

Panel Data

Erick F Molina

Summer, 2024

ITAM

This presentation is based on the references listed in the course syllabus and on notes by Arturo Aguilar, Mauricio Romero, Laura Schechter, Shaun McRae, Kirill Borusyak, among others. Any remaining errors are my own.

What are panel data?

- Panel data contain observations on multiple entities (e.g., individuals, states, firms), where each entity is observed at two or more points in time.
- Examples:
 - Data on OECD member countries, each observed for more than 20 years.
 - Data on Mexico's 32 states, observed in 1990, 2000, and 2010.
 - Data on 12,000 Mexican households, observed for five consecutive quarters.

Notation in panel data

- A double subscript distinguishes entities and time periods:
 - i : indexes the entity. We say there are N entities.
 - t : indexes time. We say there are T time periods.
- Example:

$$y_{it} = \alpha + \beta X_{it} + \epsilon_{it} \quad (1)$$

- where y_{it} is the dependent variable, X_{it} are the regressors, and ϵ_{it} is the error term.

Why use panel data?

- Advantages:
 - They control for variables that differ across entities but are constant over time, reducing omitted variable bias (OVB).
 - They provide more data points, increasing degrees of freedom and efficiency.
 - They can account for unobserved heterogeneity.

The key idea is: if an omitted variable does not change over time, then changes in Y over time cannot be attributed to that omitted variable.

Balanced vs. unbalanced panels

- **Balanced panel:**
 - All variables are observed for all entities and all time periods.
 - The number of observations is $N \times T$.
- **Unbalanced panel:**
 - Some variables are missing for at least one entity or time period.
 - Common in empirical economic settings.

First-Differences Estimator

Suppose the following panel-data model where we observe only 2 periods:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \gamma A_i + \varepsilon_{it} \quad (2)$$

where:

- y_{it} is the wage of individual i at time t .
- x_{it} is the level of education.
- A_i is **unobserved** ability, constant over time.
- ε_{it} is the error term.

Deriving the first-differences estimator

- Consider two time periods $t = 1$ and $t = 2$.
- The model at time $t = 1$:

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \gamma A_i + \varepsilon_{i1}$$

- The model at time $t = 2$:

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \gamma A_i + \varepsilon_{i2}$$

- Ability A_i does not vary over time but does vary across individuals, creating bias if it is not controlled for.

Deriving the first-differences estimator

- Subtracting the equation at $t = 1$ from the equation at $t = 2$:

$$\begin{aligned}y_{i2} - y_{i1} &= \beta_0 - \beta_0 \\&\quad + \beta_1(x_{i2} - x_{i1}) \\&\quad + \gamma(A_i - A_i) \\&\quad + (\varepsilon_{i2} - \varepsilon_{i1})\end{aligned}$$

- Ability A_i drops out because it is constant over time:

$$y_{i2} - y_{i1} = \beta_1(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \tag{3}$$

Deriving the first-differences estimator

- We estimate β_1 using ordinary least squares (OLS) on the first-differenced model.
- This $\hat{\beta}_1$ is our **first-differences estimator**, also denoted $\hat{\beta}_{FD}$.
- Differencing removes bias from time-invariant omitted variables such as A_j .

Advantages and limitations of the first-differences estimator

Advantages:

- Controls for omitted variables that are constant over time.
- Simple to implement and understand.
- Can be applied when there are only two time periods.

Limitations:

- Does not control for omitted variables that vary over time.
- Can increase the variance of the estimator if differences in the errors are large.

Fixed Effects

Motivation for the fixed effects model

- We continue with the idea of eliminating bias caused by unobserved variables that are constant over time.
- You can view this as an extension of first differences to more than two periods.
- Suppose we start from a panel-data model with T periods:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \gamma A_i + \varepsilon_{it} \quad (4)$$

Deriving the fixed effects model

- We begin by taking time averages for each variable, for each individual:

$$\bar{y}_i = \frac{1}{n} \sum_{t=1}^T y_{it}$$

$$\bar{x}_i = \frac{1}{n} \sum_{t=1}^T x_{it}$$

$$A_i = \frac{1}{n} \sum_{t=1}^T A_i$$

$$\bar{\varepsilon}_i = \frac{1}{n} \sum_{t=1}^T \varepsilon_{it}$$

Deriving the fixed effects model

- Using these averages, we obtain:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \gamma A_i + \bar{\epsilon}_i \quad (5)$$

- Then we subtract the averaged equation from the original equation:

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \quad (6)$$

- Note that the unobserved term A_i is also eliminated.
- Estimating by OLS, the resulting $\hat{\beta}_1$ is our fixed effects estimator, also known as the **within estimator** $\hat{\beta}_{FE}$.

