Econometrics **August 18, 2026**

Panel Data Techniques

Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

Contents

1 Overview of Panel Data **2**

2 Pooled OLS **2**

2.1 Limitations of Pooled OLS 3

3 Fixed Effects Model **3**

3.1 FE - Least Squares Dummy Variable Model 4

3.2 FE - Within Group Estimator 4

3.3 First Difference Method (FD) 5

4 Random Effects Model **5**

4.1 Estimation of REM 6

4.2 Advantages and Disadvantages of REM 7

4.3 Hypothesis Testing 7

5 Choosing between Fixed Effects and Random Effects **7**

5.1 The Hausman Test 8

5.2 Practical Considerations 9

6 Panel Data Techniques in Practice **9**

7 Panel Data Regression in R **11**

1 Overview of Panel Data

The panel data structure is based on the number of time periods (T) and entities (N) in the dataset.

Table 1: Comparison of Short and Long Panels

Feature	Short Panels	Long Panels
Number of entities (N)	Large	Small
Number of time periods (T)	Small	Large
Relationship between N and T	$N > T$	$T > N$
Common use	Microeconomic studies	Macroeconomic studies

2 Pooled OLS

Pooled Ordinary Least Squares (OLS) treats panel data as a single cross-section, ignoring time and entity dimensions. For example, consider wages and education:

$$\text{Wage}_{it} = \beta_0 + \beta_1(\text{Education}_{it}) + \beta_2(\text{Work Experience}_{it}) + \varepsilon_{it} \tag{1}$$

where i denotes individual/entity and t time. Pooled OLS stacks all observations and estimates a standard OLS regression. The estimator is:

$\hat{\beta} = (X'X)^{-1}X'y$

where X is the stacked matrix of independent variables (including a constant), and y is the dependent variable.
Further $\beta_0, \beta_1, \beta_2$ are constant and same for all cross-sectional units across all times. There is no distinction between cross-sectional units, i.e., Individual 1 & Individual 2 are the same.

Mathematically, $\varepsilon_{it} \sim$ i.i.d $N(0, \sigma_\varepsilon^2)$

Thus, Pooled OLS is appropriate when:

1. No individual-specific effects (e.g., unobserved ability uncorrelated with education).
2. No time-specific effects (e.g., macroeconomic conditions affect all equally).
3. Homoscedasticity and no autocorrelation.

Usefulness of Pooled OLS: Can serve as a baseline or be valid if the restrictive assumptions hold. It provides a simple starting point for panel data analysis.

2.1 Limitations of Pooled OLS

Issue with Pooled OLS

In Pooled OLS, we lump individual/entity level (possible) heterogeneity into the error. This could mean ε_{it} is correlated with individual observed characteristics. $\text{Corr}(X_{it}, \varepsilon_{it}) \neq 0$, thus making $\hat{\beta}$ biased and inconsistent.

In our wage example, this means that factors like *innate ability* or *motivation*, which might influence wages, are very plausibly related to education or experience. Highly motivated students gain education and experience as well as high wages. This challenges simple pooled OLS estimates. Similarly, for instance, if more able individuals tend to acquire more education, the estimated effect of education on wages will be biased.

