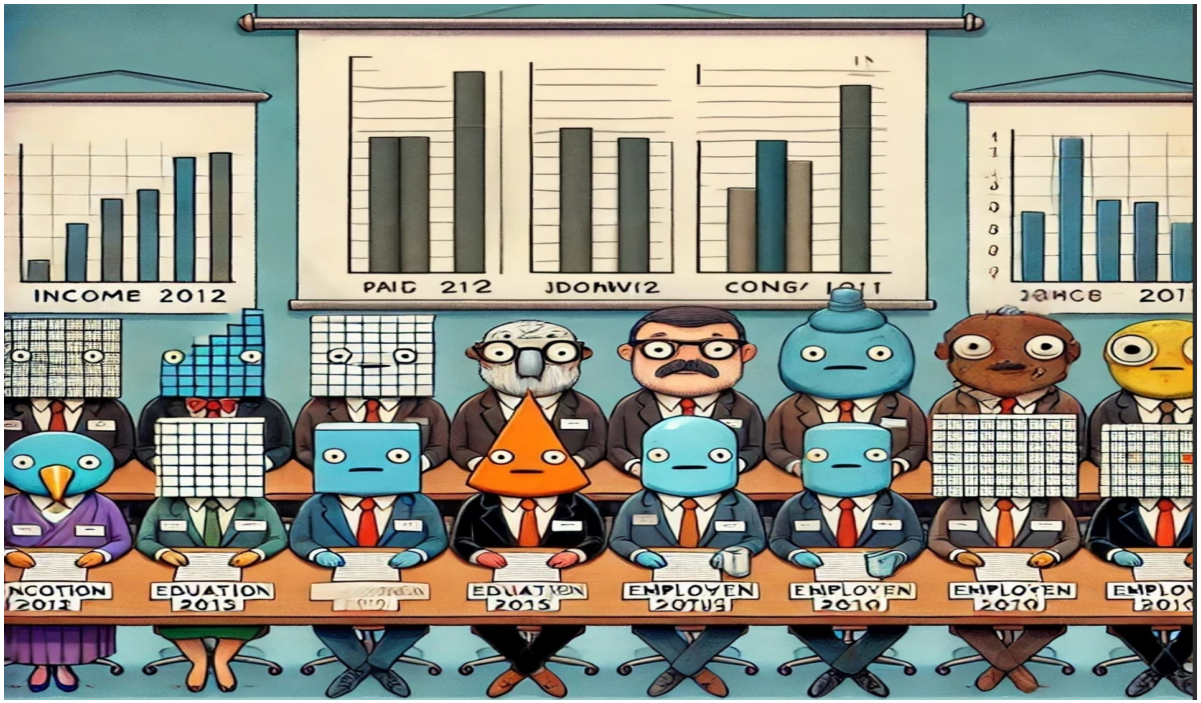
Fundamentals of Econometrics

Lecture 8: Pooling Cross Sections across Time: Simple Panel Data Methods



Section 1

Simple Panel Data Methods

Panel Data

- Panel data is a combination of time series and cross-sectional data.
- It consists of observations on multiple entities (individuals, firms, countries, etc.) over multiple time periods.
- To gather a panel data set, we can either:
  - Collect data on the same individuals over time (e.g., a survey of the same individuals at different points in time).
  - Collect data on different individuals at different points in time (e.g., a survey of different individuals at different points in time).

Independently Pooled Cross Sections

- Collection of independent, random samples from the same population at multiple periods of time.

Some Data Sources

- The Panel Study of Income Dynamics (PSID): Collected by the University of Michigan
- National Longitudinal Surveys (NLS): Collected by the Bureau of Labor Statistics
- Medical Expenditure Panel Survey (MEPS): Collected by the Agency for Healthcare Research and Quality
- National Health and Nutrition Examination Survey (NHANES): Collected by the National Center for Health Statistics at the CDC
- American Community Survey (ACS): Collected by the U.S. Census Bureau
- Current Population Survey (CPS): Collected by the Bureau of Labor Statistics
- American Time Use Survey (ATUS): Collected by the Bureau of Labor Statistics

Using Pooled Cross Sections Data

- We want to pool all cross-sections over time into one single data set

Advantages:

1. Increased sample size: By pooling data from multiple cross-sections, we can increase the sample size, which can lead to more precise estimates.
2. Improve generalizability: By pooling data from different time periods, we can improve the generalizability of our results to the population.
3. Can be used to estimate the effect of a policy change or event on a population (natural experiment).

