Panel Data & Fixed Effects

Statistic Modeling & Causal Inference | Oswald & Ramirez-Ruiz

Agenda

- Lecture review
  - Cross sectional and temporal data
  - Panel Data
  - Fixed Effects estimation
  - LSDV estimation

- Handling panel data in R

Panel Data

- Gathering multiple observations over time as powerful way to achieve causal identification
- **Cross-sections:** Samples of different units measured at the same point in time.
- **Time series:** The same subject or unit measured at different points in time.
- **Panel data:** Multiple units over multiple points in time.
- **Balanced panels**, a.k.a. longitudinal data, record *the same* individuals over time in waves and all individuals are measured at all points in time.

Two dimensions of panel data

1. Cross section

Measure different units at one point of time

CSU vote shares

Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...

Static group comparison
We compare treatment and control groups after treatment

$$Y_{i1}^0 = \theta_i^0 + \delta_1 + v_{i1}^0$$
$$Y_{i1}^1 = \tau + \theta_i^1 + \delta_1 + v_{i1}^1$$

Observations are simultaneous. Therefore, time **period effects (δ) cancel out.**

Assumptions to estimate Treatment effect (τ):

- **Exogeneity:** The idiosyncratic error is independent of treatment.
- **Random effects:** unobserved differences in units are independent of treatment.

(Potential outcomes of control group are the same as the counterfactual potential outcomes for those being treated.)

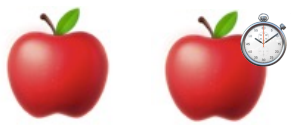
Two dimensions of panel data

1. Temporal

Measure one unit at different points in time

CSU vote shares

Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...



Longitudinal comparison
We compare Y for one unit before and after treatment

$$Y_{i0}^0 = \delta_0^0 + \theta_i + v_{i0}^0$$
$$Y_{i1}^1 = \tau + \delta_1^1 + \theta_i + v_{i1}^1$$

Because we are comparing a unit to itself, **unit-specific effects (θ) cancel out.**

Assumptions to estimate Treatment effect (τ):

- **Exogeneity:** The idiosyncratic error is independent of treatment. The effects of transitory forces cancel out over time.
- **Temporal stability:** no impact of unobserved time-varying factors. In the absence of treatment, there would be no change in the mean of Y.

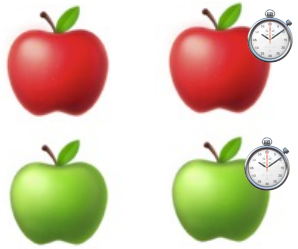
(Average potential outcome for the unit observed does no change in time)

Panel Data

Measure of **many** units,
at **multiple** different points in time.

CSU vote shares

Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...



With panel data we can **relax certain assumptions.**
We do not need to assume:

- Exogeneity: Because theta is constant with time-series data.
- Temporal stability: Because delta is constant in cross-sectional data.

$$E[Y_{i1}^1 - Y_{i1}^0] - E[Y_{i0}^1 - Y_{i0}^0] = \tau + E[\epsilon_{i1}^1 - \epsilon_{i1}^0] - E[\epsilon_{i0}^1 - \epsilon_{i0}^0]$$

To identify tau, this term must be zero.
For that to happen, the error terms of treatment and control groups do not need to be the same in each time period, they only need to change in parallel across time.

Fixed Effects

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \underbrace{\theta_i + \delta_t + \upsilon_{it}}_{\epsilon_{it}}$$

Treatment indicator

ϵ_{it} Error term of the observation of one unit at one point in time

ϵ_{it}

- θ_i Captures unit fixed effects: unmeasured characteristics of the units that *don't change in time* and do affect the outcome (y). Think of the size of a city, climate, location, gender.
- δ_t Captures time fixed effects: effects that take place at a certain time period but affect the outcome (y) of all units simultaneously. Think of a global economic shock, changes in national government.
- υ_{it} The 'idiosyncratic' error (classical error) that contains factors that are both specific to unit **and** time.

→ With panel data we can cancel out both unit and time fixed effects, even if we cannot observe or measure the variables involved.
→ We only need to be care about confounding variables that vary **both** by unit **and** time period.

Estimation 1 – "de-meaning"/FE

θ_i

Given that the unit fixed effects are constant in time, we can remove their influence by subtracting the across time average outcome of each unit, from the y value of each observation.

δ_t

Likewise, given that time fixed effects are constant across units, we can remove their effect by subtracting the across unit average outcome of each period, from the y value of each observation.

For group 1, $\overline{Y}$=20. For group 2, $\overline{Y}$=35.

		Y raw	θ_i partialled out
Group 1	I_{11}	35	15
	I_{12}	15	-5
	I_{13}	10	-10
Group 2	I_{21}	43	8
	I_{22}	40	5
	I_{23}	19	-16

$\overline{Y}_{period\ 1} = 39 \mid \overline{Y}_{period\ 2} = 27.5 \mid \overline{Y}_{period\ 3} = 14.5.$

		Y raw	δ_t partialled out
Period 1	I_{11}	35	-4
	I_{21}	43	4
Period 2	I_{12}	15	-12.5
	I_{22}	40	12.5
Period 3	I_{13}	10	-4.5
	I_{23}	19	4.5

If we estimate our regression model using this "de-meaned" equation, we are left with the FE (fixed effects) estimator: a model in which all the confounders that don't vary over time and units just drop out.

Estimation 2 – LSDV

- Least Squares Dummy Variables
- A second way to estimate fixed effects is to create dummy variables indicating the **unit of every observation / time of every observation**

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \underbrace{\beta_2 U_1 + \beta_3 U_2 + \cdots + \beta_i U_{i-1}}_{\text{unit dummies}} + v_{it}$$

- **Two-way fixed effects:** Add both dimensions as dummy variables (i-1 unit dummies, t-1 time dummies) to regression model → generalization of DiD
- In this case, do not include covariates that don't change over time and in the model, or variables that only change over time but not across units!

Further Resources

For any coding issues – Stackoverflow

Hertie's Data Science Lab – Research Consulting