Lecture 18: Dependent Errors & Mixed Effects Models

Spring 2025

Lecture Overview

- Key Topics
  - Understanding dependency in error terms
  - Implications for statistical inference
  - Mixed effects models: theory and practice
  - Fixed vs. random effects: when to use each

Linear Models: Key Assumptions

Standard linear model formulation:

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

Core assumptions:

- **Linearity**: $E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + \sum_k b_k x_k$
- **Independent Errors**: ε_i is independent of ε_j for $i \neq j$
- **Homoscedasticity**: Error ε_i has constant variance and is independent of X_i

Secondary assumption:

- **Normality**: Often assumed that $\varepsilon_i \sim N(0, \sigma^2)$ for inference

Today's focus: What happens when the independence assumption is violated?

Group Discussion: Heteroskedasticity

When errors are heteroskedastic (unequal variance across observations):

- What examples have you encountered in your field?
- What statistical problems arise from heteroskedasticity?
  - Inefficient parameter estimates
  - Invalid standard errors
  - Incorrect confidence intervals and hypothesis tests

Dependent Errors: The Problem

Identifying Dependent Errors

Linear model with potentially dependent errors:

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

Key violation we're examining today

Independent Errors Assumption: ε_i should be independent of ε_j for all $i \neq j$

Dependency creates correlation structure: $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$ for some $i \neq j$

Question: What real-world scenarios might create dependent errors?

Example 1: Repeated Measures Design

Research Question: Effect of red wine consumption on cholesterol levels

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

- **Study design**: Each participant (i) experiences all three treatments (k):
  - No wine consumption
  - Moderate wine consumption
  - High wine consumption
- **Dependency source**: Error term can be decomposed as

$$\varepsilon_{i,k} = \underbrace{\text{Individual baseline}_i}_{\text{Subject-specific effect}} + \underbrace{\delta_{i,k}}_{\text{Random noise}}$$

- Observations from the same individual will have correlated errors!

Example 2: Clustered Observations

Research Question: Impact of credit access on food insecurity in Africa

$$\text{FoodInsecurity}_i = b_0 + b_1 \text{CreditAccess}_i + \varepsilon_i$$

- **Data structure**: Individuals (i) nested within geographical regions
- **Dependency source**: Error term includes regional factors

$$\varepsilon_i = \underbrace{\text{Regional factors (drought, conflict)}}_{\text{Shared across individuals in same region}} + \underbrace{\delta_i}_{\text{Individual variation}}$$

- Observations from the same region will have correlated errors!

Example 3: Hierarchical Data

Research Question: Relationship between study time and test scores

$$\text{TestScore}_i = b_0 + b_1 \text{StudyTime}_i + \varepsilon_i$$

- **Data structure**: Students (i) nested within classrooms/teachers
- **Dependency source**: Error term includes teacher quality

$$\varepsilon_i = \underbrace{\text{Teacher quality}}_{\text{Shared across students with same teacher}} + \underbrace{\delta_i}_{\text{Student-specific}}$$

- Students with the same teacher will have correlated errors!

Why Dependent Errors Matter

Core issues with dependent errors

When errors are dependent, standard OLS procedures lead to:

- **Incorrect standard errors**
  - Usually underestimated (sometimes dramatically)
  - Leading to falsely narrow confidence intervals
- **Invalid hypothesis tests**
  - Inflated Type I error rates (rejecting true nulls)
  - Potentially misleading conclusions
- **Loss of statistical efficiency**
  - Dependent observations contribute less information
  - Equivalent to having a smaller effective sample size

Important distinction: Dependent predictors (X variables) are accounted for in OLS. It's dependent *errors* that cause problems.

