | (0.02) | (0.02) | (0.39) | (0.44) |
| Observations | 18,457 | 14,742 | 12,888 | 10,541 | 2,226 | 1,673 |

1(a): Illinois

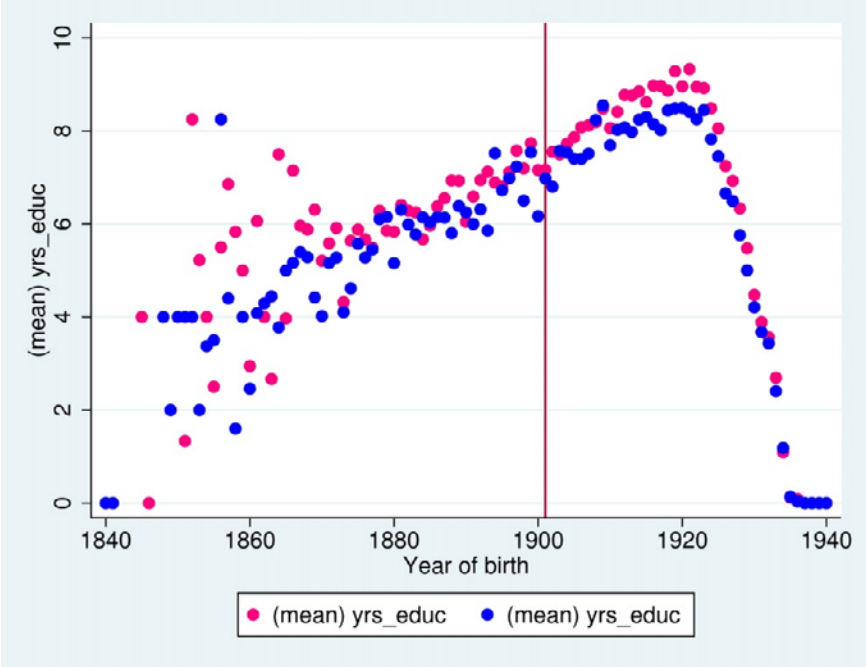
(a) Dependent Variable is Years of Education



