Panel Data Fundamentals

Introduction to Longitudinal Data Analysis with R

Dr Clemens Jarnach

University of Oxford

2025–26

What is Panel Data?

- A dataset that tracks the **same units** (individuals, firms, countries) **over time**
- Also called **longitudinal data** or **cross-sectional time-series** data
- Combines two dimensions:
  - **Cross-sectional**: variation across units i
  - **Time-series**: variation over time t

Note

Each observation is indexed by both a unit i and a time period t: (y_{it}, x_{it})

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**
  - More **efficient estimation** (greater variability)

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**
  - More **efficient estimation** (greater variability)
  - Analysis of **dynamic processes**

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**
  - More **efficient estimation** (greater variability)
  - Analysis of **dynamic processes**
  - Identification of effects **not visible in cross-sections**

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**
  - More **efficient estimation** (greater variability)
  - Analysis of **dynamic processes**
  - Identification of effects **not visible in cross-sections**
  - Use of **micro-level data** (better measurement)

Why Panel Data? Key Advantages

- Combines **cross-sectional + time-series variation**
- Enables:
  - Control for **individual heterogeneity**
  - More **efficient estimation** (greater variability)
  - Analysis of **dynamic processes**
  - Identification of effects **not visible in cross-sections**
  - Use of **micro-level data** (better measurement)
  - Cross-dimensional inference (time $\leftrightarrow$ individuals)

The Core Problem: Unobserved Heterogeneity

Consider the model:

$$y_{it} = \alpha + \beta x_{it} + \gamma z_i + u_{it}$$

- z_i is an **unobserved**, time-invariant characteristic
- If z_i is correlated with $x_{it} \Rightarrow$ **omitted variable bias**
- OLS estimates of β are **inconsistent**

Warning

Key condition for consistency: z_i must be uncorrelated with regressors and the error term.

Intuition: Agricultural Example

Suppose farm output depends on:

Variable	**Status**
Labour x_{it}	Observed
Soil quality z_i	Unobserved

The problem:

- Farmers likely adjust labour based on their soil quality
- x_{it} is correlated with z_i
- $\Rightarrow$ **OLS becomes inconsistent**

The General Panel Data Model

Absorb unobserved heterogeneity into an individual-specific intercept:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}$$

- α_i captures all **time-invariant unobserved factors** for unit i
- The goal: **remove or control for** α_i

The General Panel Data Model

Absorb unobserved heterogeneity into an individual-specific intercept:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}$$

- α_i captures all **time-invariant unobserved factors** for unit i
- The goal: **remove or control for** α_i

This is the starting point for both **Fixed Effects** and **Random Effects** models.

Three Strategies for Eliminating α_i

Strategy 1: First-Difference (FD) Estimator

Idea: Subtract last period's equation from the current period:

$$\Delta y_{it} = \beta \, \Delta x_{it} + \Delta u_{it}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$

Advantages

- Eliminates α_i completely
- Estimable via pooled OLS

Disadvantages

- Loses first time period
- Can amplify measurement error

Strategy 2: Least Squares Dummy Variables (LSDV)

Approach: Include a dummy variable d_i for each unit:

$$y_{it} = \sum_i \alpha_i d_i + \beta x_{it} + u_{it}$$

Strategy 2: Least Squares Dummy Variables (LSDV)

Approach: Include a dummy variable d_i for each unit:

$$y_{it} = \sum_i \alpha_i d_i + \beta x_{it} + u_{it}$$

- Estimates a **separate intercept** α_i for each unit

Strategy 2: Least Squares Dummy Variables (LSDV)

Approach: Include a dummy variable d_i for each unit:

$$y_{it} = \sum_i \alpha_i d_i + \beta x_{it} + u_{it}$$

- Estimates a **separate intercept** α_i for each unit
- $\hat{\beta}$ is $\sqrt{NT}$-consistent

Strategy 2: Least Squares Dummy Variables (LSDV)

Approach: Include a dummy variable d_i for each unit:

$$y_{it} = \sum_i \alpha_i d_i + \beta x_{it} + u_{it}$$

- Estimates a **separate intercept** α_i for each unit
- $\hat{\beta}$ is $\sqrt{NT}$-consistent
- Individual effects estimated with $\sqrt{T}$-consistency

Strategy 2: Least Squares Dummy Variables (LSDV)

Approach: Include a dummy variable d_i for each unit:

$$y_{it} = \sum_i \alpha_i d_i + \beta x_{it} + u_{it}$$

- Estimates a **separate intercept** α_i for each unit
- $\hat{\beta}$ is $\sqrt{NT}$-consistent
- Individual effects estimated with $\sqrt{T}$-consistency
- **Drawback:** N additional parameters $\rightarrow$ loss of degrees of freedom; computationally costly for large N

Strategy 3: Fixed Effects (Within) Estimator

Idea: Remove α_i by **demeaning** each variable over time:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \qquad \tilde{x}_{it} = x_{it} - \bar{x}_i$$

This yields the **within-transformed** model:

$$\tilde{y}_{it} = \beta\, \tilde{x}_{it} + \tilde{u}_{it}$$

Tip

α_i cancels out because $\bar{\alpha}_i = \alpha_i$. Only **within-unit variation over time** identifies β.

Fixed Effects: Properties and Limitations

Advantages

- Eliminates α_i without dummy variables
- Computationally efficient
- Consistent even when α_i is correlated with x_{it}
- Standard approach in applied work

Limitations

- Cannot estimate effects of **time-invariant** variables (e.g. gender, country)
- Relies solely on **within-unit** variation
- Less efficient when α_i is uncorrelated with x_{it}

FD vs LSDV vs Within Estimator

Method	**Removes α_i?**	**Parameters**	**Notes**
First-difference	Yes	None added	Loses one period
LSDV	Yes	N dummies	Costly for large N
Within (FE)	Yes	None added	Standard choice

FD vs LSDV vs Within Estimator

Method	**Removes α_i?**	**Parameters**	**Notes**
First-difference	Yes	None added	Loses one period
LSDV	Yes	N dummies	Costly for large N
Within (FE)	Yes	None added	Standard choice

All three are **algebraically equivalent** under strict exogeneity — the within estimator is preferred for computational efficiency.

Key Takeaways

1. Panel data solves **unobserved heterogeneity** by exploiting **within-unit variation over time**
2. The core challenge: unobserved α_i correlated with regressors $\Rightarrow$ OLS bias
3. Three equivalent strategies: **first-differencing**, **LSDV**, **within (demeaning)**
4. **Fixed Effects = most widely used practical method**

Note

Next: when should we use **Random Effects** instead, and how do we choose?