Natural or Quasi-Experiments

- A natural experiment is when an exogenous shock occurs to a system and affects individual behavior.
- We have a group of individuals affected by the shock and a group of individuals that are not. So this is similar in principle to a laboratory experiment where there is a treatment group (affected by shock) and control group (not affected by shock).
- A quasi-experiment is when a researcher uses a natural experiment to estimate the effect of a treatment on an outcome.

Two requirements: - Two time periods (one before and one after the policy change) - Two groups (treatment and control)

Natural Experiment Framework

Goal: To determine differences between traetment and control groups due to an exogenous shock.

1. Pool the data from the two time periods.
2. Include a dummy variable for time and group.
   - d2 = 1 if obs occurs after event
   - dT = 1 if obs occurs in treatment group
3. Include additional variables and an interaction term between the two dummy variables

$$y = \beta_0 + \underbrace{\delta_0 d2}_{\substack{\text{Controls for} \\ \text{unobserved} \\ \text{changes affecting} \\ \text{both groups}}} + \underbrace{\beta_1 dT}_{\substack{\text{Controls for} \\ \text{initial difference} \\ \text{between groups}}} + \delta_1 d2 \cdot dT + \underbrace{\text{other factors}}_{\substack{\text{Controls for} \\ \text{observable differences} \\ \text{between treatment} \\ \text{and control group}}} + u$$

Group	Before	After	After - Before
Control	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\delta}_0$	$\hat{\delta}_0$
Treatment	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\delta}_0 + \hat{\beta}_1 + \hat{\delta}_1$	$\hat{\delta}_0 + \hat{\delta}_1$
Treatment - Control	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\delta}_1$	$\hat{\delta}_1$

Example: Garbage Incinerator

- Blacksburg is considering the location of a new garbage incinerator. We want to conduct an analysis of the impact of the incinerator on property values.
- We could consider a town similar to Blacksburg where an incinerator was built. We will use Boston housing data from Kiel and McCain (1995). We have data from 1971 and in 1981 (when the incinerator was built).
- Similar town with:
  - 30,000 residents
  - Small college of 2,000 students
  - 25 miles from the nearest city

```
l.before <- lm(rprice ~ nearinc ,data = kielmc, subset = year == 1978)
l.after <- lm(rprice ~ nearinc ,data = kielmc, subset = year == 1981)

stargazer(l.before, l.after, font.size = "scriptsize",
        title = "Pooled Cross Sections", column.labels = c("Before", "After"),
        omit.stat = c("f", "ser", "adj.rsq"), header = FALSE)
```

Table 2: Pooled Cross Sections

	Dependent variable:	
	rprice	
	Before	After
	(1)	(2)
nearinc	−18,824.000***	−30,688.000***
	(4,745.000)	(5,828.000)
Constant	82,517.000***	101,308.000***
	(2,654.000)	(3,093.000)
Observations	179	142
R^2	0.082	0.165
Note:	*p<0.1; **p<0.05; ***p<0.01	

Garbage Incinerator Example

- What is the effect of the garbage incinerator on housing prices?
- Being close to an incinerator depresses prices, but location was one with lower prices to begin with.
- The effect of the garbage incinerator on housing prices is given by the difference in the coefficients of the `nearinc` variable in the two models.

$$\hat{\delta} = -30688.274 - (-18824.370) = \boxed{-11863.903}$$

$\hat{\delta}$ is known as the **difference-in-difference (DiD)** estimator. It can be expressed as the difference over time in the average difference in housing prices between the two groups (near and far from the incinerator):

$$\hat{\delta} = \left(\overline{rprice}_{81,near} - \overline{rprice}_{81,far}\right) - \left(\overline{rprice}_{78,near} - \overline{rprice}_{78,far}\right)$$

DiD Estimator in Regressions

$$rprice_{it} = \beta_0 + \delta_0 after + \beta_1 nearinc + \delta_1 after \cdot nearinc + u_{it}$$

