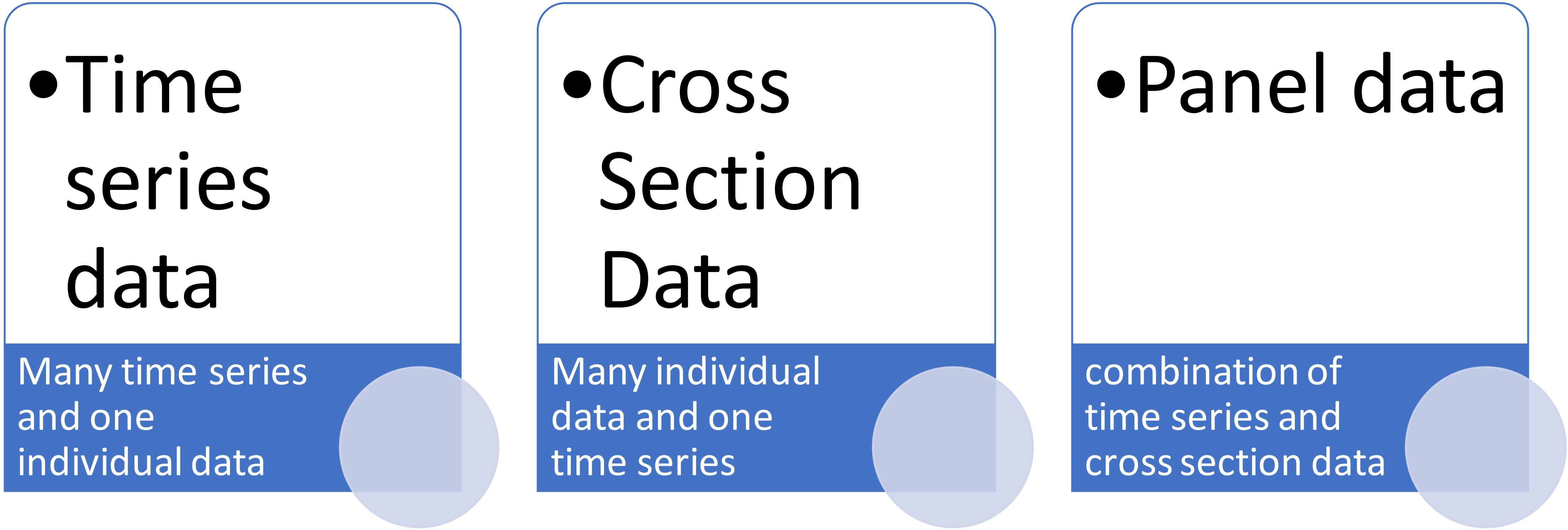
PANEL DATA ANALYSIS

BASIC AND APPLICATIONS

OUTLINE

- Introduction
- The meaning of Panel Data
- Example of Panel Data
- Advantages and Disadvantages of Panel Data
- Fixed Effect Model
- Random Effect Model
- Which model to choose? FE or RE? Hausman Test

The meaning of Panel Data



- Time series data

Many time series and one individual data

- Cross Section Data

Many individual data and one time series

- Panel data

combination of time series and cross section data

The meaning of Panel Data

- Refer to the combination of time series and cross data where the observation consist of several individual and time series. It can be

1. Individual
2. Household
3. Industry
4. City
5. State or District
6. Country

The time series of the panel data can be : daily data, weekly, monthly, quarterly and yearly

Example of Panel Data

Short and Wide (large N and short t)

Long and Narrow (Small N and long T)

Long and Wide (Large N and Long t)

How idealThe panel data?

Short and Wide

- Panel Study of Income Dynamics (PSID) has followed approximately 8,000 families since 1968.3 The U.S. Department of Labor conducts National Longitudinal Surveys (NLS) such as NLSY79, "a nationally representative sample of 12,686 young men and women who were 14–22 years old when they were first surveyed in 1979 These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis." Such data sets are "wide" and "short," because *N* is much, much larger than *T*.
- Using panel data sets of this kind we can account for unobserved individual differences, or **heterogeneity**. Furthermore, these data panels are becoming long enough so that dynamic factors, such as spells of employment and unemployment, can be studied. These very large data sets are rich in information, and require the use of considerable computing power.

