Tutorial 6: Introduction to Panel Data

Hammad Shaikh

March 4, 2021

Types of Data

4. Repeated Cross sectional: $(x_i, y_i)_{i=1}^{I}$ in T & $(x_j, y_j)_{j=1}^{J}$ in T'
   - Ex. Canadian census data

1. Cross sectional: same time period measured for different units
   - $\{(x_i, y_i) : \underbrace{i=1, \dots, N}_{\text{multiple units}}\}$ in a single period
     - Ex. Class grades for a given assignment
2. Time series: same unit measured at different time periods
   - $\{(x_t, y_t) : t=1, 2, \dots, T\}$ for a single unit
   - Ex. Canadian GDP over years
3. Panel data: range of units over time periods
   - $\{(x_{it}, y_{it}) : \underbrace{i=1,2,\dots,N,\ t=1,\dots,T}_{NT \text{ data points}}\}$
     - Ex. Daily Covid-19 cases for all Canadian Provinces

Panel Data Example

Wide format

Prov	Educ2000	Educ2001	...
ON			
Al			
⋮			

10

Table: Educational Attainment in Canada

Province	HS Graduation Rate	Years of Education	Year
Ontario	70	13	2000
⋮	⋮	⋮	⋮
Ontario	86.5	16	2018
⋮	⋮	⋮	⋮
Alberta	55	10	2000
⋮	⋮	⋮	⋮
Alberta	70	14	2018

Long format because of year index

*Note: panel data can be wide or long format

- What are the variables? HS Grad. and Years of Educ.
- What is the time period? Years from 2000-2018
- What is the unit of observation? Provinces

Regressions with panel data

Data: $\{(x_{it}, y_{it}) : i=1, \ldots, N, t=1, \ldots, T\}$

$\hat{\beta}_1$ unbias if $Cov(x_{it}, \varepsilon_{it}) = 0$

- Pooled regression: $y_{it} = \beta_0 + \beta_1 x_{it} + \epsilon_{it}$

Ignoring the panel data structure

↳ Not leveraging either the time or unit index

- Individual fixed effects: $y_{it} = \beta_0 + \beta_1 x_{it} + \overbrace{\alpha_i + u_{it}}^{\varepsilon_{it}}$

↳ α_i represents any variables that affect y_{it} but are time-invariant and can differ accross units

↳ Ex. Student's inate academic ability

Ex. Laws in a province in short time

First difference estimator

unit-level fixed effect

- Want to estimate: $y_{it} = \beta_0 + \beta_1 x_{it} + \underbrace{\alpha_i}_{} + u_{it}$

  ↳ Controls for any time-invariant unit level variables that affect outcome ⇒ Can handle endog. $Cov(x_{it}, \alpha_i) \neq 0$

- First difference (FD) estimator: OLS of Δy_{it} on Δx_{it}

  $\underbrace{y_{it} - y_{it-1}}_{\Delta y_{it}} = \beta_1 \underbrace{(x_{it} - x_{it-1})}_{\Delta x_{it}} + \Delta u_{it} \Rightarrow \hat{\beta}_1$ unbias if $Cov(\Delta x_{it}, \Delta u_{it}) = 0$

  ↳ No longer need to worry about $Cov(x_{it}, \alpha_i) \neq 0$

Difference - in - Differences (DD) for Policy Analysis

Video: `https://www.youtube.com/watch?v=V07MKhud-y0`

Control units that never get policy
Treatment units that eventually get policy

- DD used when have a pre-post and control-treatment setting (Environment)

  Before & after policy

- Want to learn impact of funding on school performance (Policy)

  ↳ Policy is educational funding

- School performance in Toronto and Ottawa in 05 and 07 (Data)

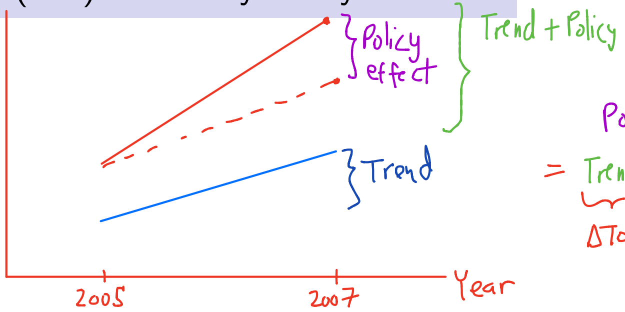
  Cities are units   Years are time periods

- Suppose in 2007 Toronto received school funding (Treatment & Control)

  ↳ Toronto is the treatment city

  Ottawa is control city

Difference - in - Differences (DD) for Policy Analysis



Policy Effect = $\underbrace{\text{Trend + Policy}}_{\Delta \text{Toronto}} - \underbrace{\text{Trend}}_{\Delta \text{Ottawa}}$

- Visualize the analysis:

Assume that performance in Ottawa is the same as Toronto
↳ "Common Trends"

- Turn visualization into a Diff.-in-Diff. regression:

$$\underbrace{Y_{ct}} = \beta_0 + \beta_1 \underbrace{Treat_c} + \beta_2 \underbrace{PostTreat_t} + \beta_3 Treat_c \times PostTreat_t \Rightarrow \hat{\beta}_3 \text{ is the policy effect}$$

c = city
t = year

$Treat_c = \begin{cases} 1, & \text{Toronto} \\ 0, & \text{Ottawa} \end{cases}$

↳ $PostTreat_t = \begin{cases} 1, & \text{Year} = 2007 \\ 0, & \text{Year } 2005 \end{cases}$

↳ D-in-D estimate

Panel data practice - Chapter 13, Question 5

$$\underbrace{S_{it}}_{\text{Savings}} = \beta_0 + \beta_1 \, Year1992_t + \beta_2 \, age_{it} + \varepsilon_{it}$$

* Panel data only include controls that vary in (i,t) index.

We want to estimate the effect of several variables on annual saving. Suppose we we have a panel data set on individuals collected on 1990, and 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.

$$\Delta S_{it} = S_{i92} - S_{i90} = \beta_1(1) + \beta_2 \underbrace{\Delta age_{it}}_{2} + \Delta\varepsilon_{it}$$

↳ Identification problem to seperate β_1 from β_2

↳ Doesn't make sense to reg. ΔS_{it} on $\underbrace{\Delta age_{it}}_{2\text{, no variation}}$