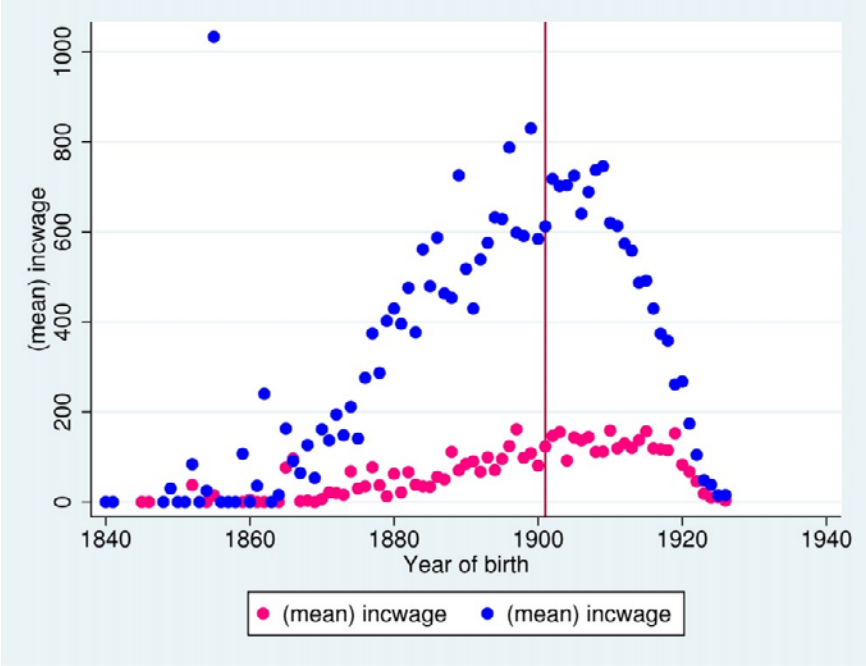

(b) Dependent Variable is Income



1(a): Louisiana
(a) Dependent Variable is Years of Education

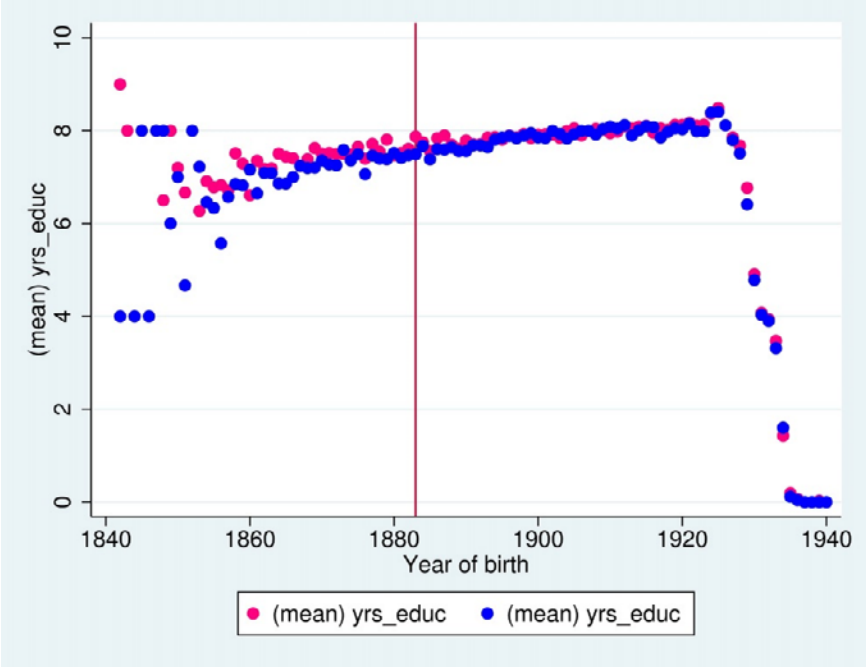



(b) Dependent Variables is Income

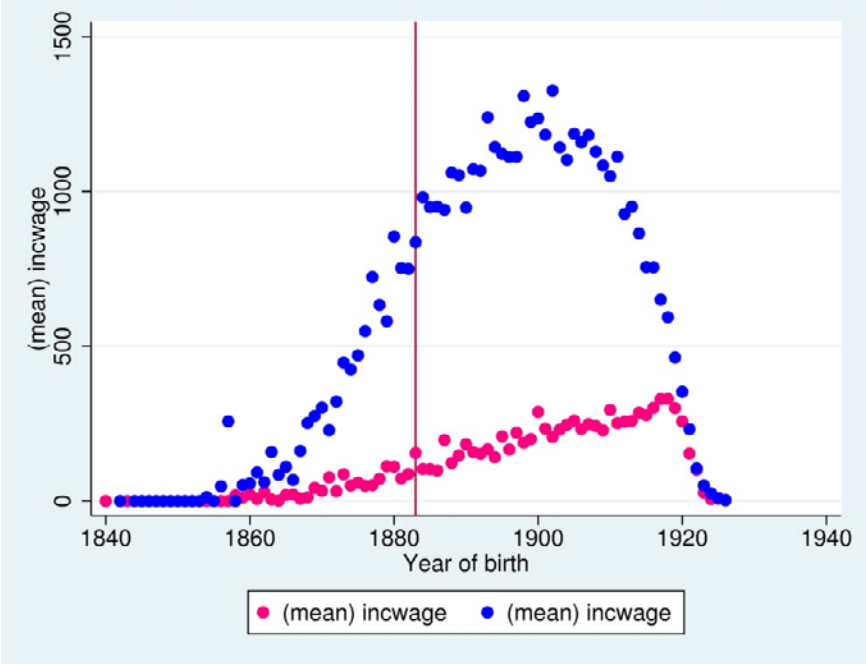


1(b): Illinois

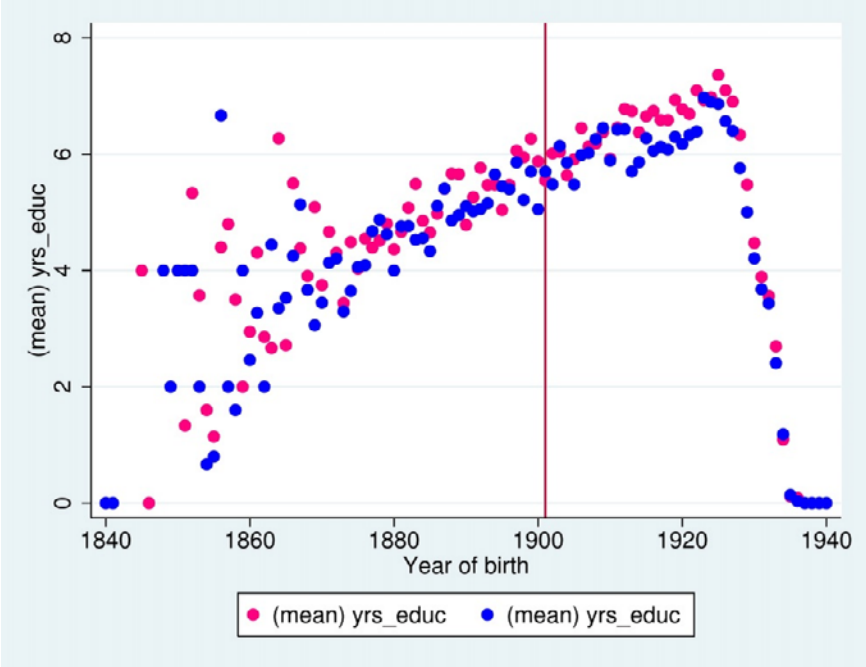
(a) Dependent Variable is Years of Education




(b) Dependent Variable is Income



1(b): Louisiana
(a) Dependent Variable is Years of Education




(b) Dependent Variable is Income