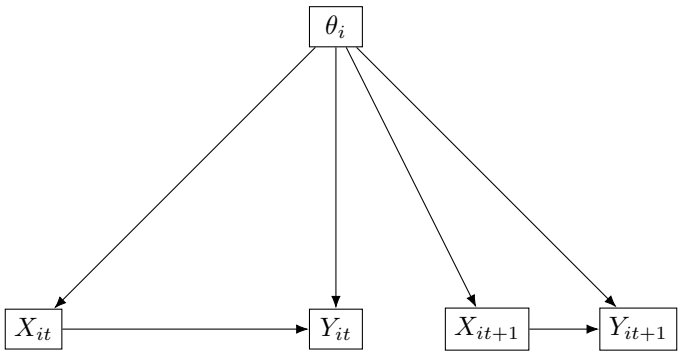
3 Fixed Effects Model

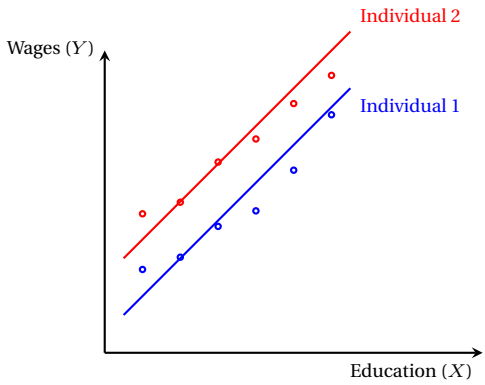
Fixed effects are particularly useful when we suspect that there are unobserved characteristics specific to each unit θ_i that might affect the dependent variable Y_{it} and are correlated with the independent variables X_{it}. These unobserved characteristics are constant over time for each unit.

$$\text{Wage}_{it} = \beta_0 + \beta_1(\text{Education}_{it}) + \beta_2(\text{Work Experience}_{it}) + \theta_i + \varepsilon_{it} \tag{2}$$

$$\text{Wage}_{it} = \beta_{0i} + \beta_1(\text{Education}_{it}) + \beta_2(\text{Work Experience}_{it}) + \varepsilon_{it} \tag{3}$$

One-way fixed effect: β_{0i}; the intercept varies for every individual/entity, but does not vary over time. If it is allowed to vary by both entity and time, it is called *two-way fixed effect*, i.e. β_{0it}.





3.1 FE - Least Squares Dummy Variable Model

How do we get the different intercepts for different individuals/entities?
We employ the least square dummy variable (*LSDV*) technique

$$\text{Wage}_{it} = \alpha_1 + \alpha_2(D_{2i}) + \alpha_2(D_{3i}) + \beta_1(\text{Education}_{it}) + \beta_2(\text{Work Experience}_{it}) + \theta_i + \varepsilon_{it} \tag{4}$$

Here the baseline intercept α_1 is the estimate for the first individual, $\alpha_1 + \alpha_2$ is the estimate for intercept for second individual; $\alpha_1 + \alpha_3$ is the estimate for intercept for the third individual.

Issues with LSDV Model

- If more dummy variables are added, it causes loss of *d.o.f.*
- LSDV does not identify the impact of time-invariant variables. All time-invariant variables like sex, ethnicity are absorbed into entity-specific intercepts.
- All heterogeneity that exists in the dependent and explanatory variables is absorbed by the entity-specific intercept.

3.2 FE - Within Group Estimator

All dependent and explanatory variables are expressed as *deviations from within a group mean.* The time-invariant characteristics are all wiped away.

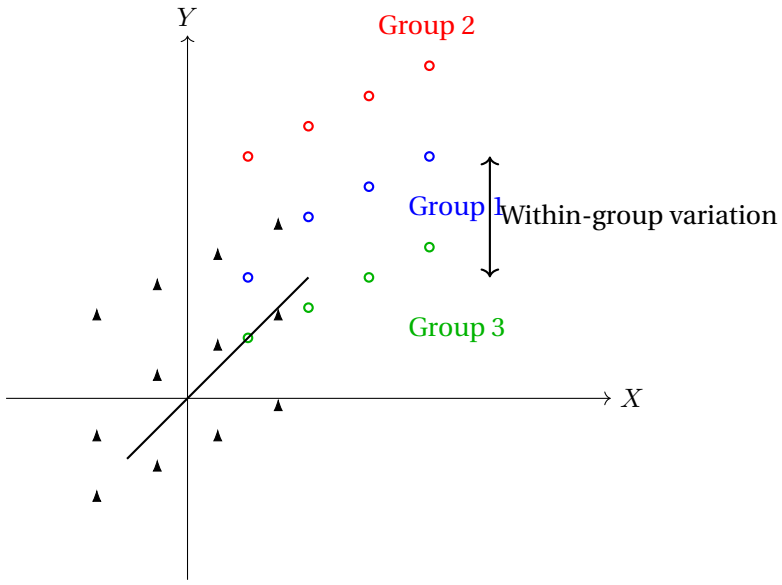
$$x_{it} = X_{it} - \overline{X_i(.)}$$

The within-group[1] FE estimation model is:

$$\text{y}_{it} = \beta_1(\text{x}_{1it}) + \beta_2(\text{x}_{2it}) + \varepsilon_{it} \tag{5}$$

[1]Note1: The within-group variables are lowercase. Note2: No intercept here.

Figure 1: Illustration of Fixed Effects Estimation



Within Group estimators are *consistent*, but *inefficient*. Because de-meaning the data removes a lot of variation, this variation is transferred to the error term, leading to greater $\mathrm{Var}(\widehat{\beta})$.

The firm-specific intercepts (indexed by i), are obtained by

$$\hat{\alpha}_i = \overline{Y_i} - \beta_1(\overline{X_{1i}}) - \beta_2(\overline{X_{2i}})$$

$\overline{Y_i}, \overline{X_{1i}}$ are entity-specific averages.

3.3 First Difference Method (FD)

An alternative to the WG-estimator

$$\Delta(X_{it}) = X_{it} - X_{it-1}$$

If explanatory variables are strictly exogenous, the first-difference estimator is unbiased.

4 Random Effects Model

The Random Effects Model (REM), also known as the Error Components Model (ECM), is a statistical model used for panel data analysis where the individual-specific effects are assumed to be random vari-

ables. It's a generalization of the standard linear regression model that accounts for unobserved heterogeneity across individuals or groups. This model is particularly useful when the number of individuals (N) is large and the number of time periods (T) is relatively small.

Model Specification

The REM can be represented as:

$$Y_{it} = \mu + X'_{it}\beta + \underbrace{\theta_i + \varepsilon_{it}}_{\nu_{it}} \tag{6}$$

where:

- Y_{it} is the dependent variable for individual i at time t.
- μ is the overall intercept.
- X_{it} is a vector of explanatory variables for individual i at time t.
- β is a vector of coefficients to be estimated.
- θ_i is the individual-specific random effect, assumed to be independently and identically distributed (i.i.d.) with mean 0 and variance σ^2_θ. This is the key difference from the fixed effects model.
- ε_{it} is the idiosyncratic error term, also assumed to be i.i.d. with mean 0 and variance σ^2_ε. It represents the error variation within each individual over time.

Assumptions:

- θ_i and ε_{it} are *independent* of each other and of the explanatory variables X_{it}.
- The errors are normally distributed: $\theta_i \sim N(0, \sigma^2_\theta)$ and $\varepsilon_{it} \sim N(0, \sigma^2_\varepsilon)$.

4.1 Estimation of REM

The parameters of the REM are typically estimated using Generalized Least Squares (GLS). The variance-covariance matrix of the error term is not diagonal due to the presence of the individual-specific effect θ_i. GLS takes this into account to produce efficient estimates.

The composite error term is $\nu_{it} = \theta_i + \varepsilon_{it}$. The variance of this composite error is:

$$\text{Var}(\nu_{it}) = \sigma^2_\theta + \sigma^2_\varepsilon \tag{7}$$

And the covariance between errors for the same individual across different time periods is:

$$\text{Cov}(\nu_{it}, \nu_{is}) = \sigma_\theta^2, \quad t \neq s \tag{8}$$

The correlation within an individual's observations is:

$$\rho = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2} \tag{9}$$

4.2 Advantages and Disadvantages of REM

Advantages:

- More efficient estimates than OLS if the assumptions are met.
- Can include time-invariant explanatory variables.
- Accounts for unobserved heterogeneity across individuals.

