CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

EFMD EPAS ACCREDITED

MSF Chula

Financial Econometrics

Lecture X

Introduction to Panel Data Analysis

Narapong Srivisal, Ph.D.

Outline

Panel Data

Models

- ① **Pooled Cross-section**
  - Structural Break
  - Chow Test
- ② **Fixed Effects**
  - First-Differenced Estimator
  - Fixed Effects Estimator
  - Between Effects Estimator
- ③ **Random Effects**
  - Random Effects Estimator

Hausman Test

relationship assumed in population

how can we make guess about parameter

Model ≠ estimator !!

Panel Data

- **Panel data analysis** deals with a dataset that contains many cross-sectional individual subjects, but each of the subjects is repeatedly sampled or observed in more than one period.
- Typically, we use subscript i to index each cross-sectional observation and subscript t to index each time period. So, each observation in a panel dataset is indexed by subscript it

1 How they store panel data

- A panel dataset, especially in STATA, can be organized in two ways in a data file: “long” and “wide”.
- Suppose we have the data Y, X_1, X_2. In **the long form**, we just need two variables to indicate individual id (i) and time (t).
- In **the wide form**, there is still an indicator for individuals, but there is no time variable t. Instead, we will need to create variables for each time period of Y, X_1, X_2

Panel Data
≠
Repeated cross-section data
e.g. survey 1st time] may not be same person
" —" 2nd " — " cannot link individual
across time unlike panel data

time-~~series~~ of many individual
doesn't need consecutive periods unlike time series → just need many periods

MSF Chula

Diff ways of organizing data

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE
AACSB ACCREDITED
EFMD EQUIS ACCREDITED
EFMD EPAS ACCREDITED

change data set to be long format

Long Format

who, which stock is this?

index time

i	t	Y	X_1	X_2
A	2010	23,000	16	2
A	2011	23,500	16	3
A	2012	24,000	16	4
B	2010	12,000	9	7
B	2011	12,000	9	8
B	2012	12,700	9	9
C	2010	20,000	16	0
C	2011	21,000	16	1
C	2012	21,000	16	2

counted as 1 obs

9 obs

$\underset{i}{w} \times \underset{time}{T}$

even unbalance not matter much

If we have balance panel (every r.v. have data at same time

Wide Format

r.v. indicates variables, no indicate time

i	$Y.2010$	$Y.2011$	$Y.2012$	$X_1.2010$	$X_1.2011$	$X_1.2012$	$X_2.2010$	$X_2.2011$	$X_2.2012$
A	23,000	23,500	24,000	16	16	16	2	3	4
B	12,000	12,000	12,700	9	9	9	7	8	9
C	20,000	21,000	21,000	16	16	16	0	1	2

t

2000

2001

⋮

2012

There's problem if : unbalanced panel

1900s $\longrightarrow$ 2020

relationship in this period is not same as

- Note that if the dataset that we have for all the individuals consists of the same number and time of periods such as the above example, we call it a **balanced-panel dataset**.
- Not all the panel datasets need to be balanced, and we can still do some analysis without always having to cut off the periods that data of some individuals are missing.
- There are mainly three linear regression **models** people consider when working with panel data:
  - **Pooled Cross-section**
  - **Fixed Effects**
  - **Random Effects**
- Each of these models has different properties and, thus, needs different estimators.

→ pool everything as same episode &also assume iid

Pooled Cross-section

- The simplest way of doing panel data analysis is to treat each data point (individual i at time t) as one observation and apply usual cross-sectional methods like OLS, IV, 2SLS to estimate: * can use OLS estimator as before

F/GLS
use OLS if no endogeneity
IV/2SLS
robust if hetero

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + U_{it}$$

- This is called **Pooled Cross-section** model or pooled cross-sectional analysis, as all the observations across periods are pooled together.
- However, in order to use pooled cross-sectional method, we need a strong assumption that the observations are independent and that the relationship between the regressand and regressors remains the same across periods and individuals.
- Or if there are some changes overtime or across individuals, we need to use explanatory variables to control for the differences.

