Problem Set 1

PPPA 6022
Due in class, on paper, February 12

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.
- This problem set uses some large data. For the Census data, I have put the full dataset up on Blackboard, and I've also put a smaller version. For the CPS, only the small one would fit.

1. Fixed Effects

For this problem, we'll use Decennial Census/American Community Survey data from IPUMS-USA for 1950 and 2010 (for 2010, the 1-year ACS). Data are available on Blackboard. The large versions, once for each year, have the years in the title (1950 and 2010); the small version is ipumscen.dta.zip. Note that analysis using the 1950 sample must use weights (perwt); for simplicity (if not correctness), please use Stata's aweights or the equivalent.

The IPUMS website is https://usa.ipums.org/usa/, and it provides detailed information on the datasets and variables.

Let's examine the effect of education on wages.

(a) Start by finding the average wage (incwage) of prime age men (25 to 64) in 1950 and 2010. Test whether these wages differ significantly across time, and present these results in a well-labeled table. Beware of missing values.

(b) Make the wages in both surveys into constant 2013 dollars. Use the all urban consumers series from the Bureau of Labor Statistics (http://www.bls.gov/cpi/data.htm, and use the "all urban consumers" row, and use the "all items" series; using the December inflation number for each year is sufficient). Update your table with these real wages.

(c) Suppose we would like to know whether husbands earn higher real wages than wives. Use a regression to estimate wages as a function of age, year, and being the husband (think about what sample you should use to do this, and explain what sample you chose. Make sure you only keep working age people.). Then re-estimate with a variety of sensible covariates. Then re-estimate with the covariates and family fixed effects (in Stata, I highly recommend areg). Then re-estimate to allow the main effect to vary between 1950 and 2010. Present these results in a table.

(d) The previous estimation included age linearly. Use two methods to relax this assumption. Interpret the results. Which method do you prefer and why?

2. Difference-in-difference

Now let's use the IPUMS –CPS. I've put this on Blackboard, but only the small sample, called ipumscps.dta.zip. Documentation for this dataset is available at https://cps.ipums.org/cps/. For the purposes of this problem set, treat each observation with equal weight. This is entirely wrong, and you should absolutely never do such a thing if you are doing a real project. Finally, beware of top-coded data!

(a) Pretend that MI, CA, AZ, NM, MN, OH, VA, KY, WV, MO, MS, GA, IA, NH, MA and ME all adopt a policy aimed at increasing wages that takes effect in 2000. For simplicity, we will focus only on employed people. We hypothesize that treatment is random conditional on age and race. Use a figure to examine the parallel pre-trend assumption (the unconditional outcome, not conditional on covariates), and show this figure (note that making a legible picture may require some summary of the data; think about the best way to summarize the data). Use the variable `incwage` for annual wages.

(b) Use a regression to test whether the treated and untreated states have similar trends before the treatment is adopted, conditional on covariates. Interpret the results of your test.

(c) Do a difference-in-difference regression to examine the effects of this policy on wages. Write the estimating equation you use. Start with a simple summary table (with standard errors) that does the same analysis, and then do a regression. What are the results? Do the two methods yield similar findings?

(d) Now suppose that the policy targeted only men. This suggests a triple difference estimation strategy. Write the estimating equation. Make a simple table that does this triple difference, and then do a regression that does the same.

(e) Explain and implement one method to correct the results from part (c) for serial correlation. For simplicity, ignore the covariates. Describe your method and present your results.