Panel Data Models

1. Characteristic of Data and Problems
2. Model with Heteroskedasticity, Autocorrelation and Cross-sectional Correlation
3. Fixed Effect Models
4. Random Effect Models

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \varepsilon_{it}$$

or $$\mathrm{y}_{it} = \mathrm{x}_{it}\beta + \varepsilon_{it}$$

Characteristic of Data and Problems

Advantage: Number of observation $N = nT$

Problems that might occur:

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

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

Pooled Ordinary Least Squared (POLS):

$$\underset{k\times 1}{\hat{\beta}} = (\underset{k\times nT}{\mathrm{x}'}\ \underset{nT\times k}{\mathrm{x}})^{-1}\ \underset{k\times nT}{\mathrm{x}'}\ \underset{nT\times 1}{\mathrm{y}}$$

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

Generalize Least Squared (GLS):

$$\underset{k\times 1}{\hat{\beta}} = \left(\underset{k\times nT}{\mathrm{x}'} \underset{nT\times nT}{\hat{V}^{-1}} \underset{nT\times k}{\mathrm{x}} \right)^{-1} \underset{k\times nT}{\mathrm{x}'} \underset{nT\times nT}{\hat{V}^{-1}} \underset{nT\times 1}{\mathrm{y}}$$

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

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.

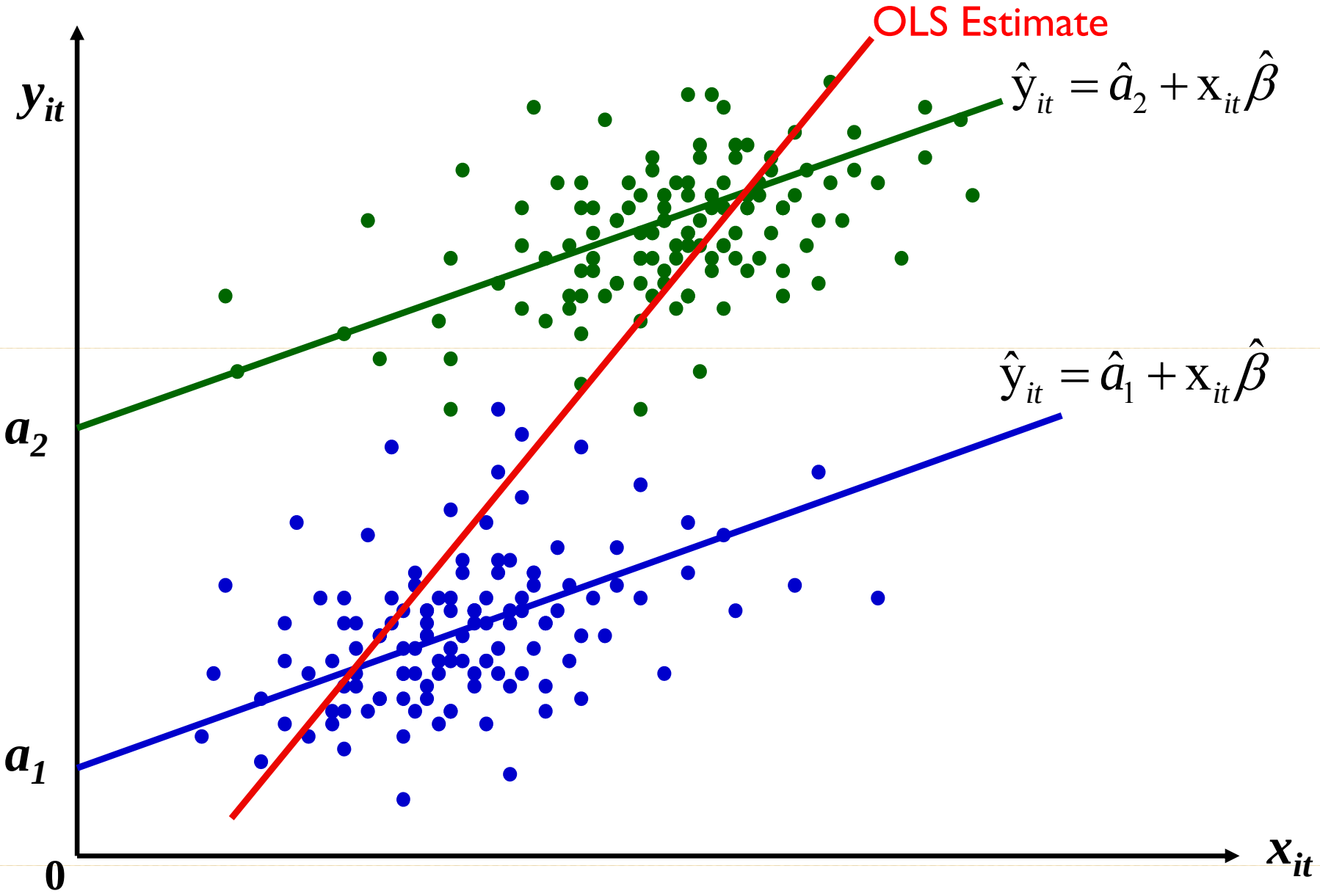
Otherwise, the data should be considered as a pooled cross-section – e.g. SES data.