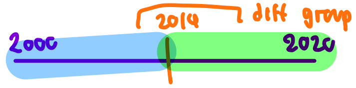
Structural Break

- Suppose we suspect that there is a structural break, i.e. a change in the relationship between Y and X variables, at time t.
- An easy way to adjust the model to control for the change is to make use of a dummy variable constructed based on t.
- E.g. A pooled cross-sectional model of wage determinants:

$$Wage_{it} = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + U_{it}$$

- Suppose we know that the government increased minimum wage in year 2012 and suspect that wage may be higher since then.
- We may define D_{it} as a dummy variable taking value 1 if year is 2012 or later and run the following regression:

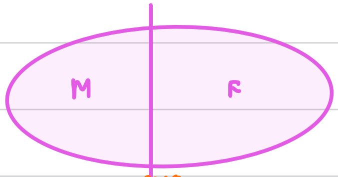
$$Wage_{it} = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \gamma D_{it} + U_{it}$$



diff group of population
as relationship change
use dummy to capture time change

Relationship change at some certain point

$$wage = \beta_0 + \beta_1 edu + \dots + u$$



$$Wage = \beta_0 + \beta_1 \cancel{M}^{\,D^{2014}} + \beta_2 edu + \beta_3 \cancel{M}^{\,D^{2014}} \cdot edu$$

- Or if we believe that the minimum wage may also affect the impact of X_1 on wage, we may include the interaction term as well:

$$Wage_{it} = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \gamma_0 D_{it} + \gamma_1 D_{it} X_{1it}$$

interaction term

- In general, for a structural break, we can allow for all the betas to be different. So, the regression model is

test γ_t with dummy

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \gamma_0 D_{it} + \gamma_1 D_{it} X_{1it} + \cdots + \gamma_k D_{it} X_{kit} + U_{it}$$

- Then, to test whether there is the structural break, we can apply the F-test to

$$H_0: \gamma_0 = \gamma_1 = \cdots = \gamma_k = 0 \text{ vs}$$

$$H_1: \gamma_1 \neq 0, \text{or } \gamma_1 \neq 0, \text{or } \ldots, \text{or } \gamma_k \neq 0$$

a kind of F-test

- This F-test is often referred to as the **Chow Test**, because it was originally proposed by Chow to test that two groups have different relationship.
- For example, Chow Test can be applied to test whether wage determinants for male and female are the same or not. In this case, the dummy will capture gender rather than pre- and post-periods.

Fixed Effects Model

- When working with panel data, there is often a concern of some unobserved hidden characteristics specific to individuals that can affect the regressand, which is called **unobserved heterogeneity**.
- If the unobserved heterogeneity is correlated with the regressors, we call it **Fixed Effects**.
- One way to capture the fixed effects is to use dummies accounting for different individuals.
- However, it is not quite a good idea to use many dummies to capture unobserved characteristics of every individual, especially when we have a lot of individuals but not so many periods.
- Instead, we can exploit the panel structure to eliminate the fixed effects and estimate the model.

not to error term

$$y_{it} = \beta X_{it} + (\alpha_i + U_{it})$$

α_i → doesn't change over time, just change across individual

it is time invariant

more than 1 unobserved HETEROGENEITY — diff across observation

& specific to each individual

* unobserved or not quantifiable

∵ no r.v. to capture or proxy for it

↓ left in error term

▷ in pool cross section → $\alpha_i = 0$

both have α_i — how it correlate with x?

FIXED EFFECT

$cov(x, \alpha) \neq 0$

has endogeneity problem, cannot use OLS

RANDOM EFFECT

$cov(x, \alpha) = 0$

OLS estimator is consistent but has serial correlation problem

MODEL: relationship in population

e.g. - culture of countries not change across time

- working environment in firm ⇓ affect on firm

Fixed Effect Model

* correlated with x
endogeneity

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + (\alpha_i + u_{it})$$

also say α_i is correlated with x or not !!!

Solution : 1) use IV or 2SLS estimator to deal with endogeneity

need to find more data to run with instrument ← hard!

2) take it out error term is by having dummy for each individual

if it's time series not cross section

who, stock is this?