Impact on Sampling Distribution: Independent Data

With independent data, the sample mean becomes a more precise estimate of the true mean as sample size increases:

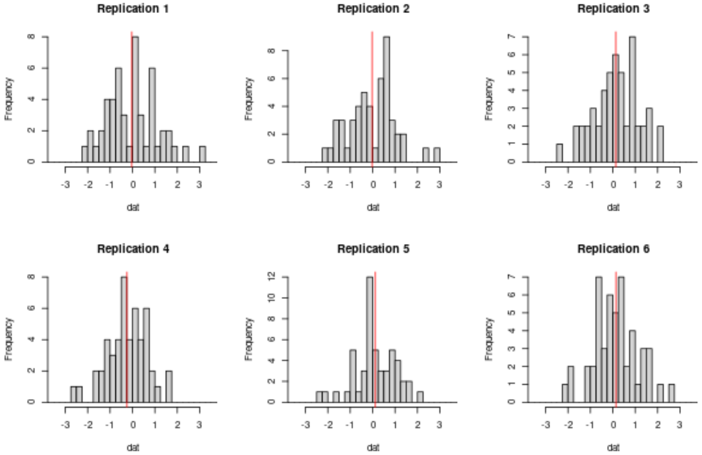

Figure: With independent errors, standard error decreases with $\sqrt{n}$

Impact on Sampling Distribution: Dependent Data

With completely dependent data, adding observations adds minimal information:

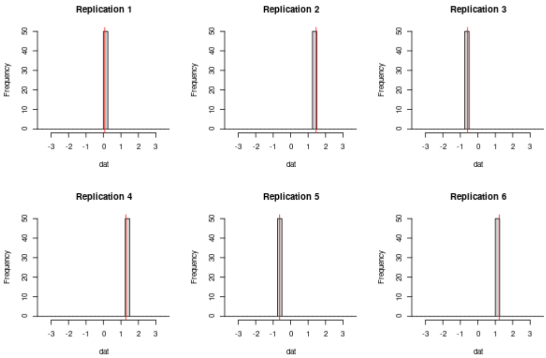

Figure: With dependent errors, standard error decreases more slowly or not at all

Extreme case: When errors are perfectly correlated, $\text{SE}(\bar{X}) = \sigma$ regardless of sample size!

Incorrect Inference with Dependent Errors

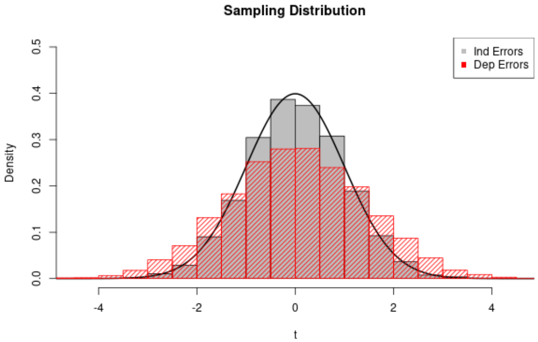

Figure: Type I error rates for nominal 5% test with dependent errors

- When setting $\alpha = .05$, the null hypothesis is rejected $\approx$ 16% of the time with dependent errors
- Three times higher than the intended 5% rate!
- Systematic bias toward finding "significant" results that aren't real

Real-World Example: Longitudinal Data

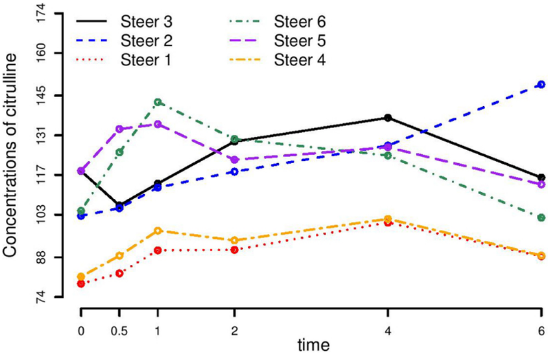

Figure: Repeated measurements on the same subjects (steers) over time (Lee et al., Frontiers in Bioscience-Landmark 2019)

- Measurements on the same subject are clearly correlated
- Treating these as independent would dramatically overstate confidence