Disadvantages:

- Inconsistent estimates if the individual effects are correlated with the explanatory variables (endogeneity).
- Requires strong distributional assumptions (normality).

4.3 Hypothesis Testing

We can test hypotheses about the coefficients β using standard t-tests or F-tests. We can also test the significance of the individual-specific variance component σ_θ^2 using a Lagrange Multiplier (LM) test.

5 Choosing between Fixed Effects and Random Effects

The Hausman test is a crucial tool in panel data analysis for deciding between the Fixed Effects (FE) model and the Random Effects (RE) model. It helps determine whether the unobserved individual effects are correlated with the explanatory variables. This correlation is the key distinction between the two models.

The choice between Fixed Effects (FE) and Random Effects depends on whether the individual effects are correlated with the explanatory variables. If they are correlated, FE is preferred. If they are uncorrelated, RE is more efficient. The *Hausman test* can be used to formally test for this correlation. A significant Hausman test suggests using FE.

The Problem: Endogeneity

The Random Effects model assumes that the individual effects (θ_i) are uncorrelated with the explanatory variables (X_{it}). If this assumption is violated (i.e., there's correlation), the RE estimates become inconsistent. The Fixed Effects model, on the other hand, allows for correlation between θ_i and X_{it} and produces consistent estimates in such cases. However, FE cannot estimate coefficients for time-invariant variables.

5.1 The Hausman Test

The Hausman test provides a formal way to test for this correlation. It compares the estimates from the consistent FE model ($\hat{\beta}_{FE}$) with the efficient (under the null) RE model ($\hat{\beta}_{RE}$).

Null Hypothesis (H_0): The individual effects are uncorrelated with the explanatory variables. RE model is consistent and efficient. ($\text{Cov}(\theta_i, X_{it}) = 0$)

Alternative Hypothesis (H_1): The individual effects are correlated with the explanatory variables. FE model is consistent; RE model is inconsistent. ($\text{Cov}(\theta_i, X_{it}) \neq 0$)

Hausman Test Statistic:

The Hausman test statistic is based on the difference between the FE and RE estimates. It's asymptotically distributed as a chi-squared distribution:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2(k) \tag{10}$$

where:

- k is the number of parameters in the model (excluding the intercept).
- $Var(\hat{\beta}_{FE})$ and $Var(\hat{\beta}_{RE})$ are the variance-covariance matrices of the FE and RE estimators, respectively. These are typically obtained from the software package used for estimation.

Decision Rule

- If the calculated Hausman statistic (H) is greater than the critical value from the chi-squared distribution with k degrees of freedom at a chosen significance level (e.g., 5%), we reject the null hypothesis.
- If we reject H_0, it suggests that the individual effects are correlated with the explanatory variables, and the Fixed Effects model is preferred.
- If we fail to reject H_0, it suggests that the individual effects are uncorrelated with the explanatory variables, and the Random Effects model is more efficient.

5.2 Practical Considerations

- The Hausman test relies on the assumptions of the RE model being met under the null hypothesis. If the RE model is misspecified for other reasons, the Hausman test may be invalid.
- Software packages (like Stata, R, etc.) provide built-in commands to perform the Hausman test, making it easy to implement in practice.
- The Hausman test is not always decisive. Sometimes the results can be borderline, or the test may lack power. It's important to consider other factors and the context of the research question when making a final decision.

6 Panel Data Techniques in Practice

Crime4 Dataset Description

The **crime4** dataset from Wooldridge contains panel data on crime rates across different counties in the United States. It includes economic and law enforcement variables such as arrest probabilities, conviction probabilities, and policing levels. The dataset spans multiple years, allowing for time-series and panel data analysis of crime determinants. Researchers use this dataset to examine the impact of deterrence measures on crime rates.

