Causal Inference (6.S059/15.C08/17.C08)

Recitation, Week 10.

Topic: Natural Experiments and Fixed Effects

Benjamín Muñoz

April 26, 2024

MIT

Massachusetts Institute of Technology

Table of contents

1/ Natural Experiments

Natural Experiments

1. Naturally occurring phenomena where the treatment is not under the analyst's control, but this condition can be characterized as ”as-if” random (Dunning, 2012).

2. Observational design in which the treatment assignment mechanism (i) is not designed or controlled by the researcher; (ii) it is unknown and unknowable by the researcher; and (iii) it is probabilistic by an external event that is outside the control of the experimental units (Titiunik, 2021).

   ⋆ Additional assumptions are needed to justify causal identification!

2/ Panel Data and Fixed Effects

Fixed Effects Set-Up

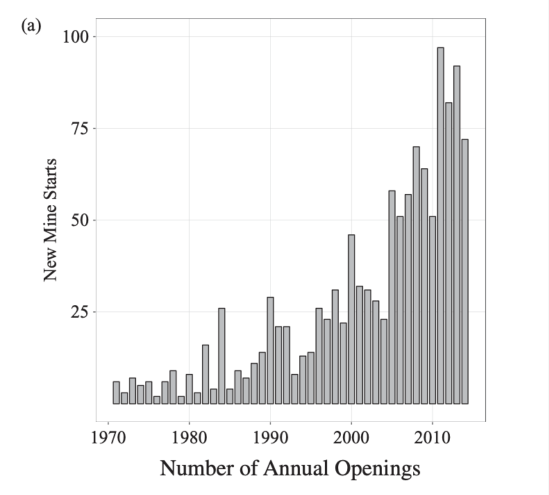
- Units are grouped into different **groups**.
- Confounders have different impact across different **groups**.
- Within each **group**, however, treatment is ignorable.

Fixed Effects Intuition

- **Causal effect can be identified within each group even in the presence of between-group confounder!**.
- How? Difference between treated and control units come from:
    1. Some groups are more likely to have treated members than others i.e. selection bias/between-group confounding.
    2. Within groups, some members are treated and some are control.
- Group fixed effects should account for all between-group variation.
- Any remaining difference between treated and control units is only due to treatment.
- Like in a block randomized experiment.

Fixed Effects Example

- Christensen (2019): How do commercial mining investments affect the likelihood of social protest in sub-Saharan Africa?
- Strategy: Fixed effects where units i are 5 km x 5 km grid-cells observed over t periods. (That's a lot of grid cells!)



Fixed Effects Example

(b)

Mine Locations

Notes: (a) displays the number of mines opened (i.e., starting production) in every year from 1970-2014; (b) maps all unique mines with start dates. Data from IntierraRMG, SNL Metals and Mining, and Mining eTrack databases.

Panel Data Set-Up

- Two dimensions: n units and t periods.
- N units across T periods $\longrightarrow$ $N \times T$ observations.
- Analogous to a block-randomized experiment where the blocks are units and treatment is randomized over periods t.
- Matrix notation:
    1. $\mathbf{y}$ is $NT \times 1$ matrix of observed outcomes.
    2. $\mathbf{X}$ is $NT \times K$ matrix of K covariates.

Fixed Effects Assumptions

- Conditional ignorability or selection on observables:

$$\{Y_{0it}, Y_{1it}\} \perp\!\!\!\perp D_{it} | X_{it}$$

- With panel data, we can relax or make that assumption more believable by conditioning on time-invariant confounders, i.e. selection on time-invariant observables:

$$\{Y_{0it}, Y_{1it}\} \perp\!\!\!\perp D_{it} | X_{it}, \alpha_i$$

  ★ Examples of time-invariant confounders?

  ★ Examples of time-variant confounders?