Real-World Example: Analysis Methods Matter

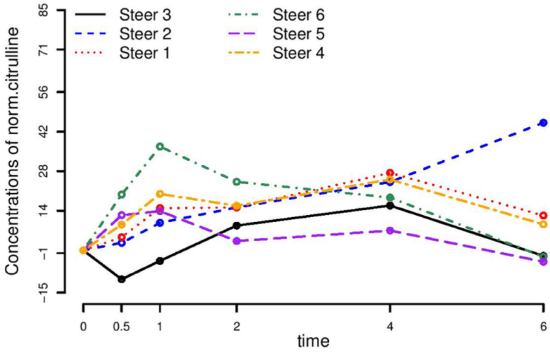

Figure: Different statistical approaches yield different conclusions (Lee et al., Frontiers in Bioscience-Landmark 2019)

- Independent analysis (left): Finds significant differences
- Accounting for dependence (right): More conservative, realistic assessment

Mixed Effects Models: A Solution

Approaches to Handling Dependent Errors

Fixed Effects Approach

- Include dummy variables for each cluster/group
- No distributional assumptions
- Estimates unique effect for each cluster
- Uses only within-cluster variation

Random Effects Approach

- Model cluster effects as random variables
- Assumes distribution (typically normal)
- "Borrows strength" across clusters
- Combines within and between-cluster variation

Mixed Effects Models: Combine fixed effects (for parameters of interest) and random effects (for sources of dependence)

Fixed Effects Approach

For our red wine example:

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

Fixed effects solution:

$$
\begin{aligned}
\text{Cholesterol}_{i,k} &= b_0 + b_1 \text{RedWine}_{i,k} + \sum_{j=2}^{n} g_j \text{Subject}_j + \delta_{i,k} \\
&= b_0 + b_1 \text{RedWine}_{i,k} + g_i + \delta_{i,k}
\end{aligned}
$$

- g_i is a fixed, unknown parameter for each individual
- No assumptions about the distribution of g_i
- Effectively creates a separate intercept for each individual

Conceptualizing Random Effects

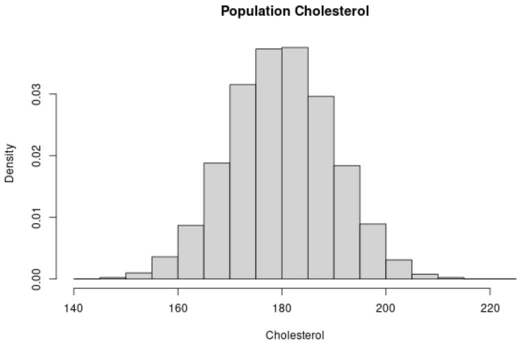

Figure: Individual baselines vary around population mean

- Each individual's baseline cholesterol can be viewed as a draw from a population distribution
- Instead of estimating each individual effect separately, we model the distribution
- Key insight: The individuals in our study represent a sample from a larger population

Random Effects Approach

For our red wine example:

$$\text{Cholesterol}_{i,k} = b_0 + b_1 \text{RedWine}_{i,k} + G_i + \delta_{i,k}$$

- G_i is a **random variable** (not a fixed parameter)
- Typically assume $G_i \sim N(0, \sigma_G^2)$
- G_i and G_j are independent for $i \neq j$
- G_i is independent of predictors $\mathbf{X}_i$
- σ_G^2 represents the variance of individual baselines in the population

Terminology

b_0 and b_1 are **fixed effects** (parameters of interest)
G_i is a **random effect** (accounts for dependence structure)
Together, they form a **mixed effects model**

Applications of Random Effects

Group Discussion: What are examples in your research field where random effects would be appropriate?

