Problem Set 3

Due Lecture 8, October 25 to Box

PPPA 8022
Fall 2023

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.
- 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.

Answer: See columns 1 and 2 of Table 1. 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 two 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

Answer: See columns 3 and 4 of Table 1. The F-tests for these instruments are in both cases quite low. The F-test value for using three instruments (column 3) is 5.6. 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 1.7. 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) Now do the second stage estimation from 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.

Answer: This regression finds that an additional year of education increases wages by a whopping 19 percent; much larger than the estimates in A&K. This coefficient is significant at the five percent level.

(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). What are your findings about education? Why are the coefficients and errors the same or not?

Answer: The coefficients using `ivregress` or `ivreg2` 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 standard errors are different at the third decimal point. In both cases the IV standard error is very slightly smaller than the OLS standard error, as we would expect.

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.

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

(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.

Answer: 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 compulsory school threshold.

Answer: I estimate

$$\begin{aligned}\text{income}_i &= \beta_0 + \beta_1\text{after}_i + \beta_2\text{year}_i + \beta_3\text{year}_i \\ &+ \beta_4\text{year}_i^2 + \beta_5\text{year}_i^3 + \\ &+ \beta_6\text{year}_i * \text{after}_i + \beta_7\text{year}_i^2 * \text{after}_i\beta_8\text{year}_i^3 * \text{after}_i + \epsilon\end{aligned}$$

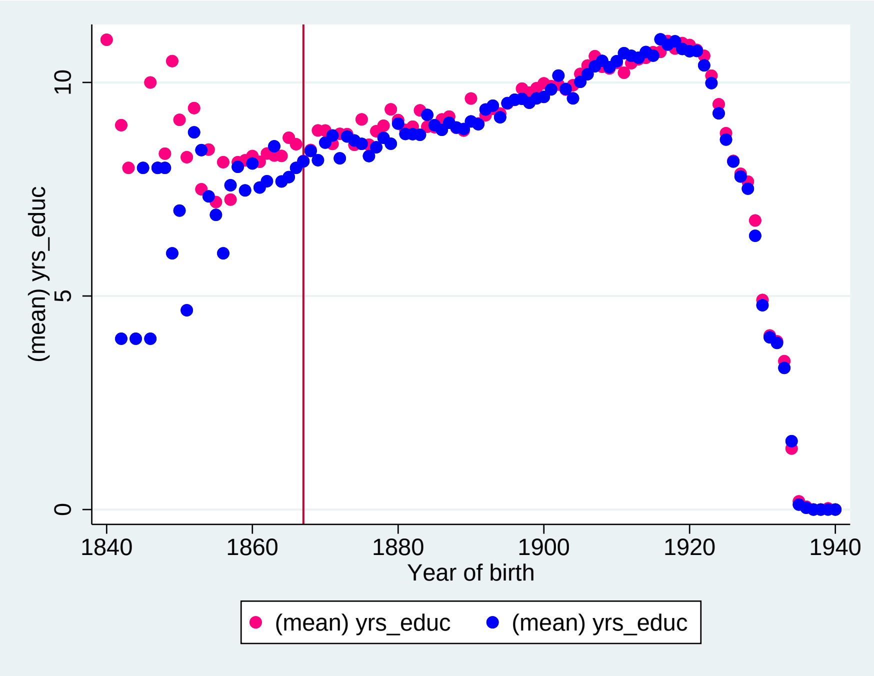
You should have some version of this specification, where you have a term for the running variable, a term for being after the discontinuity, and the running variable interacted with being after the discontinuity.

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

Answer: See table for part (e).

