# Advanced Econometrics
## Lecture 10: Models for Panel Data

Eduard Brüll

Fall 2025

Lecture 10

# Advanced Econometrics



**Literature:** Wooldridge Ch. 9–10; Baltagi (2021); Deaton (1985)

# 10.1: Introduction and Motivation

# What is Panel Data?

## Definition

Panel data (also called **longitudinal data**) consist of **observations on multiple entities (individuals, firms, regions)** over **multiple time periods**.

- Combines features of both:
  - **Cross-sectional data**: different entities, one point in time
  - **Time-series data**: one entity, many time periods
- Typical examples:
  - Wage histories of workers observed yearly
  - Firm-level production and investment over time
  - Regional unemployment rates across years

# Example: Typical Structure of Panel Data

| Respondent ID | Year | Gender | Gross Income (EUR) | Net Income (EUR) |
|---|---|---|---|---|
| 101 | 2016 | Male | 38,500 | 26,400 |
| 101 | 2017 | Male | 39,200 | 27,100 |
| 101 | 2018 | Male | 40,000 | 27,800 |
| 101 | 2019 | Male | 41,500 | 28,900 |
| 101 | 2020 | Male | – | – |
| 102 | 2017 | Female | 32,000 | 22,300 |
| 102 | 2018 | Female | 33,400 | 23,200 |
| 102 | 2019 | Female | 35,100 | 24,400 |
| 102 | 2020 | Female | 35,900 | 25,000 |
| 102 | 2021 | Female | 36,200 | 25,400 |
| 103 | 2016 | Male | 45,600 | 30,800 |
| 103 | 2017 | Male | 47,000 | 32,000 |
| 103 | 2018 | Male | – | – |
| 103 | 2019 | Male | – | – |
| 103 | 2020 | Male | 46,200 | 31,600 |
| 103 | 2021 | Male | 47,800 | 32,800 |

- Each row is one **observation**: respondent-year combination.
- You can have time-varying and time-fixed variables
- Missings depend on attrition behaviour of respondents

# Advantages and Disadvantages of Panel Data

| Advantages | Disadvantages |
| --- | --- |
| + Control for unobserved, time-invariant heterogeneity | – Attrition and missing observations |
| + Allows to capture **dynamics** and lagged effects | – Complex estimation and potential **serial correlation** |
| + More efficient estimation (more variation, less collinearity) | – **Costly** and time-consuming data collection (if based on surveys) |
| + Enable causal inference using within-entity variation | – Measurement error may be amplified by differencing |

# General Model Setup

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

- $i = 1, \ldots, N$ individuals, $t = 1, \ldots, T$ time periods
- $\mathbf{x}_{it}$: observed explanatory variables
- $c_i$: unobserved, time-invariant individual effect
- $\varepsilon_{it}$: idiosyncratic error term

## Key Question

Is $c_i$ **correlated** with $\mathbf{x}_{it}$?

- If no $\rightarrow$ Random Effects
- If yes $\rightarrow$ Fixed Effects

# 10.2: Pooled Regression and Error Dependence

# The Pooled OLS Model

## The True Model

$$y_{it} = \mathbf{x}_{it}'\beta + c_i + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

- $c_i$: unobserved, time-invariant component (individual effect)
- $\varepsilon_{it}$: idiosyncratic error
- The **pooled model** ignores the panel structure and estimates

$$y_{it} = \mathbf{x}_{it}'\beta + u_{it}, \quad u_{it} = c_i + \varepsilon_{it}$$

**Intuition:** All observations are treated as one large cross-section.

# Assumptions for Pooled OLS

**The Pooled OLS Model:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad u_{it} = c_i + \varepsilon_{it}.$$

**Required Assumptions for Consistency:**

POLS1: **Linearity** The model is linear in parameters:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it}.$$

POLS2: **Identifiability** The regressor matrix $\boldsymbol{X}$ has full column rank $(K+1)$, ensuring $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists.

POLS3: **Strict Exogeneity**

$$\mathbf{E}[u_{it} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}] = 0$$

which requires in particular that

$$\mathbf{E}[c_i \mid \boldsymbol{x}_{it}] = 0 \qquad \text{and} \qquad \mathbf{E}[\varepsilon_{it} \mid \boldsymbol{x}_{it}] = 0.$$

*This is the critical assumption: if $c_i$ is correlated with $\boldsymbol{x}_{it}$, pooled OLS is inconsistent.*

POLS4: **Independent Sampling Across Individuals** The $\{(\boldsymbol{x}_{it}, y_{it})\}$ are independent across $i$, while serial correlation within $i$ is allowed.

POLS5: **Finite Moments** $\mathrm{var}(\boldsymbol{x}_{it}'u_{it}) < \infty$ for all $i, t$.

# Conditional Expectation of $\widehat{\beta}_{POLS}$

**Population model:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad \mathbf{E}[\varepsilon_{it}|\boldsymbol{x}_{it}, c_i] = 0.$$

The population OLS estimator is

$$\widehat{\beta}_{POLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Taking expectations conditional on $\boldsymbol{X}$:

$$\mathbf{E}[\widehat{\beta}_{POLS}|\boldsymbol{X}] = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbf{E}[\boldsymbol{c}|\boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbf{E}[\varepsilon|\boldsymbol{X}].$$

*The last term vanishes by* $\mathbf{E}[\varepsilon_{it}|\boldsymbol{x}_{it}] = 0$.

# Deriving the Bias Term

Hence,

$$\mathbf{E}[\widehat{\boldsymbol{\beta}}_{POLS}|\boldsymbol{X}] = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbf{E}[\mathbf{c}|\boldsymbol{X}].$$

In expectation over $\boldsymbol{X}$:

$$\mathbf{E}[\widehat{\boldsymbol{\beta}}_{POLS}] = \boldsymbol{\beta} + \mathbf{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{c}\right].$$

- If $\mathbf{E}[c_i|\boldsymbol{x}_{it}] = 0$ for all $i, t$, the second term is zero
  $\Rightarrow \widehat{\boldsymbol{\beta}}_{POLS}$ is unbiased.
- If $\mathbf{E}[c_i|\boldsymbol{x}_{it}] \neq 0$, this expectation induces bias.

## Example: Wage Equation

Suppose

$$\text{wage}_{it} = \beta_1\,\text{educ}_{it} + c_i + \varepsilon_{it},$$

where $c_i$ captures unobserved innate ability. If more able individuals also obtain more education, then

$$\mathbf{E}[c_i \mid \text{educ}_{it}] > 0 \quad \Rightarrow \quad \widehat{\beta}_{POLS} > \beta_1.$$

Pooled OLS overstates the return to education.

