Panel Data Models

After finish this session, you should understand:

- Characteristics of Panel Data.
- Panel data model with problems (concerning Var-Cov of error term) and how to solve.
- Concept of Fixed (specific) effects – Endogeneity Biased.
- First Difference vs Mean Differences (FE) Estimation methods.
- Fixed effects vs Random effects models
- Hausman test

Panel Data Models

1. Characteristic of Data
2. Panel Least Squares
3. Fixed Effects Models
4. Random Effects Models
5. Advantage of Panel Data Models

Characteristic of Data

Panel Data = Cross-section + Time Series

Panel data mostly have more cross-section than time-series.

Advantage: Number of observation $N = nT$

Problems that might occur:

1. Heteroskedasticity
2. Autocorrelation
3. Cross-sectional Correlation

Characteristic of Data

Example:

	id	time	y	mtb	size	prof	tang	risk
1	1	1	.7579437	1.096261	18.84586	.0880269	.9177378	.0024175
2	1	2	.7428093	1.100189	18.87402	.0756492	.9061694	.0001755
3	1	3	.722805	1.051447	18.81043	.1115731	.9261793	.0018045
4	1	4	.5370298	1.297697	18.97547	.1124983	.8920739	.0022579
5	1	5	.5549089	1.139629	19.10716	.0981166	.8918622	.0003683
6	1	6	.5262893	1.054489	19.01132	.1045869	.9001783	.0015503
7	1	7	.4031329	1.181099	19.07588	.1209047	.8695951	.0032206
8	2	1	.1214492	2.170097	19.75014	.1962469	.089559	.0247711
9	2	2	.2805595	1.694395	20.20486	.1025133	.1056814	.001609
10	2	3	.3319133	1.422119	20.50326	.1768626	.1110374	.0116141
11	2	4	.145835	2.524653	20.5944	.2298589	.1113212	.0271847
12	2	5	.0856535	3.04502	20.64201	.2727948	.7193236	.0375855
13	2	6	.0739597	3.087652	20.55656	.2425512	.7179744	.0314487
14	2	7	.1260392	2.220203	20.49912	.2015931	.7302911	.0188894

True Panel vs Pooled Cross-section

Researchers mostly use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension.

More precisely, it is only data following the same cross-section units over time.

Otherwise, the data should be considered as a pooled cross-section – e.g. SES data.

Pooled Cross-sections

Researchers pool cross sections just to get bigger sample sizes.

The main purposes are to investigate the effect of time and to test whether relationships have changed over time.

Time Series vs Cross Section Data

Time Series Data

- Different time analysis – compare across different times (e.g. technology change, impact from crisis)
- Properties of time-series data – time-series data problems.

Cross-Sectional Data

- Differences among cross sectional unit – compare across different units (or firms) (e.g. scale of economy)
- Heteroscedasticity – different size
- Inconsistency of the data

Pooled Data vs Panel Data

Pooled Data

- Increase number of observation
- Treated each obs. as one cross-sectional unit.
- Similar to cross-sectional analysis

Panel Data

- Larger number of observation
- Treated data as same firm or same period
- Combine both time-series and cross-sectional analysis

Panel Data Models – no problem

The general model:

$$Y_{it} = X_{it}\beta + \varepsilon_{it}$$

Variance-Covariance Matrix (no problem):

$$V = E\left(\varepsilon\varepsilon'\right) = \begin{bmatrix} \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & \cdots & \sigma_{1n}\Omega_{1n} \\ \sigma_{21}\Omega_{21} & \sigma_{22}\Omega_{22} & \cdots & \sigma_{2n}\Omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}\Omega_{n1} & \sigma_{n2}\Omega_{n2} & \cdots & \sigma_{nn}\Omega_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 I & 0 & \cdots & 0 \\ 0 & \sigma^2 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 I \end{bmatrix}$$

Estimation Method

Panel Least Squared (POLS):

$$\underset{k\times 1}{\hat{\beta}} = (\underset{k\times nT}{X'}\ \underset{nT\times k}{X})^{-1}\ \underset{k\times nT}{X'}\ \underset{nT\times 1}{Y}$$

However, there will be just only one estimated equation model for all n cross-sectional units.