Figure 1: Illinois

(a) Education



(b) Income

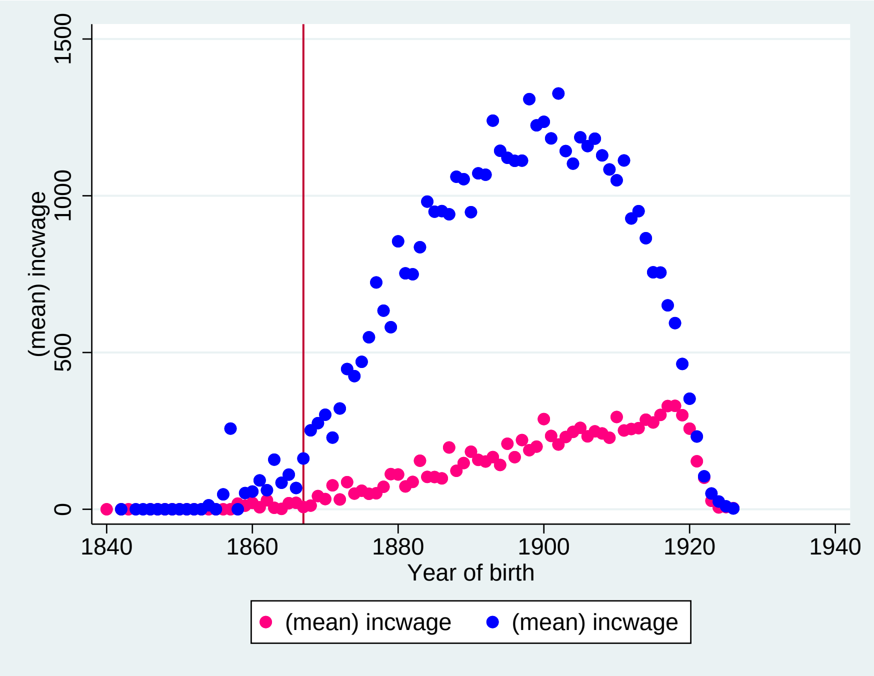
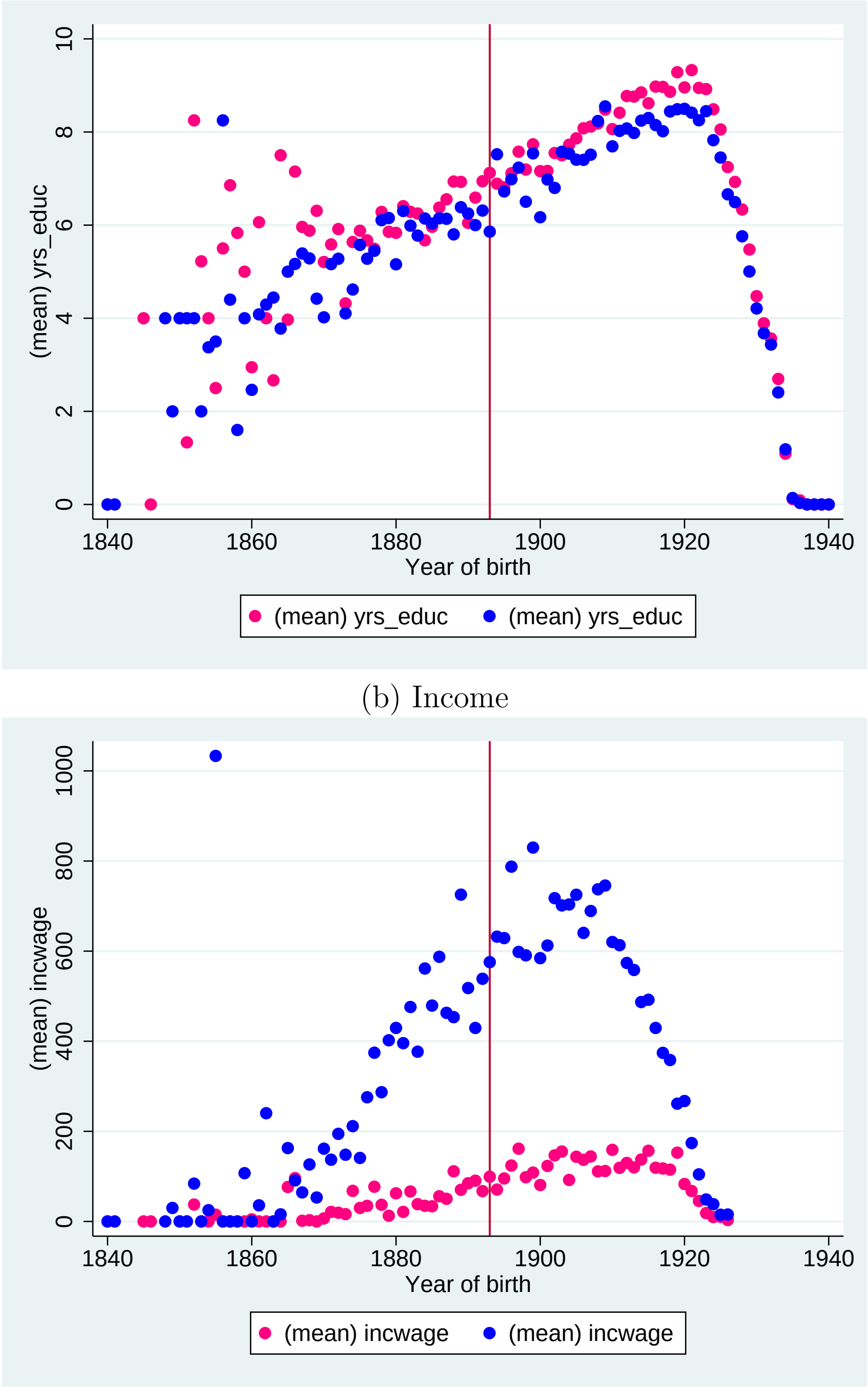