# Error Dependence

Even if $c_i$ is uncorrelated with $\boldsymbol{x}_{it}$ and pooled OLS is consistent, the composite error $u_{it} = c_i + \varepsilon_{it}$ still violates the usual OLS assumption of independent errors.

**Error Correlation:**

$$\mathbf{cov}(u_{it}, u_{is}) = \mathbf{var}(c_i) \quad \text{for } t \neq s.$$

- Errors are **correlated within** each individual (the cluster).
- Correlation within cluster = fewer *“effective”* observations.

**Across Individuals:**

$$\mathbf{cov}(u_{it}, u_{js}) = 0 \quad \text{for } i \neq j.$$

**Implication:** OLS coefficients remain unbiased if $\mathbf{E}[c_i | \boldsymbol{x}_{it}] = 0$, but inference based on classical SEs is invalid.

$\Rightarrow$We need **cluster-robust standard errors**

# Cluster-Robust Standard Errors

**Can we also allow for correlation within each individual?**
For *N* individuals observed over *T* periods, the composite error covariance has a **block-diagonal** structure.

Here is an example error covariance matrix for $T = 2$:

$$\Omega = \begin{bmatrix} \varepsilon_{11}^2 & \varepsilon_{11}\varepsilon_{12} & 0 & \cdots & 0 & 0 \\ \varepsilon_{11}\varepsilon_{12} & \varepsilon_{12}^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \varepsilon_{23}^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \varepsilon_{N1}^2 & \varepsilon_{N1}\varepsilon_{N2} \\ 0 & 0 & 0 & \cdots & \varepsilon_{N1}\varepsilon_{N2} & \varepsilon_{N2}^2 \end{bmatrix}$$

# Cluster-Robust Variance Estimator

**Idea:** Treat each cluster (e.g. individual) as a "super-observation" with correlated residuals inside.
**Setup:** Let there be $R$ clusters, and let $X_r$, $\hat{\varepsilon}_r$ denote the stacked regressors and residuals for cluster $r$.

## Estimator

$$\widehat{\mathrm{Var}}_{\mathsf{CR}}(\hat{\beta}) = (X'X)^{-1} \left( \sum_{r=1}^{R} X_r' \hat{\varepsilon}_r \hat{\varepsilon}_r' X_r \right) (X'X)^{-1}.$$

- **Bread:** $(X'X)^{-1}$ **Filling:** $\sum_r X_r' \hat{\varepsilon}_r \hat{\varepsilon}_r' X_r$
- Allows for arbitrary correlation and heteroskedasticity **within** each cluster $r$.
- Assumes independence **across** clusters.
- Asymptotics rely on the number of clusters $R \to \infty$.
- **Special case:** if each cluster has one observation ($T_r = 1$), this collapses to White's heteroskedasticity-robust variance.

# Finite-Sample Adjustment for Cluster-Robust SEs

**Problem:** With few clusters, cluster-robust standard errors are **too small** (downward-biased variance estimates)

**Finite-sample correction:**

$$\widehat{\mathrm{Var}}_{\mathsf{CR,adj}}(\hat{\beta}) = \frac{R}{R-1}\,\frac{n-1}{n-K}\,(X'X)^{-1}\left(\sum_{r=1}^{R} X_r'\hat{\varepsilon}_r\hat{\varepsilon}_r'X_r\right)(X'X)^{-1}.$$

- The first term $\frac{R}{R-1}$ corrects for the small number of clusters
- The second term $\frac{n-1}{n-K}$ adjusts for degrees of freedom in $\hat{\beta}$
- The correction fades as $R \to \infty$

## When $R$ is very small ($< 30$)

Cluster-robust SEs remain unreliable even with this correction.
$\Rightarrow$ use **wild cluster bootstrap** for valid inference.

# Multi-Way Cluster-Robust Inference (Cameron-Gelbach-Miller, 2011)

**Why multi-way clustering?**

- Errors may be correlated along more than one dimension: regions and years, workers and firms, schools and cohorts, etc.
- One-way clustering understates uncertainty when dependence is multi-dimensional.

**Key idea**

- Let $A$ and $B$ denote two cluster dimensions.
- The CGM variance estimator combines the one-way cluster variances and subtracts the overlap:

$$\widehat{\mathrm{Var}}(\hat{\beta})_{\mathrm{CGM}} = \widehat{V}_A + \widehat{V}_B - \widehat{V}_{A\cap B}.$$

- $V_A$: cluster-robust variance by cluster $A$
- $V_B$: cluster-robust variance by cluster $B$
- $V_{A\cap B}$: variance clustered at the intersection of $A$ and $B$

**Assumptions**

- Arbitrary correlation of errors within each cluster dimension.
- Independence across clusters in each dimension.
- Requires a sufficiently large number of clusters in each dimension.

# When Do Cluster-Robust SEs Work Well?

**Cluster-Robust SEs work well when:**

- There are **many independent clusters** ($R \gtrsim 30$–$50$).
- Clusters are **roughly similar in size** and none dominates the sample.
- Correlation is mainly **within clusters**, not across them.
- Treatment or variation of interest occurs **at the cluster level** (e.g., a region-specific policy shock).
- Serial correlation or shared shocks are confined to cluster boundaries.

# When Do Cluster-Robust SEs Do Not Work Well!

**Cluster-Robust SEs are unreliable when:**

- The number of clusters is small (few firms, regions, etc.).
- A few clusters contain a large fraction of all observations.
- Clusters themselves are correlated (e.g. spillovers or common shocks across regions).
- Panel length $T$ is very small and intra-cluster dependence is strong.
- Errors follow persistent time-series processes (e.g. unit roots).

**One remedy when the number of clusters is small**

Use **wild cluster bootstrap** inference for robust $p$-values and confidence intervals.

# Sidenote: The Wild Cluster Bootstrap

## What is a bootstrap?

A **bootstrap** estimates the sampling variation of an estimator (like $\hat{\beta}$) by simulating what would happen if we could repeatedly resample our data. Since we only have one dataset, we reuse model residuals to create many "pseudo-samples."

