Introduction to Panel Data

Concepts, Advantages, and Methods

Alessandro Di Nallo

Max Planck Institute for Demographic Research

February 11, 2025

Overview

Definition and Structure

Definition:

- Panel (longitudinal) data contain observations of multiple *entities* over multiple *time periods*.
- Usually organized in *long* format: each row is an entity-time combination.

Examples:

- Household surveys repeated annually (e.g., income, consumption).
- Firm-level accounting data tracked over multiple years.
- Countries' macroeconomic indicators measured over decades.

Motivation

Advantages of Panel Data:

- **Rich Information**: Combines cross-sectional and time-series aspects.
- **Control for Unobserved Heterogeneity**: Potential to reduce omitted variable bias for time-invariant factors.
- **Dynamic Analysis**: Examine trajectories, transitions, or life-course events (e.g., unemployment spells).
- **Increased Efficiency**: More observations often mean more precise estimates.

Motivation

Advantages of Panel Data:

- **Rich Information**: Combines cross-sectional and time-series aspects.
- **Control for Unobserved Heterogeneity**: Potential to reduce omitted variable bias for time-invariant factors.
- **Dynamic Analysis**: Examine trajectories, transitions, or life-course events (e.g., unemployment spells).
- **Increased Efficiency**: More observations often mean more precise estimates.

Common Research Questions:

- Do policy changes (e.g., minimum wage) affect employment over time?
- How does marriage or divorce impact wages over an individual's career?
- Can we link firm R&D spending to future profits more reliably?

Within vs. Between Variation

Panel data let us separate *within* (over time) from *between* (across units) variation, leading to more nuanced analyses.

Within Variation (Time-Series Dimension):

- Captures how an individual (person, firm, region) changes over time.
- Example: A person's annual wage pre- and post-marriage.

Between Variation (Cross-Section Dimension):

- Captures differences across entities at a given point in time.
- Example: Comparing wages across different individuals in a single year.

Method 1: Pooled OLS

- Treats all entity-time observations as one big cross-section.
- **Pros:** Simple to implement and interpret.
- **Cons:** Ignores panel structure; can be biased if unobserved heterogeneity is correlated with regressors.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_{it}, \quad \nu_{it} = \alpha_i + \epsilon_{it}$$

Issue: If α_i is correlated with x_{it}, OLS estimates will be biased.

Method 1: Pooled OLS

- Treats all entity-time observations as one big cross-section.
- **Pros:** Simple to implement and interpret.
- **Cons:** Ignores panel structure; can be biased if unobserved heterogeneity is correlated with regressors.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_{it}, \quad \nu_{it} = \alpha_i + \epsilon_{it}$$

Issue: If α_i is correlated with x_{it}, OLS estimates will be biased.

Example: Imagine that y_{it} is wage and x_{it} is education. ν_{it} includes: job experience, age, etc., which may be correlated with education.

Method 2: Fixed Effects (FE) Models

- **Key Idea:** Control for all time-invariant entity-specific characteristics by focusing on **within** variation.
- Each entity acts as its own control.
- Time-invariant covariates drop out (cannot be estimated).

Within-Transformation:

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i).$$

Pros: Eliminates bias from unobserved time-invariant heterogeneity: $\alpha_i - \bar{\alpha}_i = 0$
Cons: Cannot estimate coefficients on variables with no time variation (e.g., country of origin).

Method 3: Random Effects (RE) Models

- **Key Assumption:** The unobserved effect α_i is uncorrelated with x_{it}.
- Allows for inclusion of time-invariant regressors.
- More efficient than FE if the assumption holds.

Pros: Can estimate the effect of variables that don't vary over time.
Cons: Potential bias if the "random effects" assumption is violated.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \epsilon_{it}, \quad \alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2).$$

Choosing Between Random Effects and Fixed Effects

Fixed Effects (FE)

- Use when you suspect that the unobserved, time-invariant characteristics (α_i) **are correlated** with your regressors.
- Each entity acts as its own control, differencing out all stable traits.
- More robust when self-selection (or endogeneity) is likely.
- **Limitations:**
  - Cannot estimate time-invariant regressors (they get differenced out).
  - May have higher standard errors if you lose a lot of variation.