Long and Wide -

- Indicating that both N and T are relatively large
- Macroeconomists who study economic growth across nations employ data that is “long” and “wide.” The Penn World Table5 provides purchasing power parity and national income accounts converted to international prices for 182 countries for some or all of the years 1950–2014, which we may roughly characterize as having both large N and large T.

Long and Narrow

- with “long” describing the time dimension and “narrow” implying a relatively small number of cross-sectional units
- A “long and narrow” panel may consist of data on several firms over a period of time.
- A classic example is a data set analyzed by Grunfeld and used subsequently by many authors.2 These data track investment in plant and equipment by $N = 11$ large firms for $T = 20$ years. This panel is narrow because it consists of only $N = 11$ firms. It is relatively “long” because $T > N$.

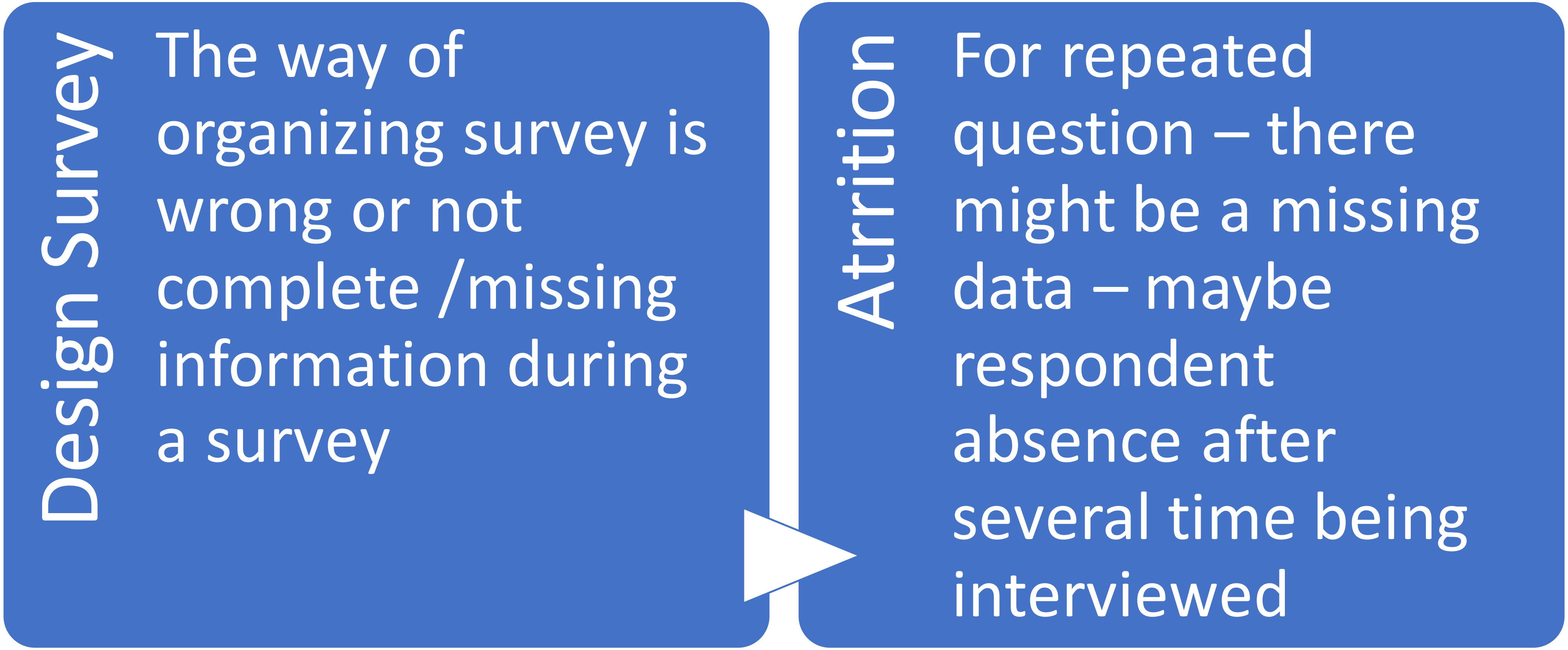
Advantages of Panel Data – I only list Three of them though there are more

Cater the heterogeneity (heterogeneity individual)

Overcome the problem of omitted variables (important variable are missing and unimportant variables are presents - not following the theory)

Minimize the problem of multicollinearity due to a lot of information and variation

Disadvantages of panel data



Design Survey

The way of organizing survey is wrong or not complete /missing information during a survey

Atrrition

For repeated question – there might be a missing data – maybe respondent absence after several time being interviewed

Panel Data – Main Issue – The unobserved Individual Heterogeneity

- Meaning – The individual heterogeneity cannot be measured, however from the panel data – the heterogeneity can be detected
- Example –
- The policy for every company and strategy is different from one another.
- The commitment of the manager in different department to achieve target in the production
- Different ability of the student to achieve higher marks

Why Panel Data anyway?