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.

Bias from Ignoring Fixed Effects



Difference-in-Differences

In medical experiment, researchers conduct their researches by control groups for their random assignment of the treatments.

Then, they can simply compare the change in outcomes across the treatments and control groups to estimate the treatment effect.

For time 1, 2, groups A, B

$$(y_{2,B} - y_{2,A}) - (y_{1,B} - y_{1,A})$$

or equivalently $(y_{2,B} - y_{1,B}) - (y_{2,A} - y_{1,A})$

This analysis is called the difference-in-differences.

Difference-in-Differences

A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well.

Consider the model:

$$y_{it} = \beta_0 + \beta_1 treatment_{it} + \beta_2 after_{it} + \beta_3 treatment_{it} \times after_{it} + \varepsilon_{it}$$

The estimated β_3 represents the difference-in-differences in the group means.

Difference-in-Differences

When we do not truly have random assignment, the regression form becomes very useful.

Additional x's can be added to the regression to control for differences across the treatment and control groups.

This is referred to as a "natural experiment" especially when a policy change is being analyzed.

Two-Period Panel Data

It is possible to use a panel just like pooled cross-sections, but we can do more than that.

Panel data can be used to address some kinds of omitted variable biased.

If we treat the omitted variables as being fixed over time, then we can set up the model as having a composite error.

Unobserved Fixed Effects

Suppose the population model is

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

Here we have added a time-constant component to the error:

$$\upsilon_{it} = a_i + \varepsilon_{it}$$

If a_i is correlated with the x's, OLS will be biased, since a_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 effect

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

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

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

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 effect, an alternative to the first differences is fixed effects estimation

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

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

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

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

$$H_0: \; a_1 = a_2 = \ldots = a_n = a$$

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

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

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

F-statistic = $\dfrac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}$

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

Random Effects

Start with the same basic model with a composite error

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

In FE model, we assumed that a_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.

$$y_{it} - \hat{\lambda}\bar{y}_i = \beta_0\left(1-\hat{\lambda}\right) + \beta_1\left(x_{it1} - \hat{\lambda}\bar{x}_{i1}\right) + \ldots + \beta_k\left(x_{itk} - \hat{\lambda}\bar{x}_{ik}\right) + \left(v_{it} - \hat{\lambda}\bar{v}_i\right)$$

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

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

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.

Other Uses of Panel Methods

It is possible to think of models where there is an unobserved fixed effect, even if we do not have true panel data.

A common example is where we think there is an unobserved family effect.

We can apply difference-in-difference to solve problem and estimate family fixed effect model.

Poolability Test

Unrestricted: $y_i = x_i \beta_i + \varepsilon_i$

Restricted: $y = x_i \beta + \varepsilon$

Assumption 1: $\varepsilon \sim N(0, \sigma^2 I_{NT})$

Chow Test can be applied

$$F = \frac{(e'e - e_1'e_1 - e_2'e_2 - \ldots - e_N'e_N)/(N-1)K'}{(e_1'e_1 + e_2'e_2 + \ldots + e_N'e_N)/N(T-K')}$$

Assumption 2: $\varepsilon \sim N(0, \sigma^2 \Omega)$

Chow test is not appropriated.

Generalized Chow test should be performed.

Additional Issues

- Nonlinear Panel Data Model
- Dynamic Panel Data Model
- Simultaneous Equation Panel Data Model
- Panel Unit Root Test
- Panel Cointegration Test