Empirical Panel Data: Lecture 2

Instructor: Chaoyi Chen
NJE & MNB

March 14, 2023

Topic 2: Data

- In econometrics, data come from one of the two sources: experiments and non-experimental observations
  - Experimental data are based on (randomized controlled) experiments designed to evaluate a treatment or policy or to investigate a causal effect.
  - Data obtained outside an experimental setting are called observational data (issued from survey, administrative records etc...)
- All of this lecture is devoted to methods for handling real-world observational data

Topic 2: Data

- Whether the data is experimental or observational, data sets can be mainly distinguished in three types:
  1. Cross-sectional data
  2. Time series data
  3. Panel data

Topic 2: Cross-sectional data

- Cross-sectional data:
  - Sample of agents taken at one point in time. Data for different entities: workers, households, firms, cities, countries, and so forth.
  - No time dimension (even if date of data collection varies somewhat across units, it is ignored).
  - Order of data does not matter!

Topic 2: Time series data

- Time series data:
  - Repeat observations on specific agents over time. Examples include stock prices, money supply, consumer price index, GDP etc,.
  - Order of data is important!
  - Observations are typically not independent over time;

Topic 2: Panel data

- Panel data (or longitudinal data):
  - Have repeat observations for the same agents in different time periods.
  - Combine cross-sectional and time series issues.
  - Present several advantages with respect to cross-sectional and time series data (depending on the question of interest!).
- Terminology and notations:
  - **Individual or cross section unit**: country, region, state, firm, consumer, individual, couple of individuals or countries
  - **Double index** : i (for cross-section unit) and t (for time)

$$y_{it} \text{ for } i = 1, \ldots, n \text{ and } t = 1, \ldots, T$$

Topic 2: Balanced and unbalanced panel

- Balanced panel: panel is said to be balanced if we have the **same** time periods, t = 1, ..,T, for each cross section observation.
- Unbalanced panel: A panel is said to be balanced if the time dimension, denoted T_i, is **specific** to each individual.

Topic 2: Balanced panel: example

Country ID	Year	GDP	CAPITAL	LABOR
1	2008	2.409	22.052	0.211
1	2009	2.442	22.048	0.204
1	2010	2.479	21.944	0.204
1	2011	2.504	22.002	0.240
2	2008	5.031	24.524	2.367
2	2009	5.047	24.809	2.385
2	2010	5.083	24.792	2.410
2	2011	5.111	24.685	2.430
3	2008	6.013	25.076	2.897
3	2009	5.952	24.817	2.916
3	2010	6.049	24.979	2.904
3	2011	6.107	25.073	2.935

Table: Balanced panel with $T = 4$ and $n = 3$

Topic 2: Unbalanced panel: example

Country ID	Year	GDP	CAPITAL	LABOR
1	2009	2.442	22.048	0.204
1	2010	2.479	21.944	0.204
1	2011	2.504	22.002	0.240
2	2008	5.031	24.524	2.367
2	2009	5.047	24.809	2.385
2	2010	5.083	24.792	2.410
2	2011	5.111	24.685	2.430
3	2006	5.887	24.914	2.894
3	2007	5.974	25.063	2.895
3	2008	6.013	25.076	2.897
3	2009	5.952	24.817	2.916
3	2010	6.049	24.979	2.904
3	2011	6.107	25.073	2.935

Table: Unbalanced panel with $T_1 = 3$, $T_2 = 4$, $T_3 = 6$, and $n = 3$

Topic 2: Panel data model

- A panel data regression model (or panel data model) is an econometric model specifically designed for panel data.
- Advantages of the panel data sets and the panel data models:
  1. Use a larger number of observations
  2. New economic questions (identification)
  3. Unobservable components. control for unobserved heterogeneity
  4. (sometimes) Easier estimation and inference

Topic 2:Panel data model specifications

- Depending on whether allowing for parameter heterogeneity, the panel data model can be mainly categorized as the following two models:
  - Homogeneous panel data model: Both slope and intercept coefficients are the **same**.
  - Heterogeneous panel data model: Slope or intercept coefficients or both are **varying** across i or t or both.

Topic 2: Homogeneous (pooled) panel data model

- Let us consider the following linear model

$$y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it},$$

  where for $i = 1, \ldots, n$ and $t = 1, \ldots, T$,

- α is a scalar and is **constant** across i and t,
- $\beta = [\beta_1, \ldots, \beta_k]'$ is a $k \times 1$ vector of parameters that is the **same** across i and t,
- $x_{it} = [x_{it,1}, \ldots, x_{it,k}]'$ is a $k \times 1$ vector of **exogenous** variables,
- ε_{it} is an error term.

Topic 2: Heterogeneous panel data model

- Let us consider the following linear model

$$y_{it} = \alpha_{it} + \beta_{it}' x_{it} + \varepsilon_{it},$$

  where for $i = 1, \ldots, n$ and $t = 1, \ldots, T$,

- α_{it} is a scalar and is **varying** across i and t,
- $\beta = [\beta_{it,1}, \ldots, \beta_{it,k}]'$ is a $k \times 1$ vector of parameters that is the **varying** across i and t,
- $x_{it} = [x_{it,1}, \ldots, x_{it,k}]'$ is a $k \times 1$ vector of **exogenous** variables,
- ε_{it} is an error term.

Topic 2: Restrictions on the regression coefficients

- The heterogeneous panel data model proposed in the previous slide is not feasible for estimation with our available data.
- Econometricians propose new models to simplify while maintaining flexibility in parameter heterogeneity.
- Different models impose restrictions on regression coefficients to simplify and account for heterogeneity.

Topic 2: Parameters are constant over time but vary over individuals

- We can assume that the parameters (including both slope coefficients and intercepts) are constant over time (no structural break, no regime switching, etc.), but can vary across individuals:

$$y_{it} = \alpha_i + \beta_i' x_{it} + \varepsilon_{it}.$$

- Caution: Only feasible to estimate if we have large enough t for **each** i!

Topic 2: Intercepts are constant over time but vary over individuals

- Instead, we can only assume intercepts vary over individuals:

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}.$$

- Constant terms α_i capture unobserved individual effects in this panel data model.
- This type of heterogeneous panel data model is the focus of this semester.
- Individual effects are an important consideration in panel data analysis.
- Ignoring individual effects *may* lead to biased or inconsistent estimates.