- If we use OLS – the OLS will be bias, because it unable to capture the individual heterogeneity from the individuals data (test pooled OLS vs LSDV –fixed effect)
- Therefore, the panel data analysis is able to capture the differences among individuals (this is individual heterogeneity) and between time period.
- The assumption is that it is not possible for every individuals to have the same characteristics. In addition, every individual can change maybe their attitude from year to year.
- How? Panel data can capture the individual heterogeneity from the error term.

Error Component Model –example for Grunfeld data

- $Invest_{it} = \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \varepsilon_{it}$
- What is error term /residual
- $\varepsilon_{it} = \leftthreetimes_i + u_{it}$ - where ε_{it} = epsilon of individual data i and time period t.
- Where $\leftthreetimes_i$ is individual heterogeneity and u_{it} is an error term – fulfil Gauss Markow teorem / CLRM
- i only for individual and not affect by time t

Determining the Panel Data Analysis

Fixed Effect Model?

Random Effect Model?

Fixed effect model

- $Invest_{it} = \beta_0 + \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \varepsilon_{it}$

$\varepsilon_{it} = \leftthreetimes_i + u_{it}$

$Invest_{it} = \beta_0 + \ \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \leftthreetimes_i + u_{it}$

$\leftthreetimes_i$ is assume as a fixed parameter can be added with the intercept such as

$Invest_{it} = (\beta_0 + \leftthreetimes_i) \ + \ \beta_1 \, mvalue_{it} + \beta_2 kstock_{it} + + u_{it}$

$Invest_{it} = \beta_{0i} \ + \beta_1 \, mvalue_{it} + \beta_2 kstock_{it} + + u_{it}$

This is known as the fixed effect model

Fixed Effect Model

- The fixed effect model capture the individual heterogeneity in the intercept value
- The estimation technique apply – LSDV – least square dummy variable)
- The fixed effect assume that the coefficient (the regression slope) has a fixed value between individuals and between time