Figure 2: Louisiana

(a) Education



(b) Income

Table 1: Regressions, Problem 2

					Question 2 (c) instruments are			
	Question 2(a)		Question 2(b): 1st stgs		quarters		quarters by year of birth	
	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*	-0.151***					
	-0.039	-0.031	-0.041					
1{birth quarter=2}	-0.098*	-0.047	-0.132**					
	-0.039	-0.031	-0.041					
1{birth quarter=3}	0.018	-0.005	-0.026					
	-0.039	-0.03	-0.04					
Predicted value,					0.119*	0.119*	0.112**	0.112**
years of education					-0.06	-0.057	-0.041	-0.039
F test: instruments	7.629	2.453	5.577	1.655				
p-value of F test	0	0.061	0.004	0.018				
R-squared	0	0	0.034	0.034	0.056	0.144	0.056	0.15
Observations	51162	71816	43163	43163	43163	43163	43163	43163

```
# delimit;

***********************************************************************************

this is the stata file that makes the answers for problem set 2

february 18, 2014
february 19, 2014
february 20, 2014
february 21, 2014
march 16, 2014
march 19, 2014
march 24, 2014
february 12, 2015
february 13, 2015
march 18, 2015
february 22, 2018
march 2, 2018
march 3, 2022
february 15, 2023

ps3_2023v01.do

***********************************************************************************;


***********************************************************************************

****** 1. iv: ipums-census data *******************************************************

***********************************************************************************;

****** A. prep stuff **********************************************************;

clear all;
pause on;
set more off;

* set todays date;
adopath ++ /home/lfbrooks/ado;
dateo;

capture log close;
log using "/groups/brooksgrp/pppa6022/2023/problem_sets/ps3/${date}_ps3_log.txt", replace;


****** B. load data ***********************************************************;

* switch for which sample we use *;
*local sample big;
local sample small;

* load the big sample *;
if "`sample'" == "big"
  {;

  * load census data, keeping only variables that are of interest *;
  use /groups/brooksgrp/census/1980census/ipums/pppa6022_spring2014/c1980_ipums_20140220;

  * take a smaller random sample *;
  gen double rand_samp = runiform();
  keep if rand_samp > 0.9;

  * save this as a junk dataset to load *;
  save /groups/brooksgrp/census/1980census/ipums/pppa6022_spring2014/c1980_ipums_20140220_small,
    replace;
  };

if "`sample'" == "small"
  {;
  * bring in the smaller data *;
  use /groups/brooksgrp/census/1980census/ipums/pppa6022_spring2014/c1980_ipums_20140220_small;
  };


****** C. set up data *******************************************************;

** keep as in A and K **;
drop if birthyr > 1959;
drop if birthyr < 1930;
* drop women *;
drop if sex == 2;

tab educ;
tab educd;

* make birth decade markers *;
gen b_decade = 0;
replace b_decade=1930 if birthyr >= 1930 & birthyr < 1940;
replace b_decade=1940 if birthyr >= 1940 & birthyr < 1950;

* make in smsa marker *;
gen in_smsa = 0;
replace in_smsa = 1 if metro == 2| metro == 3| metro == 4;

* make marital status indicator *;
gen married = 0;
replace married = 1 if marst == 1| marst == 2;

* make quarterly dummies to match a & k's table *;
tab birthq, gen(bqdum);

* make birth year dummies for 1930s *;
tab birthyr if b_decade == 1930, gen(bydum);

* make log of wages *;
* this is going to effectively drop people with zero wages *;
* this is quite a few observations *;
gen ln_incwage = ln(incwage);

* make interaction of quarter*birth year *;
local eqdum "";
forvalues q=1/3
  {;
  forvalues y=1/9
    {;
    gen bdum_q`q'y`y' = bqdum`q'*bydum`y';
    local eqdum "`eqdum' bdum_q`q'y`y' =";
    };
  };

* make age squared *;
gen age2 = age*age;

* make age with quarters *;
gen ageq = age + (birthqtr-1)*0.25;
tab ageq;
* and a squared version *;
gen ageq2 = ageq*ageq;

** 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;
replace yrs_educ = 9 if educ == 3;
