**Problem Set 2**
Due Class 5, October 4

PPPA 8022
Fall 2023

Some overall instructions

- Please use a do-file (or its SAS or SPSS or R 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 need to transfer to another format, use StatTransfer or contact me. For loading these data in R, I recommend the `haven` package, with `read_dta()`.
- 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.
- Data are
  - Decennial Census data
    - Large: 1950 and 2010
    - Small: 1950 and 2010
    - Be in touch if you would like csv versions of these files
  - Current Population Survey (CPS)
    - Stata format
    - CSV

1. Interpreting Indicator Variables

For this problem, we'll use Decennial Census/American Community Survey data from IPUMS-USA for 1950 and 2010 (for 2010, the 1-year American Community Survey), linked above.

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) Find the income from wages prime age men (`sex == 1`, ages 25 to 64) in 1950 and 2010. Use the variable `incwage`. I have re-coded top coded values (99999) to missing.

(b) Use summary statistics (not a regression, but means) and a t-test to test whether the average income from wages from (a) are significantly different in 1950 and 2010. Present your results in a well-labeled table that shows the averages and the $t$ value (feel free to combine tables across steps if that is helpful). Beware of missing values. Write a sentence or two to interpret your table.

(c) Use a regression to do the same test as in (b). Write the regression equation you're estimating. Estimate the equation using software. Report the results in a well-labeled table.

(d) Explain how you can combine estimated regression coefficients from (c) to generate one of the means you found in (a).

(e) Now re-do the means in (a) using the constant 2022 dollar income (`rl_wage`). Report

this income in a well-labled table.[1]

(f) Suppose we would like to know whether the average husband earns higher real income from wages than the average wife.

1. Write a regression equation that we could use to test this hypothesis. Let $X_{i,t}$ represent covariates, $H_{i,t}$ to indicate 1 if the person is a husband, and $Y_{i,t}$ represent the outcome.

2. Use Stata to estimate the regression from (f.1) using covariates age and year. Think about what sample you should use to do this, and explain what sample you choose. Report results in a well-labeled table. Make sure you only keep working age people.

3. Modify the regression equation from (f.1.) to allow the relationship between being a husband and earnings to vary between 1950 and 2010.

4. Estimate the regression from (f.3), report the results in a well-labled table and write a sentence that interprets the estimated coefficient that tells us whether the relationship between being a husband and earnings is different in 2010 than in 1950.

Your tables need only include the relevant coefficients; do not report information on all coefficients.

(g) The previous estimation included age linearly. Use the estimation for (f)(1) and use a method that relaxes the linear assumption on age. Report the results in a table. Write a few sentences that interpret the results, comparing with part (e).

[1]For your information, here is how I adjust for inflation:

- Go to the Bureau of Labor Statistics (http://www.bls.gov/cpi/data.htm, and choose "all urban consumers" row and the "top picks" column.
- From the following window, choose the "US city average, All items" and choose "retrieve data," at the bottom.
- Download the data using the xls icon, making sure you're grabbing the relevant years; see the selection at top.
- Use the December inflation number for each year (this is not exactly correct, but it is sufficient for this example).
- To inflation adjust
  - re-scale the inflation adjustment so that it is 1 in 2022
  - to do this, divide the 2022 value by each year's value
  - this gives 1 in 2022, numbers $> 1$ in years before 2022 and numbers $< 1$ in years after 2022
  - this new ratio is the adjustment factor
  - multiply the adjustment factor by the values (e.g., wage) I make into constant dollars

2. Difference-in-difference

Now let's use the IPUMS–CPS; data are linked above. 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, focus only on employed people for this entire question. Use the variable `incwage` for annual wages. Create a figure that examines the parallel pre-trend assumption.

Hints on how to create this figure:

- Sketch yourself what this graph should look like
- Then ask "what summary statistics do I need to make this graph?"
- Create the summary statistics
- Plot the summary statistics

Write a few sentences that interpret the figure.

(b) We hypothesize that treatment is random conditional on age and race. Use a regression to test whether the treated and untreated states have similar trends before the treatment is adopted, conditional on covariates. Look at Lecture 4 for our discussion of trends. Report the results in a table, and write a few sentences that interpret the results of your test.

Some hints: You'll need to limit the sample to only pre-treatment data. You'll also need to create a time trend variable. Then regress the outcome of interest on this time trend variable interacted with treatment.

(c) Do a summary statistics version of a difference-in-difference estimate (no covariates for this question). Create a table with means and standard errors from which you can calculate the single and double differences (don't worry about calculating the errors for these differences).

(d) Do a difference-in-difference regression that parallels the summary statistics in part (c), meaning that it has no covariates. Write the estimating equation you use. You should get the same result as in (c). If you don't get the same result as in (c), you are doing something wrong.