Lecture 23: Fixed Effects Models

Spring 2025

Logistics

- Continuing on Dependent Errors
- Today: Fixed Effects Models as a solution

Final Project

- Theoretical questions will be very close to assessments
- Practical work on a real-world databse
- Pick a database from https://www.kaggle.com/datasets
- Your database should have a continuous target (linear regression)
- Your objective will be:
  - Compute all the sanity check to verify if assumptions hold
  - Apply procedures for assumption violations
  - Do variable selection / hypothesis testing
  - Explain what is the impact of each feature on the target.
- Objective: Have a github repo you will be able to showcase to future potential employers.
- Date: Full instructions released on May 5th. You will have until Sunday 11th to complete it and send your github repo link and pdf file by email (nbb45@cornell.edu).

Recap

Commonly used linear model where ε_i is an error term:

$$Y_i = b_0 + \sum_{k=1}^{p} b_k X_{i,k} + \varepsilon_i$$

Recap

Commonly used linear model where ε_i is an error term:

$$Y_i = b_0 + \sum_{k=1}^{p} b_k X_{i,k} + \varepsilon_i$$

Assumptions of the model:

- Linear function: $E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + \sum_k b_k x_k$
- Independent Errors: ε_i is independent of ε_j where i and j denote different observations
- Homoscedasticity: The error ε_i has mean 0 and is independent of X_i

Less important assumption:

- Normality: Sometimes, we assume that $\varepsilon_i \sim N(0, \sigma^2)$

The Problem of Dependent Errors

- We saw in the last lectures that dependent errors can arise from:
  - Repeated measures on the same individuals
  - Shared unmeasured factors across observations
  - Observations clustered within groups (regions, classrooms, etc.)

The Problem of Dependent Errors

- We saw in the last lectures that dependent errors can arise from:
  - Repeated measures on the same individuals
  - Shared unmeasured factors across observations
  - Observations clustered within groups (regions, classrooms, etc.)
- Consequences of ignoring dependent errors:
  - Incorrect sampling distribution
  - Underestimated standard errors
  - Inflated Type I error rates (rejecting true null 16% of time at $\alpha = .05$)
  - Potentially decreased power

Fixed Effects Models

Introduction to Fixed Effects

- Fixed effects models are a way to address dependent errors
- Key idea: Account for the source of dependence in the model
- Instead of trying to model the correlation structure of errors, we include variables that absorb the dependence

Introduction to Fixed Effects

- Fixed effects models are a way to address dependent errors
- Key idea: Account for the source of dependence in the model
- Instead of trying to model the correlation structure of errors, we include variables that absorb the dependence
- Used widely in:
  - Panel data analysis
  - Studies with repeated measurements
  - Multi-level data (e.g., students within classrooms)

Fixed Effects

Suppose this is the true model:

$$\begin{aligned}\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{red wine}_{i,k} + \varepsilon_{i,k} \\ &= b_0 + b_1 \text{red wine}_{i,k} + \underbrace{\text{baseline cholesterol}_i + \delta_{i,k}}_{\varepsilon_{i,k}}\end{aligned}$$

- We use **cluster** to refer to the observations which are dependent

Fixed Effects

Suppose this is the true model:

$$\begin{aligned}\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{red wine}_{i,k} + \varepsilon_{i,k} \\ &= b_0 + b_1 \text{red wine}_{i,k} + \underbrace{\text{baseline cholesterol}_i + \delta_{i,k}}_{\varepsilon_{i,k}}\end{aligned}$$

- We use **cluster** to refer to the observations which are dependent
- We could use a categorical variable that encodes belonging to a cluster

$$\begin{aligned}\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{red wine}_i + \sum_{Z=2}^{n} g_Z \text{ClusterZ}_i + \varepsilon_{i,k} \\ &= b_0 + b_1 \text{red wine}_{i,k} + g_i + \varepsilon_{i,k}\end{aligned}$$

where $\text{ClusterZ}_i = 1$ if individual i belong to cluster Z and is 0 otherwise

Fixed Effects

Suppose this is the true model:

$$\begin{aligned}\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{red wine}_{i,k} + \varepsilon_{i,k} \\ &= b_0 + b_1 \text{red wine}_{i,k} + \underbrace{\text{baseline cholesterol}_i + \delta_{i,k}}_{\varepsilon_{i,k}}\end{aligned}$$