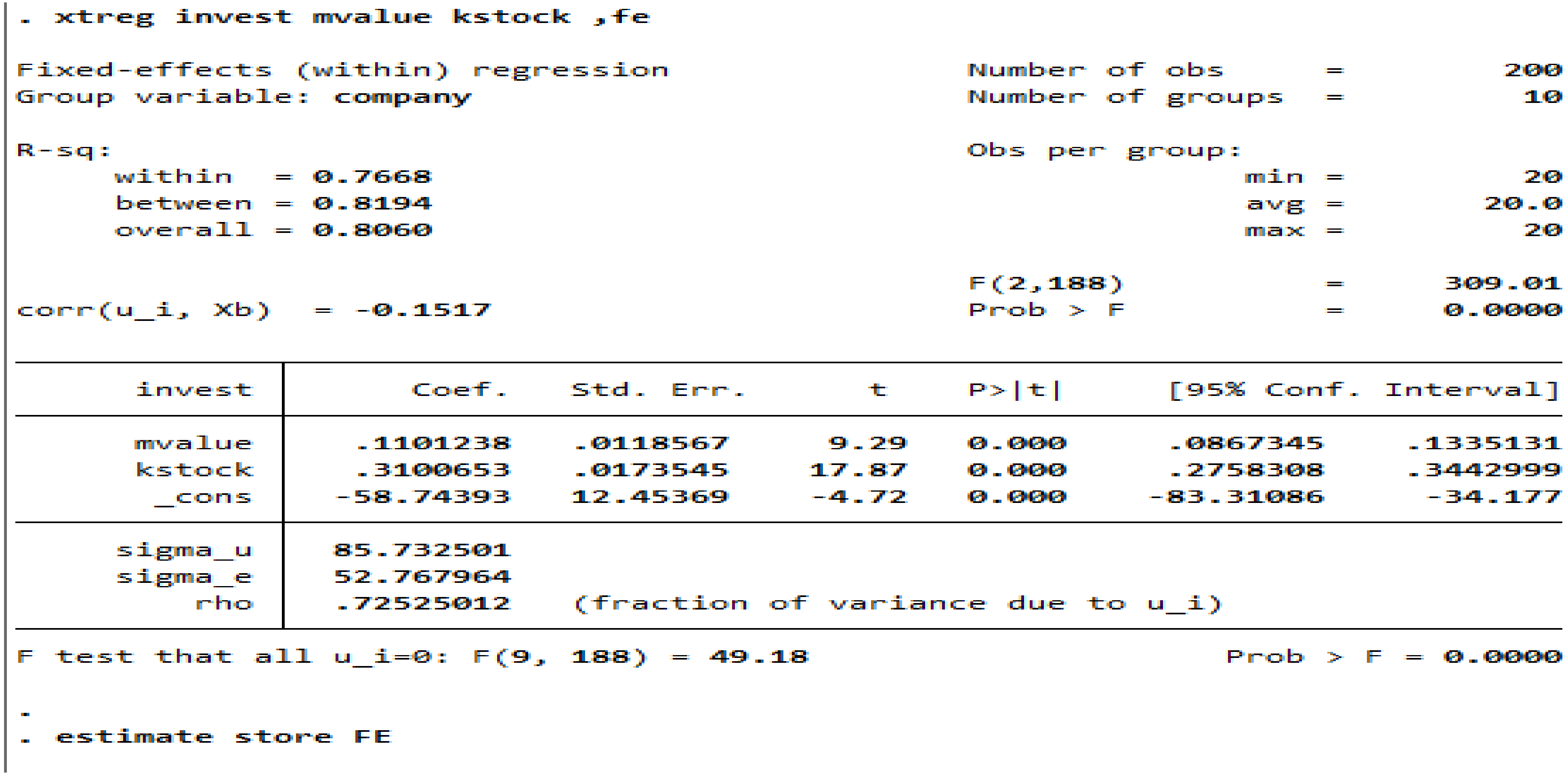
$$Invest_{it} = \beta_0 + \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + \varepsilon_{it}$$

if we have 4 firms = 3 dummy variables

A notes about fixed effect estimation-within estimator

- A fixed group model examines group differences in intercept.
- The LSDV for this fixed model need to create as many dummy variables as the number of entities or subjects.
- When many dummies are needed, the within effect model is useful since it uses transformed variables without creating dummies.
- Because “within” estimation does not involve dummy variables, thus will make the model has larger degree of freedom, smaller MSE, and smaller SE of parameter than those of LSDV.
- This estimation does not report individual dummy coefficients and we need to compute them if really need.
- Also notice that the usual reported in the within effect model is incorrect.
- In Stata we use xtreg command

Example of Fixed effect estimation, xtreg within estimation (balance panel data)



```
. xtreg invest mvalue kstock ,fe

Fixed-effects (within) regression               Number of obs      =       200
Group variable: company                         Number of groups   =        10

R-sq:                                           Obs per group:
     within  = 0.7668                                          min =        20
     between = 0.8194                                          avg =      20.0
     overall = 0.8060                                          max =        20

                                                F(2,188)           =    309.01
corr(u_i, Xb)  = -0.1517                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1101238   .0118567     9.29   0.000     .0867345    .1335131
      kstock |   .3100653   .0173545    17.87   0.000     .2758308    .3442999
       _cons |  -58.74393   12.45369    -4.72   0.000    -83.31086     -34.177
-------------+----------------------------------------------------------------
     sigma_u |  85.732501
     sigma_e |  52.767964
         rho |  .72525012   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(9, 188) = 49.18                     Prob > F = 0.0000

.
. estimate store FE
```

Random Effect Model

- The main purpose to add dummy variables in the fixed effect model is to let us know the unknown real model due to the heterogeneity
- However, the drawback of LSDV is the degree of freedom will be decrease if we keep adding the dummy variable (increase in i)
- Therefore, the estimated parameter is not efficient.
- This problem can be overcome applying the Random Effect Model where the model estimate the panel data in a such where the error has a relationship with the time t and also the individual i.