Biological Sciences

- Genetic studies with family clusters
- Repeated measurements on same organism
- Plots within experimental fields

Social Sciences

- Students within classrooms/schools
- Voters within districts
- Repeated surveys of same individuals

Medical Research

- Patients within hospitals
- Repeated measures in clinical trials
- Multicenter studies

Economics

- Individuals within households
- Longitudinal income data
- Firms within industries

Covariance Structure in Mixed Effects Models

For the model $Y_i = b_0 + \sum_k b_k X_{i,k} + G_{Z_i} + \varepsilon_i$ where Z_i indicates cluster membership:

Mean structure:

$$E(Y_i \mid \mathbf{X}_i, Z_i) = b_0 + \sum_k b_k X_{i,k}$$

Covariance structure:

$$\begin{aligned}
\mathrm{Cov}(Y_i, Y_j \mid \mathbf{X}) &= E[(Y_i - E(Y_i))(Y_j - E(Y_j))] \\
&= E[(G_{Z_i} + \varepsilon_i)(G_{Z_j} + \varepsilon_j)] \\
&= E(G_{Z_i} G_{Z_j}) + E(G_{Z_i}\varepsilon_j) + E(G_{Z_j}\varepsilon_i) + E(\varepsilon_i\varepsilon_j) \\
&= E(G_{Z_i} G_{Z_j}) \\
&= \begin{cases} \sigma_G^2 & \text{if } Z_i = Z_j \text{ (same cluster)} \\ 0 & \text{if } Z_i \neq Z_j \text{ (different clusters)} \end{cases}
\end{aligned}$$

Key insight: The model explicitly accounts for within-cluster correlation

Visualizing Mixed Effects Models

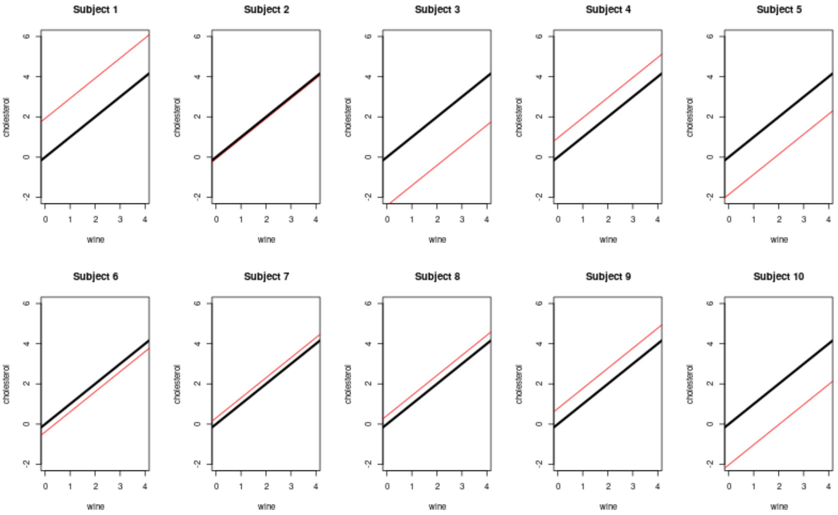

Figure: Random intercept model: Each cluster has its own intercept drawn from a distribution

Fixed vs. Random Effects: Which to Choose?

When to use Fixed Effects

- Primary interest is in specific cluster differences
- Limited number of clusters
- Many observations per cluster
- Potential correlation between cluster effects and predictors
- No need to generalize beyond observed clusters

When to use Random Effects

- Interest in population-level inference
- Many clusters, few observations per cluster
- Cluster effects independent of predictors
- Need for more statistical efficiency
- Want to estimate variance components
- Need to predict effects for new clusters

Hausman test: Formal test for deciding between fixed and random effects

Advantages of Mixed Effects Models

- **Improved precision:** By imposing structure on cluster effects ($G_Z \sim N(0, \sigma_G^2)$), we gain statistical efficiency
- **Valid inference:** Accounts for dependence structure in the data
- **Variance component estimation:** Quantifies between-cluster vs. within-cluster variability
- **Flexible modeling:** Can handle unbalanced designs, missing data
- **Prediction for new clusters:** Can predict outcomes for clusters not in original data
- **Longitudinal analysis:** Natural framework for repeated measures data

Critical assumption: Random effects G_Z are independent of predictors $\mathbf{X}$
(If violated, can lead to biased fixed effect estimates)