Long Format

i	t	Y	X_1	X_2	D^A	D^B
A	2010	23,000	16	2	1	0
A	2011	23,500	16	3	1	0
A	2012	24,000	16	4	1	0
B	2010	12,000	9	7	0	1
B	2011	12,000	9	8	0	1
B	2012	12,700	9	9	0	1
C	2010	20,000	16	0	0	0
C	2011	21,000	16	1	0	0
C	2012	21,000	16	2	0	0

$$y_{it} = \beta_0 + \alpha^A D_1^A + \alpha^2 D_1^B + \beta_1 X_{1it} + \beta_2 X_{2it} + u_{it}$$

3) kill α from equation

First difference estimator

remain same regardless period

$$y_{it} = \cancel{\beta_0} + \beta_1 X_{1it} + \beta_2 X_{2it} + \cancel{\alpha_i} + u_{it}$$

$$- \quad y_{it-1} = \cancel{\beta_0} + \beta_1 X_{1it-1} + \beta_2 X_{2it-1} + \cancel{\alpha_i} + u_{it-1}$$

$$\Delta y_{it} = \beta_1 \Delta X_{1it} + \beta_2 \Delta X_{2it} + \Delta u_{it}$$

no endogeneity anymore :)

$y_{it} - y_{it-1}$

reg dy dx

try to kill α_i ⇒ α_i is same regardless time → average still α_i

Fixed Effect Estimator

way to get β estimate → reg

$$y_{it} = \cancel{\beta_0} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cancel{\alpha_i} + u_{it}$$ ⇒ fixed effect model

$$\bar{y}_i = \cancel{\beta_0} + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \cancel{\alpha_i} + \bar{u}_i$$

$$y_{it} - \bar{y}_i = \beta_1 (x_{1it} - \bar{x}_{1i}) + \beta_2 (x_{2it} - \bar{x}_{2i}) + \text{error}$$

who, stock is this?

Long Format

$\bar{y}_i$	i	t	Y	X_1	X_2	y_{t-1}	Δy_t
23.5 k	A	2010	23,000	16	2	-	-
23.5k	A	2011	23,500	16	3	23 k	500
23.5k	A	2012	24,000	16	4	23.5 k	500
12,233.33	B	2010	12,000	9	7	-	-
12,233.33	B	2011	12,000	9	8	12 k	0
12,233.33	B	2012	12,700	9	9	12 k	700
	C	2010	20,000	16	0	-	-
	C	2011	21,000	16	1	20k	1000
	C	2012	21,000	16	2	21k	0

ไม่มีข้อมูล 2009

lose observation → First Diff → $\cancel{N \times T}$ $N \times (T-1)$

lose as there is 1 lag

↳ degree of freedom

$N(T-1) - \#\text{beta}$

Fixed Effect estimator → NT obs

how many variable you can fill in

↳ degree of freedom

use in t-test $\boxed{N(T-1) - \#\text{beta}}$

bcs we have to calculate mean

eg. 5 stocks

Return _ _ _ _ _

five stocks to fill number

5 stocks & mean = 2

_ _ _ _ _

can fill this

lose freedom to fill this

must add specific number that make mean = 2

First-Differenced Estimator

- Mathematically, we can write the **Fixed Effects model** as

$$Y_{it} = \beta_0 + \beta X_{it} + \delta d_t + \boldsymbol{\alpha_i} + U_{it}$$

where $\boldsymbol{cov(\alpha_i, X_{it}) \neq 0}$.

- Since the fixed effects of each individual is constant through time, and, for each individual, we have data of more than one period. Then, a simple way to eliminate the fixed effects is to use the first difference.

- Consider first the case that we have n individuals and $T = 2$ periods, and we can let D_t be the dummy for the second period:

$$Y_{it_1} = \beta_0 + \beta X_{it_1} + \alpha_i + U_{it_1}; \;\; t = t_1$$

$$Y_{it_2} = \beta_0 + \beta X_{it_2} + \delta D_t + \alpha_i + U_{it_2}; t = t_2$$

- Then, first differencing gives us:

$$Y_{it_2} - Y_{it_1} = \delta + \beta\left(X_{it_2} - X_{it_1}\right) + (U_{it_2} - U_{it_1})$$

- Now, if we have the usual assumptions that U_{it} is the idiosyncratic error that is uncorrelated with the regressors, we can use the OLS to estimate the model by regressing ΔY_{it} on ΔX_{it}
- This is called the **First Differenced Estimator**
- Note: t_1, t_2 do not need to be consecutive periods.
- Note: if X_{it_2} contains Y_{it_1}, we will have an endogeneity problem because Y_{it_1} is correlated with U_{it_1} which is a part of the error term in the first difference estimation.

