Panel Data Analysis

Aims

- Introduce the distinctive features of panel data.
- Review some panel data methods commonly used in finance, economics, and accounting.
- Present the advantages (and limitations) of panel data, and consider what sort of questions panel data can(not) address.
- Show how to handle and describe panel data.
- Introduce the basic estimation techniques for panel data (linear and non-linear).
- Discuss how to choose (and test for) the right technique for the question being addressed.

Structure

Basics

- What type of data one might encounter (Data DNA)
- Stata ice-breaker

Panel Data

- What and whys?
- Handling panel data in Stata – some basic commands.
- Within and between variation
- Understanding Fixed and Random Effects

Dynamic linear models (continuous variables)

- Arellano & Bond and Blundell & Bond estimators

Discrete variables

- binary response variables
- Ordered response models

What and Why?

- **What:**
- Panel data are a form of longitudinal data, involving regularly repeated observations on the same individuals
- **Individuals** may be people, households, firms, countries, etc
- **Repeat observations** are typically different time periods
- **Why**:
- Repeated observations on individuals allow for possibility of isolating effects of unobserved differences between individuals
- We can study dynamics
- The ability to make causal inference is enhanced by temporal ordering

BUT don't expect too much…

- Variation between firms (or people) usually far exceeds variation over time for a firm

  $\Rightarrow$ a panel with T waves doesn't give T times the information of a cross-section

- Variation over time may not exist for some important variables or may be inflated by measurement error
- We still need very strong assumptions to draw clear inferences from panels: sequencing in time does *not* necessarily reflect causation

The Basic Data Structure

$$
\begin{array}{r}
\\
\text{Individual 1} \\
\\
\\
\\
\text{Individual 2} \\
\\
\\
\\
\text{Individual N} \\
\\
\\
\end{array}
\begin{pmatrix}
x_{111} & x_{211} & .. & x_{K11} \\
x_{112} & x_{212} & .. & x_{K12} \\
.. & .. & .. & .. \\
x_{11T} & x_{21T} & .. & x_{K1T} \\
x_{121} & x_{221} & .. & x_{K21} \\
x_{122} & x_{222} & .. & x_{K22} \\
.. & .. & .. & .. \\
x_{12T} & x_{22T} & .. & x_{K2T} \\
x_{1N1} & x_{2N1} & .. & x_{KN1} \\
x_{1N2} & x_{2N2} & .. & x_{KN2} \\
.. & .. & .. & .. \\
x_{1NT} & x_{2NT} & .. & x_{KNT}
\end{pmatrix}
\begin{array}{l}
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T} \\
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T} \\
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T}
\end{array}
$$

Some terminology

- A **balanced panel** has the same number of time observations (T) on each of the n individuals
- An **unbalanced panel** has different numbers of time observations (T_i) on each individual
- A **compact panel** covers only consecutive time periods for each individual – there are no "gaps"
- **Attrition** is the process of drop-out of individuals from the panel, leading to an unbalanced and possibly non-compact panel
- A **short panel** has a large number of individuals but few time observations on each
- A **long panel** has a long run of time observations on each individual, permitting separate time-series analysis for each
- We consider only short panels in this seminar

Panel and time variables

- Use `tsset` to tell Stata which are panel and time variables:

```
. tsset id year
```

- Note that `tsset` automatically sorts the data accordingly.

Our dataset

- Sample size: 79,558 (bank-year) obs
- Sample dimensions:

Time span	Cross-section
2001	9598
2002	9349
2003	9168
2004	8965
2005	8819
2006	8666
2007	8525
2008	8322
2009	8146