- `county`: County identifier
- `year`: Year (81 to 87)
- `crmrte`: Crimes committed per person
- `prbarr`: 'Probability' of arrest
- `prbconv`: 'Probability' of conviction
- `prbpris`: 'Probability' of prison sentence
- `avgsen`: Average sentence, days
- `polpc`: Police per capita
- `density`: People per sq. mile
- `taxpc`: Tax revenue per capita
- `west`: =1 if in western N.C.
- `central`: =1 if in central N.C.
- `urban`: =1 if in SMSA
- `pctmin80`: Percent minority, 1980
- `wcon`: Weekly wage, construction
- `wtuc`: Weekly wage, transportation, utilities, communication
- `wtrd`: Weekly wage, wholesale, retail trade
- `wfir`: Weekly wage, finance, insurance, real estate
- `wser`: Weekly wage, service industry
- `wmfg`: Weekly wage, manufacturing
- `wfed`: Weekly wage, federal employees
- `wsta`: Weekly wage, state employees
- `wloc`: Weekly wage, local government employees
- `mix`: Offense mix: face-to-face/other

- `pctymle`: Percent young male
- `d82`: =1 if year == 82
- `d83`: =1 if year == 83
- `d84`: =1 if year == 84
- `d85`: =1 if year == 85
- `d86`: =1 if year == 86
- `d87`: =1 if year == 87
- `lcrmrte`: log(crmrte)
- `lprbarr`: log(prbarr)
- `lprbconv`: log(prbconv)
- `lprbpris`: log(prbpris)
- `lavgsen`: log(avgsen)
- `lpolpc`: log(polpc)
- `ldensity`: log(density)
- `ltaxpc`: log(taxpc)
- `lwcon`: log(wcon)
- `lwtuc`: log(wtuc)
- `lwtrd`: log(wtrd)
- `lwfir`: log(wfir)
- `lwser`: log(wser)
- `lwmfg`: log(wmfg)
- `lwfed`: log(wfed)
- `lwsta`: log(wsta)
- `lwloc`: log(wloc)
- `lmix`: log(mix)
- `lpctymle`: log(pctymle)
- `lpctmin`: log(pctmin)
- `clcrmrte`: lcrmrte - lcrmrte[_n-1]
- `clprbarr`: lprbarr - lprbarr[_n-1]
- `clprbcon`: lprbconv - lprbconv[_n-1]
- `clprbpri`: lprbpri - lprbpri[t-1]
- `clavgsen`: lavgsen - lavgsen[t-1]
- `clpolpc`: lpolpc - lpolpc[t-1]
- `cltaxpc`: ltaxpc - ltaxpc[t-1]
- `clmix`: lmix - lmix[t-1]

7 Panel Data Regression in R

```r
    ```{r paneldata}
library(wooldridge) # for wooldridge intro ecotrix datasets
library(plm)        # For panel data models
library(modelsummary) # for aggregating various model results


# Load a panel dataset from Wooldridge
data('crime4') # Example panel dataset on crime rates
# View first few rows
head(crime4)
# Check structure of the dataset
str(crime4)
# Summarize key variables
summary(crime4)

# Define the panel data structure
crime_panel <- pdata.frame(crime4, index = c("county", "year")) # Panel data format

# Run a Pooled OLS regression (Baseline)
pooled_ols <- lm(crmrte ~ prbarr + prbconv + prbpris + polpc, data = crime_panel)

# Fixed Effects Model
fe_model <- plm(crmrte ~ prbarr + prbconv + prbpris + polpc,
                data = crime_panel, model = "within")

# Random Effects Model
re_model <- plm(crmrte ~ prbarr + prbconv + prbpris + polpc,
                data = crime_panel, model = "random")

# Display results
models <- list("Pooled OLS" = pooled_ols, "Fixed Effects" = fe_model, "Random Effects" = re_model)
modelsummary(models, stars = TRUE, gof_omit = "IC|Log")

# Perform Hausman Test to choose between FE and RE models
phtest(fe_model, re_model)

```
```

References

The End