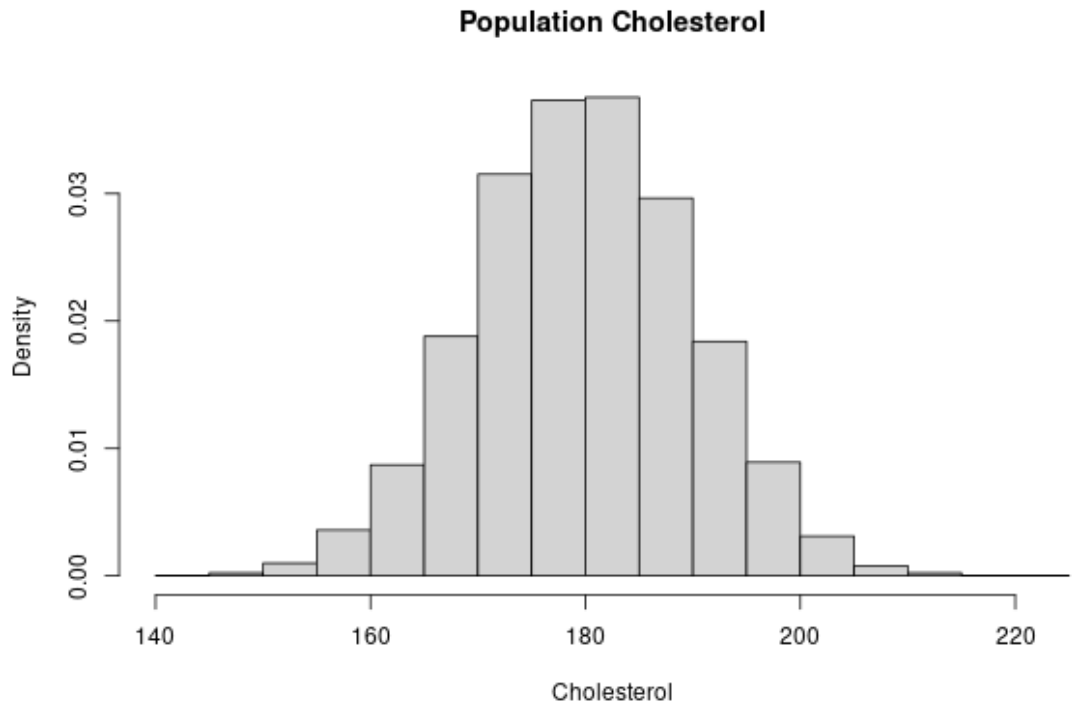
- We use **cluster** to refer to the observations which are dependent
- We could use a categorical variable that encodes belonging to a cluster

$$\begin{aligned}\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{red wine}_i + \sum_{Z=2}^{n} g_Z \text{ClusterZ}_i + \varepsilon_{i,k} \\ &= b_0 + b_1 \text{red wine}_{i,k} + g_i + \varepsilon_{i,k}\end{aligned}$$

  where $\text{ClusterZ}_i = 1$ if individual i belong to cluster Z and is 0 otherwise
- This is called a **fixed effect** because g_Z is a fixed (but unknown) number
- We make no assumptions on g_Z and $\hat{g}_Z$ is completely determined by the data

Samples from a population

- In some settings, the clusters you observe come from a larger population
- Each g_Z can be thought as being drawn from some distribution
- We might even assume a distribution for the g_Z



Types of Fixed Effects Models

1. **Individual Fixed Effects**
   - Controls for time-invariant individual characteristics
   - Example: Unobserved student ability in education studies

Types of Fixed Effects Models

1. **Individual Fixed Effects**
   - Controls for time-invariant individual characteristics
   - Example: Unobserved student ability in education studies
2. **Time Fixed Effects**
   - Controls for factors that vary over time but affect all individuals
   - Example: Economic conditions in different years

Types of Fixed Effects Models

1. **Individual Fixed Effects**
   - Controls for time-invariant individual characteristics
   - Example: Unobserved student ability in education studies
2. **Time Fixed Effects**
   - Controls for factors that vary over time but affect all individuals
   - Example: Economic conditions in different years
3. **Two-way Fixed Effects**
   - Includes both individual and time fixed effects
   - Example: Student-year panel with both student and year effects

Types of Fixed Effects Models

1. **Individual Fixed Effects**
   - Controls for time-invariant individual characteristics
   - Example: Unobserved student ability in education studies
2. **Time Fixed Effects**
   - Controls for factors that vary over time but affect all individuals
   - Example: Economic conditions in different years
3. **Two-way Fixed Effects**
   - Includes both individual and time fixed effects
   - Example: Student-year panel with both student and year effects
4. **Group/Cluster Fixed Effects**
   - Controls for unobserved factors at group level
   - Example: School fixed effects in student data

Example: Panel Data Model

A panel data model with fixed effects:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \gamma_t + \varepsilon_{it}$$

Where:

- y_{it} is the outcome for individual i at time t
- x_{it} is the predictor of interest
- α_i is the individual fixed effect (captures all time-invariant individual characteristics)
- γ_t is the time fixed effect (captures all individual-invariant time effects)
- ε_{it} is the error term (now assumed independent)

Example: Educational Achievement

Suppose we're studying the effect of study time on test scores:

$$\text{Score}_{ict} = \beta_0 + \beta_1 \text{StudyTime}_{ict} + \alpha_i + \gamma_c + \delta_t + \varepsilon_{ict}$$

Where:

- Score_{ict} is the test score for student i in class c at time t
- StudyTime_{ict} is hours spent studying
- α_i is the student fixed effect (student ability)
- γ_c is the classroom fixed effect (teacher quality)
- δ_t is the time fixed effect (test difficulty)

Example: Educational Achievement

Suppose we're studying the effect of study time on test scores:

$$\text{Score}_{ict} = \beta_0 + \beta_1 \text{StudyTime}_{ict} + \alpha_i + \gamma_c + \delta_t + \varepsilon_{ict}$$

Where:

- Score_{ict} is the test score for student i in class c at time t
- StudyTime_{ict} is hours spent studying
- α_i is the student fixed effect (student ability)
- γ_c is the classroom fixed effect (teacher quality)
- δ_t is the time fixed effect (test difficulty)

Without fixed effects, errors would be dependent because:

- Same student measured multiple times
- Students in same classroom share teacher effects

Implementation of Fixed Effects

Method 1: Dummy Variable Approach

- Include a dummy variable for each cluster (except one reference cluster)
- Straightforward but can be computationally intensive with many clusters

Implementation of Fixed Effects

Method 1: Dummy Variable Approach

- Include a dummy variable for each cluster (except one reference cluster)
- Straightforward but can be computationally intensive with many clusters

Method 2: Within Transformation (Demeaning)

- Subtract the cluster mean from each variable:

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

- Removes all time-invariant effects (including fixed effects)
- Regress $\tilde{y}_{it}$ on $\tilde{x}_{it}$

Implementation of Fixed Effects

Method 1: Dummy Variable Approach

- Include a dummy variable for each cluster (except one reference cluster)
- Straightforward but can be computationally intensive with many clusters

Method 2: Within Transformation (Demeaning)

- Subtract the cluster mean from each variable:

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

- Removes all time-invariant effects (including fixed effects)
- Regress $\tilde{y}_{it}$ on $\tilde{x}_{it}$

Method 3: First Differences

- Take the difference between consecutive time periods:

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

- Also removes time-invariant effects

When to Use Fixed Effects

- Use fixed effects when:
  - Observations are naturally clustered
  - You suspect unmeasured cluster-level factors affect the outcome
  - You're interested in within-cluster variation

When to Use Fixed Effects

- Use fixed effects when:
  - Observations are naturally clustered
  - You suspect unmeasured cluster-level factors affect the outcome
  - You're interested in within-cluster variation
- Common applications:
  - Students within classrooms
  - Patients within hospitals
  - Repeated measures on same subjects
  - Observations within geographic regions
  - Panel data with multiple time periods

Advantages of Fixed Effects

- Controls for all time-invariant confounders (observed or unobserved)
- Reduces omitted variable bias
- Accounts for clustered error structure
- Allows correlation between unobserved effects and predictors
- Improves quality of causal inference

Advantages of Fixed Effects

- Controls for all time-invariant confounders (observed or unobserved)
- Reduces omitted variable bias
- Accounts for clustered error structure
- Allows correlation between unobserved effects and predictors
- Improves quality of causal inference

Example: Estimating returns to education

- Ability bias: Unobserved ability affects both education and earnings
- With panel data, individual fixed effects control for time-invariant ability

Limitations of Fixed Effects

- Cannot estimate effects of variables that don't vary within clusters
  - Example: Can't estimate effect of gender with individual fixed effects

Limitations of Fixed Effects

- Cannot estimate effects of variables that don't vary within clusters
  - Example: Can't estimate effect of gender with individual fixed effects
- Requires sufficient within-cluster variation
  - If predictors barely vary within clusters, estimates will be imprecise

Limitations of Fixed Effects

- Cannot estimate effects of variables that don't vary within clusters
  - Example: Can't estimate effect of gender with individual fixed effects
- Requires sufficient within-cluster variation
  - If predictors barely vary within clusters, estimates will be imprecise
- Consumes many degrees of freedom with many clusters
  - Can reduce statistical power

Limitations of Fixed Effects

- Cannot estimate effects of variables that don't vary within clusters
  - Example: Can't estimate effect of gender with individual fixed effects
- Requires sufficient within-cluster variation
  - If predictors barely vary within clusters, estimates will be imprecise
- Consumes many degrees of freedom with many clusters
  - Can reduce statistical power
- Not suitable for all types of dependent errors
  - Doesn't address time-varying correlation structures