Model with Heteroskedasticity, Autocorrelation and Cross-sectional Correlation

$$Y_{it} = X_{it}\beta + \varepsilon_{it}$$

Variance-Covariance Matrix:

$$V = E\left(\varepsilon\varepsilon'\right) = \begin{bmatrix} \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & \cdots & \sigma_{1n}\Omega_{1n} \\ \sigma_{21}\Omega_{21} & \sigma_{22}\Omega_{22} & \cdots & \sigma_{2n}\Omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}\Omega_{n1} & \sigma_{n2}\Omega_{n2} & \cdots & \sigma_{nn}\Omega_{nn} \end{bmatrix}$$

$$\Omega_i = \begin{bmatrix} 1 & \rho_i & \rho_i^2 & \cdots & \rho_i^{T-1} \\ \rho_i & 1 & \rho_i & \cdots & \rho_i^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1 \end{bmatrix}$$

Estimation Method

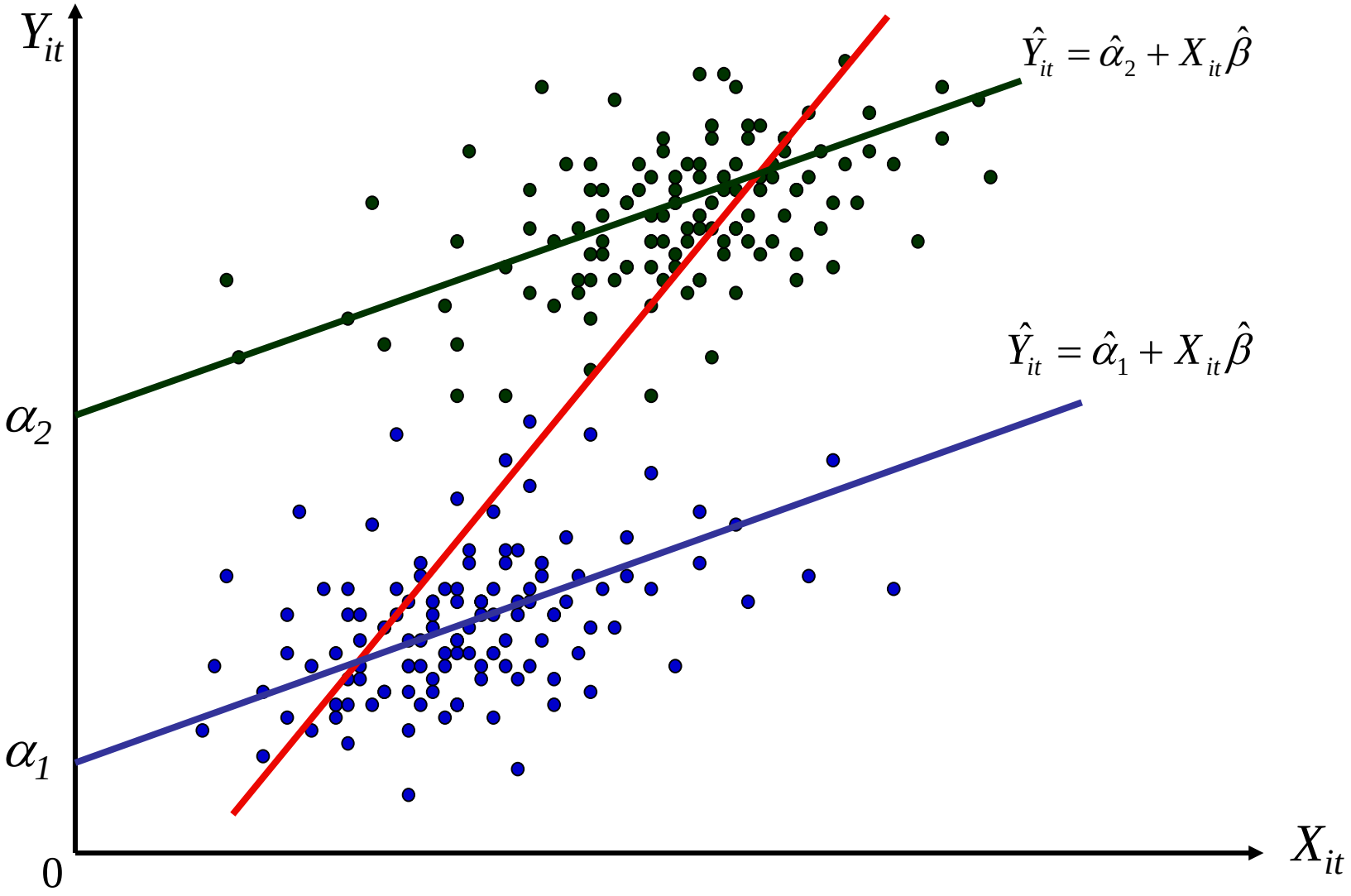
Panel Generalize Least Squared (PGLS):

