ECON3500 Lab 7: Difference in differences

Materials

- `banks.dta`
- `nsly_marijuana.dta`
- Do-file template `econ3500_lab_template.do`

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

What is panel data?

Up to now, we've worked with **cross-sectional** data — one observation per person (or state, or county) at a single point in time. In this lab, we'll work with **panel data** (also called longitudinal data), where we observe the *same* individuals or units across *multiple* time periods.

Panel data lets us control for characteristics of each unit that don't change over time — even ones we can't directly measure — by comparing each unit to *itself* over time. This is the key idea behind **fixed effects** models.

What is difference-in-differences?

Difference-in-differences (DiD) is a method for estimating causal effects when one group is exposed to a treatment and another is not. The idea: compare how the outcome changed over time for the **treatment group** vs. the **control group**. The first difference removes time-invariant characteristics of each group; the second difference removes common time trends. What's left is the estimated treatment effect — *if* the two groups would have trended the same way absent the treatment.

Key commands

command	description
`xtset panelvar timevar`	Declare your data as a panel (e.g., `xtset id year`)
`xtreg y x, fe`	Panel regression with fixed effects on `panelvar`
`xtreg y x, fe cluster(panelvar)`	Same, with clustered standard errors
`i.varname`	Add fixed effects for every value of `varname`
`xi: reg y i.varname`	Same as above, but works with string variables
`areg y x, absorb(varname)`	Absorb fixed effects (estimated but not reported)

Using `xtset` and `xtreg`

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

General format: `xtset panelvar timevar`

After declaring your panel with `xtset`:

- Use `xtreg` instead of `regress` for panel regression. Everything else proceeds as normal.
- Add `,fe` to estimate a fixed effects model, where the fixed effects are the `panelvar` variable you declared.
- Add `cluster(panelvar)` to cluster standard errors at the panel level (accounts for correlation within units over time).

For example: `xtreg income education i.year, fe cluster(id)` regresses income on education with individual fixed effects (from `xtset`) and year fixed effects (from `i.year`), clustering standard errors at the individual level.

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 absorb(bpl)
```

1. `xi:` — this prefix is necessary for adding `i.` variables if the variables are in string form. 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 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` because you use up degrees of freedom.

Workflow overview

1. Load a dataset and start your log file.
2. Explore the data structure (`describe`, `browse`, `tab`).
3. For Part A: Calculate the DiD estimator by hand, then estimate it as a regression.
4. For Part B: Declare your panel data and estimate fixed-effects models.
5. Compare results across specifications and interpret.
6. Answer the worksheet questions.

Lab 7 Worksheet

What do I submit?

- Your written-up answers to exercise questions (1)–(18). This can be typed or written out then scanned (or photographed), in any reasonable format.
- The do-file(s) you created that run this analysis
- A log file that contains the results from this exercise.

Part A: Difference-in-differences

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

Data context Mississippi is split between two Federal Reserve Districts. During the early years of the Great Depression, 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.

Each row in `banks.dta` represents a Federal Reserve district in a given year. The dataset is small — use `browse` to see the full thing.

Variables (Part A)

variable	meaning	notes
`district`	Federal Reserve district	6 or 8
`year`	year	
`bib`	number of banks in business	outcome variable

Tip: use `describe` and `browse` to confirm the variable names in your dataset.

Questions *Use robust standard errors in all regressions.*

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 choosing (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 year on the x-axis.
    - Include only the years 1930 and 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 four actual values clearly.

Hint: The counterfactual line shows what *would* have happened to District 8 if it had followed the same trend as District 6. To draw it: start from District 8's 1930 value and apply the same change that District 6 experienced between 1930 and 1931.

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?

Hint: Use `browse` or `list if year == 1930 | year == 1931` to see the values you need.

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

5. Now, generate the following variables:
   - `treat`: a binary variable equal to 1 for District 8 and 0 otherwise
   - `post`: a binary variable equal to 1 for the year 1931 or greater
   - `treatXpost = treat*post`

Hint: Use `tab district` and `tab year` to check the values before generating your variables. For example:

```
gen treat = district == 8
gen post = year >= 1931
gen 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.

Reminder: You can restrict the sample *within* a regression using `if` without dropping data:

```
regress bib treat post treatXpost if year == 1930 | year == 1931, robust
```

7. Now estimate the same regression (same variables), but remove the sample restriction so all years are included. 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 income using the National Longitudinal Survey of Youth 1997 Cohort (NLSY97).

Data context Each row in `nsly_marijuana.dta` is an individual-year observation from the NLSY97 — the same people surveyed across multiple years. This is **panel data**: we observe the same individuals over time, which lets us control for time-invariant individual characteristics (like innate ability or family background) using fixed effects.

Variables (Part B)

variable	meaning	notes
`id`	individual identifier	use with `xtset`
`year`	survey year (1997–2011)	use with `xtset`
`income`	total wage and salary income	
`marij`	used marijuana in past year	1 = yes, 0 = no
`gender`	gender	1 = male, 2 = female
`race`	race/ethnicity	4 categories (use `tab race` to see labels)

Questions

9. Now switch to the second dataset. Open `nsly_marijuana.dta` in your do-file.
10. If starting a new do-file, set your working directory and start a log. (You can also continue in the same do-file from Part A.)
11. How many individuals are in the data? How many years are they observed?

Hint: Try `codebook id` to see the number of unique individuals, and `tab year` to see which years are in the data.

12. Estimate a regression of whether marijuana use (`marij`) affects income, with no additional controls. Report your results in equation form.
13. Estimate a regression of whether marijuana use affects income, but add any controls you deem important (from the relatively limited selection available — use `describe` to see what's there). There is no single correct answer — use your judgment and explain your choices. How do the results change? Report your results in equation form.
14. 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.

    Your model should include:

    - Individual-level fixed effects (these come from `xtreg ... , fe`)
    - Year-level fixed effects (add `i.year` to your regression)
    - Clustered standard errors at the individual level

Step by step:

```
xtset id year
xtreg income marij i.year, fe cluster(id)
```

Clustering standard errors at the `id` level accounts for the fact that observations from the same person across years are not independent.

15. What is the coefficient on `marij`? What is the interpretation?
16. After adding fixed effects, should you include controls for gender and race/ethnicity to reduce omitted variable bias? Why or why not?

Think about it: What happens to a variable that *never changes* within an individual when you include individual fixed effects?

17. How do your results in question 14 using fixed effects compare to your results in questions 12 and 13? Why do they differ?
18. Name one specific factor that would create omitted variable bias in the pooled OLS regressions (questions 12–13) but is controlled for by fixed effects.