Pooled OLS versus Fixed Effects

- Pooled OLS: $y_{it} = \boldsymbol{x}_{it}^T \boldsymbol{b} + \tau D_{it} + \nu_{it}$
  - D_{it} captures whether grid-cell i has an active mine in year t.
  - Ignoring panel structure creates composite error: $\nu_{it} = \epsilon_{it} + \alpha_i$.
  - Unbiasedness requires that ν_{it} is not correlated with past, current, and future $\boldsymbol{x}_{it}^T$ and D_{it} ...and that seems unlikely!
  - Example: Whether or not i has a mine (D_{it}) may depend on time-invariant characteristics (α_i) such as the likelihood of natural resources underground.
- Fixed Effects Model: $y_{it} = \alpha_i + \boldsymbol{x}_{it}^T \boldsymbol{b} + \tau D_{it} + \epsilon_{it}$
  - Interpretation of α_i ? All aspects of grid-cell i that do not vary over time, like natural resource wealth.

How do I estimate FEs?

- LSDV estimation: Think of α_i as coefficients on unit dummy variables you include in OLS.

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}^T \boldsymbol{b} + \tau D_{it} + \epsilon_{it}$$

- Interpretation of α_i ? Unit-specific intercept, i.e. the mean of y_{it} for grid-cell i when all other variables are 0.
- Problem? Really slow in R/Python. (Think about the # of grid-cells!).
- "Within" estimation: Demean variables with the mean value for each unit i, i.e. $\bar{y}_i$ and $\bar{\boldsymbol{x}}_i$ or the mean across time for each i.

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}^T \boldsymbol{b} + \tau \ddot{D}_{it} + \epsilon_{it}$$

★ You should normally cluster your SEs, as the FE will not control for all of the intra-cluster correlation.

Back to the Assumptions

- Selection on time-invariant unobservables: $\{Y_{0it}, Y_{1it}\} \perp\!\!\!\perp D_{it}|X_{it}, \alpha_i$.
- Alternatively stated as strict exogeneity (conditional on the unobserved effect):

$$\mathbb{E}[\epsilon_{it}|D_{i1}, \ldots, D_{iT}, \boldsymbol{x}_{i1}, \ldots, , \boldsymbol{x}_{iT}, \alpha_i] = 0, t = 1, 2, \ldots, T.$$

- When is strict exogeneity violated?
    1. Time-varying omitted variables: Mining and protests are increasing overtime due to unobserved factor Z_{it}
        - Violation: Presence of companies from certain countries (like China?) changes over t.
        - What about general time trends? That's effectively a time-varying omitted variable. See next slide!

Back to the Assumptions

2. Simultaneity: D_{it} is not related to what happened in the past, i.e. $D_{it} \not\Longrightarrow Y_{it} \not\Longrightarrow D_{it+1}$
    - Violation: Mining companies decide whether to keep mining based on past protest activity.

Time Fixed Effects

- **Intuition:** Strict exogeneity is often violated due to common shocks affecting all units that are correlated with x_{it}.

$$y_{it} = \alpha_i + \delta_t + \boldsymbol{x}_{it}^T \boldsymbol{b} + \tau D_{it} + \epsilon_{it}$$

- Models a "common shock" in each time period.
- Example: Commodity shocks or general economic trends.
- Alternatively known as "Two-Way" Fixed Effects.

FEs estimation in R/Python

1. `plm` from the `plm` package:
   - model = ”within” to avoid LSDV estimation.
   - Enter your FE id with the index option.
   - Cluster your SEs using vcovBK() and coeftest()
2. `lfe` from the `felm` package.
3. `lm_robust` from the `estimatr` package.
   -
   - Enter your FE id with the `fixed_effects` option.
4. `feols` from the `fixest` package.
   - Enter your FE id after the | operator in the formula.
   - Flexible options for multiway FEs + multiway clustering.

Python Code

Python Code

```
from linearmodels.panel import PanelOLS   # FE models

### Setting multiindex for panel data
panel_df = data.set_index(['firm_id', 'year'])

### Run a Linear model with Firm Fixed Effects
fe_reg = PanelOLS.from_formula('roce ~ zmanagement + sic + lsales + lemp + EntityEffects', data=panel_df).fit()
print(fe_reg.summary.as_text())

### Model with firm FE
fe_reg2 = PanelOLS.from_formula('roce ~ zmanagement + ever_family_ceo + EntityEffects',
drop_absorbed=True, data=panel_df).fit()
print(fe_reg2.summary.as_text())
```

- See this link.
- Other options: `PyFixest` is a Python implementation of `fixest`. Another good library is `pyhdfe`.