replace yrs_educ = 10 if educ == 4;
replace yrs_educ = 11 if educ == 5;
replace yrs_educ = 12 if educ == 6;
replace yrs_educ = 13 if educ == 7;
replace yrs_educ = 14 if educ == 8;
replace yrs_educ = 15 if educ == 9;
replace yrs_educ = 16 if educ == 10;
replace yrs_educ = 17 if educ == 11;

** other sample limits **;
gen reg_keep = 1;
replace reg_keep = 0 if incwage < 0;
* keep only positive weeks worked and positive wage and salary -- but i didnt get weeks worked
*;

****** D. Table 1, first two rows ***************************************************;

sort b_decade;
by b_decade: summ yrs_educ;

regress yrs_edu bqdum1-bqdum3 if b_decade == 1930;
test bqdum1=bqdum2=bqdum3=0;
estadd scalar f_qs_eq = r(F);
estadd scalar p_qs_eq = r(p);
estimates store t1a;

regress yrs_edu bqdum1-bqdum3 if b_decade == 1940;
test bqdum1=bqdum2=bqdum3=0;
estadd scalar f_qs_eq = r(F);
estadd scalar p_qs_eq = r(p);
estimates store t1b;


****** E. first stage ***************************************************************;

* set covariates *;
* change age to yob and change age to age in quarters *;
local covs "in_smsa married i.region i.birthyr ageq ageq2";

* quarters only *;
xi: regress yrs_educ `covs' if b_decade == 1930 & ln_incwage != .;
estimates store fs1nbq;
xi: regress yrs_educ `covs' bqdum1-bqdum3 if b_decade == 1930 & ln_incwage != .;
test bqdum1=bqdum2=bqdum3;
estadd scalar f_qs_eq = r(F);
estadd scalar p_qs_eq = r(p);
predict xhat1, xb;
estimates store fs1;

* quarters * year of birth *;
xi: regress yrs_educ `covs' if b_decade == 1930 & ln_incwage != .;
estimates store fs2nbq;
xi: regress yrs_educ `covs' bdum_q*y* if b_decade == 1930 & ln_incwage != .;
predict xhat2, xb;
test `eqdum'=0;
estadd scalar f_qs_eq = r(F);
estadd scalar p_qs_eq = r(p);
estimates store fs2;


****** F. second stage **************************************************************;

* just quarters as instruments *;
regress ln_incwage `covs' xhat1 if b_decade == 1930 & ln_incwage != .;
estimates store ss1a;
ivregress 2sls ln_incwage `covs' (yrs_educ=bqdum1-bqdum3) if b_decade == 1930 & ln_incwage != .;
estimates store ss1b;

* quarters*years of birth *;
regress ln_incwage `covs' xhat2 if b_decade == 1930 & ln_incwage != .;
estimates store ss2a;
ivregress 2sls ln_incwage `covs' (yrs_educ=bdum_q*y*) if b_decade == 1930 & ln_incwage != .;
estimates store ss2b;

****** G. output results **************************************************************;

*estout *
```

```
  using "/home/lfbrooks/pppa6022/2018/stataout/problem_set_2/${date}_prob2_ivregs.txt",
  replace
  varwidth(12) varlabels(_cons Constant)
  cells(b(star fmt(%12.3f)) se(par fmt(%12.3f)))
  stats(r2 N f_qs_eq p_qs_eq, fmt(%9.3f %9.0g %9.3f %9.3f) labels("R-squared" "Observations" "F
test: instruments" "p-value of F test"));

estimates clear;


************************************************************************************

****** 2. regression discontinuity *************************************************

************************************************************************************;


****** A. load and clean data ******************************************************;

clear all;
set more off;
pause on;

dateo;

* these data are created in
  /home/lfbrooks/pppa6022/2015/stataprg/problem_set_2/usa_00006.do;
use /home/lfbrooks/pppa6022/2015/datasets/ipums1940_20150212;

** 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;
replace yrs_educ = 9 if educ == 3;
replace yrs_educ = 10 if educ == 4;
replace yrs_educ = 11 if educ == 5;
replace yrs_educ = 12 if educ == 6;
replace yrs_educ = 13 if educ == 7;
replace yrs_educ = 14 if educ == 8;
replace yrs_educ = 15 if educ == 9;
replace yrs_educ = 16 if educ == 10;
replace yrs_educ = 17 if educ == 11;

** replace incwage missing code with missing *;
replace incwage = . if incwage == 999999;
summ incwage, detail;


****** B. make regression discontinuity charts *************************************;

preserve;