First Difference with T>2

- Now, suppose that we have $T = 3$ periods.

$$Y_{it} = \beta_0 + \beta X_{it} + \delta_2 d2_t + \delta_3 d3_t + \alpha_i + U_{it}$$

where $d2, d3$ are the dummy variables for the second and third periods respectively.

- We can still use the first different method:

$$\Delta Y_{it} = \beta \Delta X_{it} + \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \Delta U_{it}$$

where Δ indicates the difference between period t and the most recent preceding period.

- Note that the variable $\Delta d2_t$ and $\Delta d3_t$ are not constant, as $\Delta d2_{t_2} = 1$; $\Delta d2_{t_3} = -1$; $\Delta d3_{t_2} = 0$; $\Delta d3_{t_3} = 1$.
- Nonetheless, we can still estimate this model using OLS but without an intercept.

- In practice, it is more convenient to instead estimate the following model which has an intercept:

$$\Delta Y_{it} = \beta \Delta X_{it} + \gamma_0 + \gamma_3 \Delta d3_t + \Delta U_{it}$$

- This model with an intercept is related to the previous model as

$$\gamma_0 = \delta_2; \gamma_3 = \delta_3 - 2\delta_2$$

- In general, when we have $T > 2$ periods, the First Differenced Estimator can be derived from using the OLS to estimate:

$$\Delta Y_{it} = \gamma_0 + \gamma_3 \Delta d3_t + \gamma_4 \Delta d4_t + \cdots + \gamma_T \Delta dT_t + \beta \Delta X_{it} + \Delta U_{it}$$

- Note: the parameter that we are mainly interested in estimating is the effect of X on Y, which is β.
- Note: the number of observations for this estimator using a balanced panel dataset is $n(T-1)$, where n is the number of individuals and T is the number of total periods in the dataset.

Fixed Effects Estimator

- Now, consider an alternative way of eliminating the fixed effects.
- Suppose the model of interest is

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

where X is the vector of all regressors, possibly including time dummies.

- Since the fixed effects α_i is time-invariant for each individual, then if we average it over time:

$$\bar{Y}_i = \beta \bar{X}_i + \alpha_i + \bar{U}_i$$

- So, instead of using the first difference, we can **use the time-demeaned method to get rid of the fixed effects**:

$$Y_{it} - \bar{Y}_i = \beta(X_{it} - \bar{X}_i) + (U_{it} - \bar{U}_i)$$

- Then, if we have the usual assumptions that U_{it} is uncorrelated with the regressors, we can use the OLS to estimate the model by regressing the time-demeaned Y on the time-demeaned X without an intercept.
- This method is called the fixed effects transformation or within transformation, because we use time-variation within each cross-sectional unit i to estimate the model.
- The estimator for β derived from this transformation is called **Fixed Effects Estimator** or **Within Estimator**.
- Note that we still have nT observations to estimate by the fixed effects estimator.
- However, when making inference, the degree of freedom is not $nT - k$, but $n(T-1) - k$ because we lose one degree of freedom for each individual i from estimating its sample mean.

FD or FE Estimator?

- Both have the same properties regarding unbiasedness and consistency.
- If the idiosyncratic error U_{it} is serially uncorrelated, then FE is more efficient than FD.
- If T is large relative to n, the data are more like time series which require stationary assumption. FE will be problematic if the assumptions, including stationary and no serial correlation, is violated. FD is like an integrated series, which can turn non-stationarity to be weak stationarity in some cases.
- If there is a measurement error in X, FE is generally better than FD because the bias declines at the rate $1/T$, whereas the bias of FD is not sensitive to T.