$$\underset{k\times 1}{\hat{\beta}} = (\underset{k\times nT}{X'}\ \underset{nT\times nT}{\hat{V}}^{-1}\ \underset{nT\times k}{X})^{-1}\ \underset{k\times nT}{X'}\ \underset{nT\times nT}{\hat{V}}^{-1}\ \underset{nT\times 1}{Y}$$

$k\times k$ $\qquad$ $k\times 1$

However, there will be just only one estimated equation model for all n cross-sectional units.

Bias from Ignoring Fixed Effects



Unobserved Fixed Effects

Suppose the population model is

$$Y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}$$

Here we have added a time-constant component to the error: $u_{it} = \alpha_i + \varepsilon_{it}$

In scalar form:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}$$

If α_i is correlated with the x's, OLS will be biased, since α_i is part of the error term.

With panel data, we can difference-out the unobserved fixed effect.

First Difference

First difference method can be used to difference-out the unobserved fixed effects

$$(1) \qquad y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}$$

$$(2) \qquad y_{it-1} = \beta_1 + \beta_2 x_{2it-1} + \cdots + \beta_k x_{kit-1} + \alpha_i + \varepsilon_{it-1}$$

$$(1)-(2) \quad y_{it} - y_{it-1} = (\beta_1 - \beta_1) + \beta_2 (x_{2it} - x_{2it-1}) + \cdots + \beta_k (x_{kit} - x_{kit-1}) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \varepsilon_{it-1})$$

$$\Delta y_{it} = \beta_2 \Delta x_{2it} + \cdots + \beta_k \Delta x_{kit} + \Delta \varepsilon_{it}$$

This model has no correlation between the x's and the error term, so no bias.

Need to be careful about organization of the data to be sure compute correct change.

Differencing with Multiple Periods

We can extend this method to more periods.

Simply difference adjacent periods.

If 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual.

Simply estimate by OLS, assuming the $\Delta\varepsilon_{it}$ are uncorrelated over time.

However, the problem with this technique is that it cannot be used in case of unbalance panel data cases.

Fixed Effects Estimation

When there is an observed fixed effects, an alternative to the first differences is fixed effects estimation

Consider the deviation from average cross-sectional group mean over time model:

$$y_{it} - \bar{y}_i = \beta_2(x_{2it} - \bar{x}_{2i}) + \cdots + \beta_k(x_{kit} - \bar{x}_{ki}) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The average of α_i will be α_i, so if you subtract the mean, α_i will be differenced out just as when doing first differences.

Fixed Effects Estimation

Example:

	id	time	y	mtb	size	prof	tang	risk
1	1	1	.7579437	1.096261	18.84586	.0880269	.9177378	.0024175
2	1	2	.7428093	1.100189	18.87402	.0756492	.9061694	.0001755
3	1	3	.722805	1.051447	18.81043	.1115731	.9261793	.0018045
4	1	4	.5370298	1.[illegible]7697	18.97547	.1124983	.8920739	.0022579
5	1	5	.5549089	1.[illegible]9629	19.10716	.0981166	.8918622	.0003683
6	1	6	.5262893	1.054489	19.01132	.1045869	.9001783	.0015503
7	1	7	.4031329	1.181099	19.07588	.1209047	.8695951	.0032206
8	2	1	.1214492	2.170097	19.75014	.1962469	.089559	.0247711
9	2	2	.2805595	1.694395	20.20486	.1025133	.1056814	.001609
10	2	3	.3319133	1.[illegible]22119	20.50326	.1768626	.1110374	.0116141
11	2	4	.145835	[illegible]4653	20.5944	.2298589	.1113212	.0271847
12	2	5	.0856535	[illegible]4502	20.64201	.2727948	.7193236	.0375855
13	2	6	.0739597	3.087652	20.55656	.2425512	.7179744	.0314487
14	2	7	.1260392	2.220203	20.49912	.2015931	.7302911	.0188894

$\bar{y}_1$

$\bar{y}_2$

Fixed Effects Estimation

If we were to do this estimation by hand, we would need to be careful because we would think that $df = NT - k$, but is $N(T-1) - k$ because we used up dfs calculating means.