**Why do we need it in clustered data?**

- When the number of clusters $R$ is small (e.g. few firms or regions), cluster-robust $t$-tests tend to be **too optimistic** (they reject the Null too often)
- The wild cluster bootstrap preserves the correlation structure **within clusters** while generating variation **across clusters**

**Core Idea:** Instead of resampling observations, we **resample the pattern of residuals** using random weights at the cluster level.

# Sidenote: The Wild Cluster Bootstrap

**Goal:** Approximate the sampling distribution of $\hat{\beta}$ under the same cluster dependence as in the data.

**How Wild Cluster Bootstrap works:**

1. Estimate the model and keep the fitted values $\hat{y}_{it}$ and residuals $\hat{\varepsilon}_{it}$
2. For each cluster $r$, randomly change the **pattern of residuals** (multiply by a random weight that averages to zero)
3. Add these modified residuals back to $\hat{y}_{it}$ to make a new pseudo-sample
4. Re-estimate the model and record the new $\hat{\beta}^{(b)}$
5. Repeat many times to see how $\hat{\beta}$ varies across $B$ simulated samples

**How to do inference after bootstraping:**
Compute bootstrap $t$-statistics or percentile intervals from $\{\hat{\beta}^{*(b)}\}_{b=1}^{B}$

# 10.3: Fixed Effects Models

# Motivation for Fixed Effects

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

- $c_i$: time-invariant, unobserved heterogeneity (e.g. ability, preferences, management quality)
- If $c_i$ is **correlated** with $\boldsymbol{x}_{it}$, pooled OLS and random effects are **biased**.
- The **fixed effects model** controls for $c_i$ by exploiting only **within-unit variation**.

## Key assumption (strict exogeneity)

$$\mathbf{E}[\varepsilon_{it} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}, c_i] = 0$$

**Interpretation:** The idiosyncratic shock $\varepsilon_{it}$ is uncorrelated with all regressors across time.

**Implication:** No feedback from current shocks to future regressors, i.e. past or current errors do not influence future $\boldsymbol{x}_{it}$.

# Least Squares Dummy Variable (LSDV)

Example with $N = 3$ and $T = 2$:

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_2 \\ \alpha_3 \end{bmatrix}$$

$$\Rightarrow \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

- Include Dummies for individuals 2 and 3. Individual 1 stays the reference.
- Each color corresponds to one individual.
- $\boldsymbol{D}$ repeats that individual's dummy over its $T$ periods.
- $\alpha$ holds the unit-specific intercepts
- $\beta_0$: Unit-specific intercept for reference category

# Stacked General Representation of LSDV

Stacking all individuals and time periods:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{D}\alpha + \varepsilon, \quad \boldsymbol{D} = \begin{bmatrix} \mathbf{1}_T & 0 & \cdots & 0 \\ 0 & \mathbf{1}_T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_T \end{bmatrix},$$

$$\text{with} \quad \mathbf{1}_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{T\times 1} \quad \text{for each individual.}$$

- $\boldsymbol{D}$ contains one column per individual, repeated $T$ times.
- LSDV explicitly estimates $\beta$ and $\alpha$ jointly.

## Problem

With large $N$, estimating thousands of dummy coefficients is computationally inefficient.

# The Within Transformation: Idea

Start again from

$$y_{it} = c_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

1. Average over all $T$ periods for individual $i$:

$$\bar{y}_i = c_i + \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i.$$

2. Subtract this mean equation from the original:

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

## Result: Within (demeaned) regression

The individual effect $i$ **drops out**. Estimate $\beta$ by OLS on demeaned variables:

$$\hat{\beta}_{\text{FE}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i \sum_t (y_{it} - \bar{y}_i - (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta})^2.$$

# From a Single Individual to Demeaning

Focus on one individual $i$ observed over $T$ periods. Consider the simplest possible regression with only an individual-specific intercept:

$$\boldsymbol{y}_i = \alpha_i \mathbf{1}_T + \varepsilon_i, \quad \mathbf{1}_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

The OLS fitted values are obtained via the **projection matrix**

$$\boldsymbol{P}_1 = \mathbf{1}_T(\mathbf{1}_T'\mathbf{1}_T)^{-1}\mathbf{1}_T' = \tfrac{1}{T}\mathbf{1}_T\mathbf{1}_T'$$

The corresponding **residual maker** is

$$\boldsymbol{M}_1 = \boldsymbol{I}_T - \boldsymbol{P}_1 = \boldsymbol{I}_T - \tfrac{1}{T}\mathbf{1}_T\mathbf{1}_T'$$

Applying $\boldsymbol{M}_1$ to any time-varying variable $\boldsymbol{x}_i$ gives

$$\boldsymbol{M}_1\boldsymbol{x}_i = \boldsymbol{x}_i - \tfrac{1}{T}\mathbf{1}_T(\mathbf{1}_T'\boldsymbol{x}_i) = \boldsymbol{x}_i - \bar{x}_i\,\mathbf{1}_T,$$

**which subtracts the individual mean from each observation.**

# Residual Maker for All Individual Dummies

In the full LSDV model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{D}\alpha + \varepsilon,$$

where

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{1}_T & 0 & \cdots & 0 \\ 0 & \mathbf{1}_T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_T \end{bmatrix}.$$

The OLS residual maker that removes all dummy effects is

$$\boldsymbol{M}_D = \boldsymbol{I}_{NT} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'.$$

Since each dummy column in $\boldsymbol{D}$ has $T$ ones and they never overlap:

$$\boldsymbol{D}'\boldsymbol{D} = T\,\boldsymbol{I}_N.$$

Hence,

$$\boldsymbol{M}_D = \boldsymbol{I}_{NT} - \tfrac{1}{T}\boldsymbol{D}\boldsymbol{D}'.$$

This matrix removes the fitted means of each individual from all variables.