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE
AACSB ACCREDITED
EFMD EQUIS ACCREDITED
EFMD EPAS ACCREDITED

(effect)

Between Estimator

- In contrast to Within Estimator, Between Estimator uses only variation across (or between) individuals, rather than across time within each individual, to estimate the relationship between X and Y.
- **Between Estimator** or **Between Effects Estimator** is the cross-sectional of the OLS regression of $\bar{Y}$ on $\bar{X}$

$$\bar{Y}_i = \beta \bar{X}_i + \alpha_i + \bar{U}_i$$

- Between estimator cannot eliminate the fixed effects α_i. Hence, the between estimator is inconsistent under the fixed effects model assumption that $cov(\alpha_i, X_{it}) \neq 0$.
- The between estimator is helpful for the case of measurement error in a regressor if the expected measurement error $E[\varepsilon]$ is zero.

$$\bar{X}_i + \bar{\varepsilon}_i \xrightarrow{p} E[X] + E[\varepsilon] = E[X]$$

approach

$$y_{it} = \beta X_{it} + \alpha_i + u_{it}$$

$\rightarrow \bar{y}_i = \beta \bar{x}_i + (\alpha_i + \bar{u}_i)$

not get consistent → α_i & $\bar{x}_i$ still correlate

Between : reg $\bar{y}_i$ $\bar{x}_i$

⇓ not consistent, still has endogeneity problem when applied to Fixed Effect model

→ if pool cross-section model → can use between → not recommend
($\alpha_i = 0$)

→ Degree of freedom
N - # beta
NT - # beta

Measurement error in X

$$y_{it} = \beta x_{it} + u_{it}$$

↑ not observed

$$\bar{\tilde{x}}_{it} = \bar{x}_{it} + \cancel{\bar{\varepsilon}_{it}} \rightarrow 0$$

↑ observed

possible

Between Estimator

⇓

reg $\bar{y}$ on $\bar{x}$ } roughly same
$\bar{y}$ on $\bar{\tilde{x}}$

⇓

help in case of measurement error

reg y $\tilde{x}$ ⇒ $y = \beta[\tilde{x} - \varepsilon] + u$

$\bar{y} = \beta \bar{\tilde{x}} + u - \beta \cancel{\bar{\varepsilon}} \rightarrow 0$

Random Effects Model

- When discussing the fixed effects, we saw α_i as causing the endogeneity problem and tried to eliminate it.
- On the contrary, the random effects concept sees α_i as "random" in the sense that it has no correlating relationship with the regressors.
- Mathematically, the **Random Effects model** is:

$$\boldsymbol{Y_{it} = \beta_0 + \beta X_{it} + (\alpha_i + U_{it})}$$

no endogeneity problem bcs α_i is not correlated with X

$$\boldsymbol{Cov(X_{it}, \alpha_i) = 0 \text{ for } t = 1, \dots, T; \ i = 1, \dots, n}$$

- Therefore, we can simply use OLS to run the regression of Y on X and get a consistent or unbiased estimator for β with regular assumptions.
- However, OLS is not efficient for this model, because the composite error terms $v_{it} = (\alpha_i + U_{it})$ are correlated across observations.
- A better estimator for this is based on the FGLS

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}$$

$\left.\begin{matrix} i,t \\ i,t-k \end{matrix}\right\}$ $cov(error_{it}, error_{it-k})$

$\alpha_i + u_{it}$ $\quad$ $\alpha_i + u_{i(t-k)}$

0 0 0

* σ^2_α

OLS not give good standard error

- Notice that there is α_i in the error term in ever period t. So, the autocorrelation for between time t and s is

$$corr(\alpha_i + U_{it}, \alpha_i + U_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_U^2}$$

- Thus, we can apply the GLS by transforming the regression before running the OLS. The transformed model is

$$\boxed{\boldsymbol{Y_{it} - \theta \bar{Y}_i = \beta_0(1-\theta) + \beta(X_{it} - \theta \bar{X}_i) + \nu_{it} - \theta \bar{\nu}_i}}$$

$$\theta = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}\right)^{\frac{1}{2}}$$

Can use STATA to transform model

if this is small ... serial correlation very small → OLS ✓

- In practice, we cannot use GLS but FGLS because the true θ is unknown but has to be estimated from consistent estimators of $\sigma_\alpha^2, \sigma_U^2$.
- The FGLS in this case is called the **Random Effects Estimator**.

