Problem Set 2

Due Class 6, February 18

PPPA 8022
Spring 2026

Some overall instructions

- 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.
- Please turn in **one pdf document** with three parts: (i) a set of written answers to these problems, (ii) a do file (or program from the statistical software of your choice), and and (iii) the output from the program of your choice. For the third, it is sufficient to copy the output into a word doc and save as a pdf. The program file should have comments that indicate the commands associated with each question.
- Use this link to submit the problem set. You need to be logged into GW email to submit. If it still gives you trouble, open an incognito window, log in and try again.
- 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. For loading these data in R, I recommend the `haven` package, with `read_dta()`. If you need another format, please let me know.
- 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

- You are both welcome and encouraged to work on this problem set with your classmates. The problem set you turn in should be your own work – both the code and the written output. If we notice exactly duplicative work, we will give zero credit to both assignments.
- You are also welcome to use AI to assist on the problem set. Make sure, however, that you understand the answer that AI gives you, as you will need the coding experience these problem sets deliver to succeed on your replication project.

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) In a well-labeled table, report the mean, standard deviation and number of observations for income from wages (`incwage`; I have re-coded top coded values (99999) to missing) for prime age men (`sex == 1`, ages 25 to 64) in 1950 and 2010.

(b) Use the values in the table from part (a) to calculate a t-test for whether 1950 average income from wages for prime-age men from (a) differs from the 2010 figure. Update your previous table to include the t value. 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) Show how you can combine your estimated regression coefficients from (c) to yield either the 1950 or 2010 mean in the table from (a).

(e) Re-do the means in (a) using the constant 2022 dollar income (Use the variable `rl_wage` I have already calculated for you.). Report this income in a well-labled table.[1]

[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.

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

This question requires you to analyze a subset of the data that contains only husbands and wives. To do this, you'll need to condition on variables such as sex, relation to the household head, and marital status. Also, watch out for missing values so that you can properly subset the data.

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. Here and elsewhere, your tables need only include the relevant coefficients; do not report information on all coefficients.

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), and report the results in a well-labled table. Write a sentence to interpret your test of whether the relationship between being a husband and earnings is different in 2010 than in 1950.

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

- 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. (This type of plot is legible only when you present grouped means.)

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) Suppose we hypothesize that treatment is random conditional on age and race.

1. Write a regression equation 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.
2. Estimate the equation from the previous step. Report the results in a table, and write a few sentences that interpret the results of your test.

   Some advice for this estimation: Limit the sample to only pre-treatment data. 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). This means find the means for the treated and untreated groups, both before and after treatment. Create a table with these means and the relevant 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.

1. Write the estimating equation you use.

2. Estimate the regression and report the results in a well-labeled table. Include the constant.

3. Show that you can combine the regression coefficients from the table you just created to add up to one of the sample means from part (c). If you don't get the same result as in (c), you are doing something wrong. The key to getting the diff-in-diff regression and the summary stats to match is to use the **exact** same sample.