# How $\boldsymbol{M}_D$ Demeans Within Each Individual

Let us write $\boldsymbol{y}$ and $\boldsymbol{X}$ as stacked blocks:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}, \qquad \boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}.$$

Because of the structure of $\boldsymbol{DD}'$, the matrix $\boldsymbol{M}_D$ acts separately on each individual's $T$ observations:

$$\boldsymbol{M}_D = \begin{bmatrix} \boldsymbol{M}^0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{M}^0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{M}^0 \end{bmatrix}, \qquad \boldsymbol{M}^0 = \boldsymbol{I}_T - \tfrac{1}{T}\mathbf{1}_T\mathbf{1}_T'.$$

Applying $\boldsymbol{M}_D$ gives:

$$\boldsymbol{M}_D\boldsymbol{y} = \begin{bmatrix} \boldsymbol{M}^0\boldsymbol{y}_1 \\ \boldsymbol{M}^0\boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{M}^0\boldsymbol{y}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 - \bar{y}_1\mathbf{1}_T \\ \boldsymbol{y}_2 - \bar{y}_2\mathbf{1}_T \\ \vdots \\ \boldsymbol{y}_N - \bar{y}_N\mathbf{1}_T \end{bmatrix}.$$

**Each group is demeaned separately: Exactly the within transformation.**

After estimating $\hat{\beta}_{\text{FE}}$, we can retrieve the individual effects:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{FE}}),$$

which simplifies to

$$\hat{\alpha}_i = \bar{y}_i - \bar{\boldsymbol{x}}_i'\hat{\beta}_{\text{FE}}.$$

- Each $\hat{\alpha}_i$ is identical to the coefficient obtained by estimating the model directly with LSDV (person dummies).
- It represents an **individual-specific intercept** capturing all **time-invariant characteristics** (e.g., ability, preferences, long-run productivity).
- While often treated as nuisance parameters, in some applications, such as **AKM regressions**, these fixed effects are **substantively interesting**: they quantify persistent individual heterogeneity and can be for example used for decompositions of wage inequality.

# Assumptions for Fixed Effects (Within Estimator)

After the within (demeaning) transformation:

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

**Required Assumptions for Consistency:**

FE1: **Linearity** The model is linear in parameters:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}.$$

FE2: **Within-Variation / Identifiability** The demeaned regressor matrix has full column rank:

$$\tilde{\boldsymbol{X}} = \boldsymbol{M}_D\boldsymbol{X} \quad \text{has rank } K,$$

so time-invariant regressors (perfectly collinear with $c_i$) cannot be identified.

FE3: **Strict Exogeneity (Given $c_i$)**

$$\mathbf{E}[\varepsilon_{it} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}, c_i] = 0 \quad \forall t.$$

*Past, current, and future regressors are uncorrelated with $\varepsilon_{it}$. Correlation between $c_i$ and $\boldsymbol{x}_{it}$ is allowed.*

FE4: **Independent Sampling Across Individuals** The $\{(\boldsymbol{x}_{it}, y_{it}, c_i)\}$ are independent across $i$; arbitrary serial correlation and heteroskedasticity within $i$ are allowed.

FE5: **Finite Moments** $\operatorname{var}\left((\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'(\varepsilon_{it} - \bar{\varepsilon}_i)\right) < \infty$ for all $i, t$.

# Asymptotics in Fixed Effects Models

- The FE estimator is OLS with many individual dummies, so the usual OLS results on **unbiasedness** still apply under strict exogeneity:
$$\mathbf{E}[\varepsilon_{it}|\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT},c_i]=0$$
- However, because errors are typically **cluster-correlated within** each individual, classical OLS efficiency and "BLUE" properties do **not** hold
  - FE remains unbiased and consistent
  - Efficient inference requires **cluster-robust standard errors**
- For asymptotics, we let the overall sample size $NT\to\infty$
  - Usually: $N\to\infty$ with fixed $T$: Realistic for micro panels
  - Averaging over many individuals yields consistent estimates of $\beta$
- **But:** each individual intercept $\alpha_i$ is based on only $T$ observations $\Rightarrow$ **incidental parameters problem**

- In nonlinear models (e.g. logit, probit), fixed effects enter the likelihood **non-additively**:

$$\Pr(y_{it} = 1 | \boldsymbol{x}_{it}, \alpha_i) = F(\alpha_i + \boldsymbol{x}_{it}' \boldsymbol{\beta})$$

- Each $\alpha_i$ must be estimated jointly with $\boldsymbol{\beta}$ from only $T$ observations
- As $N \to \infty$ with fixed $T$:
  - The number of nuisance parameters $\alpha_i$ grows with $N$
  - This error propagates into $\hat{\boldsymbol{\beta}}$, producing a **bias of order** $1/T$
- Result: $\hat{\boldsymbol{\beta}}_{FE}$ is **biased and inconsistent** for fixed $T$

## Implications

- Linear FE: consistent despite incidental parameters
- Nonlinear FE (logit, probit): inconsistent unless $T \to \infty$

# First-Differencing as an Alternative to Demeaning

**Model:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

## First-Difference Transformation

Subtract the previous period's observation:

$$\Delta y_{it} = y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1})$$

$\Rightarrow$ $c_i$ drops out since it is time-invariant.

**Estimator:**

$$\widehat{\boldsymbol{\beta}}_{\mathsf{FD}} = \arg\min_{\boldsymbol{\beta}} \sum_i \sum_{t=2}^{T} (\Delta y_{it} - \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta})^2$$

# Relation Between First Differences and Demeaning

- Both Demeaning and FD remove unobserved heterogeneity $c_i$.
- They differ in the transformation used:
  - **Demeaning (Within):** subtracts the time mean $\bar{y}_i$.
  - **First-Differencing:** subtracts the previous period $y_{i,t-1}$.
- When $T = 2$, the two estimators are **identical**.
- For $T > 2$, they generally differ because Demeaning uses all time periods, while FD only uses $(T - 1)$ differences.

# When Do First-Difference and FE Estimates Differ?

**If errors are serially uncorrelated:**

$$\varepsilon_{it} \sim \text{i.i.d.} \Rightarrow \hat{\beta}_{\text{FE}} \text{ is more efficient.}$$

**If errors are persistent (e.g. AR(1)):**

$$\varepsilon_{it} = \rho\varepsilon_{i,t-1} + u_{it},$$

then differencing helps remove part of the serial correlation in $\varepsilon_{it}$.

**Rule of thumb:**

- **FE** preferred for i.i.d. errors and longer panels.
- **FD** preferred for short panels or dynamic models.

# Error Structure and Practical Issues with FD

## Error Process After Differencing

Even if $\varepsilon_{it}$ is i.i.d.,

$$\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} \Rightarrow \mathbf{cov}(\Delta\varepsilon_{it}, \Delta\varepsilon_{i,t-1}) = -\mathbf{var}(\varepsilon_{it})$$

$\Rightarrow$ FD errors are **MA(1)**.

**Consequences:**

- OLS on $\Delta y_{it}$ is still consistent, but not efficient.
- Standard errors must allow for MA(1) correlation.
- Differencing magnifies measurement error.

$\Rightarrow$ In practice, FE (demeaning) is usually preferred unless $T$ is very small.