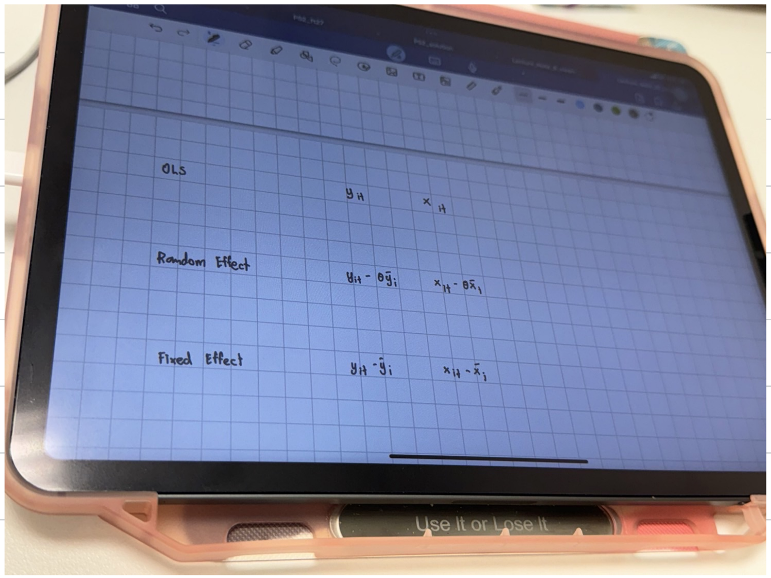
endogeneity ?? * ส่งวันพฤ

(think about regressand & regressor & error term

$$y = \beta x + \alpha + u$$

Estimator	I. Pool cross-section ($\alpha = 0$)	II. Fixed effect ($cov(\alpha, x) \neq 0$)	III. Random effect ($cov(\alpha, x) = 0$)
OLS	✓	X	
FGLS			
Between		X	
First diff		✓	
Fixed Eff		✓	
Random Eff			✓
2SLS/IV			
MLE			

OLS	y_{it}	x_{it}
	θ close to 0 ↑	
Random effect	$y_{it} - \theta \bar{y}_i$	$x_{it} - \theta \bar{x}_i$
	θ close to 1 ↓	
Fixed effect	$y_{it} - \bar{y}_i$	$x_{it} - \bar{x}_i$



cannot add it to first diff & fixed effect _ no. fi between

$\theta = 0.2$

who's stock is this?

Long Format

$X_1 - 0.2\bar{X}_1$	i	t	Y	X_1	X_2	$X_{1_{t-1}}$	$\bar{X}_1$	$X_{1_t} - \bar{X}_i$
12.8	A	2010	23,000	16	2	-	16	0
12.8	A	2011	23,500	16	3	16	16	0
12.8	A	2012	24,000	16	4	16	16	0
7.2	B	2010	12,000	9	7	-	9	0
7.2	B	2011	12,000	9	8	9	9	0
7.2	B	2012	12,700	9	9	9	9	0
	C	2010	20,000	16	0	-	16	0
	C	2011	21,000	16	1	16	16	0
	C	2012	21,000	16	2	16	16	0

- Notice that the Random Effects and Fixed Effects Estimators are derived from similar methods of demeaning the regressand and the regressors. However, the sample mean in the RE is weighted by θ.
- If $\theta = 0$, then the RE becomes pooled cross-sectional estimator. This is the case when σ_α^2 is small, i.e. α is relatively unimportant factor that explains variation in Y. However, it is not usually the case in practice to get $\hat{\theta}$ close to zero.
- If $\theta = 1$, then the RE is the same as the FE. In general, we may see RE and FE produce similar estimates when T is large, as $\hat{\theta}$ usually gets to 1.

Fixed or Random Effects?

cannot estimated by FE

- If there is no variation in X_i across time, e.g. years of education, we cannot estimate the effect of X on Y by using FE, because $X_{it} - \bar{X}_i$ is zero. However, it is fine to use RE or pooled OLS.
- However, RE is based on the strong assumption that α_i is uncorrelated with the regressors, which is hard to justify in some cases.
- For example, if you want to estimate the return on education by using years of education as a regressor in RE, it may not be sensible to assume that the individual unobserved heterogeneity like talent, ability, or family are uncorrelated with how much ones get education.
- In practice, some people use all the three methods (RE, FE, and pooled OLS) and compare the results.
- Theoretically, we should conduct a hypothesis test to see whether the data fit RE or FE model better