To get more info use xtdescribe

```
. xtdescribe

      id:  9, 14, ..., 91363                                 n =      10627
    year:  2001, 2002, ..., 2009                             T =          9
           Delta(year) = 1 unit
           Span(year)  = 9 periods
           (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       2       6         9         9       9       9

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------
      7201    67.76   67.76 |  111111111
       345     3.25   71.01 |  1........
       337     3.17   74.18 |  11111....
       329     3.10   77.27 |  111......
       318     2.99   80.27 |  1111.....
       310     2.92   83.18 |  111111...
       297     2.79   85.98 |  11.......
       279     2.63   88.60 |  1111111..
       182     1.71   90.32 |  .....1111
      1029     9.68  100.00 | (other patterns)
 ---------------------------+-----------
     10627   100.00         |  XXXXXXXXX
```

Variation of the dependent variable and the regressors

- See word file
- Main concepts:
- overall variation
- Between variation
- Within variation

Between- and within-group variation

Define the individual-specific or group mean for any variable, *e.g.* y_{it} as:

$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it}$$

y_{it} can be decomposed into 2 orthogonal components:

$$y_{it} - \bar{y} = \left(y_{it} - \bar{y}_i\right) + \left(\bar{y}_i - \bar{y}\right)$$
$$= \text{within} \quad + \quad \text{between}$$

where

$$\bar{y} = \sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{it} \Big/ \sum_{i=1}^{n} T_i$$

Corresponding decomposition of sum of squares:

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(y_{it} - \bar{y}\right)^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(y_{it} - \bar{y}_i\right)^2 + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\bar{y}_i - \bar{y}\right)^2$$

or:

$$T_{yy} = W_{yy} + B_{yy}$$

Between- and within-group variation `xtsum`

- Stata contains a 'canned' routine, `xtsum`, that summarises within and between variation.
- But it does not give an exact decomposition:
  - Converts sums of squares to variance using different 'degrees of freedom' so they are not comparable
  - Reports square root (i.e. standard deviation) of these variances
  - Documentation is not very clear!
- But useful as a good approximation.

Develop an error components model

$$\boxed{y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}}$$

Explanatory variables

Normally distributed error -

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

$$u_{it} \sim N(0, \sigma_u^2)$$

Constant across individuals

Composite error term

Treatment of individual effects

Then two options for treatment of individual effects:

- Fixed effects – assume λ_i are constants
- Random effects – assume λ_i are drawn independently from some probability distribution