Most statistical software can estimate fixed effects.

This method is also identical to including a separate intercept for every individual.

First Differences vs Fixed Effects

First Differences and Fixed Effects will be exactly the same when $T = 2$.

For $T > 2$, the two methods are different.

Probably we might see fixed effects (within) estimation more often than first differences – probably more because it is easier than that it is better.

Fixed effects easily implemented for unbalanced panels, not just balanced panels.

Test for Fixed Effects

$$H_0: \ \alpha_1 = \alpha_2 = \ldots = \alpha_n = \alpha$$

Unrestricted Model: $y_{it} = \alpha_i + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}$

Restricted Model: $y_{it} = \alpha + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}$

Unrestricted-Restricted F-test or Chi-squares test can be applied.

F-statistic = $\dfrac{(RSS_R - RSS_{UR})/n}{RSS_{UR}/(N-k-1)}$

Chi-squares Test = $2(L_{UR} - L_R)$

Random Effects

Start with the same basic model with a composite error

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}$$

In FE model, we assumed that α_i was correlated with the x's, but what if they are not correlated?

OLS would be consistent in that case, but composite error will be serially correlated

Random Effects

To estimate the model, researchers need to transform the model and do GLS to solve the problem and make correct inferences.

End up with a sort of weighted average of OLS and Fixed Effects – use quasi-demeaned data.

$$y_{it} - \hat{\lambda}\bar{y}_i = \beta_1\left(1-\hat{\lambda}\right) + \beta_2\left(x_{2it} - \hat{\lambda}\bar{x}_{2i}\right) + \ldots + \beta_k\left(x_{kit} - \hat{\lambda}\bar{x}_{ki}\right) + \left(u_{it} - \hat{\lambda}\bar{u}_i\right)$$

where: $u_{it} = (1-\hat{\lambda})\alpha_i + (\varepsilon_{it} - \hat{\lambda}\bar{\varepsilon}_i)$ is iid.

$$\lambda = 1 - \left[\frac{\sigma_\varepsilon}{\sqrt{\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)}}\right]$$

Random Effects

If $\lambda = 1$, then this is just the fixed effects estimator.

If $\lambda = 0$, then this is just the OLS estimator.

Thus, the bigger the variance of the unobserved effect, the closer the model is to fixed effect.

The smaller the variance of the unobserved effect, the closer it is to OLS.

Fixed or Random Effects

In most cases, fixed effects model seem to be more appropriated, since it is more likely that unobserved variables are correlated with the independent variables x's.

To determine whether to apply fixed effects or random effects, researchers sometimes apply Hausman test.

If Hausman test is rejected, FE should be applied. If not, it should be RE.

