Econometrics for Public Policy II
Spring 2022

Problem Set 1

Due Class 3, January 26

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'd like non-Stata datasets, please let me know.

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.
Please turn in a set of written answers to these problems, as well as a do file (or program from the statistical software of your choice). The program file should have comments that indicate the commands associated with each question.

1. Summary statistics

   a. By year, find the average of

      - population
      - log of population
      - share white
      - share black
      - share women age 25+ with education of some college or more
      - share men age 25+ with education of some college or more

   b. 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`. If you use Stata's `collapse` to generate these outputs, note that you can use Stata's `outsheet` command to output the data you've created to a .csv or .txt file. You can open a csv or txt file in Excel and create a reasonable-looking table very quickly.

2. Matching Data

   a. How many counties are in both the 1950 and 2010 datasets? (It may be helpful to make a indicator variable (0/1) for having an observation in a given year and merge the two dataset. You can use Stata's `tab` command to see a cross-tab of two indicator variables; this table will tell you the answer to the first three parts of this question.)

   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. Make a panel dataset from 1950 and 2010, meaning a dataset that has one observation per county and year, where most counties have two observations, one for 1950 and one for 2010. Stata's `append` command stacks one dataset on top of another.

   b. Regress log of population on the four share variables you created above and a fixed effect for year = 2010. (For this and the next question, it is sufficient to have the results in the log; for future problem sets you will need to make a regression table, but you do not need one here.)

   c. Repeat the previous regression with state fixed effects

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

   e. Report how much a one standard deviation change in that share impacts population.

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 holds.

c. Write the formula of coefficients and their estimated values to show equality (this last step is all you need to present; your program should have the regression estimation in it).

How to Turn This In
Go to the "for students" google folder, and then the subfolder "problem set 1." Name your file "problem_set_1_sn[secret number].pdf". So if my secret number is three, I should name my file "problem_set_1_sn3.pdf".