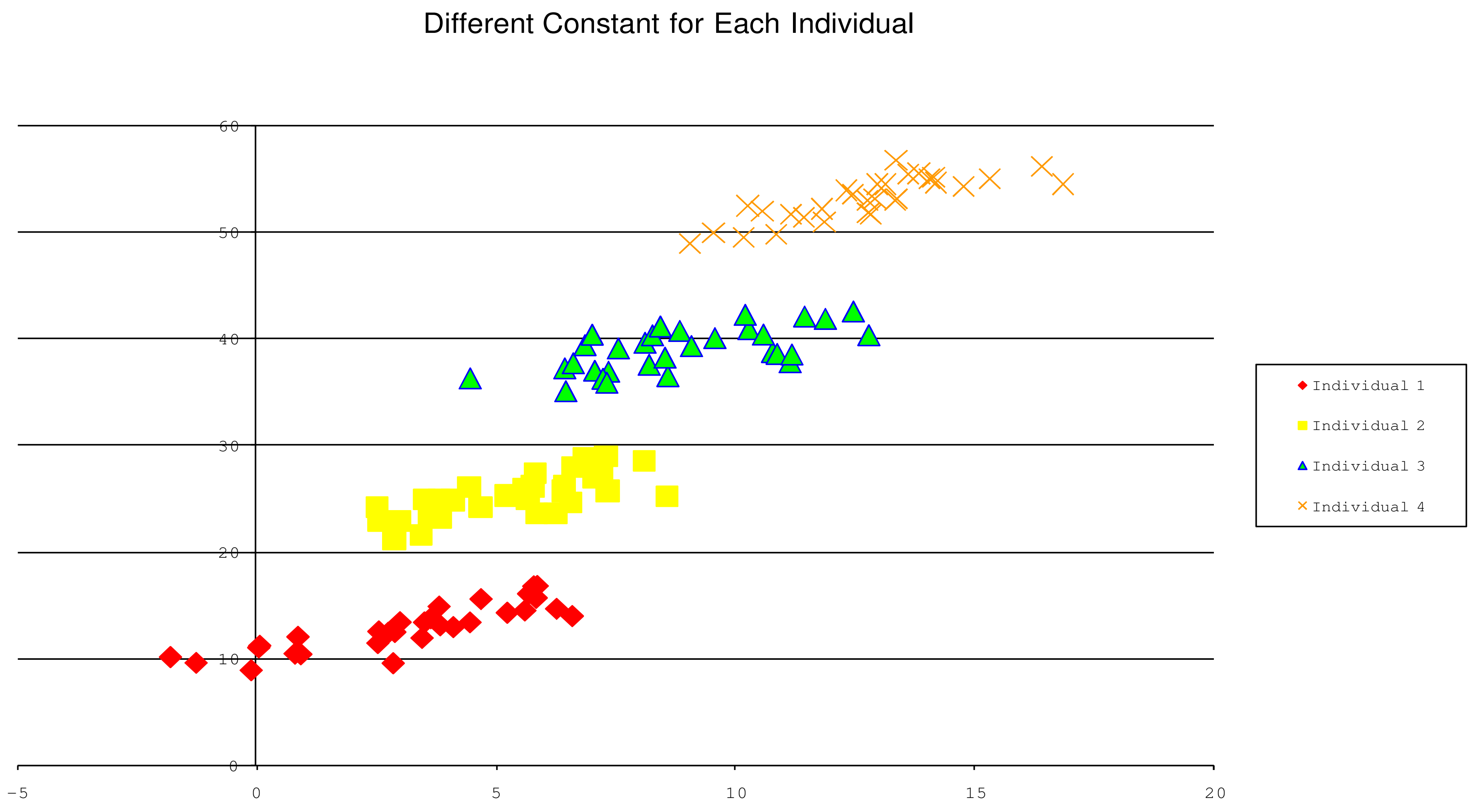
The Fixed Effects Model

Treat λ_i as a constant for each individual

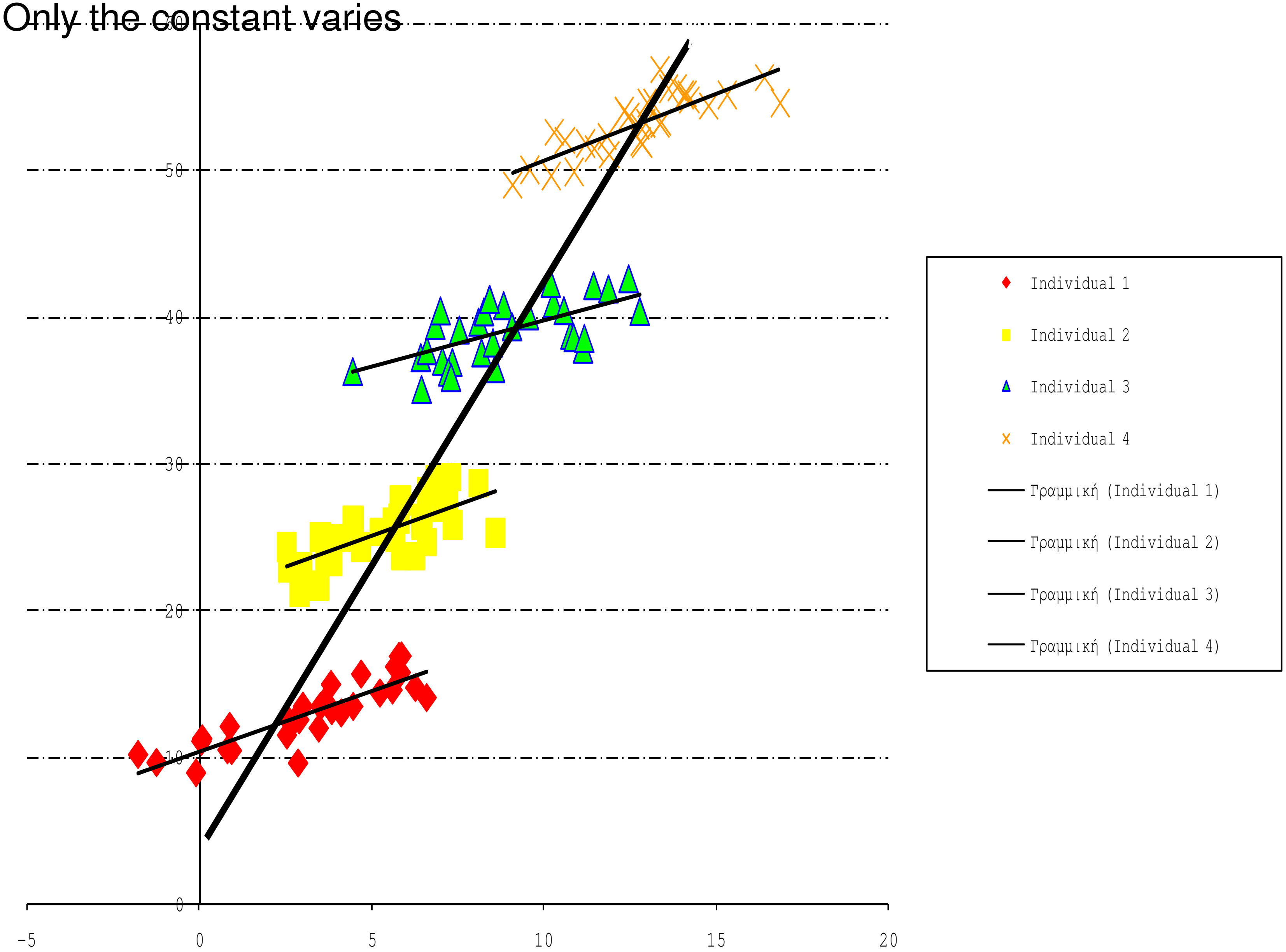
$$y_{it} = (\beta_0 + \lambda_i) + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + u_{it}$$

λ now part of constant – but varies by individual

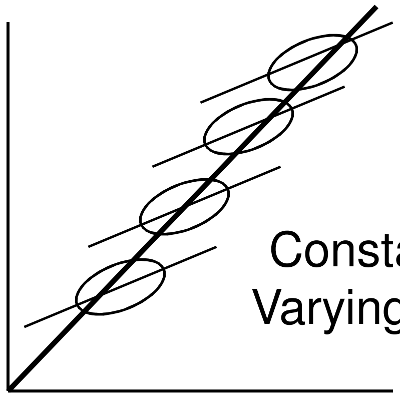
Graphically this looks like:



And the slope that will be estimated is BB rather than AA
Note that the slope of BB is the same for each individual
Only the constant varies