* find average years of education and wage by birthplace state birth year and sex *;
sort bpl birthyr sex;
collapse (mean) yrs_educ incwage, by(bpl birthyr sex);
sort bpl birthyr sex;
save /home/lfbrooks/pppa6022/2015/datasets/${date}_mnbpl, replace;

* find average years of education and wage but w/o college or more *;
restore;
preserve;
sort bpl birthyr sex;
collapse (mean) yrs_educ_noc=yrs_educ incwage_noc = incwage if yrs_educ < 12, by(bpl birthyr
sex);
sort bpl birthyr sex;

* merge two datasets together *;
merge 1:1 bpl birthyr sex using /home/lfbrooks/pppa6022/2015/datasets/${date}_mnbpl;

* set up for graphs *;
```

```
graph set eps orientation landscape;

* program to run graphs *;
capture program drop graphit;
program define graphit;

syntax, stab(string) stnum(string) styear(string) stname(string) dv(string);

  graph twoway
    (scatter `dv' birthyr if bpl == `stnum' & sex == 2,  mcolor(pink) xline(`styear'))
    (scatter `dv' birthyr if bpl == `stnum' & sex == 1 , mcolor(blue)),
    xsize(11)
    ysize(8.5)
    ;
  graph export
    "/groups/brooksgrp/pppa6022/2023/problem_sets/ps3/${date}_`stab'_`dv'_vs_yob.pdf", replace;

end;


** set year the policy starts *;
local layear = 1893;
local ilyear = 1867;

** 1(a) charts for overall regression discontinuity **;

graphit, stab(la) stnum(22) styear(`layear') stname(Louisiana) dv(yrs_educ);
graphit, stab(la) stnum(22) styear(`layear') stname(Louisiana) dv(incwage);

*graphit, stab(md) stnum(24) styear(1902) stname(Maryland) dv(yrs_educ);
*graphit, stab(md) stnum(24) styear(1902) stname(Maryland) dv(incwage);

graphit, stab(il) stnum(17) styear(`ilyear') stname(Illinois) dv(yrs_educ);
graphit, stab(il) stnum(17) styear(`ilyear') stname(Illinois) dv(incwage);

** 1(b) charts omitting college graduates ***;

graphit, stab(la) stnum(22) styear(`layear') stname(Louisiana) dv(yrs_educ_noc);
graphit, stab(la) stnum(22) styear(`layear') stname(Louisiana) dv(incwage_noc);

*graphit, stab(md) stnum(24) styear(1902) stname(Maryland) dv(yrs_educ_noc);
*graphit, stab(md) stnum(24) styear(1902) stname(Maryland) dv(incwage_noc);

graphit, stab(il) stnum(17) styear(`ilyear') stname(Illinois) dv(yrs_educ_noc);
graphit, stab(il) stnum(17) styear(`ilyear') stname(Illinois) dv(incwage_noc);


****** C. do regression discontinuity regressions ****************************;

capture program drop rdreg;
program define rdreg;

syntax, stnum(string) styear(string);

  * need to make a flexible time before, flexible time after *;
  * if the policy is passed in year x, if affects people born in year y or later,
    where x = y + 16, so y = x - 16;
  * make dummy for being after the policy *;
  gen after = 0;
  replace after = 1 if birthyr >= `styear' - 15;

  * make X-c as in Lee and Lemieux *;
  gen yr = (birthyr + 16) - `styear';
  * make X-c * after *;
  gen yr_after = yr*after;

  * make similar values for squared and cubed *;
  forvalues j=2/3
    {;
    gen yr_`j' = yr^`j';
    gen yr_`j'_after = yr_`j' * after;
    };

  * regression *;
  regress incwage after yr yr_2 yr_3 yr_after yr_2_after yr_3_after
    if bpl == `stnum';
  estimates store r`stnum'_1;

  * restricted to people who are 20+ in 1940 *;
  regress incwage after yr yr_2 yr_3 yr_after yr_2_after yr_3_after
    if bpl == `stnum' & birthyr < 1920;
  estimates store r`stnum'_2;

  * output regressions *;

  * drop variables ill want to create again *;
  drop after yr yr_*;

end;

* bring back the person-level dataset we started with *;
restore;

* run the rd regressions *;
rdreg, stnum(22) styear(`layear');
*rdreg, stnum(24) styear(1902);
rdreg, stnum(17) styear(`ilyear');

* output regression results *;
estout *
  using "/groups/brooksgrp/pppa6022/2023/problem_sets/ps3/${date}_prob2_rdregs.txt",
  replace
  varwidth(12) varlabels(_cons Constant)
  cells(b(star fmt(%12.3f)) se(par fmt(%12.3f)))
  stats(r2 N, fmt(%9.3f %9.0g %9.3f %9.3f) labels("R-squared" "Observations"));

estimates clear;

log close;
```