Hausman Test

- Hausman (1978) proposed a hypothesis testing to see if the data fits RE or FE better.
- The idea is to assume that the assumption of RE is valid. That is $cov(X_{it}, \alpha_i) = 0$. Under this hypothesis, both FE and RE estimators are consistent under regular assumption; so, the parameters estimated by each of the estimators should be the same.

fail to reject — Random Effect Model   reject — Fixed Effect Model

$$H_0: \beta_{RE} = \beta_{FE} \quad vs \quad H_1: \beta_{RE} \neq \beta_{FE}$$

- He also showed that under the null, $\hat{\beta}_{RE}$ is efficient, both $\hat{\beta}_{FE}, \hat{\beta}_{RE}$ are asymptotically normal, and

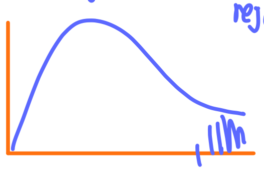
$$Var(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})$$

- So, the test statistic for the regression with k regressors is

chi-squared

low p-value → reject
high → fail to reject

$$\xi_H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' \{Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})\}^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_k$$



MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

STATA:

declare that data is panel data

get database from stata online

```
. webuse nlswork
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
```

change to long format

```
. xtset idcode year
       panel variable:  idcode (unbalanced)
        time variable:  year, 68 to 88, but with gaps
                delta:  1 unit
```

r.v. tells i

r.v. tells t

```
. su idcode year ln_w grade age ttl_exp tenure race
```

Variable	Obs	Mean	Std. Dev.	Min	Max
idcode	28,534	2601.284	1487.359	1	5159
year	28,534	77.95865	6.383879	68	88
ln_wage	28,534	1.674907	.4780935	0	5.263916
grade	28,532	12.53259	2.323905	0	18
age	28,510	29.04511	6.700584	14	46
ttl_exp	28,534	6.215316	4.652117	0	28.88461
tenure	28,101	3.123836	3.751409	0	25.91667
race	28,534	1.303392	.4822773	1	3

Between Effects Estimator

create dummy variable for each category

between effect

```
. xtreg ln_w grade age ttl_exp tenure i.race, be
```

NT

\# of individuals

```
Between regression (regression on group means)  Number of obs      =     28,099
Group variable: idcode                          Number of groups   =      4,697

R-sq:                                           Obs per group:
     within  = 0.1371                                         min =          1
     between = 0.4339                                         avg =        6.0
     overall = 0.3189                                         max =         15
```

unobserved heterogeneity

α

```
                                                F(6,4690)          =     599.20
sd(u_i + avg(e_i.))=  .3197248                  Prob > F           =     0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.069644	.0020313	34.29	0.000	.0656617	.0736263
age	-.0057459	.0011033	-5.21	0.000	-.0079088	-.003583
ttl_exp	.0284016	.0021193	13.40	0.000	.0242468	.0325563
tenure	.0288536	.0023147	12.47	0.000	.0243157	.0333915
race						
black	-.0545226	.0105788	-5.15	0.000	-.0752621	-.0337831
other	.1217347	.0427118	2.85	0.004	.0379995	.2054699
_cons	.7091377	.0344501	20.58	0.000	.6415992	.7766761

β se t p-value 95% CI

subtract mean of individual to kill α

Fixed Effects Estimator

fixed effect

```
. xtreg ln_w grade age ttl_exp tenure i.race, fe
```

level of educ remain the same

value doesn't change across time

ignore this doesn't have any meaning

```
                                                F(3,23399)        =    1315.26
corr(u_i, Xb)  = 0.1651                         Prob > F          =     0.0000
------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |          0  (omitted)
         age |  -.0030427   .0008644    -3.52   0.000    -.0047369   -.0013484
     ttl_exp |    .029036   .0014505    20.02   0.000      .026193     .031879
      tenure |   .0116574   .0009249    12.60   0.000     .0098444    .0134704
             |
        race |
      black  |          0  (omitted)
      other  |          0  (omitted)
             |
       _cons |   1.547951   .0181798    85.15   0.000     1.512317    1.583584
-------------+----------------------------------------------------------------
     sigma_u |   .3751722
     sigma_e |  .29556813
         rho |  .61703248   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(4696, 23399) = 7.64                 Prob > F = 0.0000
```