# Two-Way Fixed Effects (Unit and Time Effects)

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \lambda_t + \varepsilon_{it}$$

- $c_i$: unit-specific effect, $\lambda_t$: time-specific effect (common shocks).

## Double Demeaning (Equivalent to LSDV)

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

OLS of $\tilde{y}_{it}$ on $\tilde{\boldsymbol{x}}_{it}$ yields $\hat{\beta}_{\text{TWFE}}$.

- Standard in Difference-in-Differences and event-study applications.
- Interpret $\hat{\beta}_{\text{TWFE}}$ as the **average within-unit over-time effect**.

# Attrition in Panel Data

**Attrition** = units drop out of the panel over time

- Panels become **unbalanced**
- Often unavoidable in survey and administrative data
- Distinguish:
  - **Random attrition**: unrelated to $(x_{it}, \varepsilon_{it})$
  - **Non-random attrition**: related to outcomes or regressors

**Why Attrition Matters:**

- Non-random attrition can violate FE assumptions!
- FE requires strict exogeneity:

$$\mathbf{E}[\varepsilon_{it} \mid x_{i1}, \ldots, x_{iT}, c_i] = 0$$

  If dropout depends on past shocks $\varepsilon_{i,t-1}$, this fails.
- **Results:** biased coefficients, misleading dynamics

# Typical Patterns of Non-Random Attrition

**Typical Patterns of Non-Random Attrition:**

- Low-income households more likely to exit survey
- Bad health $\rightarrow$ higher dropout in health panels
- Firms exit survey panels when close to bankruptcy

**Practical Approaches to deal with non-random attrition:**

- **Inverse probability weighting** (IPW)
  - Estimate dropout probability $\hat{p}_{it}$
  - Reweight: $w_{it} = 1/\hat{p}_{it}$
- **Sensitivity checks** Compare results on:
  - balanced vs. unbalanced panel
  - early-dropout vs. long-stayers

# 10.4: Random Effects Models and the Hausman Test

# Random Effects: Intuition

## Motivation

The **Fixed Effects (FE)** model removes $c_i$ completely. But what if $c_i$ is **not correlated** with $\boldsymbol{x}_{it}$?

- Then we **lose efficiency** by eliminating between-unit variation.
- We can treat $c_i$ as a random draw from a population:

$$c_i \sim \text{i.i.d.}(0, \sigma_c^2)$$

- In that case, $y_{it}$ contains two random noise components:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \underbrace{c_i}_{\text{shared across } t} + \underbrace{\varepsilon_{it}}_{\text{idiosyncratic}}$$

- The Random Effects (RE) model exploits both:
  - **Within variation:** over time for a given $i$
  - **Between variation:** across different $i$

# Random Effects Model in Detail

**Model:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

- $c_i$ is a random effect: $c_i \sim \text{i.i.d.}(0, \sigma_c^2)$
- $\varepsilon_{it} \sim \text{i.i.d.}(0, \sigma_\varepsilon^2)$
- Independent across $i$, uncorrelated with $\boldsymbol{x}_{it}$

**Error structure:**

$$\Omega_i = \mathbf{E}[\boldsymbol{u}_i\boldsymbol{u}_i'] = \sigma_\varepsilon^2 \boldsymbol{I}_T + \sigma_c^2 \mathbf{1}_T\mathbf{1}_T'$$

$\Rightarrow$ serial correlation within $i$: $\mathbf{cov}(u_{it}, u_{is}) = \sigma_c^2$ for $t \neq s$

## Estimation

Because errors are correlated, OLS is inefficient. We can use **Generalized Least Squares (GLS)**:

$$\hat{\beta}_{RE} = (\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}\boldsymbol{y}.$$

In practice: **Feasible GLS (FGLS)** estimates $\sigma_c^2$ and $\sigma_\varepsilon^2$ first.

# Assumptions for Random Effects

**The Random Effects Model:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad u_{it} = c_i + \varepsilon_{it},$$

with random intercept $c_i$ treated as part of the composite error.

**Required Assumptions for Consistency:**

RE1: **Linearity** The model is linear in parameters.

RE2: **Identifiability** The regressor matrix has full column rank. Time-invariant regressors can be estimated under RE.

RE3: **Strict Exogeneity**

$$\mathbf{E}[\varepsilon_{it} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}, c_i] = 0 \quad \forall t.$$

RE4: **Orthogonality of Random Effects**

$$\mathbf{E}[c_i \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}] = 0.$$

*Key condition: if $c_i$ and $\boldsymbol{x}_{it}$ are correlated, RE is inconsistent.*

RE5: **Independent Sampling Across Individuals** Observations are independent across $i$; within-unit correlation arises through $c_i$.

RE6: **Finite Moments** $\mathrm{var}(u_{it} \mid \boldsymbol{x}_{it}) < \infty$, and $\mathrm{var}(c_i) < \infty$.

# Quasi-Demeaning: Intuition

- FE removes all between variation (demeans fully).
- RE wants a compromise: only **partial demeaning**.

$$y_{it} - \theta \bar{y}_i = (\boldsymbol{x}_{it} - \theta \bar{\boldsymbol{x}}_i)' \boldsymbol{\beta} + (\varepsilon_{it} - \theta \bar{\varepsilon}_i)$$

where $\theta \in [0, 1]$ controls how much within variation is used.

## Interpretation of $\theta$

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_c^2 + \sigma_\varepsilon^2}}$$

- $\theta = 1$: full demeaning $\Rightarrow$ Fixed Effects
- $\theta = 0$: no demeaning $\Rightarrow$ Pooled OLS
- $0 < \theta < 1$: partial demeaning $\Rightarrow$ Random Effects

The transformation gives:

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}_{it}' \boldsymbol{\beta} + \tilde{\varepsilon}_{it},$$

which can be estimated by OLS.

# Quasi-Demeaning: Technical Details

**Start from:**

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}.$$

**Transform both sides:**

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

The constant component $c_i$ partially cancels:

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}'_{it}\boldsymbol{\beta} + (1-\theta)c_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i).$$

**Variance of transformed error:**

$$\mathbf{var}(\tilde{u}_{it}) = \sigma^2_\varepsilon(1-\theta)^2 + \sigma^2_c(1-\theta)^2.$$

## Feasible GLS (FGLS) Estimator

Estimate $\sigma^2_c$ and $\sigma^2_\varepsilon$ from residuals, compute $\theta$, transform variables, and estimate by OLS:

$$\hat{\beta}_{RE} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}.$$

# Hausman Test: Intuition

**Goal**

Decide whether the RE assumption $\mathbf{E}[c_i|\boldsymbol{x}_{it}] = 0$ is valid.

