Lecture 24: Random Effects Models

Spring 2025

Outline

Introduction to Random Effects

Linear Model Review

Recall the standard linear model:

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

Standard assumptions:

- 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

But what happens when observations are not independent?

Problem: Dependent Observations

- Many real-world datasets have dependent observations
- Examples:
  - Repeated measurements on the same individual
  - Students within classrooms within schools
  - Households within neighborhoods within cities
  - Measurements over time for the same subject
- These dependencies violate the independence assumption
- Ignoring dependency leads to:
  - Biased standard errors
  - Invalid hypothesis tests
  - Incorrect confidence intervals

Example: Repeated Measures

Consider a study measuring the effect of red wine consumption on cholesterol:

$$\text{Cholesterol}_{i,k} = b_0 + b_1 \text{red wine}_{i,k} + \varepsilon_{i,k}$$

Where:

- i denotes individual
- k denotes measurement occasion
- Each individual has multiple measurements

The error term can be decomposed:

$$\varepsilon_{i,k} = \text{baseline cholesterol}_i + \delta_{i,k}$$

Measurements from the same individual are dependent due to the shared baseline!

Solutions for Dependent Data

Three main approaches to handle dependent observations:

1. **Fixed effects models**
   - Add dummy variables for each cluster
2. **Random effects models** (focus of this lecture)
   - Model cluster effects as random variables
3. **Clustered standard errors**
   - Adjust standard errors to account for within-cluster correlations

Random Effects Models

Fixed vs. Random Effects: Conceptual Difference

Fixed Effects

- Separate parameter for each cluster
- Parameters are fixed but unknown constants
- No assumptions about distribution
- Estimates completely determined by data
- Each cluster has its own intercept

Random Effects

- Cluster effects are random variables
- Drawn from a probability distribution
- Typically assumed to be normally distributed
- Shrink estimates toward the mean
- Model the variance of the cluster effects

Mixed Effects Model Specification

A basic mixed effects model:

$$Y_i = b_0 + \sum_k b_k X_{i,k} + G_{Z_i} + \varepsilon_i$$

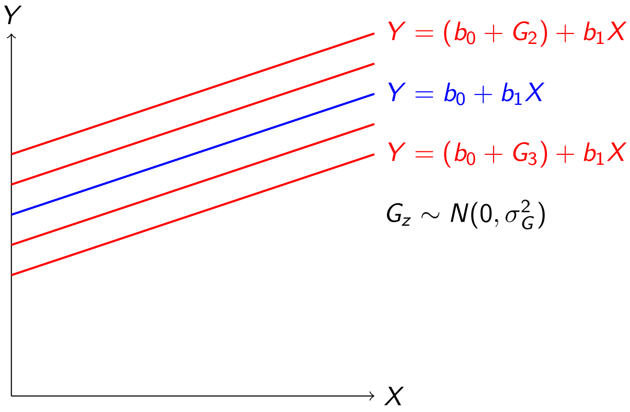
Where:

- b_0, b_k: Fixed effects (same as regular regression)
- G_{Z_i}: Random effect for cluster Z_i
- Z_i: Cluster to which observation i belongs
- ε_i: Individual error term

Assumptions:

- $G_z \sim N(0, \sigma_G^2)$
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
- G_z is independent of ε_i
- G_z is independent of the covariates X

Visualizing Random Intercepts



Each cluster has a different intercept but the same slope, with intercepts drawn from a normal distribution.

Covariance Structure in Random Effects Models

Consider the random intercept model:

$$Y_i = b_0 + \sum_k b_k X_{i,k} + G_{Z_i} + \varepsilon_i$$

Where:

- $G_{Z_i} \sim \mathcal{N}(0, \sigma_G^2)$: random intercept for group Z_i
- $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$: independent error

Then, the conditional covariance between any two observations is:

$$\mathrm{cov}(Y_i, Y_j \mid \mathbf{X}) = \begin{cases} \sigma_G^2 & \text{if } Z_i = Z_j \quad \text{(same group)} \\ 0 & \text{if } Z_i \neq Z_j \quad \text{(different groups)} \end{cases}$$

Key Insight: Random intercepts induce correlation *within* groups, but not *between* groups.

Intra-class Correlation (ICC)

The **intra-class correlation coefficient** (ICC) quantifies the similarity of responses within the same group:

$$\text{ICC} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\varepsilon^2}$$

Where:

- σ_G^2: variance *between groups* (random effect)
- σ_ε^2: variance *within groups* (residual error)

Interpretation:

- $\text{ICC} = 0$: No within-group correlation (no clustering)
- $\text{ICC} = 1$: Perfect within-group correlation (identical values within groups)
- Higher ICC $\Rightarrow$ stronger clustering effect

Use in Practice: A large ICC justifies using random effects to model group-level variability.

