Version: November 10, 2021 Fall 2021
EC200: Econometrics and Applications

Lab 7: Difference in differences and panel data

Objectives

There are two separate parts to this lab - a set of data for working with difference-in-differences models, and another set for working with fixed-effects models

By the end of this lab, you should be able to complete the following tasks in Stata:

- Estimate and interpret difference-in-differences models
- Estimate panel data models using dummy variables
- Interpret panel data models

Key commands

Using `xtset` and `xtreg`:

The `xtset` command will tell Stata that you have panel data. For example, if you have state and year data, then you would enter `xtset state year`, or whatever the appropriate variable names are.

General format: `xtset panelvar timevar`

After declaring your panel with `xtset`:

- Ask Stata to do a panel-regression by using xtreg instead of regress. Everything else proceeds as normal.
- You have to tell Stata that you want to estimate a fixed effects model, so you add ,fe as an "option"

For example, something like this: `xtreg income education,fe` would regress income on education, and include fixed effects, where the fixed effects are the `panelvar` variable you declared.

Adding other fixed effects:

You can add fixed effects to a model more generally with the `i.` prefix or `areg`. A few examples:

```
xi: reg income i.educ i.bpl, robust
reg income i.educ i.bpl, robust

areg income i.educ,robust abosrb(bpl)
```

1. `xi:` this prefix is necessary for adding `i.` variables if the variables are in string form. Of course, you can use it any time! You can also use it to do fancier interactions with fixed effects, like `xi: reg income i.educ*i.bpl, robust`
2. You can exclude the prefix and just do `i.var` to create a bunch of indicator variables so long as your variable is *numeric*
3. You can use `areg` to "absorb" a set of fixed effects - they will not be reported in your output, but they will be estimated. This method is less efficient than `xtreg` becuase you're using up degrees of freedom.

Exercises

Part A: Differences-in-differences

This part of the lab looks at a simple difference-in-differences model based on Richardson and Troost (2009).[1]

Mississippi is split between two Federal Reserve Districts. During the early years of the Great Depression, the each district took a different approach to bank runs. The Sixth District increased lending,while the Eighth District responded by restricting lending to threatened banks. We look at the impact of these policies on bank survival rates using difference-in-differences.

1. Start a new do-file and change directory to your working directory .
2. In your do-file, start a log and open `banks.dta`
3. Using pencil & paper or electronic means of your choosin (ie, you don't need to do this in Stata), plot a graph of the number of banks in business, by district, by year. Plot number of banks in business on the y axis and the year on the x axis. Include only the years 1930 & 1931. Draw separate lines for the numbers of banks in District 6 and District 8. Draw a dotted "counterfactual" line based on your understanding of the change in bank policies. Mark all (4) actual values clearly
4. First, we're going to calculate a difference-in-difference estimator by hand between 1930 and 1931. Using the `browse` command, fill in x values from the following table:

Number of banks in business			
District	1930	1931	1931-1930
District 6	x	x	x
District 8	x	x	x
District 8 - District 6	x	x	x

What is the difference-in-difference estimator?

[1] Based on Chapter 5 of *Mastering 'Metrics*

5. Now, generate a variable `treat`, a binary variable equal to 1 for District 8 and 0 otherwise, and a variable `post`, a binary variable equal to 1 for the year 1931 or greater. Generate `treatXpost = treat*post`.
6. Using the above variables, estimate the impact of looser lending restrictions on the number of banks using a difference-in-difference estimator, restricting the sample to 1930 and 1931. Write your estimates, in equation form.
7. Now estimate the same regression, but use all years. What is the overall impact of looser lending restrictions on bank survival? Write your estimates, in equation form.
8. State clearly the assumption needed to interpret these difference-in-difference estimators as causal.

Part B - Fixed effects

Next, we're going to look at the relationship between marijuana use and employment based on the National Longitudinal Survey of Youth 1997 Cohort (NSLY97).

1. Download the NSLY97 data files and save them to your working directory.
2. Start a new do-file and change directory to your working directory .
3. In your do-file, start a log and open `nsly_marijuana.dta`
4. How many individuals are in the data? How many years are they in the data?
5. Estimate a regression of whether marijuana use affects income, with no additional controls. Report your results in equation form.
6. Estimate a regression of whether marijuana use affects income, but add any controls you deem important (from the relatively limited selection I provide). How do the results change? Report your results in equation form.
7. One way to estimate fixed effects models is to use `xtreg` with the `,fe` option. Use `xtset` to tell Stata you have panel data. Then, estimate a fixed-effects regression of whether marijuana use affects income, with no additional controls. Include both individual-level and year-level fixed effects. Cluster your standard errors at the individual (id) level.
8. What is the coefficient on marijuana usage? What is the interpretation?
9. After adding fixed effects, should you include controls for gender and race/ethnicity to reduce omitted variable bias? Why or why not?
10. How do your results in part 9 using fixed effects compare to your results in parts 5 and 6? Why do they differ?
11. Name one specific factor that would create omitted variable bias in this regression.