- **If true:** both FE and RE are consistent, but RE is more efficient.
- **If false:** RE is inconsistent, only FE is consistent.

**Idea:** Compare $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ — if they differ substantially, $c_i$ must be correlated with $\boldsymbol{x}_{it}$.

$$H = (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})'[\mathrm{Var}(\hat{\boldsymbol{\beta}}_{FE}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{RE})]^{-1}(\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})$$

- $H \sim \chi^2_K$
- If $H$ is large $\Rightarrow$ reject $H_0$: use FE.
- If $H$ is small $\Rightarrow$ fail to reject: use RE.

# Hausman Test: Implementation and Interpretation

**Null Hypothesis:**

$$H_0 : \mathbf{E}[c_i | \boldsymbol{x}_{it}] = 0 \quad \Rightarrow \text{RE consistent \& efficient.}$$

**Alternative:**

$$H_1 : \mathbf{E}[c_i | \boldsymbol{x}_{it}] \neq 0 \quad \Rightarrow \text{RE inconsistent, prefer FE.}$$

**Steps:**

1. Estimate $\hat{\boldsymbol{\beta}}_{FE}$ and $\hat{\boldsymbol{\beta}}_{RE}$.
2. Compute variance matrices $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{FE})$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{RE})$.
3. Evaluate test statistic $H$ as above.
4. Compare to $\chi^2_K$ critical value.

**Intuition:**

- If $c_i$ correlates with $\boldsymbol{x}_{it}$, FE and RE estimates diverge.
- The difference between them is evidence against RE.

# Limitations of the Classical Hausman Test

## Context

The Hausman test is widely taught, but in practice, it often performs poorly, especially with modern robust inference methods.

- The test statistic

$$H = (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})'[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{FE}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{RE})]^{-1}(\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})$$

  assumes:
    - consistent and positive definite covariance difference
    - homoskedastic, independent errors
    - correct model specification in both FE and RE
- In small or unbalanced panels:
    - $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{FE}) - \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{RE})$ may not be invertible
    - Test becomes numerically unstable or yields negative *p*-values
- With clustered or heteroskedastic errors:
    - The classical $\chi^2$ reference distribution is invalid
    - Cluster-robust SEs make *H* undefined

# Why Pure Random Effects Are Rarely Used Today

## In theory:

RE is efficient and elegant under $\mathbf{E}[c_i|\boldsymbol{x}_{it}] = 0$.

- In practice, this assumption is rarely credible:
  - Individual heterogeneity ($c_i$) almost always correlates with regressors.
  - Especially in labor, firm, and regional data, where unobserved traits drive both $\boldsymbol{x}_{it}$ and $y_{it}$.
- The "efficiency gain" is small compared to potential bias.
- Modern applied work prioritizes **robust identification** over efficiency:
  - **Fixed Effects (FE)** dominates for causal inference

## Bottom Line

In empirical work: **Start with FE**, use RE only if the identifying assumptions are clearly defensible.

# 10.5: Mundlak's Approach

# Equivalent to Fixed Effects: Mundlak's Approach

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

- Problem: $\mathbf{E}[c_i|\boldsymbol{X}_i] \neq 0$ and we cannot (or do not want to) use FE.
- **Example:** Non-linear models (logit, probit) suffer from incidental parameter bias.
- **Idea:** Capture the correlated part of $c_i$ with a control function.

## Assumed Relationship

$$\mathbf{E}[c_i|\boldsymbol{X}_i] = \bar{\boldsymbol{x}}_i'\gamma \quad + \boldsymbol{z}_i'\delta$$

$\bar{\boldsymbol{x}}_i$ contains group means of all time-varying regressors, and $\boldsymbol{z}_i$ are time-invariant regressors.

# Mundlak's Approach: The Augmented Model

Substituting $\mathbf{E}[c_i|\boldsymbol{X}_i]$ into the structural equation gives

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \bar{\boldsymbol{x}}_i'\gamma + \boldsymbol{z}_i'\boldsymbol{\delta} + \varepsilon_{it} + (c_i - \mathbf{E}[c_i|\boldsymbol{X}_i]).$$

- Estimate by Random Effects GLS: The residual $(c_i - \mathbf{E}[c_i|\boldsymbol{X}_i])$ is now uncorrelated with $\boldsymbol{x}_{it}$.
- $\boldsymbol{\beta}$ has the same interpretation as in FE.
- $\gamma$ captures correlation between $c_i$ and the regressors.

# Equivalence to Fixed Effects

- Even if $\mathbb{E}[c_i|\boldsymbol{X}_i] = \bar{\boldsymbol{x}}_i'\gamma$ does not hold exactly, Mundlak's estimator yields
$$\hat{\boldsymbol{\beta}}_{Mundlak} = \hat{\boldsymbol{\beta}}_{FE}.$$
- Proven via partitioned regression algebra.
- Only interpretation of $\gamma$ changes: Slopes $\beta$ are identical.
- Hence, FE can be seen as a special case of RE with $\bar{\boldsymbol{x}}_i$ included.

## Key Takeaway

Mundlak's approach bridges FE and RE: it makes RE robust to correlation between $c_i$ and $\boldsymbol{x}_{it}$.

# Interpreting the Coefficients

- $\gamma$ measures how much of the individual effect $c_i$ is explained by long-term differences in the regressors.
- If $\gamma = 0 \Rightarrow c_i$ uncorrelated with $\boldsymbol{x}_{it} \Rightarrow$ RE valid.
- Time-invariant regressors ($\boldsymbol{z}_i$) can enter and be estimated, unlike in FE.
- Two complementary interpretations:
  - **Correlational:** $\gamma$ reflects the association between unobserved traits and the average regressor levels.
  - **Between-effects:** $\gamma$ captures how individuals with persistently higher $\bar{\boldsymbol{x}}_i$ differ in their average $y_i$:
    A "long-run" or "between-individual" effect.

# Mundlak's Approach and the Hausman Test

- Hausman test compares FE vs. RE under $H_0 : \mathbf{E}[c_i|\boldsymbol{X}_i] = 0$.
- Mundlak's specification nests RE as a special case:

$$H_0 : \gamma = 0.$$

- If $H_0$ rejected $\to$ RE inconsistent $\to$ prefer FE (or Mundlak model itself).
- Advantage: same test, no covariance matrix inversion, works with robust SEs.