Limitations of Fixed Effects

- Cannot estimate effects of variables that don't vary within clusters
  - Example: Can't estimate effect of gender with individual fixed effects
- Requires sufficient within-cluster variation
  - If predictors barely vary within clusters, estimates will be imprecise
- Consumes many degrees of freedom with many clusters
  - Can reduce statistical power
- Not suitable for all types of dependent errors
  - Doesn't address time-varying correlation structures
- Magnifies measurement error bias
  - Within-transformation can increase noise-to-signal ratio

Testing for Dependent Errors

1. **Intraclass Correlation (ICC)**
   - Measures proportion of variance due to clustering
   - $\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$
   - Higher ICC indicates stronger clustering effect

Testing for Dependent Errors

1. **Intraclass Correlation (ICC)**
   - Measures proportion of variance due to clustering
   - $\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$
   - Higher ICC indicates stronger clustering effect
2. **F-test for Fixed Effects**
   - Tests joint significance of fixed effects
   - H_0: All fixed effects coefficients equal zero

Testing for Dependent Errors

1. **Intraclass Correlation (ICC)**
   - Measures proportion of variance due to clustering
   - $\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$
   - Higher ICC indicates stronger clustering effect
2. **F-test for Fixed Effects**
   - Tests joint significance of fixed effects
   - H_0: All fixed effects coefficients equal zero
3. **Durbin-Watson Test**
   - For time series autocorrelation

Testing for Dependent Errors

1. **Intraclass Correlation (ICC)**
   - Measures proportion of variance due to clustering
   - ICC $= \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$
   - Higher ICC indicates stronger clustering effect
2. **F-test for Fixed Effects**
   - Tests joint significance of fixed effects
   - H_0: All fixed effects coefficients equal zero
3. **Durbin-Watson Test**
   - For time series autocorrelation
4. **Breusch-Pagan Lagrange Multiplier Test**
   - Tests for random effects
   - H_0: No significant difference across clusters

Real-World Example

Card (1993): Returns to Education Using College Proximity

https://davidcard.berkeley.edu/papers/geo_var_schooling.pdf

Research question: What is the causal effect of education on earnings?

- Problem: Unobserved ability affects both education and earnings
- Solution: Use proximity to college as instrument + fixed effects for family background

Real-World Example

Card (1993): Returns to Education Using College Proximity

https://davidcard.berkeley.edu/papers/geo_var_schooling.pdf

Research question: What is the causal effect of education on earnings?

- Problem: Unobserved ability affects both education and earnings
- Solution: Use proximity to college as instrument + fixed effects for family background

Model:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{education}_i + \gamma \text{family}_i + \varepsilon_i$$

- Fixed effects for family background control for shared genetic and environmental factors

Summary: Fixed Effects Models

- Fixed effects address dependence in errors by modeling cluster-level effects
- Implementation options:
  - Dummy variables
  - Within transformation (demeaning)
  - First differences

Summary: Fixed Effects Models

- Fixed effects address dependence in errors by modeling cluster-level effects
- Implementation options:
  - Dummy variables
  - Within transformation (demeaning)
  - First differences
- Key advantages:
  - Controls for unobserved cluster-level confounders
  - Reduces omitted variable bias
  - Accounts for dependence in error structure

Summary: Fixed Effects Models

- Fixed effects address dependence in errors by modeling cluster-level effects
- Implementation options:
  - Dummy variables
  - Within transformation (demeaning)
  - First differences
- Key advantages:
  - Controls for unobserved cluster-level confounders
  - Reduces omitted variable bias
  - Accounts for dependence in error structure
- Key limitations:
  - Cannot estimate effects of cluster-invariant variables
  - Requires within-cluster variation
  - May reduce power

Summary: Fixed Effects Models

- Fixed effects address dependence in errors by modeling cluster-level effects
- Implementation options:
  - Dummy variables
  - Within transformation (demeaning)
  - First differences
- Key advantages:
  - Controls for unobserved cluster-level confounders
  - Reduces omitted variable bias
  - Accounts for dependence in error structure
- Key limitations:
  - Cannot estimate effects of cluster-invariant variables
  - Requires within-cluster variation
  - May reduce power
- Alternative: Random effects (more efficient but requires stronger assumptions)

Questions for Discussion

1. In your field of study, what are some examples where fixed effects models would be appropriate?
2. What are the trade-offs between using fixed effects versus including observed control variables?
3. When might random effects be preferable to fixed effects?
4. How would you determine if dependent errors are a concern in your analysis?