Lecture#11
Panel Data

Cross – section

- Many observation units.

- Each observed just once.

- Examples:

1. Student grades in the nth year of studies.
2. Customer decision(s) during a single shopping trip.
3. Firm’s bids in a procurement auction.

Time series

- Same phenomenon for the same unit observed many times at different points in time.

- Examples:

1. Inflation at the monthly level for a country.
2. Stock market index by minute during a day.
3. Electricity prices at 12.00 for 400 days in a row.

Panel data

- Observe same units several times.

- Examples:

1. Individuals annual income and jobs for t years in the Finnish job market.
2. Finnish firms' accounting information since 2000.
3. Prices and sold quantities for each car type on sale in Finland 2000 – 2015.

Panel data

- Formally, one observes Y_{it}, X_{it} for
- units i=1, …, n, and
- periods t=1,…,T

Balanced vs. unbalanced

- Panel data is ***balanced*** if all units are observed for the same time periods.

- Panel data is ***unbalanced*** if this is not the case.

- Examples:

1. Firm panel data unbalanced because firms are born and die.
2. Customer panel data unbalanced because customers appear and disappear.

What does panel data bring to the table?

- In a cross-section, the only source of variation is ***across observation*** units.
- In time-series, the only source of variation is changes ***over time***.
- Panel data combines these.
- Gil: same theatres' prices (and other) observed over several years.

What does panel data bring to the table?

- Consider the univariate regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

What does panel data bring to the table?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

- With enough time-series data, you could estimate this separately for each observation unit.

$$Y_{it} = \beta_{0i} + \beta_{1i} X_{it} + u_{it}$$

What does panel data bring to the table?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

- With enough observation units, you could estimate this separately for each time period.

$$Y_{it} = \beta_{0t} + \beta_{1t} X_{it} + u_{it}$$

What does panel data bring to the table?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

- Or you could decide on some combination.
- Why? To **reduce bias** & **increase precision** of your parameter estimates.
- Is there any reason to think the effect of X on Y varies over time?
- Is there reason to think the effect of X on Y varies across observation units?

What does panel data bring to the table?

- The **Fixed effects** panel data estimator

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + u_{it}$$

- Example: Effect of R&D (=X) on productivity (=Y).

- What is the interpretation of β_{0i}?

- Firms have different productivity levels even when they invest the same amount in R&D.

What does panel data bring to the table?

- The **Fixed effects** panel data estimator

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + u_{it}$$

- It is natural to see the FE estimator as a generalization of the cross-section regression that you would (have) run.

2 – period example

- Imagine you observe customers in 2 time periods and know how much advertising they are subjected to.
- You are interested in the amount of sales that ads generate.
- For simplicity, let’s assume you have randomized the ads.

2 – period example

1. β_{0i} disappear.

2. ➔ they could be correlated with u_{it}.

3. Note what variation (”within-variation”) is left to identify the parameters.

4. ➔ Needed: changes w/in an observation unit in both X and Y.

2 – period example

5. If no variation left, then ”everything” explained by β_{0i}.

6. Famous example: Firm level R&D.

7. Problem: dummy variables.

Example: Gil's data

theaterid	year	highprice	dhighprice	vi	dvi
1	1945	0.600		1	
1	1946	0.607	0.007	1	0
1	1953	0.720	-0.173	0.555556	-0.44444
1	1954	0.876	0.156	0	-0.55556
1	1955	0.976	0.101	0	0
3	1954	1.000		0	
3	1955	1.000	0.000	0	0
4	1945	0.729		0	
4	1946	0.681	-0.048	0	0
4	1947	0.704	0.023	0	0
4	1948	0.700	-0.004	0	0
4	1953	1.098	0.398	0	0
4	1954	1.102	0.005	0	0
4	1955	1.114	0.012	0	0

Dummy variable approach

- Add a dummy variable for each ***observation unit***:

$$Y_{it} = \beta_1 X_{it} + \beta_{01} D_1 + \beta_{02} D_2 \ldots + \beta_{0n} D_n + u_{it}$$

- These are analogous to other dummy variables, almost.

- The differences: what happens to #variables when n increases?

Dummy variable approach

- Number of variables should not be a fcn of the number of observation units.

- Remedy:

1. (First) differencing.

2. Taking deviations from means (software do this).

Gil univariate – different estimators

Variable	ols	fd	dum	fe
vi	-0.147		-0.231	-0.231
	0.014		0.016	0.023
	0.000		0.000	0.000
dvi		0.003		
		0.016		
		0.853		
N	2685	2292	2685	2685
r2	0.029	0.000	0.67	0.084

Time Fixed effects

- The **Fixed effects** panel data estimator with time FE is

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + \beta_t + u_{it}$$

FE assumptions

- A1: conditional distribution of u has mean zero given **X**.

$$\mathrm{E}[u_{it}|\boldsymbol{X}_{it}, \alpha_i] = 0$$

- A2: $\boldsymbol{X_{it}}$, Y_{it} i = 1, …, n and t = 1, …, T are i.i.d.
- A3: $\boldsymbol{X}_{it}$ and Y_{it} have nonzero finite *fourth* moments.
- A4: No perfect multicollinearity.
- A5: the errors for a given obs. unit are uncorrelated over time conditional on the observables

$$cov[u_{it}, u_{is}|\boldsymbol{X}_{it}, \alpha_i] = 0 \text{ for } t \neq s$$

- A5: the errors are uncorrelated over time conditional on the observables.

- Let’s use the R&D example.

- The ”shock” that leads to high (low) productivity today disappears and the new ”shock” tomorrow is uncorrelated.

- What could be a shock to productivity? E.g.,

1. A new idea that gets implemented (and e.g. decreases waste).

2. A new product that is introduced (and sells well at a high price).

- Some shocks are not transitory (i.e., they affect Y over many periods).