Statistics II

Week 9:

Panel data and fixed effects

Lecture Review

Panel Data

As we introduced last week, a powerful way to achieve causal identification is to gather multiple observations over time.

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.

Two components of the panel data setup

(Remember slides from last week)

1. Cross section

Measure of different units, at one point 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
...	...	...	...

Static group comparison:
We compare the T and C groups after treatment.

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

Because observations are **simultaneous**, time period effect is the same and cancels out.

To estimate *tau*, we need to make strong **assumptions:**

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.)

That is assuming a lot...

Two components of the panel data setup

Temporal

Measure of 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 the 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 are the same across time and are ruled out.

To estimate tau, we need to make strong **assumptions**:

Exogeneity: The idiosyncratic error is independent of treatment. The effects of transitory forces cancel out over time.

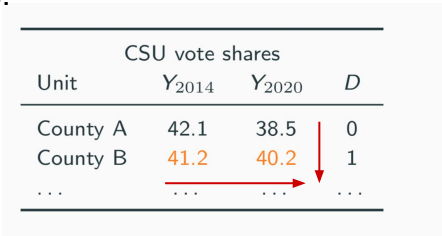
Temporal stability: no impact of unobserved time-varying factors. In the absence of treatment, there would be no change in the mean of y.

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

That is also assuming a lot...

Panel Data

Measure of **many** (and the same) units, at **many** 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 **<u>do not</u>** need to assume

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

We rely on a weaker assumption:

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

To identify *tau,* ____ must be zero. For that to happen, the error terms of T and C *do not need to be the same* in each time period, they only need to change in parallel across time.

(Remember DiD parallel trends assumption)

Fixed Effects

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

The error term of the observation of one unit at one point in time

In a panel data setup we can decompose the error term in:

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; also gender or racial identity.
- 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 fixed effects, even if we cannot observe or measure the variables involved.

We would only need to be careful for confounding variables that vary **both** by unit **and** time period.

Estimation (1/2)

Fixed effects – or “de-meaning”

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

		Y raw	θ_i partialled out
i_1	I_{11}	35	15
	I_{12}	15	-5
	I_{13}	10	-10
i_2	I_{21}	43	8
	I_{22}	40	5
	I_{23}	19	-16

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

		Y raw	δ_t partialled out
t_1	I_{11}	35	-4
	I_{21}	43	4
t_2	I_{12}	15	-12.5
	I_{22}	40	12.5
t_3	I_{13}	10	-4.5
	I_{23}	19	4.5

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \theta_i + \upsilon_{it}$$
$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\theta}_i + \bar{\upsilon}_i$$

Subtract

$$Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + (\upsilon_{it} - \bar{\upsilon}_i)$$

If we estimate our regression model using this time-demeaned equation, we are left with the **FE estimator**: a model where all the confounders that don’t vary over time just drop out.

Estimation (2/2)

Least Squares Dummy Variables (LSDV)

A second way to estimate fixed effects is to create dummy variables indicating the unit of every observation, and include them in the regression equation.

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

! In this case, do not include covariates that don't change over time in the model. They are already accounted for.

Two-way fixed effects: We can also add dummy variables for every time period observed to account for δ_t , unobserved time-specific effects. (This is a generalization of DiD).

! In this case, do not include covariates that don't change over time in the model, or variables that only change over time but not across units.

FE estimator and LSDV are mathematically identical

Questions?