Choosing Between Random Effects and Fixed Effects

Fixed Effects (FE)

- Use when you suspect that the unobserved, time-invariant characteristics (α_i) **are correlated** with your regressors.
- Each entity acts as its own control, differencing out all stable traits.
- More robust when self-selection (or endogeneity) is likely.
- **Limitations:**
  - Cannot estimate time-invariant regressors (they get differenced out).
  - May have higher standard errors if you lose a lot of variation.

Random Effects (RE)

- Use when you assume that α_i is **uncorrelated** with your regressors.
- More efficient (smaller standard errors) if the assumption holds.
- Allows estimation of coefficients on time-invariant variables.
- **Limitations:**
  - Potentially biased if there is *any* correlation between α_i and regressors.
  - Hausman test often used to compare FE vs. RE results.

Method 4: Dynamic Panel Models

- Useful when past values of the dependent variable (y_{it-1}) affect current y_{it}.
- Example: Arellano-Bond GMM estimators deal with endogeneity introduced by lagged dependent variables.
- More complex, but address key questions about "state dependence" (e.g., past unemployment influencing current unemployment).

Data Management

Key Steps:

- Reshape data into *long format*: each row = (entity, time).
- Check for **inconsistencies**: missing IDs, repeated time points, outliers.
- Address **missing data**: panel attrition is common; consider multiple imputation or weighting if appropriate.

Estimation in Software

R (some common packages):

- `plm`: `plm(..., model = "within")`, `model = "random"`.
- `fixest`: `feols()` for fixed effects.
- `lme4`, `nlme` for mixed (hierarchical) models.

Stata:

- `xtset id time`; `xtreg y x, fe`; `xtreg y x, re`.
- `xtabond` for Arellano-Bond dynamic panel.

Python:

- `linearmodels` package by Kevin Sheppard.

Practical examples

Research Question:

- How does the birth of a first (biological) child affect mothers' life satisfaction?

Inspired by:

- Ludwig & Brüderl (2021) replicate/adapt that approach using the **German Family Panel (pairfam)**.
- Myrskylä & Margolis (2014), who studied parental happiness in SOEP and BHPS.

Data: The German Family Panel (pairfam)

pairfam Overview:

- A nationwide longitudinal study started in 2008/09.
- $\sim$ 12,000 respondents from 3 birth cohorts (1971–73, 1981–83, 1991–93).
- Annual follow-up interviews (Waves 1–11 used, covering 2008–2019).

Key Strengths for This Study:

- Measures both **birth events** and **life satisfaction** prospectively.
- Large sample size, repeated observations enable within-person (FE) designs.

Sample Construction

- **Include** only women who had not given birth before the first pairfam wave (i.e., “never-treated” at baseline).
- **Require** at least two observations in pairfam.
- For **first-time mothers**, censor observations at second pregnancy/birth.
- **Final Sample:**
  - 2,982 women
  - 505 experienced a first birth during the panel
  - Total of 19,996 person-years

Age Range: 15–47 (after recoding a few outliers)

Outcome & Variables

Outcome: Life Satisfaction

- Question: "All in all, how satisfied are you with your life at the moment?"
- 11-point scale (0 = very dissatisfied, 10 = very satisfied).

Treatment Variable: First Birth

- Derived from `nkidsbio` (number of biological children).
- **Time since birth**: 0 (birth year) up to 9 years after birth.
- Group 4+ years into a single category for parsimony.

Controls:

- Age (dummies for 16–47, ref = 15)
- Relationship status (LAT, cohab, marriage, single)
- Subjective health (5-point scale)
- Pregnancy dummy

Estimation: Fixed-Effects Regressions

Why Fixed Effects (FE)?

- Removes time-invariant confounders (personality, stable traits).
- Focuses on **within-person** variation pre- vs. post-birth.

Specifications (Impact Functions):

1. **Step impact**: single parameter for a permanent jump post-birth.
2. **Quadratic impact**: immediate effect + polynomial time trend.
3. **Dummy impact**: separate dummies for each year-since-birth (0, 1, 2, ..., 9).

Cluster-robust standard errors used (Stata 16.1).

Key Idea: An immediate and permanent jump in the outcome once the event (birth) occurs.