Random effect model

$$Invest_{it} = \beta_0 + \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \varepsilon_{it}$$

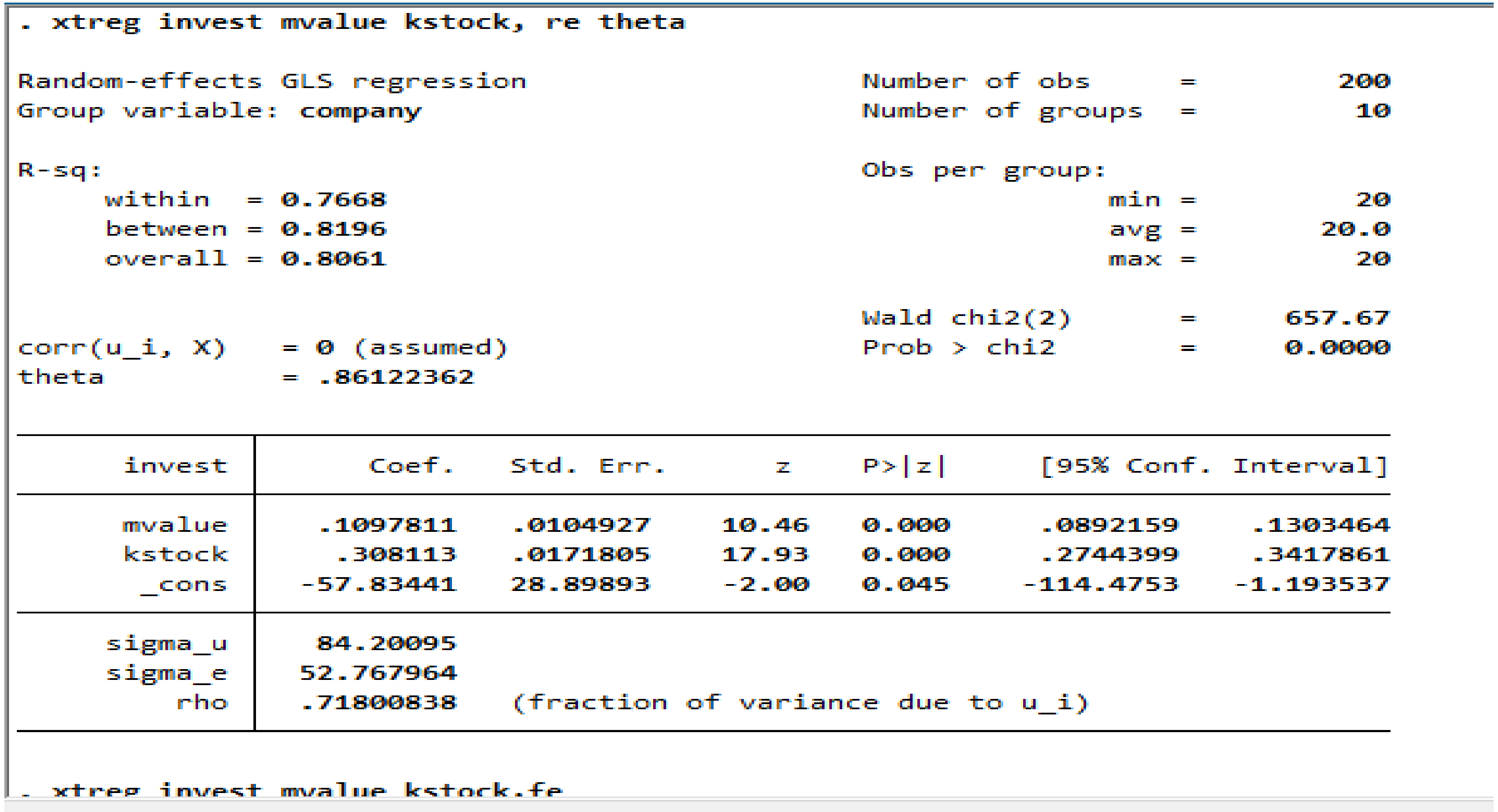
$$\varepsilon_{it} = \leftthreetimes_i + u_{it}$$

$\leftthreetimes_i$ is assume as part of the error term or random elements and:

$$Invest_{it} = \beta_0 + \beta_1 mvalue_{it} + \beta_2 kstock_{it} + \leftthreetimes_i + u_{it}$$