*

```
. estimates store FE /* store estimates to use in Hausman Test */
```

you name it

Fixed Effect Estimator → $\overset{0}{\beta_0} + \overset{\alpha_A}{\alpha_i}$ doesn't mean anything

is not identified

STATA try to normalize this

$$y_{it} = \cancel{\beta_0} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cancel{\alpha_i} + u_{it}$$

$$\bar{y}_i = \cancel{\beta_0} + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \cancel{\alpha_i} + \bar{u}_i$$

$$y_{it} - \bar{y}_i = \beta_1 (x_{1it} - \bar{x}_{1i}) + \beta_2 (x_{2it} - \bar{x}_{2i}) + \text{error}$$

← ดูอันนี้

Fixed Effects Estimator using OLS and dummies

r.v. captures individual

```
. reg ln_w age ttl_exp tenure i.idcode
```

create dummy as regressor

```
matsize too small
    You have attempted to create a matrix with too many rows or columns or attempted to
    fit a model with too many variables.  You need to increase matsize; it is currently
    400.  Use set matsize; see help matsize.

    If you are using factor variables and included an interaction that has lots of
    missing cells, either increase matsize or set emptycells drop to reduce the required
    matrix size; see help set emptycells.

    If you are using factor variables, you might have accidentally treated a continuous
    variable as a categorical, resulting in lots of categories.  Use the c. operator on
    such variables.
r(908);
```

error
เพราะ
individuals
มากไป

MSF Chula

CHULALONGKORN BUSINESS SCHOOL FLAGSHIP FOR LIFE AACSB ACCREDITED EFMD EQUIS ACCREDITED EFMD EPAS ACCREDITED

want STATA to do dummy variables as control

Fixed Effects Estimator using OLS and dummies

```
. areg ln_w age ttl_exp tenure, absorb(idcode)

Linear regression, absorbing indicators           Number of obs     =     28,101
                                                  F(   3,  23399)   =    1315.26
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.6813
                                                  Adj R-squared     =     0.6173
                                                  Root MSE          =     0.2956

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0030427   .0008644    -3.52   0.000    -.0047369   -.0013484
     ttl_exp |    .029036   .0014505    20.02   0.000      .026193     .031879
      tenure |   .0116574   .0009249    12.60   0.000     .0098444    .0134704
       _cons |   1.547925   .0181797    85.15   0.000     1.512291    1.583558
-------------+----------------------------------------------------------------
      idcode |   F(4698, 23399) =      7.637   0.000          (4699 categories)
```

Random Effects Estimator

random effect

```
. xtreg ln_w grade age ttl_exp tenure i.race, re
```

α

```
                                                Wald chi2(6)      =    7468.75
corr(u_i, X)    = 0 (assumed)                   Prob > chi2       =     0.0000
------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0723646    .001857    38.97   0.000     .0687249    .0760044
         age |  -.0044626   .0006658    -6.70   0.000    -.0057675   -.0031577
     ttl_exp |     .03052   .0011405    26.76   0.000     .0282846    .0327554
      tenure |   .0136254   .0008514    16.00   0.000     .0119567     .015294
             |
        race |
      black  |  -.0561311   .0103136    -5.44   0.000    -.0763455   -.0359168
      other  |   .1028286   .0425718     2.42   0.016     .0193895    .1862678
             |
       _cons |   .6711433   .0286437    23.43   0.000     .6150026     .727284
-------------+----------------------------------------------------------------
     sigma_u |  .27513121
     sigma_e |  .29556813
         rho |  .46423555   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store RE   /* store estimates to use in Hausman Test */
```

consistent estimator under both H_0, H_1 ← fixed effect model

consistent under H_0 only ← random effect model

has to estimate by FE & RE then store value first → then refer these to Hausman test

Hausman Test

```
. hausman FE RE
```

← put this before RE

	---- Coefficients ---- (b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
age	-.0030427	-.0044626	.0014199	.0005513
ttl_exp	.029036	.03052	-.001484	.0008962
tenure	.0116574	.0136254	-.001968	.0003615

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 62.04
Prob>chi2 = 0.0000

↳ reject H_0 → fixed effect model / cannot use random effect estimator
$cov(\alpha, x) \neq 0$
Between
ols