- The differential effect of being in the location **and** after the incinerator was built is given by δ_1.
- The DiD estimator is the difference in the coefficients of `nearinc` in the two models.
- If houses sold before and after the incinerator was built were systematically different, further explanatory variables should be included.
  - **This will also reduce the error variance and thus standard errors.**

DiD Estimator in Regressions

Table 3: Pooled Cross Sections

	Dependent variable:		
	rprice		
	(1)	(2)	(3)
y81	18,790.000***	21,321.000***	13,928.000***
	(4,050.000)	(3,444.000)	(2,799.000)
nearinc	−18,824.000***	9,398.000*	3,780.000
	(4,875.000)	(4,812.000)	(4,453.000)
y81:nearinc	−11,864.000	−21,920.000***	−14,178.000***
	(7,457.000)	(6,360.000)	(4,987.000)
Constant	82,517.000***	89,117.000***	13,808.000
	(2,727.000)	(2,406.000)	(11,167.000)
Other Controls	No	Age, AgeSq	Full Set
Observations	321	321	321
R^2	0.174	0.414	0.660
Note:			*p<0.1; **p<0.05; ***p<0.01

```r
l.difference <- lm(rprice ~ y81 + nearinc + y81:nearinc,
                   data = kielmc)

l.difference2 <- lm(rprice ~ y81 + nearinc + y81:nearinc +
                      age + agesq,
                    data = kielmc)

l.difference3 <- lm(rprice ~ y81 + nearinc + y81:nearinc +
                      age + agesq + intst + land + area +
                      rooms + baths,
                    data = kielmc)

stargazer(l.difference, l.difference2, l.difference3,
        font.size = "scriptsize",
        title = "Pooled Cross Sections",
        # Keep only y81, nearinc, y81:nearinc, "Constant")
        keep = c("Constant", "y81", "nearinc", "y81:nearinc"),
        add.lines = list(c("Other Controls", "No", "Age, AgeSq", "Full Set")),
        omit.stat = c("f", "ser", "adj.rsq"), header = FALSE)
```

Model in Logs

Table 4: Pooled Cross Sections

	Dependent variable:		
	log(rprice)		
	(1)	(2)	(3)
y81	0.193***	0.220***	0.162***
	(0.045)	(0.037)	(0.028)
nearinc	−0.340***	0.007	0.032
	(0.055)	(0.052)	(0.047)
y81:nearinc	−0.063	−0.185***	−0.132**
	(0.083)	(0.068)	(0.052)
Constant	11.300***	11.400***	7.650***
	(0.031)	(0.026)	(0.416)
Other Controls	No	Age, AgeSq	Full Set
Observations	321	321	321
R^2	0.246	0.509	0.733
Note:			*p<0.1; **p<0.05; ***p<0.01

Model in Logs

```
l.difference.log <- lm(log(rprice) ~ y81 + nearinc + y81:nearinc,
                       data = kielmc)
l.difference2.log <- lm(log(rprice) ~ y81 + nearinc + y81:nearinc +
                          age + agesq,
                        data = kielmc)
l.difference3.log <- lm(log(rprice) ~ y81 + nearinc + y81:nearinc +
                          age + agesq + lintst + lland + larea +
                          rooms + baths,
                        data = kielmc)
stargazer(l.difference.log, l.difference2.log, l.difference3.log,
          font.size = "scriptsize",
          title = "Pooled Cross Sections",
          # Keep only y81, nearinc, y81:nearinc
          keep = c("Constant", "y81", "nearinc", "y81:nearinc"),
          add.lines = list(c("Other Controls", "No", "Age, AgeSq", "Full Set")),
          omit.stat = c("f", "ser", "adj.rsq"), header = FALSE)
```

Adding Additional Control Groups

- An implicit assumption of the DiD estimator is that the treatment and control groups are similar in all respects except for the treatment.
- We assume that potential trends in the outcome, y, would trend at the same rate in the absence of the treatment.
- This is known as the **parallel trends assumption**.
- If this assumption is violated, the DiD estimator will be biased.
- We can add an additional control group (say a state that has not received the treatment but has a similar trend pre-treatment) to the model.
- Leads us to a **triple difference** estimator (**Difference-in-Difference-in-Difference**).