- Before birth (t ¡ 0), happiness is at a baseline level.
- At the event time $t = 0$, happiness shifts **upward** by a fixed amount.
- After birth, the outcome remains at this new "permanent" level.

```
Happiness (Y)
    ^
    |                          After birth
    | (Jump)      +---------- (new permanent level)
    |             |
    |             |
    |-------------+---------------------> Time
              t=0 (Birth)
```

Quadratic Impact: Immediate Effect + Polynomial Trend

Key Idea:

- Right after the event (e.g., birth), the outcome jumps by some amount.
- Over time, the effect follows a **quadratic** pattern (a parabola) rather than a single flat line.

Simple Model:

$$Y_{it} = \alpha_i + \underbrace{\beta_0 D_{it}}_{\text{immediate jump}} + \underbrace{\beta_1 (D_{it} \cdot t) + \beta_2 (D_{it} \cdot t^2)}_{\text{quadratic trend}} + \gamma X_{it} + \epsilon_{it}.$$

```
Happiness (Y)
    ^
    |             Quadratic shape after birth
    | (Jump)
    |     +----  <-- immediate effect
    |        \
    |         \
    |----------+---------------- Time -->
             t=0 (Event)
```

Dummy Impact: Flexible, Time-Varying Effects

Key Idea:

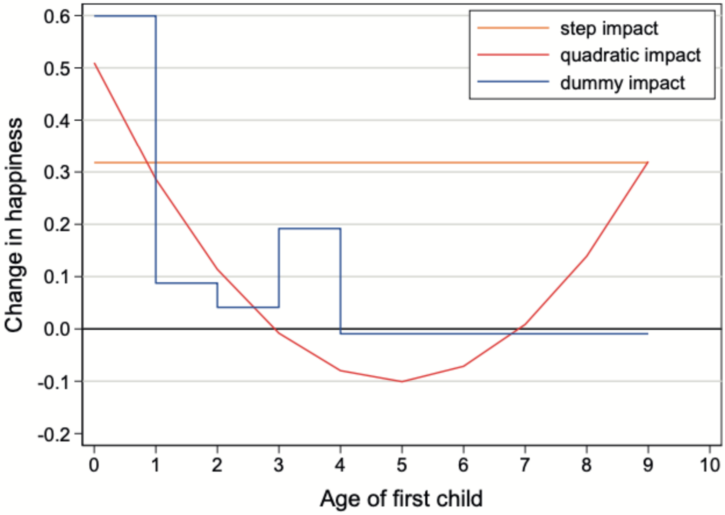
- After an event occurs, each period post-event can have a **different** effect on the outcome.
- **Multiple dummies**: Year 0, Year 1, Year 2, etc., each gets its own coefficient.

$$Y_{it} = \alpha_i \ + \ \sum_{k=0}^{K} \beta_k \, D_{it}^k \ + \ \gamma X_{it} \ + \ \varepsilon_{it},$$

- $D_{it}^k = 1$ if $\text{TimeSinceEvent} = k$, else 0.

```
Outcome change
    ^
    |  Year0
    |     |XX (Beta0)
    |     |    Year1
    |     |      |XXXX (Beta1)
    |     |           Year2
    |     |             |X (Beta2)
    +-----+-----------------------> Time
         t=0  1   2   3   ...
```

Main Results: Comparing Step vs. Quadratic vs. Dummy



Main Results: Comparing Step vs. Quadratic vs. Dummy

Step Impact Function:

- $+0.32$ points on the 0–10 scale *on average*
- Implies a **permanent shift** post-birth

Quadratic Impact Function:

- $+0.51$ immediate jump
- Then a steep decline around 3–5 years later (negative dip), followed by a rebound

Dummy Impact Function:

- $+0.60$ in the first year after birth
- **Rapid fade-out** to near 0 by the second year
- Essentially zero after year 1

Short-Lived "Baby Effect" and Negative Weighting Bias

True Average Effect:

- Dummy approach suggests an average ˜0.086 after birth (when averaging across all post-birth years).

Step Function Overestimates (+0.32):

- *Why?* FE "down-weights" late post-treatment observations in a staggered design, over-focusing on early periods.
- This phenomenon is the **negative weighting bias**.

Conclusion:

- The real effect is **strong but short-lived**.
- Step or quadratic models can give **misleading** long-term estimates.

