Problem Set 1

PPPA 8022
Due in class, on paper, February 15

Some overall instructions

- Please use a do-file (or its SAS or SPSS equivalent) for this work. Do not program interactively. It may seem faster at first, but it is inevitably slower.
- 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. Do not present statistical software output.
- This problem set uses some large data. For the Census data, I have posted full dataset as well a smaller version; use whichever you prefer. For the CPS, we are using a random sample.
- If the question is insufficiently clear, explain the assumptions you made to reach your final estimates.

1. Fixed Effects

For this problem, well use Decennial Census/American Community Survey data from IPUMS-USA for 1950 and 2010 (for 2010, the 1-year ACS). Data are linked on the handouts page. The large versions, which are one file per year, have the years in the title (1950 and 2010); the small version is ipumscen.dta.zip.

For purposes of **this problem set only** we will not use any survey-defined weights. This is totally wrong and you should never do it when you really analyze a dataset. We are doing it here so that 1(b) does not become extremely difficult.

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 average wages are significantly different in 1950 and 2010, and present these results in a well-labeled table. Beware of missing values. You can – but need not – do a regression to do this.

(b) Replicate these results on means by using grouped data, properly weighted. (For an example, see MHE Table 3.1.3.) The grouping variable is your choice. I used education (so, average wages by years of education), but you could use age or any other variable for which you can create discrete categories.

(c) 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.

(d) Suppose we would like to know whether the average husband earns higher real wages than the average wife. 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 effect of being a husband to vary between 1950 and 2010. Present all results in one table and interpret the coefficients in each regression, explaining why they change.

(e) 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. The small sample (a random sample of the full dataset) is called ipumscps.dta.zip and is linked on the handouts page. 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 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 statistics table (with standard errors) that shows both differences and then the difference in difference (as in the

Milligan paper). Now run a regression that is parallel to this estimate. Present both results in one or two tables and interpret the magnitudes.

(d) Now suppose that the policy targets only men. This suggests a triple difference estimation strategy. Write the estimating equation. Make a summary statistics table that does this triple difference, and then do a regression that does a parallel analysis. Present results in a table and describe them.

(e) Explain and implement one method to correct the results from part (c) for serial correlation. For simplicity, do not use covariates. Describe your method and present your results.