Computational Approaches

- **Restricted Maximum Likelihood (REML)**: Standard approach for estimating mixed models
  - Developed by Charles Henderson at Cornell Animal Science (1948-1976)
  - Reduces bias in variance component estimation compared to ML
  - Adjusts for uncertainty in fixed effect estimation
- **Henderson's contributions**:
  - Developed Best Linear Unbiased Prediction (BLUP)
  - Revolutionized animal breeding programs
  - Methods now standard across many scientific fields

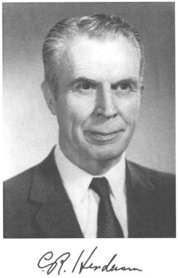

Figure: Charles Henderson, Cornell University

Summary: Key Points

1. **Dependent errors** arise when observations share unmeasured factors affecting the outcome
2. Ignoring dependence leads to **invalid statistical inference** (typically overconfidence)
3. **Mixed effects models** provide a flexible framework for handling dependent data by:
   - Modeling cluster-specific effects as random variables
   - Explicitly accounting for within-cluster correlation
   - Combining fixed effects (parameters of interest) with random effects (source of dependence)
4. **Choice between fixed and random effects** depends on:
   - Research question and inference goals
   - Data structure and sample sizes
   - Whether cluster effects correlate with predictors

Appendix: Random Slopes Models

We can extend mixed models by allowing slopes to vary randomly too:

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

- $G_{Z_i} \sim N(0, \sigma_G^2)$ is the random intercept
- $H_{Z_i,k} \sim N(0, \sigma_H^2)$ is the random slope for predictor k
- Can model correlation between random intercepts and slopes

Covariance structure:

$$\mathrm{Cov}(Y_i, Y_j \mid \mathbf{X}) = \begin{cases} \sigma_G^2 + \sigma_H^2 x_{i,k} x_{j,k} & \text{if } Z_i = Z_j \\ 0 & \text{if } Z_i \neq Z_j \end{cases}$$

Implication: Correlation between observations from same cluster depends on predictor values

Appendix: Random Slopes Visualization

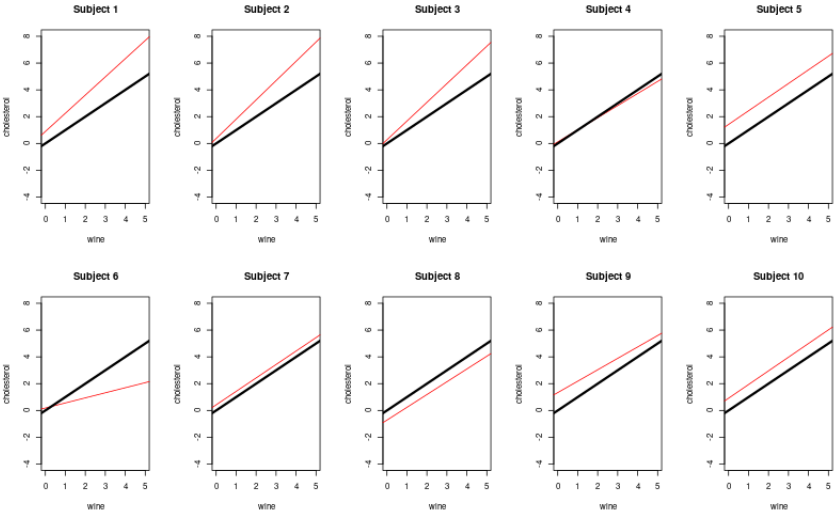

Figure: Random slopes model: Both intercepts and slopes vary by cluster

Appendix: Simulation Results - Balanced Design

$$Y_{i,k} = b_1 X_{i,k,1} + G_i + \varepsilon_{i,k}$$

Setup:

- 40 clusters, 2 observations per cluster (total n=80)
- True $b_1 = 1$
- $G_i \sim N(0,1)$ (random cluster effect)
- $\varepsilon_{i,k} \sim N(0,1)$ (independent error)