Fixed or Random Effects

Hausman Test: $H_0 : \beta_{RE} = \beta_{FE}$

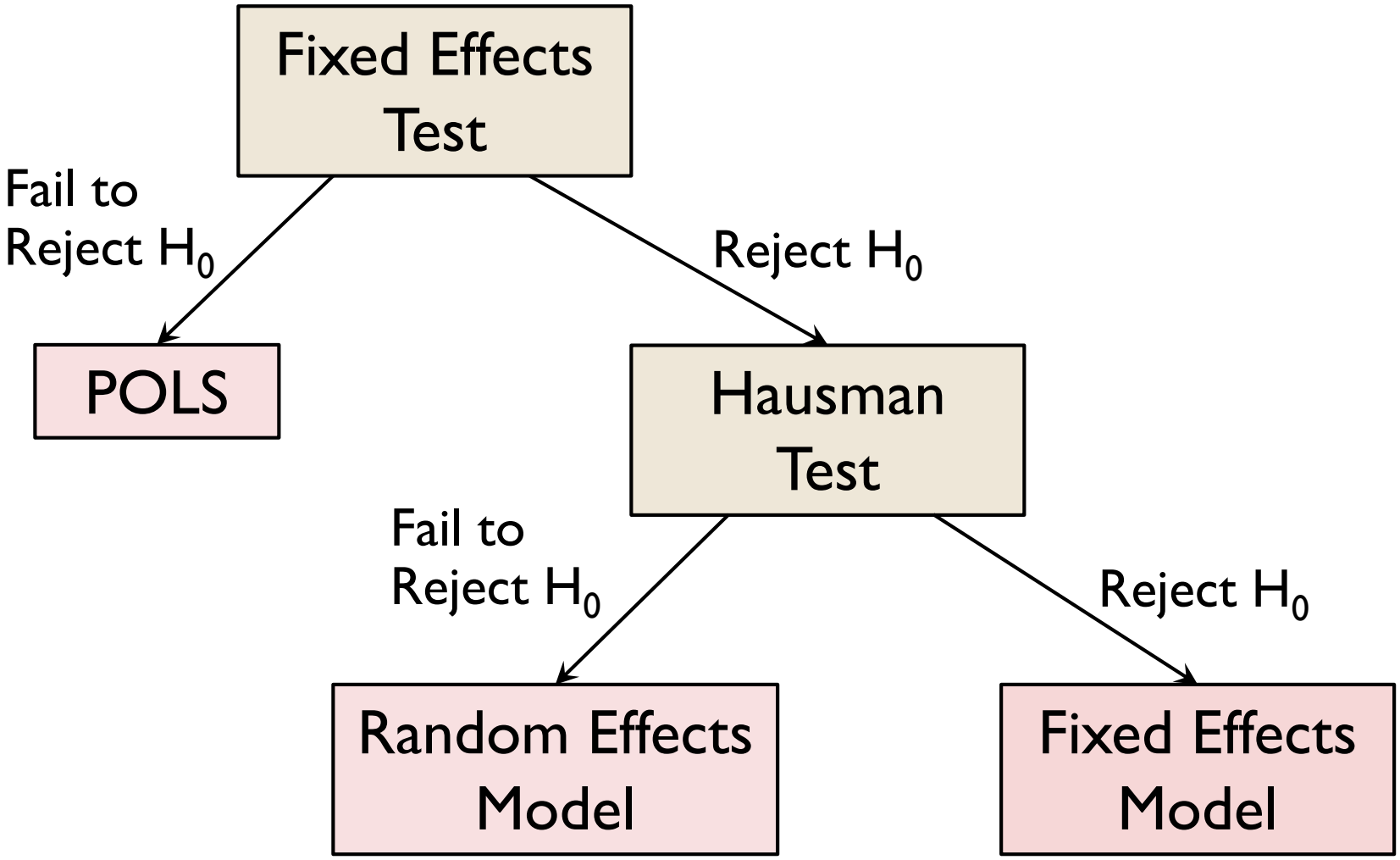
$$Hausman\ Statistic = \left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right)' \left(\hat{V}_{RE} - \hat{V}_{FE}\right)^{-1} \left(\hat{\beta}_{RE} - \hat{\beta}_{FE}\right) \sim \chi^2_{(k-1)}$$

where:

- $\hat{\beta}_{RE}$ is coefficient vector of RE.
- $\hat{\beta}_{FE}$ is coefficient vector of FE.
- $\hat{V}_{RE}$ is covariance matrix of RE.
- $\hat{V}_{FE}$ is covariance matrix of RE.

If Hausman test is rejected, FE should be applied. If not, it should be RE.

Panel Data Analysis



Advantages of Appling Panel Data

1. Addition of observations.
2. Additional aspects – Both cross-sectional aspect (differences according to individual or firm) and time-series aspect (differences according to time).
3. Solving ***endogeneity bias*** from unobservable fixed effects – especially from cross-sectional aspect (individual or firm specific characteristics).

Other Issues

How to test and correct for heteroscedasticity and serial correlation in the errors?

How to estimate standard errors robust to both problems?

Additional issue to think of whether both cross-section and time series data should be applied as panel data:

Whether the data should be pooled –Panel Poolability (Chow) Test?

Panel Poolability Test

$$H_0: \ \gamma_1 = \gamma_2 = \cdots = \gamma_n = 0$$

Unrestricted: $y_{it} = X_{it}\beta + \sum_{i=1}^{n} D_{it} X_{it} \gamma_i + \varepsilon_{it}$

Restricted: $y_{it} = X_{it}\beta + \varepsilon_{it}$

Chow-similar dummy F-test can be applied.

Rejection of the null hypothesis means data cannot be pooled together as panel, thus, models should be separated.

Fail to reject implies panel data model can be applied.