## Interpretation

Mundlak's model provides a regression-based alternative to the Hausman test.

# 10.6: Repeated Cross-Sections and Pseudo-Panels

# From Panel to Repeated Cross-Sections

## Definition

A **repeated cross-section** contains observations from different individuals or firms in each time period.

- Unlike a true panel, we do **not follow the same entities** over time.
- Typical in large household surveys such as:
  EU-SILC, CPS, Mikrozensus
- We can still analyze **aggregate or average changes over time**.

**Structure:**

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_t + \varepsilon_{it}, \qquad i = 1, \ldots, n_t, \; t = 1, \ldots, T.$$

Each $t$ has a new sample.

# Estimating Changes Over Time

## Model with Time Dummies

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \sum_{t=2}^{T} \delta_t D_t + \varepsilon_{it}$$

- The $\delta_t$'s measure **average differences across years**, controlling for $\boldsymbol{x}_{it}$.
- Estimate by pooled OLS on the stacked repeated cross-sections.
- Allows testing for structural change:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T$$

  versus year-specific coefficients.

## Example

Compare average log wages across survey years after controlling for education, age, and region.

# Pseudo-Panel Approach (Deaton, 1985)

## Idea

Construct a "synthetic" panel by grouping repeated cross-sections into **cohorts** that are stable over time (e.g. birth years, education levels, gender).

- Define cohorts $g = 1, \dots, G$ (e.g. people born 1970–74).
- For each year $t$, compute **cohort means**:

$$\bar{y}_{gt} = \frac{1}{n_{gt}} \sum_{i \in g,t} y_{it}, \qquad \bar{\boldsymbol{x}}_{gt} = \frac{1}{n_{gt}} \sum_{i \in g,t} \boldsymbol{x}_{it}.$$

- Treat $(\bar{y}_{gt}, \bar{\boldsymbol{x}}_{gt})$ as panel observations for cohort $g$.
- Estimate:

$$\bar{y}_{gt} = \bar{\boldsymbol{x}}'_{gt}\boldsymbol{\beta} + c_g + \varepsilon_{gt}.$$

**Interpretation:** The cohort means behave as if we had followed a representative individual from each cohort through time.

# Pseudo-Panels: Assumptions and Properties

- Cohorts must be **large and time-invariant**: membership should not change substantially over time.
- The individual errors $\varepsilon_{it}$ average out within cohort:

$$\mathbf{E}[\varepsilon_{gt}] \approx 0 \quad \text{for large } n_{gt}.$$

- Sampling noise introduces **measurement error** in $\bar{y}_{gt}$ and $\bar{\boldsymbol{x}}_{gt}$.
- As $n_{gt} \to \infty$, this measurement error vanishes.

## Econometric Implication

With sufficiently large cohorts, standard FE or RE methods can be applied to cohort-level data to estimate within-cohort dynamics.

# Trade-Offs of the Pseudo-Panel Approach

| Advantages | Limitations |
| --- | --- |
| + Allows analysis of long-run dynamics when true panels unavailable | − Loss of micro-level variation (since aggregation to cohort means) |
| + Reduces bias from composition changes across years | − Cohort definitions must be stable and sufficiently large |
| + Enables fixed effects at cohort level | − Measurement error from finite-sample cohort means |
| + Particularly useful for long survey series (e.g. CPS, EU-SILC, LFS) | − Limited ability to track individual transitions |

# 10.7: Diagnostics and Robust Inference

# Serial Correlation in Panel Models

## Why It Matters

If $\varepsilon_{it}$ is serially correlated, even cluster-robust SEs may underestimate true sampling variation.

- Typical in micro panels (wages, firm productivity, health outcomes)
- Violates classical OLS/GLS assumption of independent errors
- Leads to over-rejection of $H_0$

**Wooldridge (2002)** test for serial correlation in FE models:

1. Estimate the FE model and save residuals $\hat{\varepsilon}_{it}$.
2. Regress $\hat{\varepsilon}_{it}$ on $\hat{\varepsilon}_{i,t-1}$.
3. Test $H_0 : \rho = 0$ (no serial correlation).

# Cross-Sectional Dependence and Robust SEs

- Common shocks (macro events, policy changes) induce correlation across $i$.
- Ignoring this inflates $t$-statistics even with unit clustering.

**Diagnostic Tests:**

- Breusch–Pagan LM test for cross-sectional dependence
- Pesaran (2004) CD test (robust for large $N$, small $T$)

**Robust Alternatives:**

- Two-way clustering (e.g. by individual and time)
- Wild cluster bootstrap for few clusters

# Why Cluster Standard Errors, and at Which Level?

## Principle

Cluster SEs whenever residuals are correlated within identifiable groups.

- Dependence often arises from
  - repeated observations on the same unit
  - treatment or policy applied to groups
  - shared environment (regions, firms, schools)
- Clustering accounts for arbitrary correlation within those groups.

**Rule of Thumb:** Cluster at the level of **treatment assignment or data dependence**. Finer clustering $\Rightarrow$ too optimistic; coarser clustering $\Rightarrow$ conservative but valid.

# When Should You Adjust SEs for Clustering?

**Abadie, Athey, Imbens, and Wooldridge (2023, QJE)**

## Key Message

Standard errors should reflect the level of effective treatment variation.

- In experiments or DiD setups, the treatment is assigned at the group level.
- Individuals within a treated group share the same shock $\Rightarrow$ correlated errors.
- Clustering must occur at that **treatment-group level**.

**Examples:**

- State-level minimum wage reform $\Rightarrow$ cluster by state
- Firm-level training program $\Rightarrow$ cluster by firm
- School-level intervention $\Rightarrow$ cluster by school

**Clustering finer (by individuals)** underestimates uncertainty.

# When All Else Fails $\rightarrow$ Randomization Inference

## Core Idea

If asymptotic or cluster-robust inference is unreliable (e.g. few clusters, strong dependence, or small samples), we can test hypotheses by **resampling the treatment assignment itself**.

- Keep the outcome data fixed.
- Randomly **permute or reassign treatment labels** according to the original experimental design.
- Re-estimate the effect for each permutation $\Rightarrow$ get the **randomization distribution** of the test statistic.
- Compare the actual estimate $\hat{\beta}_{\text{obs}}$ to this distribution. The *p*-value is the share of permutations where $|\hat{\beta}^{(b)}| \geq |\hat{\beta}_{\text{obs}}|$.

