Panel Data Models

Ani Katchova

Panel Data Models Overview

- Panel data characteristics, panel data types
- Variation types (overall, within, and between variation)
- Panel data models (pooled model, fixed effects model, and random effects model)
- Estimator properties (consistency and efficiency)
- Estimators (pooled OLS, between, fixed effects, first differences, random effects)
- Tests for choosing between models (Breusch-Pagan LM test, Hausman test)

Panel Data Models

Panel data model examples

- Labor economics: effect of education on income, with data across time and individuals.
- Economics: effects of income on savings, with data across years and countries.

Panel data characteristics

- Panel data provide information on individual behavior, both across individuals and over time – they have both cross-sectional and time-series dimensions.
- Panel data include *N* individuals observed at *T* regular time periods.
- Panel data can be balanced when all individuals are observed in all time periods ($T_i = T$ for all *i*) or unbalanced when individuals are not observed in all time periods ($T_i \neq T$).
- We assume correlation (clustering) over time for a given individual, with independence over individuals.
    - Example: the income for the same individual is correlated over time but it is independent across individuals.

Panel data types

- Short panel: many individuals and few time periods (we use this case in class)
- Long panel: many time periods and few individuals
- Both: many time periods and many individuals

Regressors

- Varying regressors x_{it}.
  - annual income for a person, annual consumption of a product
- Time-invariant regressors $x_{it} = x_i$ for all t.
  - gender, race, education
- Individual-invariant regressors $x_{it} = x_t$ for all i.
  - time trend, economy trends such as unemployment rate

Variation for the dependent variable and regressors

- Overall variation: variation over time and individuals.
- Between variation: variation between individuals.
- Within variation: variation within individuals (over time).

Id	Time	Variable	Individual mean	Overall mean	Overall deviation	Between deviation	Within deviation	Within deviation (modified)
i	t	x_{it}	$\bar{x}_i$	$\bar{x}$	$x_{it} - \bar{x}$	$\bar{x}_i - \bar{x}$	$x_{it} - \bar{x}_i$	$x_{it} - \bar{x}_i + \bar{x}$
1	1	9	10	20	-11	-10	-1	19
1	2	10	10	20	-10	-10	0	20
1	3	11	10	20	-9	-10	1	21
2	1	20	20	20	0	0	0	20
2	2	20	20	20	0	0	0	20
2	3	20	20	20	0	0	0	20
3	1	25	30	20	5	10	-5	15
3	2	30	30	20	10	10	0	20
3	3	35	30	20	15	10	5	25

Individual mean $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$

Overall mean $\bar{x} = \frac{1}{NT}\sum_i \sum_t x_{it}$

Overall variance $s_O^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x})^2$

Between variance $s_B^2 = \frac{1}{N-1}\sum_i (\bar{x}_i - \bar{x})^2$

Within variance $s_W^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x}_i)^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x}_i + \bar{x})^2$

The overall variation can be decomposed into between variation and within variation.

$$s_O^2 \approx s_B^2 + s_W^2$$

- Time-invariant regressors (race, gender, education) have zero within variation.
- Individual-invariant regressors (time, economy trends) have zero between variation.
- We need to check the data to see if the between or within variation is larger for each variable.

Panel data models

- Panel data models describe the individual behavior both across time and across individuals. There are three types of models: the pooled model, the fixed effects model, and the random effects model.

Pooled model

- The pooled model specifies constant coefficients, the usual assumptions for cross-sectional analysis.

$$y_{it} = \alpha + \mathbf{x}_{it}^{'}\beta + u_{it}$$

- This is the most restrictive panel data model and is not used much in the literature.

Individual-specific effects model

- We assume that there is unobserved heterogeneity across individuals captured by α_i.
  - Example: unobserved ability of an individual that affects wages.
- The main question is whether the individual-specific effects α_i are correlated with the regressors. If they are correlated, we have the fixed effects model. If they are not correlated, we have the random effects model.

Fixed effects model (FE)

- The FE model allows the individual-specific effects α_i to be correlated with the regressors x.
- We include α_i as intercepts.
- Each individual has a different intercept term and the same slope parameters.

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + u_{it}$$

- We can recover the individual specific effects after estimation as:

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\beta}$$

  In other words, the individual-specific effects are the leftover variation in the dependent variable that cannot be explained by the regressors.
- Time dummies can be included in the regressors x.

Random effects model (RE)

- The RE model assumes that the individual-specific effects α_i are distributed independently of the regressors.
- We include α_i in the error term.
  Each individual has the same slope parameters and a composite error term $\varepsilon_{it} = \alpha_i + e_{it}$.

$$y_{it} = \mathbf{x}_{it}^{'}\beta + (\alpha_i + e_{it})$$

  Here $var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_e^2$ and $cov(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2$
  so $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_e^2)$
- Rho is the interclass correlation of the error. Rho is the fraction of the variance in the error due to the individual-specific effects. It approaches 1 if the individual effects dominate the idiosyncratic error.

Panel data estimators

- The panel data models can be estimated with several estimators.
- The estimators differ based on whether they consider the between or within variation in the data.
- Their properties (consistency) differ based on which model is appropriate.

Estimator properties

- We prefer estimators that are consistent and efficient. We check for consistency first and then for efficiency.

Consistency

- The distribution of $\hat{\beta}_n$ collapses on β as n becomes large:

$$plim\ \hat{\beta}_n = \beta$$

- Consistency is established based on the law of large numbers.
- If an estimator is consistent, more observations will tend to provide more precise and accurate estimates.

