Topic 3: Panel Models: Statistical Foundations

UNIVERSITY OF STIRLING

The Basic Data Structure

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

Linear model that reflects structure of the panel

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

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

Normally distributed error -

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

Individual Effect

Composite error term

UNIVERSITY OF STIRLING

Unobserved Heterogeneity (λ_i)

- Individual characteristics not all observed
  – e.g. enthusiasm, willingness to take risks
- Vary across individuals – described as unobserved heterogeneity
- If these influence the variable of interest, and are correlated with observed variates, then the estimated effects of these variables will be biased due to omitted variables
- Also sometimes described as a “latent” variable

What if we omit a variable that should be in the regression?

Population relationship: $$y_i = \beta_1 + \beta_{21} x_{2i} + \beta_{32} x_{3i} + \varepsilon_i$$

Estimated relationship: $$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + e_i$$

We have wrongly omitted the explanatory variable x_3 from the relationship we have estimated.

Does this matter?

Yes - it can cause **omitted variables bias**.

Omitted Variable Bias

The omitted variable x_3 will have an *indirect* effect through x_2 on the dependent variable. The size of this effect depends on the relationship between x_2 and x_3.

If x_2 and x_3 are unrelated, then omission of x_3 will have no impact on the estimate of the coefficient on x_2 .

If they are related, the estimate of the coefficient on x_2 is biased because x_2 is now doing two tasks:

- having its own independent effect on y
- proxying the effect of x_3 on y

UNIVERSITY OF STIRLING

Return to the panel model ..

$$\boxed{y_{it} = \beta_0 + \beta_1 x_{1it} + \varepsilon_{it}}$$

Only one explanatory variable, education

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

Observable

Ability $\leftarrow$ Unobservable

UNIVERSITY OF STIRLING

Covariance between x_1 and ε

- If ability and education correlated, then the assumption of no covariance between regressor and error term breaks down
- Estimates of the parameter β_1 will be biased
- Because ability is not included as a separate regressor
- But there is no option to include it - it is not observable

Dealing with Unobserved Heterogeneity

Two options:

- fixed effects – assume λ_i are constants
- random effects – assume λ_i are drawn independently from some probability distribution

The Fixed Effects Model

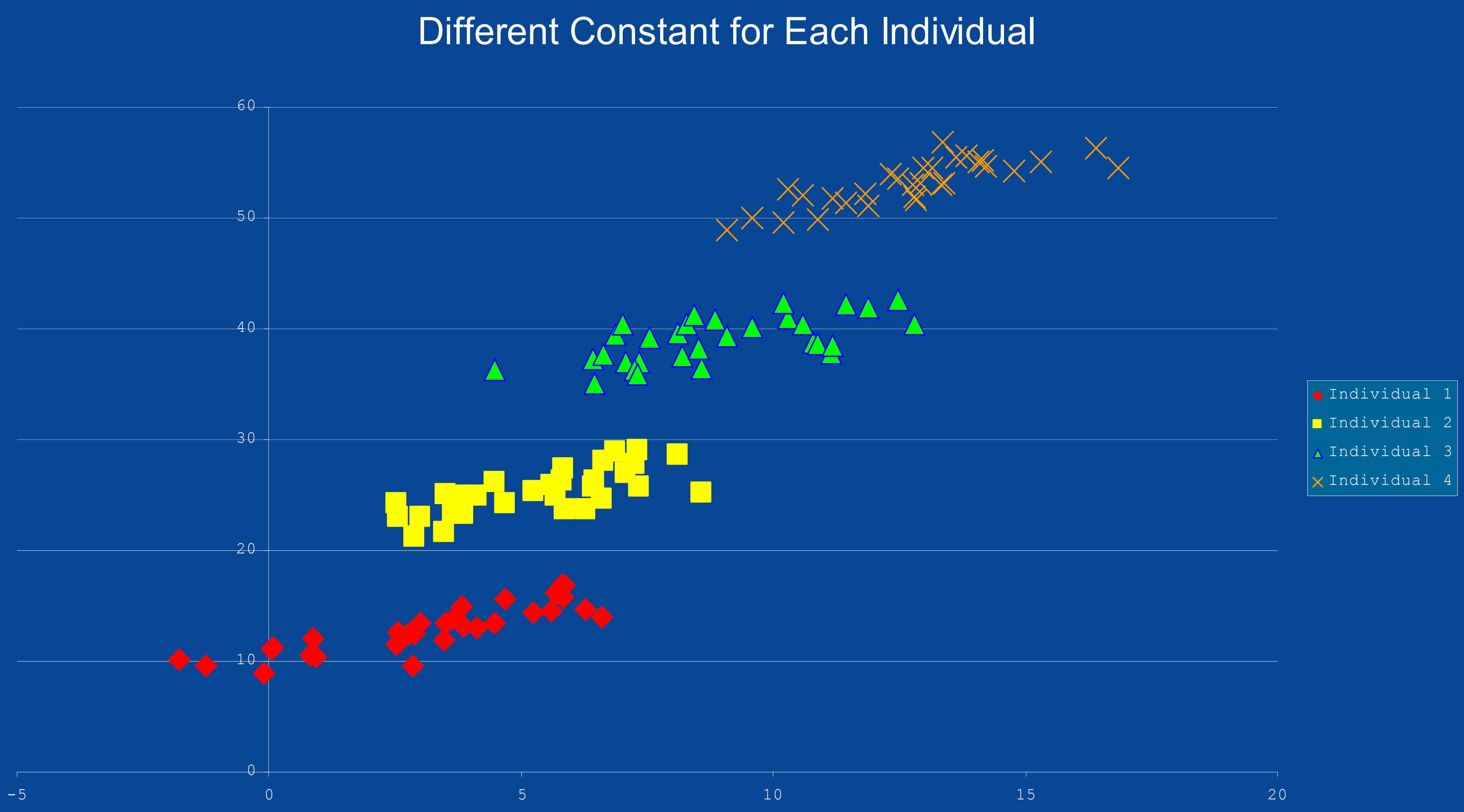
Treat λ_i as a **constant** for each individual