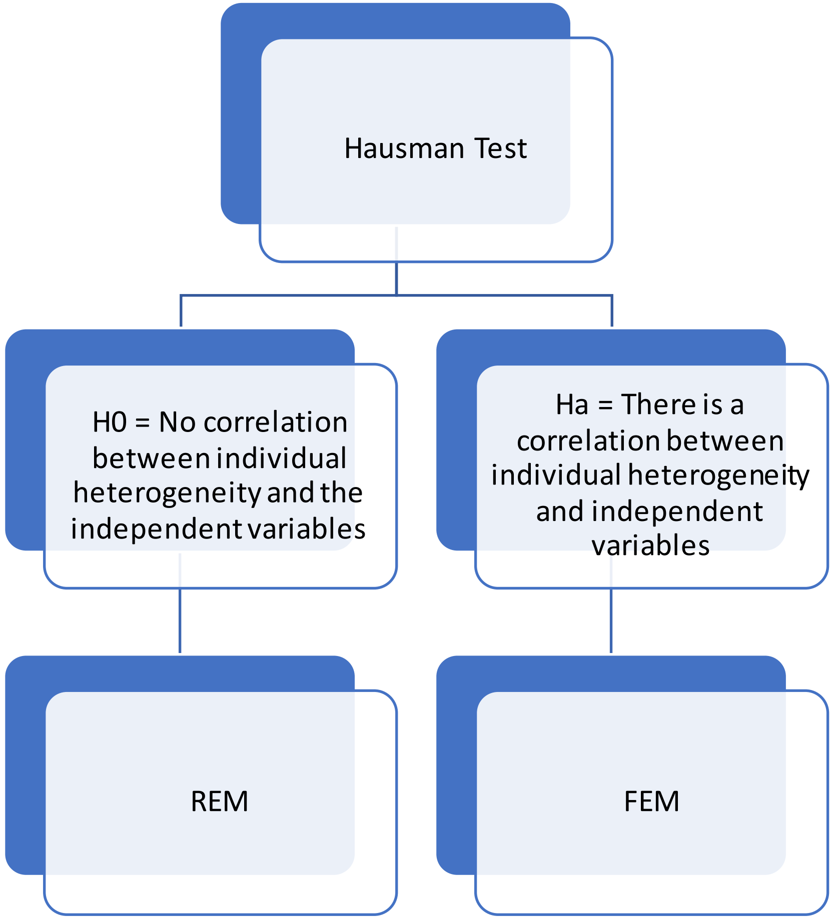
The meaning of REM is originated from the overall error or the combination of cross section u_{it} and individual heterogeneity $\leftthreetimes_i$

Random effect examples (between estimation)-balance panel data **xtreg** with **re** option to produce FGLS estimates



```
. xtreg invest mvalue kstock, re theta

Random-effects GLS regression                   Number of obs      =       200
Group variable: company                         Number of groups   =        10

R-sq:                                           Obs per group:
     within  = 0.7668                                          min =        20
     between = 0.8196                                          avg =      20.0
     overall = 0.8061                                          max =        20

                                                Wald chi2(2)       =    657.67
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
theta          = .86122362

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1097811   .0104927    10.46   0.000     .0892159    .1303464
      kstock |    .308113   .0171805    17.93   0.000     .2744399    .3417861
       _cons |  -57.83441   28.89893    -2.00   0.045    -114.4753   -1.193537
-------------+----------------------------------------------------------------
     sigma_u |   84.20095
     sigma_e |  52.767964
         rho |  .71800838   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. xtreg invest mvalue kstock,fe
```

Which one to choose? Random effect or Fixed effect model?



Hausman Test

H0 = No correlation between individual heterogeneity and the independent variables

Ha = There is a correlation between individual heterogeneity and independent variables

REM

FEM

Hausman Test

- H null : No correlation between $\leftthreetimes_i$ and independent variables – REM
- H alternative : correlation between $\leftthreetimes_i$ and independent variables – FEM