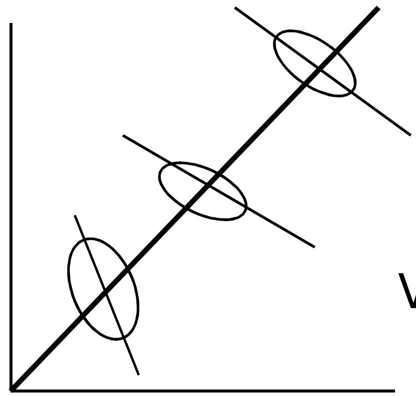


Possible Combinations of Slopes and Intercepts



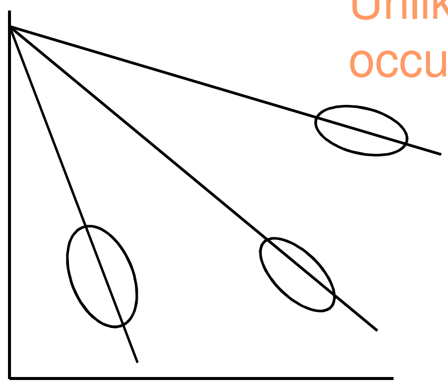
The fixed effects model

Constant slopes
Varying intercepts



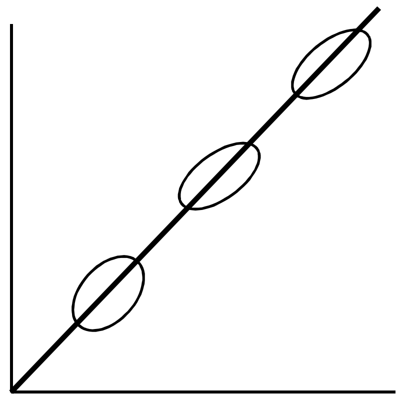
Separate regression for each individual

Varying slopes
Varying intercepts



Unlikely to occur

Varying slopes
Constant intercept



The assumptions required for this model are unlikely to hold

Constant slopes
Constant intercept

Constructing the fixed-effects model - eliminating unobserved heterogeneity by taking first differences

Original equation

$$y_{it} = \beta_0 + \lambda_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + u_{it}$$

Lag one period and subtract

$$\begin{aligned} y_{it} - y_{it-1} = {} & \beta_0 + \lambda_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + u_{it} \\ & - \beta_0 - \lambda_i - \beta_1 x_{1it-1} - \beta_2 x_{2it-1} - \ldots - \beta_k x_{kit-1} - u_{it-1} \end{aligned}$$

Constant and individual effects eliminated

$$\begin{aligned} y_{it} - y_{it-1} = {} & \beta_1 \left(x_{1it-1} - x_{1it-1} \right) + \beta_2 \left(x_{2it} - x_{2it-1} \right) + \ldots \\ & + \beta_k \left(x_{kit} - x_{kit-1} \right) + \left(u_{it} - u_{it-1} \right) \end{aligned}$$

Transformed equation

$$\Delta y_{it} = \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it}$$

An Alternative to First-Differences:

Deviations from Individual Means

$$\Delta y_{it} = \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it}$$

Applying least squares gives the first-difference estimator – it works when there are two time periods.

More general way of "sweeping out" fixed effects when there are more than two time periods - *take deviations from individual means*.

Let $x_{1i.}$ be the mean for variable x_1 for individual i, averaged across all time periods. Calculate means for each variable (including y) and then subtract the means gives:

$$y_{it} - \bar{y}_{i.} = \beta_0 - \beta_0 + \lambda_i - \bar{\lambda}_{i.} + \beta_1 \left(x_{1it} - \bar{x}_{1i.} \right) + \ldots + \beta_k \left(x_{kit} - \bar{x}_{ki.} \right) + u_{it}$$

The constant and individual effects are also eliminated by this transformation

Estimating the Fixed Effects Model

Take deviations from individual means and apply least squares – fixed effects, LSDV or "within" estimator

$$y_{it} - \bar{y}_{i.} = \beta_1 \left(x_{1it} - \bar{x}_{1i.} \right) + \ldots + \beta_k \left(x_{kit} - \bar{x}_{ki.} \right) + u_{it}$$

It is called the "within" estimator because it relies on variations within individuals rather than between individuals. Not surprisingly, there is another estimator that uses only information on individual means. This is known as the "between" estimator. The Random Effects model is a combination of the Fixed Effects ("within") estimator and the "between" estimator.

Three ways to *estimate* β

$$y_{it} = \beta' x_{it} + \varepsilon_{it}$$ overall

$$y_{it} - \bar{y}_{i.} = \beta'\left(x_{it} - \bar{x}_{i.}\right) + \varepsilon_{it} - \bar{\varepsilon}_{i.}$$ within

$$\bar{y}_{i.} = \beta' \bar{x}_{i.} + \bar{\varepsilon}_{i.}$$ between

The overall estimator is a weighted average of the "within" and "between" estimators. It will only be *efficient* if these weights are correct.

