Introduction to Panel Data Econometrics

Jakub Mućk
SGH Warsaw School of Economics

Course outline

Literature

Basic:

1. Baltagi B. H., (2014), *Econometric Analysis of Panel Data*, 5th edition, Wiley.
2. Wooldridge J. M., (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, The MIT Press.

Additional :

1. Angrist J. D. and Pischke J.-S., (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.
2. Arellano M., (2004), *Panel Data Econometrics. Advanced Texts in Econometrics*, Oxford University Press.
3. Baltagi B. H. (ed.), (2015), *The Oxford Handbook of Panel Data*, Oxford University Press.
4. Cameron C. A. and Trivedi P. K., (2006),*Microeconometrics: Methods and Applications*, Cambridge University Press.
5. Hsiao Ch., (2014), *Analysis of Panel Data*, 3rd edition, Cambridge University Press.
6. Pesaran H. (ed.), (2015), *Time Series and Panel Data Econometrics*, Oxford University Press.

1. Panel data - basic definitions, characteristics, etc.
2. Linear static model - common types. Fixed and random effects approaches.
3. Fixed effects estimation. Random effects models.
4. Hausman test. The between estimator.
5. Two-way error component models.
6. Heteroskedasticity, serial correlation and cross-sectional dependence in static models. GLS estimation.
7. Endogeneity. Instrumental variables (IV) regression. The Hausman-Taylor estimator.
8. Dynamic panel data models. The FD (first differences) estimator. The Nickell's bias. The Anderson-Hsiao estimator.
9. Estimation of dynamic models. Generalized Method of Moments (GMM). The Arellano-Bond estimator and a system estimator.
10. Heterogeneous panels. Seemingly unrelated regression. Swamy's random coefficient model. The Mean Group estimator. The Common Correlated Effects Mean Group estimator.
11. Panel unit root tests. Panel cointegration tests.
12. Nonstationary panels. Panel VAR.
13. Limited dependent variable. The FE logit. The RE binary outcome models.
14. The FE and RE Poisson models. The RE tobit model.
15. Estimating average treatment effects (ATE). Difference-in-differences (DID).

- **Office hours:** MS Teams
- **E-mail:** jmuck@sgh.waw.pl.
- **WWW:** http://web.sgh.waw.pl/~jmuck/
  $\Longrightarrow$ teaching
  $\Longrightarrow$ Econometrics of Panel Data.
- **Software:** Stata (+ R).
- **Final grades:**
  - Homeworks.

Panel Data

Panel of data consists of a group of cross-section units (people, firms, states, countries) that are observed over the time:

- Cross-section:
  y_i where
  $i \in \{1, \ldots, N\}$.
- Time series:
  y_t where
  $t \in \{1, \ldots, T\}$.
- Panel data:
  y_{it} where
  $i \in \{1, \ldots, N\}$
  $t \in \{1, \ldots, T\}$.

In general,

- N - the cross-sectional dimension.
- T - the time dimension.

We might describe panel data using T and N:

- **long/short** describes the time dimension (T);
- **wide/narrow** describes the cross-section dimension (N);

For example: panel with relatively large N and T: long and wide panel.

- In a **balanced** panel, each individuals(unit) has the same number of observation.
- **Unbalanced** panel is a panel in which the number of time series observations is different across units.

Advantages of Panel Data (Baltagi, 2014)

- Controlling for **individual heterogeneity**.
- Panel data offer more informative data, more variability, less collinearity among the dependent variables, more degrees of freedom and more efficiency in estimation.
- Identification and measurement of effects that are simply not detectable in pure cross-section or pure time-series data.
- Testing more complicated behavioral models than purely cross-section or time-series data.
- Reduction in biases resulting from aggregation over firms or individuals.
- Overcome the problem of nonstandard distributions typical of unit roots tests $\Longrightarrow$ macro panels.

Limitations of Panel Data

SGH

- **Design and data collection problems:**
  - coverage;
  - nonresponse;
  - frequency of interviewing;
- **Distortions of measurement errors**
- **Selectivity problems:**
  - self-selectivity;
  - nonresponse;
  - attrition;
- **Short T.**
- **Cross-sectional dependence.**

- Classical example
  Agricultural Cobb-Douglas production function. Consider the following model:

$$y_{it} = \beta x_{it} + u_{it} + \eta_i \tag{1}$$

  - y_{it} – the log output.
  - x_{it} – the log of a variable input;
  - η_i – an farm-specific input that is constant over time, e.g., soil quality.
  - u_{it} – a stochastic input that is outside framer's control, e.g., rainfalls.
  - β - the technological parameter.

- Classical example
  Agricultural Cobb-Douglas production function. Consider the following model:

$$y_{it} = \beta x_{it} + u_{it} + \eta_i \tag{1}$$

  - y_{it} – the log output.
  - x_{it} – the log of a variable input;
  - η_i – an farm-specific input that is constant over time, e.g., soil quality.
  - u_{it} – a stochastic input that is outside framer's control, e.g., rainfalls.
  - β - the technological parameter.

- **An example in which panel data does not work**
  Returns to education. Consider the following model:

$$y_{it} = \alpha + \beta x_{it} + u_{it} \tag{2}$$

  - y_{it} – the log wage;
  - x_{it} – years of the full-time education;
  - β – returns to education.

  In addition:

$$u_{it} = \eta_i + \varepsilon_{it} \tag{3}$$

  where η_i stands for the **unobserved individual** *ability*.
  Problem: x_{it} lacks of time variation.

Pooled OLS estimator

- **Pooled model** is one where the data on different units are pooled together with **no assumption on individual differences**:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \tag{4}$$

  where
  - y_{it} – the dependent variable;
  - x_{kit} – the $k-th$ explanatory variable;
  - u_{it} – the error/disturbance term;
  - β_0 – the intercept;
  - $\beta_1, \ldots, \beta_k$ – the structural parameters;
- Note that the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are the same for all unit (do not have i or t subscript).

SGH

- **Pooled model** is one where the data on different units are pooled together with **no assumption on individual differences**:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \tag{4}$$

  where
  - y_{it} – the dependent variable;
  - x_{kit} – the $k-th$ explanatory variable;
  - u_{it} – the error/disturbance term;
  - β_0 – the intercept;
  - β_1, ..., β_k – the structural parameters;
- Note that the coefficients β_0, β_1, ..., β_k are the same for all unit (do not have i or t subscript).

Pooled OLS estimator

SGH

Assumptions (for linear pooled model):

$$\mathbb{E}(u) = 0 \quad (5)$$

$$\mathbb{E}(uu') = \sigma_u^2 I \quad (6)$$

$$rank(X) = K + 1 < NT \quad (7)$$

$$\mathbb{E}(u|X) = 0 \quad (8)$$

- (8): X is nonstochastic and is not correlated with u.
- (6): the error term (u) is not autocorrelated and homoscedastic.
- (8) $\Longrightarrow$ strictly exogeneity of independent variables.

Gauss-Markov Theorem

If (5)-(8)and are satisfied then $\hat{\beta}^{POOLED}$ is BLUE (the best linear unbiased estimator).

- The general assumption in pooled regression on the error terms are very strong or even unrealistic.
- **The lack of correlation between errors corresponding to the same individuals**.
- Let us relax the above assumption:

$$\text{cov}(u_{i,t}, u_{i,s}) \neq 0 \tag{9}$$

- Then we have problem of both autocorrelation and heteroskedasticity.
- The OLS estimator is still consistent but the standard errors are incorrect.
- We might use the **clustered/robust standard errors**. Here, the time series for each individual are clusters.