Efficiency

- Efficiency (minimum variance) is usually established relative to specific classes of estimators.
  - Example: OLS is efficient (minimum variance) among the class of linear, unbiased estimators (Gauss-Markov Theorem).
  - Maximum likelihood (given correct distributional assumptions) is asymptotically efficient among consistent estimators.

Pooled OLS estimator

- The pooled OLS estimator uses both the between and within variation to estimate the parameters.
- The pooled OLS estimator is obtained by stacking the data over *i* and *t* into one long regression with *NT* observations and estimating it by OLS:

$$y_{it} = \alpha + \mathbf{x}_{it}^{'}\beta + (\alpha_i - \alpha + e_{it})$$

- If the true model is the pooled model and the regressors are uncorrelated with the error terms, the pooled OLS regressor is consistent.
- If the true model is fixed effects then the pooled OLS regressor is inconsistent.
- We need to have panel-corrected standard errors.

Between estimator

- The between estimator only uses the between variation (across individuals).
- It uses the time averages of all variables.
  - If an individual has a work experience of 9, 10, and 11 years measured over 3 periods then the average experience is 10.
- This is an OLS estimation of the time-averaged dependent variable on the time-averaged regressors for each individual.

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\beta + (\alpha_i - \alpha + \bar{e}_i)$$

- The number of observations is *N*. The time variation is not considered and the data are collapsed with one observation per individual.
- This estimator is seldom used because the pooled and RE estimators are more efficient.

Within estimator or fixed effects estimator

- The within estimator uses the within variation (over time).
- It uses time-demeaned variables (the individual-specific deviations of variables from their time-averaged values).

- If an individual has a work experience of 9, 10, an 11 years measured over 3 periods, the average experience is 10. So the time-demeaned values are -1, 0, and 1.

- This is an OLS estimation of the time-demeaned dependent variable on the time-demeaned regressors.

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + (e_{it} - \bar{e}_i)$$

Some software packages estimate:

$$y_{it} - \bar{y}_i + \bar{y} = \alpha + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i + \bar{\mathbf{x}})'\beta + (e_{it} - \bar{e}_i + \bar{e})$$

- The number of observations is *NT*.
- The individual-specific effects α_i cancel out.
- Here, α is the average of the individual effects.
- A limitation of the within estimator is that time-invariant variables are dropped from the model and their coefficients are not identified.
  - A female/male will have values of 1/0 for the female dummy variable, so the values minus the mean values (calculated over time) for each individual will be zero.
  - If we are interested in the effects of time-invariant variables, we need to consider different models (OLS or between estimators).

First-differences estimator

- The first-difference estimator uses the one-period changes for each individual.
- It uses first-differenced variables (the individual-specific one-period changes for each individual).
    - If an individual has a work experience of 9, 10, and 11 years measured over 3 periods then the first difference experience are missing (.), 1, and 1.
- This is an OLS estimation of the one-period changes of the dependent variable on the one-period changes in the regressors.

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (e_{it} - e_{i,t-1})$$

- The number of observations is $N(T\text{-}1)$. We lose the first observation for each individual because of differencing.
- The individual-specific effects α_i cancel out.
- A limitation of the first-differences model is that time-invariant variables are dropped from the model and their coefficients are not identified.

Random effects estimator

- This is an OLS estimation of the transformed model:

$$y_{it} - \hat{\lambda}\overline{y_i} = \left(1 - \hat{\lambda}\right)\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)'\beta + \nu_{it}$$

$$\nu_{it} = \left(1 - \hat{\lambda}\right)\alpha_i + (e_{it} - \hat{\lambda}\bar{e}_i)$$

$$\lambda = 1 - \sigma_e \Big/ \sqrt{\sigma_e^2 + \sigma_\alpha^2}$$

- The number of observations is *NT*.
- The individual-specific effects α_i are in the error term.
- Note that $\hat{\lambda} = 0$ corresponds to pooled OLS and $\hat{\lambda} = 1$ corresponds to the within (fixed effects) estimator.
- The random effects estimates are a weighted average of the between and within estimates.
- The random effects estimator is fully efficient under the random effects model.

Models and estimators

Estimator/true model	Pooled model	Random effects model	Fixed effects model
Pooled OLS estimator	Consistent	Consistent	Inconsistent
Between estimator	Consistent	Consistent	Inconsistent
Within or fixed effects estimator	Consistent	Consistent	Consistent
First differences estimator	Consistent	Consistent	Consistent
Random effects estimator	Consistent	Consistent	Inconsistent

- The fixed effects estimator will always give consistent estimates, but they may not be the most efficient.
- The random effects estimator is inconsistent if the appropriate model is the fixed effects model.
- The random effects estimator is consistent and most efficient if the appropriate model is random effects model.

Choosing between fixed and random effects

Breusch-Pagan Lagrange Multiplier test

- This is a test for the random effects model based on the OLS residual.
- Test whether σ_u^2 or equivalently $cor(u_{it}, u_{is})$ is significantly different from zero.
- If the LM test is significant, use the random effects model instead of the OLS model.
- We still need to test for fixed versus random effects.

Hausman test

- The random effects estimator is more efficient so we need to use it if the Hausman test supports it. If it does not support it, use the fixed effects model.
- Hausman test tests whether there is a significant difference between the fixed and random effects estimators.
- The Hausman test statistic can be calculated only for the time-varying regressors.
- The Hausman test statistics is:

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

- It is chi-square distributed with degrees of freedom equal to the number of parameters for the time-varying regressors.
- If the Hausman test is insignificant use the random effects.
- If the Hausman test is significant use the fixed effects.