The *random effects* estimator uses the **correct weights**.

Stata output: within-group regression

```
. xtreg noi size1 risk1 cap, fe

Fixed-effects (within) regression               Number of obs      =     68125
Group variable: id                              Number of groups   =     10131

R-sq:  within  = 0.0798                         Obs per group: min =         1
between = 0.0008                                avg =        6.7
overall = 0.0181                                max =          8
F(3,57991)          =   1675.46
corr(u_i, Xb)  = -0.1275                        Prob > F           =    0.0000


noi         Coef.   Std. Err.      t    P>t     [95% Conf. Interval]

size1   -.2288585   .0195904   -11.68   0.000    -.2672557   -.1904613
risk1    .0111155   .0006681    16.64   0.000     .0098061    .0124249
cap      1.800679    .027743    64.91   0.000     1.746302    1.855055
_cons   -4.448611   .2699042   -16.48   0.000    -4.977625   -3.919598

sigma_u   1.9432036
sigma_e   1.3609366
rho    .67091573   (fraction of variance due to u_i)

F test that all u_i=0:     F(10130, 57991) =    10.37       Prob > F = 0.0000
```

Stata output: between-group regression

```
. xtreg noi size1 risk1 cap, be

Between regression (regression on group means)  Number of obs      = 68125
Group variable: id                              Number of groups   = 10131

R-sq:  within  = 0.0623                         Obs per group: min = 1
between = 0.0438                                avg =        6.7
overall = 0.0372                                max =        8

F(3,10127)          =      154.64
sd(u_i + avg(e_i.))=  1.851404    Prob > F          =      0.0000


noi        Coef.    Std. Err.      t P>t     [95% Conf.       Interval]

size1    .1613118   .0137935    11.69      0.000      .1342738        .1883498
risk1   -.0093633   .0011373    -8.23      0.000     -.0115926        -.007134
cap       2.404782   .1536881    15.65     0.000      2.103523        2.706041
_cons   -10.12686   .6474204   -15.64      0.000     -11.39593        -8.857783

------------------------------------------------------------------------------
```

The Random Effects Model

Original equation

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}$$

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \lambda_i + u_{it}$$

Remember $\varepsilon_{it} = \lambda_i + u_{it}$ λ_i now part of error term

This approach might be appropriate if observations are representative of a sample rather than the whole population. This seems appealing.

The Variance Structure in Random Effects

In random effects, we assume the λ_i are part of the composite error term ε_{it}. To construct an efficient estimator we have to evaluate the structure of the error and then apply an appropriate generalised least squares estimator to find an efficient estimator. The assumptions must hold if the estimator is to be efficient. These are:

$$E(u_{it}) = E(\lambda_i) = 0; \qquad E(u_{it}^2) = \sigma_u^2;$$

$$E(\lambda_i^2) = \sigma_\lambda^2; \qquad E(u_{it}\lambda_i) = 0 \quad for\, all\, i,t$$

$$E(\varepsilon_{it}^2) = \sigma_u^2 + \sigma_\lambda^2 \quad t = s; \qquad E(\varepsilon_{it}\varepsilon_{is}) = \sigma_\lambda^2, \quad t \neq s;$$

$$and$$

$$E(x_{kit}\lambda_i) = 0\; for\, all\, k,t,i$$

This is a crucial assumption for the RE model.
It is necessary for the consistency of the RE model,
but not for FE. It can be tested with the Hausman test.