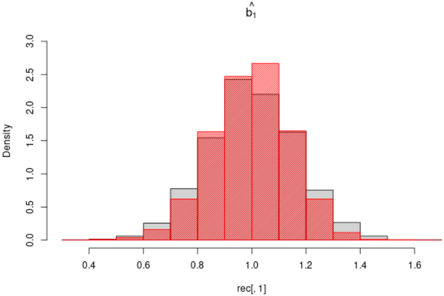

Figure: Both fixed effects (SD=0.16) and random effects (SD=0.14) models produce unbiased estimates with proper coverage of 95% CIs

Appendix: Simulation Results - Few Large Clusters

$$Y_{i,k} = b_1 X_{i,k,1} + G_i + \varepsilon_{i,k}$$

Setup:

- 5 clusters, 16 observations per cluster (total n=80)
- True $b_1 = 1$
- $G_i \sim N(0, 1)$ (random cluster effect)
- $\varepsilon_{i,k} \sim N(0, 1)$ (independent error)

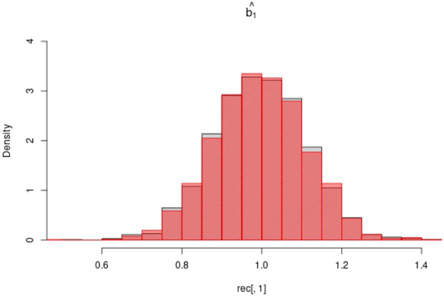

Figure: With few large clusters, both approaches perform similarly (SD=0.13) with proper coverage

Appendix: When Random Effects Assumptions Are Violated

$$Y_{i,k} = b_1 X_{i,k,1} + G_i + \varepsilon_{i,k}$$

Setup:

- 20 clusters, 2 observations per cluster (total n=40)
- True $b_1 = 1$
- $G_i \sim N(0, 1)$ but **depends on** X (violating key assumption)
- $\varepsilon_{i,k} \sim N(0, 1)$ (independent error)

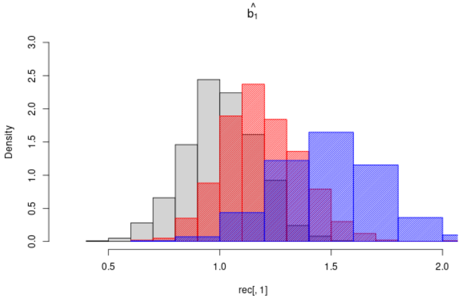

Figure: Fixed effects model maintains proper coverage (95%), while random effects model coverage drops to 88%

Appendix: Clustered Standard Errors

Alternative approach: Use regular regression but adjust standard errors

- **Key idea:** Allow for arbitrary correlation structure within clusters
- Keep the same point estimates as OLS but correct the variance
- **Sandwich formula:**

$$\mathrm{Var}(\hat{\mathbf{b}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

- $\mathbf{W}$ represents the error correlation structure

Advantages

- Robust to misspecification of correlation structure
- No assumption about random effects distribution
- Simple implementation in most statistical software

Disadvantages

- Requires large number of clusters (rule of thumb: 50+)
- Less efficient than correctly specified mixed models
- Cannot estimate variance components

Appendix: Cluster Correlation Structures

Common correlation structures for clustered data:

1. **Compound symmetry (exchangeable):**

$$W_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \text{ but in same cluster} \\ 0 & \text{if } i, j \text{ in different clusters} \end{cases}$$

2. **Autoregressive:** For time series or spatial data

$$W_{i,j} = \rho^{|i-j|} \text{ for observations in same cluster}$$

3. **Distance-based:** For spatial data

$$W_{i,j} = f(\text{dist}(i,j))$$

e.g., $f(\text{dist}(i,j)) = \frac{1}{\text{dist}(i,j)^2}$ or $f(\text{dist}(i,j)) = e^{-\lambda \cdot \text{dist}(i,j)}$

Software implementation: Specify correlation structure in mixed models or use cluster-robust standard errors