# Caveats and Takeaways for Randomization Inference

**Strengths:**

- Provides **exact finite-sample inference** under the known treatment assignment mechanism.
- Does not rely on large-$N$ or large-$T$ asymptotics.
- Particularly valuable when:
  - The number of clusters is small, or
  - Cross-sectional dependence is strong.

**Limitations:**

- Can be **overly conservative**, especially with few possible reassignments.
- Tests validity only under the experimental assignment, not under broader sampling variation.

**Takeaway:** Use randomization inference as a **robust fallback** when cluster-robust SEs or bootstrap methods are unreliable. It validates inference by returning to the core question: Would we see this effect if treatment were randomly reassigned?

# 10.8: Comparison and Empirical Guidelines

# The "Within vs. Between" Confusion

## Two Sources of Variation

- **Within variation:** Changes in $x_{it}$ for the same individual over time.
- **Between variation:** Differences in $\bar{x}_i$ across individuals.

- Fixed Effects (FE) uses only within variation
  $\Rightarrow$ identifies how changes within a unit affect $y$.
- Random Effects (RE) and Pooled OLS mix within and between variation.
- Misinterpreting FE estimates as cross-sectional effects
  $\Rightarrow$ classic "within/between confusion."

## Example

FE: Does a firm's productivity rise when it invests more than its own average?
Between: Are firms that invest more on average more productive?

# Decomposing Variation: Within vs Between $R^2$

Total variation in $y_{it}$ can be decomposed as:

$$y_{it} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{it} - \bar{y}_i)$$

$$\Rightarrow \mathrm{Var}(y_{it}) = \underbrace{\mathrm{Var}(\bar{y}_i)}_{\text{between variation}} + \underbrace{\mathbf{E}_i[\mathrm{Var}(y_{it}|i)]}_{\text{within variation}}.$$

## Three $R^2$ Measures

- **Within $R^2$:** Fit of the demeaned (FE) regression. Uses only $y_{it} - \bar{y}_i$ and $x_{it} - \bar{x}_i$.
- **Between $R^2$:** Fit across $\bar{y}_i$ and $\bar{x}_i$.
- **Overall $R^2$:** Fit of pooled OLS ignoring $c_i$.

High within-$R^2$ $\Rightarrow$ model explains time variation within units.
High between-$R^2$ $\Rightarrow$ model explains cross-unit level differences.

# Pooled OLS: Mixing Within and Between Variation

Pooled OLS estimates

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it}, \quad u_{it} = c_i + \varepsilon_{it}.$$

Decompose each regressor and outcome into mean and deviation:

$$x_{it} = \bar{x}_i + (x_{it} - \bar{x}_i), \qquad y_{it} = \bar{y}_i + (y_{it} - \bar{y}_i).$$

Substituting gives:

$$y_{it} = \bar{x}_i'\boldsymbol{\beta} + (x_{it} - \bar{x}_i)'\boldsymbol{\beta} + c_i + \varepsilon_{it}.$$

## Implication

Pooled OLS combines:

$$\beta_{\text{POLS}} = w_W\, \beta_{\text{within}} + w_B\, \beta_{\text{between}},$$

where weights $w_W, w_B$ depend on $\mathrm{Var}(x_{it} - \bar{x}_i)$ and $\mathrm{Var}(\bar{x}_i)$.

$\Rightarrow$ If $c_i$ correlated with $\bar{x}_i$, the between component biases $\beta_{\text{POLS}}$.

# Random Effects: A Weighted Combination

Random Effects quasi-demeans the data:

$$y_{it} - \theta \bar{y}_i = (\boldsymbol{x}_{it} - \theta \bar{\boldsymbol{x}}_i)' \boldsymbol{\beta} + (\varepsilon_{it} - \theta \bar{\varepsilon}_i),$$

where

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_c^2 + \sigma_\varepsilon^2}}.$$

## Interpretation

- $\theta = 1 \Rightarrow$ full demeaning $\rightarrow$ FE estimator
- $\theta = 0 \Rightarrow$ no demeaning $\rightarrow$ Pooled OLS
- $0 < \theta < 1 \Rightarrow$ partial demeaning $\rightarrow$ RE estimator

$$\hat{\boldsymbol{\beta}}_{\text{RE}} = w_W \, \hat{\boldsymbol{\beta}}_{\text{within}} + (1 - w_W) \, \hat{\boldsymbol{\beta}}_{\text{between}},$$

where $w_W$ increases with $\theta$.

**Hence:** RE blends within and between information, approaching FE as intra-cluster correlation grows.

# Choosing the Right Model: Practical Guidelines

| Model Type | When to Use |
|---|---|
| **Pooled OLS** | Quick benchmark; only valid if $c_i$ uncorrelated with regressors and no serial dependence. |
| **Fixed Effects (FE)** | Default for causal inference with unobserved, time-invariant heterogeneity. Removes all between-unit variation. |
| **First Differences (FD)** | Short panels ($T = 2$-$3$), or dynamic settings with serially correlated errors. |
| **Random Effects (RE)** | Appropriate only if $\mathbf{E}[c_i \mid \boldsymbol{x}_{it}] = 0$ is credible; more efficient but rarely defensible empirically. |
| **Mundlak Model** | Hybrid: retains RE efficiency while controlling for correlation between $c_i$ and regressors. |
| **Pseudo-Panels / Repeated Cross-Sections** | When true panels unavailable; use large, stable cohorts to recover within-group dynamics. |

**Always:** Cluster standard errors at the appropriate level, test for serial and cross-sectional dependence, and prefer identification over efficiency.

# Conclusion and Outlook: Toward Modern Diff-in-Diff

- Panel estimators (FE, FD) form the foundation for modern causal inference with treatment variation over time.
- Classical two-way FE DiD models:

$$y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it}$$

  estimate an average treatment effect under **parallel trends**.
- Recent literature challenges this simple setup:
  - Treatment timing heterogeneity $\Rightarrow$ **TWFE bias**
  - Event-study estimators and group-time average treatment effects
  - **New DiD estimators:** Callaway-Sant'Anna (2021), Sun-Abraham (2021), de Chaisemartin-D'Haultfoeuille (2020)

## Next Lecture

**Lecture 11: The New Difference-in-Differences Literature**

Dynamic treatment effects, staggered adoption, and identification beyond parallel trends.