The Variance Structure in Random Effects

Derive the T by T matrix that describes the variance structure of the ε_{it} for individual *i*. Because the randomly drawn λ_i is present each period, there is a correlation between each pair of periods for this individual.

$$\varepsilon_i' = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots \varepsilon_{iT}); \text{ then } E(\varepsilon_i \varepsilon_i') =$$

$$\begin{bmatrix} \sigma_u^2 + \sigma_\lambda^2 & \sigma_\lambda^2 & \sigma_\lambda^2 & \sigma_\lambda^2 \\ \sigma_\lambda^2 & \sigma_u^2 + \sigma_\lambda^2 & & \sigma_\lambda^2 \\ \sigma_i^2 & & .. & .. \\ \sigma_\lambda^2 & \sigma_\lambda^2 & .. & \sigma_u^2 + \sigma_\lambda^2 \end{bmatrix} = \sigma_u^2 I + \sigma_\lambda^2 ee' = \Omega$$

where $e' = (111\ldots..1)$ is a unit vector of size T

Random Effects (GLS Estimation)

The Random Effects estimator has the standard generalised least squares form summed over all individuals in the dataset i.e.

$$\hat{\beta}_{RE} = \left[\sum_{i=1}^{N} (X_i' \Omega^{-1} X_i) \right]^{-1} \sum_{i=1}^{N} X_i' \Omega^{-1} y_i$$

Where, given Ω from the previous slide, it can be shown that:

$$\Omega^{-1/2} = \frac{1}{\sigma_u} \left(I_T - \frac{\theta}{T} ee' \right) \text{ where } \theta = 1 - \frac{\sigma_u}{\sqrt{T\sigma_\lambda^2 + \sigma_u^2}}$$

Relationship between Random and Fixed Effects

The random effects estimator is a weighted combination of the “within” and “between” estimators. The “between” estimator is formed from:

$$\hat{\beta}_{RE} = \Psi \hat{\beta}_{Between} + (I_K - \Psi)\hat{\beta}_{Within}$$

Ψ depends on θ in such a way that if $\theta \rightarrow 1$ then the RE and FE estimators coincide. This occurs when the variability of the individual effects is large relative to the random errors.

$\theta \rightarrow 0$ corresponds to OLS (because the individual effects are small relative to the random error).

Random or Fixed Effects?

For random effects:

- Random effects are efficient
- Why should we assume one set of unobservables fixed and the other random?
- Sample information more common than that from the entire population?
- Can deal with regressors that are fixed across individuals

Against random effects:

Likely to be correlation between the unobserved effects and the explanatory variables. These are assumed to be zero in the random effects model, but in many cases we might expect them to be non-zero. This implies **inconsistency** due to omitted-variables in the RE model. In this situation, fixed effects is inefficient, but still consistent.

The Hausman Test

A test for the independence of the λ_i and the x_{kit}.

The covariance of an efficient estimator with its difference from an inefficient estimator should be zero. Thus, under the null hypothesis we test:

$$W = (\beta_{\mathrm{RE}} - \beta_{FE})'\hat{\Sigma}^{-1}(\beta_{\mathrm{RE}} - \beta_{FE}) \sim \chi^2(k)$$

If W is significant, we should not use the random effects estimator.

Can also test for the significance of the individual effects

feasible GLS estimates

```
. xtreg noi size1 risk1 cap, re theta

Random-effects GLS regression     Number of obs       =      68125
Group variable: id                 Number of groups    =      10131
R-sq:  within  = 0.0767             Obs per group: min =      1
between = 0.0139 avg =    6.7
overall = 0.0403 max =    8
Wald chi2(3)        =      4981.38
corr(u_i, X)   = 0 (assumed)      Prob > chi2         =      0.0000
```

theta -------------------

min	5%	median	95%	max
0.3857	**0.5178**	**0.7346**	**0.7346**	**0.7346**

noi	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
size1	.0034773	.0112073	0.31	0.756	-.0184885	.0254432
risk1	.0066515	.000574	11.59	0.000	.0055265	.0077764
cap	1.873067	.0272287	68.79	0.000	1.8197	1.926435
_cons	-7.277254	.1789141	-40.67	0.000	-7.627919	-6.926589

```
sigma_u   1.7481621
sigma_e   1.3609366
rho   .62264352   (fraction       of variance due to u_i)
```