Assumptions of the fixed effects model

- Standard OLS assumptions.
- No perfect multicollinearity.
- Homoskedasticity and no serial correlation in the errors:

$$E(\varepsilon_{it}\varepsilon_{js}) = \begin{cases} \sigma^2 & \text{if } i = j \text{ and } t = s \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Random Effects

What is the random effects estimator?

- The random effects model assumes that differences across individuals are random and uncorrelated with the regressors included in the model.
- It is used when we expect that individual-specific variation is not correlated with the explanatory variables.

Random effects model

- Suppose the following model:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \mu_i + \varepsilon_{it} \quad (8)$$

- where:
 - y_{it} is the dependent variable for individual i at time t .
 - x_{it} is the independent variable.
 - μ_i is the entity-specific random effect. Suppose $\mu_i \sim N(0, \sigma_\mu^2)$.
 - ε_{it} is the idiosyncratic error term. Suppose $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$.

GLS transformation (Generalized Least Squares)

- To obtain the random effects estimator, we apply the GLS transformation to address serial correlation:

$$y_{it}^* = y_{it} - \theta \bar{y}_i$$

$$x_{it}^* = x_{it} - \theta \bar{x}_i$$

- where $\theta = 1 - \sqrt{\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_\mu^2}}$.
- The transformed equation becomes:

$$y_{it}^* = \beta_0(1 - \theta) + \beta_1 x_{it}^* + \epsilon_{it}^*$$

- This can be estimated using ordinary least squares (OLS).

- The GLS estimator of β is:

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

- where X^* and y^* are the transformed matrices of regressors and the dependent variable.

- The variances σ_μ^2 and σ_ϵ^2 can be estimated using regression residuals.
- An estimator of σ_μ^2 is:

$$\hat{\sigma}_\mu^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 - \frac{1}{T} \sum_{i=1}^N (\bar{\hat{u}}_i)^2$$

- An estimator of σ_ϵ^2 is:

$$\hat{\sigma}_\epsilon^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\hat{u}_{it} - \bar{\hat{u}}_i)^2$$

Advantages and limitations of the random effects model

Advantages:

- Allows inclusion of time-invariant variables.
- More efficient if the model assumptions hold.
- Provides estimates for the effect of entity-specific variables that do not vary over time.

Limitations:

- The validity of the model depends on the assumption that the random effects are uncorrelated with the regressors.
- If this assumption fails, the estimator will be biased and inconsistent.

The Between Estimator

The between estimator

- Recall the time-averaged equation we obtained in the fixed effects model:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + A_i + \bar{\epsilon}_i \quad (9)$$

- If we estimate this equation, we obtain $\hat{\beta}_1$, the **between estimator**, denoted $\hat{\beta}_{Be}$.
- Define $b = A_i + \bar{\epsilon}_i$.

Variance and efficiency of the estimator

The variance of the *between estimator* simplifies when the panel is balanced:

$$V_{Be} = \text{var}[\hat{\beta}_{Be}|X] = \left(\sum_{i=1}^N X_i X_i' \right)^{-1} \left(\sigma_A^2 + \frac{\sigma_\varepsilon^2}{T} \right) \quad (10)$$

Under the random effects assumption, the *between estimator* is unbiased, but less efficient than the random effects estimator:

$$\hat{\beta}_{Be} \implies \text{unbiased} \quad (11)$$

$$\hat{\beta}_{GLS} \implies \text{more efficient} \quad (12)$$

The main application of the *between estimator* is to estimate σ_A^2 .
Consider:

$$\sigma_b^2 = \sigma_A^2 + \frac{\sigma_\varepsilon^2}{T} \quad (13)$$

A natural estimator for σ_b^2 is:

$$\hat{\sigma}_b^2 = \frac{1}{N - k} \sum_{i=1}^N \hat{v}_i^2 \quad (14)$$

Constructing the estimator

Using the fixed effects estimator to obtain $\hat{\sigma}_\varepsilon^2$ and the *between estimator* to obtain $\hat{\sigma}_b^2$, we can derive an estimator for σ_A^2 :

$$\hat{\sigma}_A^2 = \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\varepsilon^2}{T} \quad (15)$$

$$\hat{\sigma}_A^2 = \max \left(0, \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\varepsilon^2}{T} \right) \quad (16)$$

Comparing within and between estimators

- **Within estimator:**
 - More efficient when fixed effects are present.
 - Controls for time-invariant omitted variables.
- **Between estimator:**
 - Less efficient when fixed effects are present.
 - Can be biased if omitted variables are correlated with the regressors.