Advanced Random Effects Models

Multi-level Random Effects

We can add multiple levels of clustering:

$$Y_i = b_0 + \sum_k b_k X_{i,k} + G_{Z_{i,1}} + G_{Z_{i,2}} + \varepsilon_i$$

Where:

- $Z_{i,1}$: First level cluster (e.g., classroom)
- $Z_{i,2}$: Second level cluster (e.g., school)
- Each level has its own variance component

Example hierarchies:

- Students within classrooms within schools
- Patients within doctors within hospitals
- Employees within departments within companies

Covariance Structure with Multi-level Random Effects

For a model with two levels of clustering:

$$Y_i = b_0 + \sum_k b_k X_{i,k} + G_{Z_{i,1}} + G_{Z_{i,2}} + \varepsilon_i$$

The covariance between observations is:

$$\mathrm{cov}(Y_i, Y_j) = \begin{cases} \sigma^2_{G,1} + \sigma^2_{G,2} & \text{if sharing both clusters} \\ \sigma^2_{G,1} & \text{if sharing only first-level cluster} \\ \sigma^2_{G,2} & \text{if sharing only second-level cluster} \\ 0 & \text{if sharing no clusters} \end{cases}$$

This creates a complex but realistic correlation structure.

Random Slopes

We can allow slope coefficients to vary across clusters:

$$Y_i = b_0 + (b_1 + H_{Z_i})X_{i,1} + G_{Z_i} + \varepsilon_i$$

Where:

- b_1: Fixed (average) effect of X_1
- H_{Z_i}: Random adjustment to the slope for cluster Z_i
- G_{Z_i}: Random intercept for cluster Z_i

Assumptions:

- $H_{Z_i} \sim N(0, \sigma_H^2)$
- H_{Z_i} and G_{Z_i} may be correlated
- Typically modeled as multivariate normal:
  $\begin{pmatrix} G_{Z_i} \\ H_{Z_i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_G^2 & \sigma_{G,H} \\ \sigma_{G,H} & \sigma_H^2 \end{pmatrix}\right)$

Estimation and Implementation

Estimation Methods

- **Restricted Maximum Likelihood (REML)**
  - Most common method
  - Less biased for variance components than ML
  - Developed by Charles Henderson at Cornell (1948-1976)
- **Bayesian estimation**
  - Allows specification of prior distributions
  - Handles small sample sizes better
  - Provides full posterior distributions

Implementation in R

```
# Install and load packages
library(lme4)
library(lmerTest)  # For p-values

# Random intercept model
model1 <- lmer(Cholesterol ~ Wine + (1|Individual),
               data = cholesterol_data)

# Random slope model
model2 <- lmer(Cholesterol ~ Wine + (Wine|Individual),
               data = cholesterol_data)

# Multi-level model (e.g., students in schools)
model3 <- lmer(Score ~ Treatment +
              (1|School) + (1|School:Class),
              data = student_data)

# Model summary
summary(model1)
```

Interpreting Model Output

```
Linear mixed model fit by REML ['lmerMod']
Formula: Cholesterol ~ Wine + (1 | Individual)

REML criterion at convergence: 423.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4563 -0.5972  0.0321  0.6245  2.3301

Random effects:
 Groups     Name        Variance Std.Dev.
 Individual (Intercept) 12.85    3.58
 Residual                4.21    2.05

Number of obs: 80, groups:  Individual, 40

Fixed effects:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  195.32      0.83   235.21  < 2e-16 ***
Wine          -1.45      0.24    -6.04  < 2e-16 ***
```

When to Use Random Effects

When to Use Random Effects

Recommended when:

- Clusters are randomly sampled from a larger population
- You aim to generalize to other clusters not in the sample
- You expect different clusters if the study were repeated
- Cluster effects are uncorrelated with covariates
- You have *many* clusters with *few* observations per cluster

Examples:

- **Education Study:** Measuring the effect of a new curriculum across 100 randomly selected schools to generalize results to all schools in the country
- **Healthcare:** Analyzing recovery times across 50 hospitals to quantify hospital-to-hospital variability in patient care
- **Multicenter Trial:** Estimating drug effectiveness in a trial conducted across many clinics, assuming clinic-specific effects are random
- **Corporate Productivity:** Modeling department-level variation in productivity across a firm with 60 small departments

When to Use Fixed Effects

Recommended when:

- Clusters are unique and not sampled from a larger population
- The identity of each cluster is substantively important
- Cluster effects may be correlated with covariates
- You have *few* clusters with *many* observations per cluster
- You're not interested in generalizing to other clusters

Examples:

- **Policy Evaluation:** Estimating the impact of tax reform on GDP in a fixed set of EU countries — inference is only about these specific countries
- **Leadership Impact:** Measuring how CEO changes affect productivity in 20 large firms over time — firm identity is critical, and CEO changes may correlate with firm covariates
- **Elite Education:** Analyzing student outcomes in five elite schools where the schools are of primary interest
- **Longitudinal Panels:** Estimating wage dynamics using repeated observations from the same individuals — controls for unobserved, time-invariant individual heterogeneity

Key Point: Fixed effects control for cluster-level confounders but do not allow generalization to new units.