within-group estimates

```
. xtreg noi size1 risk1 cap, fe

Fixed-effects (within) regression               Number of obs      =     68125
Group variable: id                              Number of groups   =     10131

R-sq:  within  = 0.0798                         Obs per group: min =         1
between = 0.0008                                       avg =       6.7
overall = 0.0181                                       max =         8

F(3,57991)          =   1675.46
corr(u_i, Xb)  = -0.1275                        Prob > F           =    0.0000


noi       Coef.   Std. Err.      t    P>t     [95% Conf. Interval]

size1   -.2288585   .0195904   -11.68   0.000    -.2672557   -.1904613
risk1    .0111155   .0006681    16.64   0.000     .0098061    .0124249
cap    1.800679    .027743    64.91   0.000    1.746302    1.855055
_cons   -4.448611   .2699042   -16.48   0.000    -4.977625   -3.919598

sigma_u   1.9432036
sigma_e   1.3609366
rho    .67091573   (fraction of variance due to u_i)

F test that all u_i=0:     F(10130, 57991) =    10.37        Prob > F = 0.0000
```

Hausman test

```
xtreg noi size1 risk1 cap, fe
estimates store fixed
xtreg noi size1 risk1 cap, re
estimates store random
hausman fixed random
                 ---- Coefficients ----
                 (b)          (B)             (b-B)     sqrt(diag(V_b-V_B))
                 fixed        random        Difference          S.E.

        size1    -.2288585     .0034773       -.2323359         .0160679
        risk1    .0111155     .0066515         .004464         .0003419
        cap      1.800679     1.873067       -.0723887         .0053173

                 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)
                 =      453.94
                 Prob>chi2 =      0.0000
```

Conclusion: we reject H_0 – so the random-effects regression is biased

Random effects ordered probit (2)

- Finally:

$$\begin{aligned}\Pr(y_{it} = J \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) &= \Pr(\mu_J < y_{it}^* \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) \\ &= 1 - \Pr(y_{it}^* \leq \mu_J \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) \\ &= 1 - \Phi(\mu_J - \mathbf{z}_i \boldsymbol{\alpha} - \mathbf{x}_{it} \boldsymbol{\beta} - u_i)\end{aligned}$$

- Check that these probabilities sum to one!
- Predicting probabilities and calculating marginal effects is done analogously to the binary RE probit.

Random effects ordered probit estimation example (xtoprobit)

```
Random-effects ordered probit regression        Number of obs      =     68125
Group variable: id                              Number of groups   =     10131

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       6.7
                                                               max =         8

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(3)       =   1000.87
Log likelihood  = -5725.3604                    Prob > chi2        =    0.0000
```

cap	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
noi	.277461	.0091586	30.30	0.000	.2595104	.2954115
risk1	-.0134609	.0009823	-13.70	0.000	-.0153862	-.0115356
size1	-.1494287	.016144	-9.26	0.000	-.1810703	-.1177872
/cut1	-7.832985	.2770601	-28.27	0.000	-8.376013	-7.289957
/cut2	-7.009726	.2600338	-26.96	0.000	-7.519383	-6.500069
/cut3	-6.530671	.2533777	-25.77	0.000	-7.027283	-6.03406
/cut4	-5.726291	.2447736	-23.39	0.000	-6.206039	-5.246544
/sigma2_u	.9987348	.10055			.8198859	1.216598

LR test vs. oprobit regression: chibar2(01) = 620.84 Prob>=chibar2 = 0.0000

Obtain predicted probabilities: predict prob*, pu0

```
. sum prob1 prob2 prob3 prob4 prob5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
prob1	68125	.0004656	.0176663	0	1
prob2	68125	.0003956	.0078075	0	.3193535
prob3	68125	.0005367	.0062296	0	.1889938
prob4	68125	.0029317	.0157829	0	.3124515
prob5	68125	.9956704	.0362212	0	1