Econometrics for Public Policy II
Spring 2026

Problem Set 1

Due before Class 3, January 28

There are two datasets for this problem set: one for 1950 and another for 2010. (You should see links in the previous sentence for data downloading.) If you are using R, use the `haven` package and the `read_dta` function to read these data. If you need another type of dataset, please let me know and I'm happy to provide.

Each dataset has one observation per US county in that year (1950 or 2010). Data come from the Decennial Cenus (1950, some 2010) and the American Community Survey (2010, which is really 2008-2012 5-year average). All variables are labeled. The census tabulates data from the individual collection at a variety of levels of geography; here we use county-level data.

The variables statefips/countyfips uniquely identify observations in each dataset. You can find definitions for the statefips and countyfips variables at this helpful webpage from the University of Missouri. and many other websites.

Use Stata's `describe` command to see the definitions of the remaining variables, or reference the definitions I give below.

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. See turn-in link at the end of this problem set.

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. Summary statistics

   a. Make a panel dataset from 1950 and 2010, meaning a dataset that has one observation per county and year. In this dataset, most counties will have two observations, one for 1950 and one for 2010. Stata's `append` command stacks one dataset on top of another. The equivalent command in R is `rbind()`.

b. By year, find the average of

- population (`cv1`)
- log of population (create yourself from `cv1`)
- share white (`s1`)
- share black (`s2`)
- share women age 25+ with education of some college or more (`s3`)
- share men age 25+ with education of some college or more (`s4`)

Use one command to find all these averages. In Stata, you can use `collapse` combined with `, by(year)`.

After using `collapse`, you can use `outsheet` to output the resulting dataset as a txt or csv. Using this output file, it should not be difficult to create a labeled table.

In R, you can use `group_by()` and `summarize()` from the `tidyverse` package. I usually use the `averages <- dt[, .(mean.pop = mean(cv1, na.rm = TRUE), by = year]} type of not`

c. Find averages of the same variables by year and state for California, Mississippi and New Jersey.

Note that California's state code is 6, and that it has a leading zero – so write it `06`. Again, use Stata's `collapse` combined with `, by(year state)` to generate these outputs. Also again, you can use Stata's `outsheet` command to output the data you've created to a .csv or .txt file.

2. Matching Data

a. How many counties are in both the 1950 and 2010 datasets?

In the previous question you created a panel dataset. To answer this question, you may prefer to make a "wide" dataset with one observation per county. It may be helpful to make a indicator variable (0/1) for having an observation in a given year in the 1950 and 2010 datasets, and merge the datasets.

In the merged dataset, you can use Stata's `tab` command to see a cross-tab of two indicator variables. For example, if your variables are called `y1950` and `y2010`, you can tell Stata to report `tab y1950 y2010`. Correctly interpreting this table will tell you the answer to the first three parts of this question.

In R, before the merge make a variable equal to one for each dataset before the merge. Suppose we call these variables `v1` and `v2`. Then use the `merge()` function and set `all = TRUE`. After the merge, set missing values of `v1` and `v2` to zero (you can use an `ifelse(is.na(df$v1) == TRUE, yes = 0, no = df$v1)` function to do this).

b. How many counties are in the 1950 dataset, but not the 2010 dataset?

c. How many counties are in the 2010 dataset, but not the 1950 dataset?

d. Investigate two counties that are in the 2010 dataset, but not the 1950 dataset. Why is this?

3. Regressions

   a. Return to the panel dataset from question 1.

   b. Regress log of population on the four share variables from question 1 and a fixed effect for year = 2010.

   For this and the next question, it is sufficient to paste the results from the log; for future problem sets you will need to make a regression table, but you do not need one here.

   In Stata, there are multiple says to create indicator variables and use them in a regression. Here are two equivalent methods:

   -
     ```
     gen y2010 == 0
     replace y2010 = 1 if year == 2010
     regress y x y2010
     ```
   - `xi:  regress y x i.y2010`
   - You can test for yourself whether these yield equivalent results

   c. Interpret the coefficient on the year indicator variable

   d. Repeat the previous regression with state fixed effects

   e. Interpret one of the share coefficients from the second regression

4. Long and Short Regressions and Omitted Variable Bias

   - From the lecture, we learned the omitted variable bias formula. Now you're going to calculate a specific example.
   - We limit our analysis just to 2010.
   - We are interested in the impact of the share of college educated men on the employment to population ratio and on the extent of omitted variable bias if we exclude the share of women who are college educated.
     - Let E_i, defined as `cv59 / cv1`, denote the employment to population ratio in county i
     - Let M_i be the share of men age 25 or above who are college educated in county i
     - Let W_i be the share of women age 25 or above who are college educated in county i

- Let's suppose that we have a "true" long equation

$$E_i = \beta_0 + \beta_l M_i + \gamma W_i + \epsilon_{l,i} \tag{1}$$

- However, we sometimes want to be lazy and estimate a "short" regression:

$$E_i = \beta_0 + \beta_s M_i + \epsilon_{s,i} \tag{2}$$

- How bad is the short regression? The omitted variable bias formula tells us that

$$\beta_s - \beta_l = \pi * \gamma \tag{3}$$

  where γ is the coefficient on W from the long regression and π is the coefficient on M from this regression that estimates the strength of the correlation between M and W:

$$W_i = \alpha + \pi M_i + \epsilon_{c,i} \tag{4}$$

a. Estimate equations 1, 2 and 4 above.

b. Use your estimated coefficients to show that the omitted variable bias formula (equation 3) holds. To do so, write the estimates for β_s and β_l, and show that their difference is equal to the product of your estimates of π and γ.

How to Turn This In

Name the pdf (see notes at the top about what to turn in) "ps1_lastname.pdf". Thus, my submission would be ps1_brooks.pdf

Turn in the problem set via this google link. Make sure you are logged in to your GW account to make the upload work.