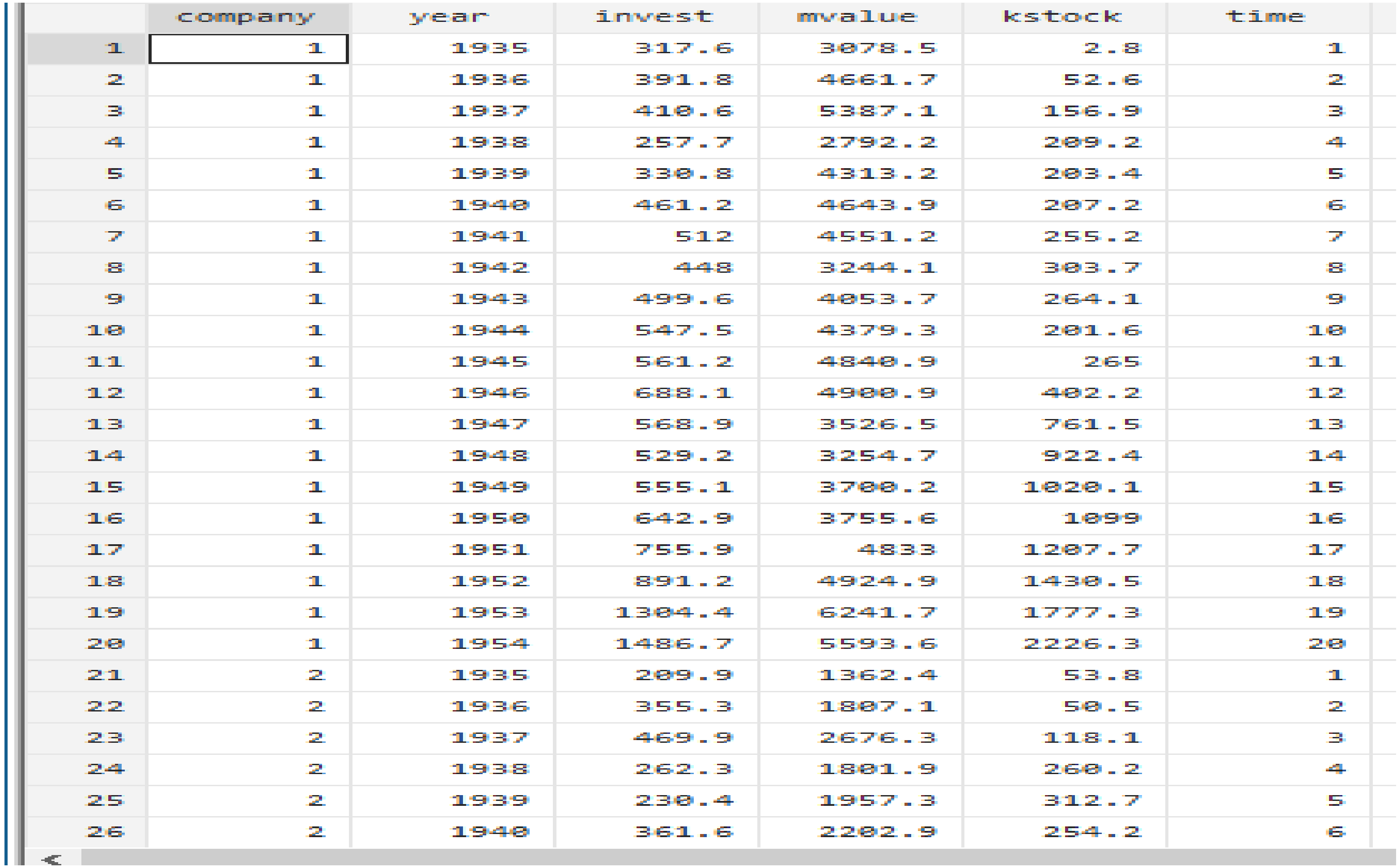
When the P value approach zero or P values < α, so :

- Do not reject H null and reject H alternative
- Mean no correlation between $\leftthreetimes_i$ and independent variables
- So the best model is the Random Effect Model

When the P value increase or p values > α , then:

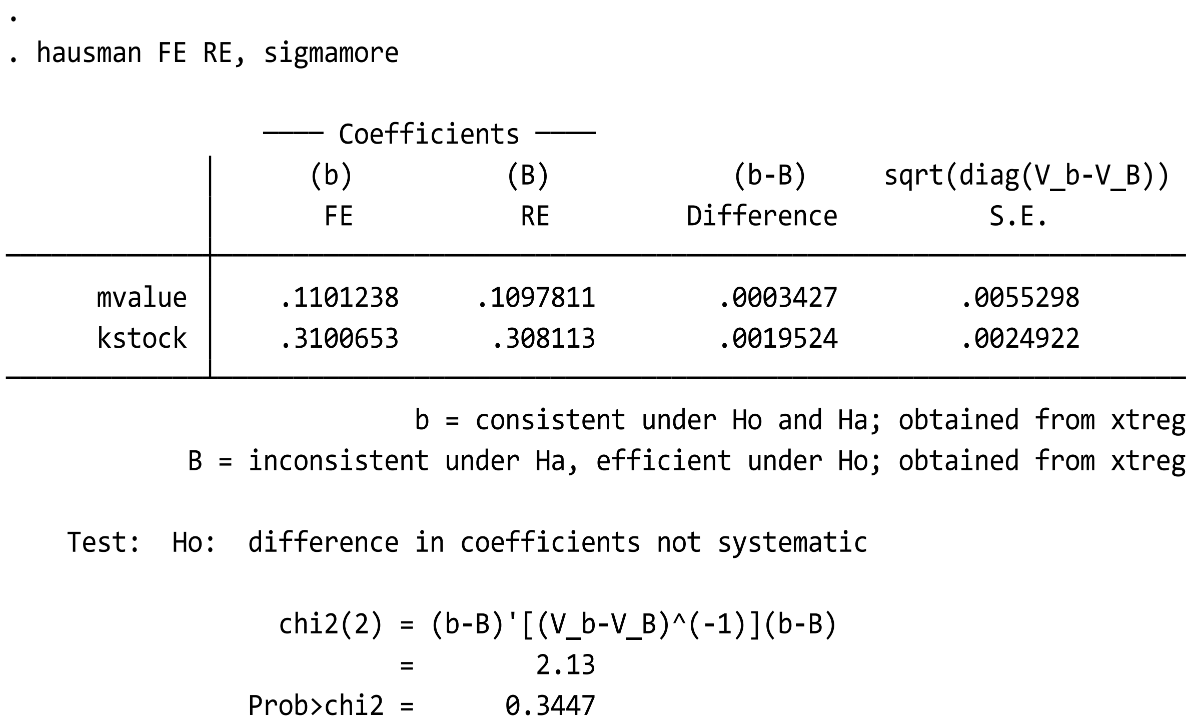
- reject H null and do not reject H alternative
- Mean there is correlation between $\leftthreetimes_i$ and independent variables
- So the best model is the Fixed Effect Model

STATA EXAMPLE - SEE DATA EDITOR



	company	year	invest	mvalue	kstock	time
1	1	1935	317.6	3078.5	2.8	1
2	1	1936	391.8	4661.7	52.6	2
3	1	1937	410.6	5387.1	156.9	3
4	1	1938	257.7	2792.2	209.2	4
5	1	1939	330.8	4313.2	203.4	5
6	1	1940	461.2	4643.9	207.2	6
7	1	1941	512	4551.2	255.2	7
8	1	1942	448	3244.1	303.7	8
9	1	1943	499.6	4053.7	264.1	9
10	1	1944	547.5	4379.3	201.6	10
11	1	1945	561.2	4840.9	265	11
12	1	1946	688.1	4900.9	402.2	12
13	1	1947	568.9	3526.5	761.5	13
14	1	1948	529.2	3254.7	922.4	14
15	1	1949	555.1	3700.2	1020.1	15
16	1	1950	642.9	3755.6	1099	16
17	1	1951	755.9	4833	1207.7	17
18	1	1952	891.2	4924.9	1430.5	18
19	1	1953	1304.4	6241.7	1777.3	19
20	1	1954	1486.7	5593.6	2226.3	20
21	2	1935	209.9	1362.4	53.8	1
22	2	1936	355.3	1807.1	50.5	2
23	2	1937	469.9	2676.3	118.1	3
24	2	1938	262.3	1801.9	260.2	4
25	2	1939	230.4	1957.3	312.7	5
26	2	1940	361.6	2202.9	254.2	6

Hausman Test = P value > α do not reject H null – choose REM



```
.
. hausman FE RE, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
      mvalue |    .1101238     .1097811        .0003427        .0055298
      kstock |    .3100653      .308113        .0019524        .0024922
------------------------------------------------------------------------------
                           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(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.13
                Prob>chi2 =      0.3447
```

Do not reject Hnull –choose RE