Take-Home Messages from the Example

- **FE design** controls for stable characteristics, crucial for causal inference in observational data.
- **Childbirth** leads to an **immediate happiness boost** (around $+0.60$), which largely disappears by year 1.
- **Step/Quadratic** approaches can *over- or under-estimate* the true effect due to **negative weighting bias**.
- **Dummy approach** is flexible, revealing a **short-lived** "baby effect."

Myrskylä & Margolis (2014): Overview

Title: *"Happiness: Before and After the Kids"* **Published:** Demography, 2014

- Motivated by **low fertility** concerns and understanding the *subjective well-being* of parents.
- Aims to see **how having children affects parental happiness** in Britain and Germany.
- Uses **longitudinal data** (British Household Panel Survey, German SOEP) and **fixed-effects regressions**.

Motivation and Research Questions

Fertility Trends:

- Declining and postponed childbearing in developed countries.
- Why are many people stopping at one or two children when they desire two or more?

Hypothesis:

- *Subjective well-being* (i.e., happiness) is a key driver of fertility behavior.
- Parents' experiences may inform decisions on having additional children (learning theory).
- Observing others' experiences can also shape timing of births (social learning).

Data and Methods

Data:

- **German SOEP**: 1984–2009, large representative panel.
- **British Household Panel Survey (BHPS)**: 1991–2008.
- Only includes *new* parents during the panel (those who had a child during observation).

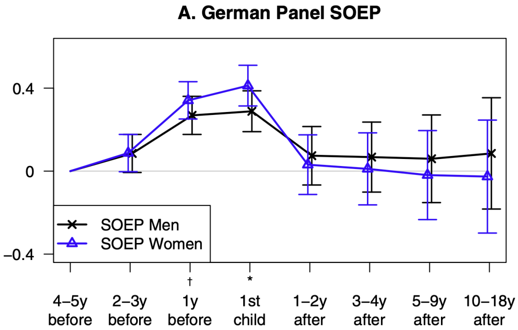
Measures of Happiness:

- SOEP: Life satisfaction scale (0–10).
- BHPS: General happiness scale (rescaled 0–10).

Estimation:

- **Fixed-Effects** regressions to control for unobserved, time-invariant factors (e.g., personality).
- Examine *years before* and *after* birth (up to 18 years).

Results



- Happiness peaks around the time of birth, especially for the first child.
- Happiness generally returns to pre-birth levels a few years after birth.

Key Findings: Overall Happiness Trajectory

- **Temporary spike** in happiness around the birth of the child.
- *Anticipation effect*: Happiness increases *before* the birth, possibly due to partnership formation or planning.
- Post-birth: Happiness often returns to pre-birth levels within a few years (consistent with psychological adaptation).

Key Findings: Overall Happiness Trajectory

OLS vs. FE comparison: FE shows a **stronger pre-birth increase** and less steep post-birth drop. Why?

Key Findings: Overall Happiness Trajectory

OLS vs. FE comparison: FE shows a **stronger pre-birth increase** and less steep post-birth drop. Why?

- When people who are inherently happier (or have unobserved traits related to well-being) are more likely to have children, OLS partially conflates stable individual differences with true before/after changes.

Key Findings: Overall Happiness Trajectory

OLS vs. FE comparison: FE shows a **stronger pre-birth increase** and less steep post-birth drop. Why?

- When people who are inherently happier (or have unobserved traits related to well-being) are more likely to have children, OLS partially conflates stable individual differences with true before/after changes.
- By eliminating these time-invariant confounders with FE, the pre-birth rise becomes more evident (because we are measuring the pure change for each person rather than averaging across people of different happiness baselines).

Key Findings: Overall Happiness Trajectory

OLS vs. FE comparison: FE shows a **stronger pre-birth increase** and less steep post-birth drop. Why?

- When people who are inherently happier (or have unobserved traits related to well-being) are more likely to have children, OLS partially conflates stable individual differences with true before/after changes.
- By eliminating these time-invariant confounders with FE, the pre-birth rise becomes more evident (because we are measuring the pure change for each person rather than averaging across people of different happiness baselines).
- Likewise, the post-birth drop appears less severe once you remove the bias introduced by stable traits that might be correlated with the timing of childbearing.