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

λ now part of constant – but varies by individual

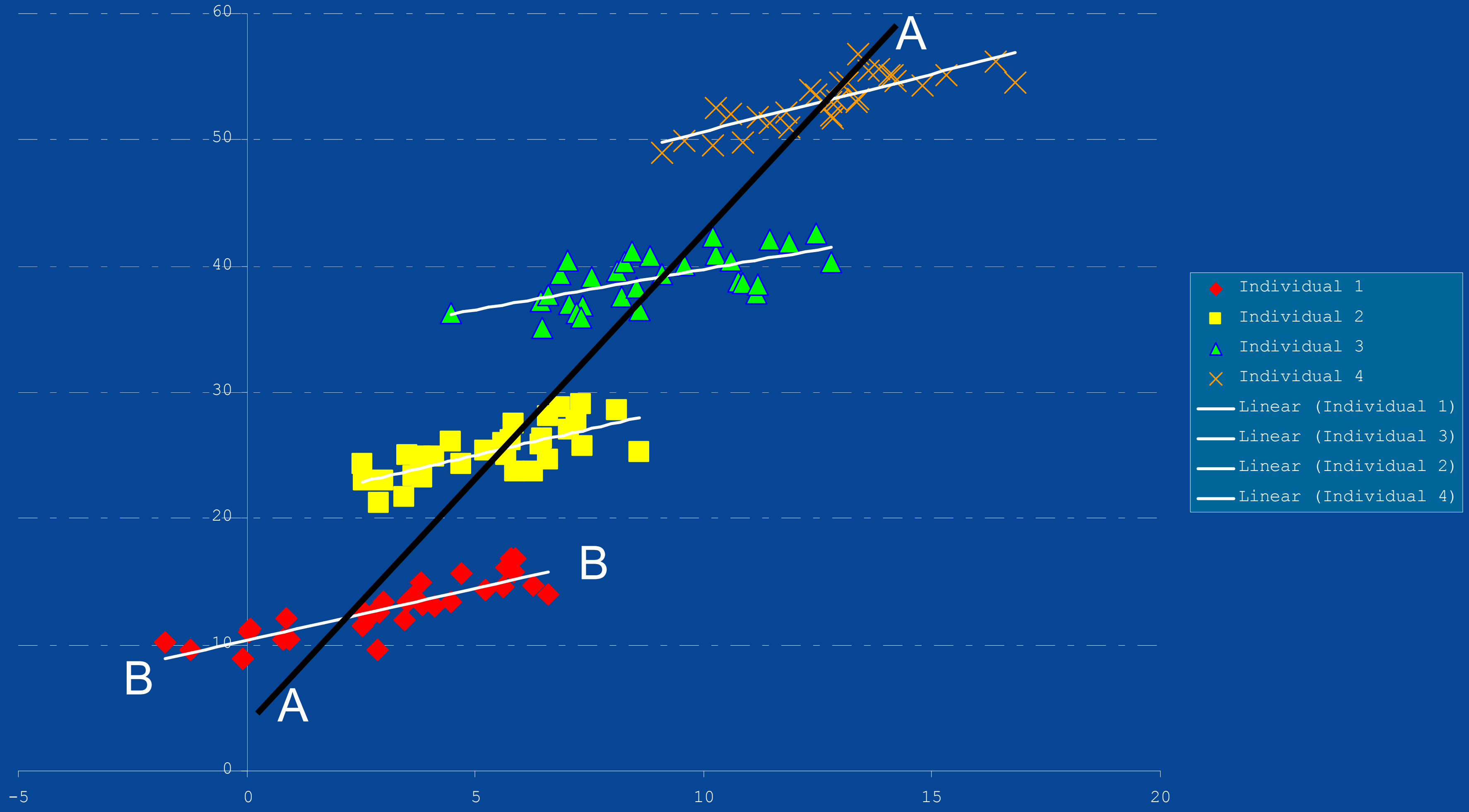
UNIVERSITY OF STIRLING

Graphically might look like:



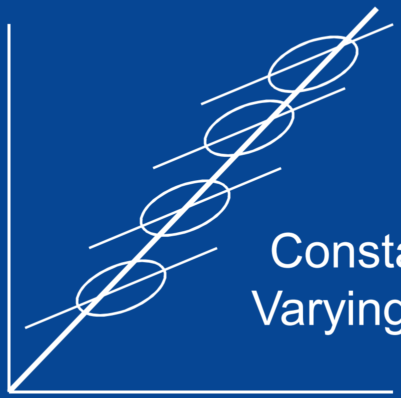
UNIVERSITY OF STIRLING

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



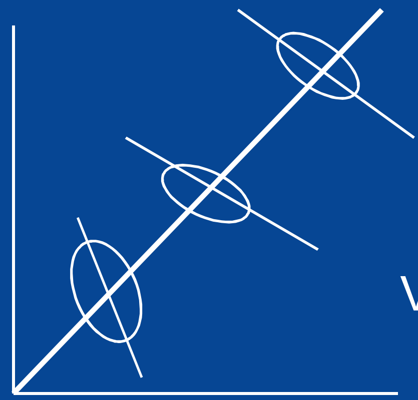
UNIVERSITY OF STIRLING

Possible Combinations of Slopes and Intercepts



The fixed effects model

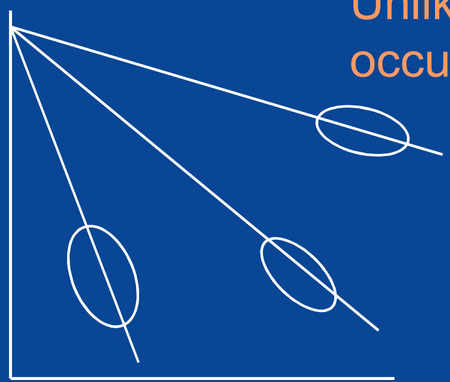
Constant slopes
Varying intercepts



Separate regression for each individual

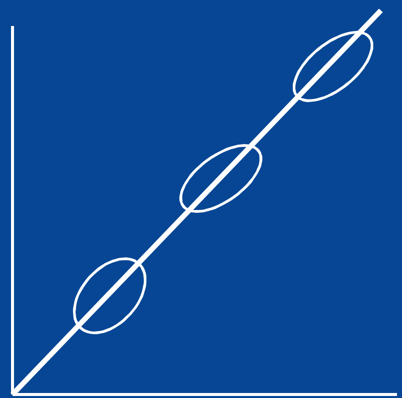
Varying slopes
Varying intercepts

Can test down to the most parsimonious model



Unlikely to occur

Varying slopes
Constant intercept



Pooled model

Constant slopes
Constant intercept

UNIVERSITY OF STIRLING

Eliminate unobserved heterogeneity by taking first differences

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

Lag one period and subtract

Original equation

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

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

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

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 to the above equation gives the first-difference estimator. But there is another way of "sweeping out" fixed effects when there are more than two time periods - taking deviations from individual means. Let $x_{1i.}$ be the mean for variable x_1 for individual i, averaged across all time periods. Calculating this transformation for each variable (including y) and subtracting the means from our initial equation gives

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

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. This gives the fixed effects, LSDV or the “within” estimator – which is consistent

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

It is called the “within” estimator because it relies on variations within individuals rather than between individuals. There is another estimator that uses 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

$$y_{it} - \bar{y}_{i.} = \beta'\left(x_{it} - \bar{x}_{i.}\right) + u_{it} - \bar{u}_{i.}$$ 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** to ensure efficiency.

UNIVERSITY OF STIRLING