Week 9: Mixed Models

Francisco Richter and Ernst Wit

Introduction to Data Science (MSc)

1 Introduction to Mixed Models

Mixed models, also known as mixed-effects models or hierarchical models, represent a fundamental advancement in statistical modeling that addresses the limitations of classical linear regression when dealing with complex data structures. These models explicitly account for correlation patterns that arise naturally in many data science applications, such as repeated measurements on the same subjects, observations clustered within groups, or hierarchical data structures where lower-level units are nested within higher-level units.

The theoretical foundation of mixed models rests on the recognition that many datasets violate the independence assumption of classical regression. When observations are correlated, standard regression methods produce inefficient estimates and incorrect standard errors, leading to invalid statistical inference. Mixed models provide a principled framework for modeling these correlation structures while maintaining computational tractability and interpretability.

The distinction between fixed and random effects forms the conceptual core of mixed modeling. Fixed effects represent population-level parameters that are constant across all subjects or clusters in the population. These are the primary parameters of scientific interest and represent systematic relationships that we wish to estimate and interpret. Random effects, in contrast, represent subject-specific or cluster-specific deviations from the population average, capturing unobserved heterogeneity that creates correlation among observations within the same group.

2 Mathematical Framework for Mixed Models

2.1 General Linear Mixed Model Specification

The general linear mixed model provides a unified framework for analyzing correlated data structures. The model can be expressed in matrix notation as a natural extension of the classical linear model, incorporating both fixed effects that apply to the entire population and random effects that vary across subjects or clusters.

Definition 2.1 (Linear Mixed Model)**.** *The general linear mixed model is specified as:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ *is the* $n \times 1$ *response vector containing all observations,* $\mathbf{X}$ *is the* $n \times p$ *design matrix for fixed effects,* $\boldsymbol{\beta}$ *is the* $p \times 1$ *vector of fixed effects parameters,* $\mathbf{Z}$ *is the* $n \times q$ *design matrix for random effects,* $\mathbf{u}$ *is the* $q \times 1$ *vector of random effects with* $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$*, and* $\boldsymbol{\epsilon}$ *is the* $n \times 1$ *vector of residual errors with* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$*.*

The covariance matrices $\mathbf{G}$ and $\mathbf{R}$ encode the correlation structure of the data. The matrix $\mathbf{G}$ captures the variability and correlation among random effects, while $\mathbf{R}$ represents the residual covariance structure. In many applications, $\mathbf{R} = \sigma^2\mathbf{I}$, assuming independent and identically distributed residual errors, though more complex structures can be accommodated.

2.2 Marginal and Conditional Distributions

The mixed model specification leads to a natural decomposition of the response variable into marginal and conditional distributions. The marginal distribution of $\mathbf{y}$ integrates over the random effects, while the conditional distribution conditions on specific values of the random effects.

Theorem 2.1 (Marginal Distribution)**.** *Under the linear mixed model, the marginal distribution of* $\mathbf{y}$ *is:*

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where the marginal covariance matrix is:

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$$

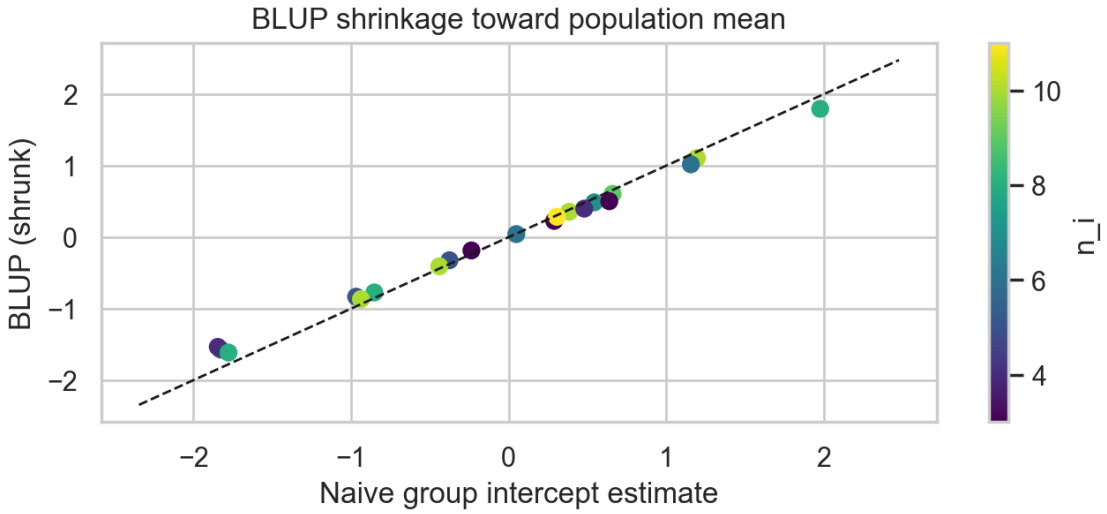

Figure 1: BLUP shrinkage: naive group intercept estimates versus shrunk BLUPs; color denotes group sample size.

This marginal covariance structure $\mathbf{V}$ captures the complex correlation patterns induced by the random effects. The term $\mathbf{Z}\mathbf{G}\mathbf{Z}^T$ represents the contribution of random effects to the overall covariance, while $\mathbf{R}$ represents the residual covariance. This decomposition allows mixed models to accommodate a wide variety of correlation structures that would be impossible to specify directly.

The conditional distribution provides insight into the behavior of individual subjects or clusters. Given specific values of the random effects $\mathbf{u}$, the conditional distribution of $\mathbf{y}$ is:

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R})$$

This conditional perspective reveals how individual subjects deviate from the population average through their specific random effects values.

3 Random Intercept Models

3.1 Theoretical Foundation

The random intercept model represents the simplest and most commonly used mixed model specification. This model assumes that subjects or clusters differ in their baseline levels but share common slopes for all covariates. The mathematical elegance of this model lies in its ability to capture between-subject heterogeneity while maintaining a parsimonious parameter structure.

For longitudinal data with n_i observations on subject i, the random intercept model is specified as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \epsilon_{ij}$$

where $i = 1, \ldots, m$ indexes subjects and $j = 1, \ldots, n_i$ indexes observations within subjects. The random intercept $u_i \sim N(0, \sigma_u^2)$ represents the subject-specific deviation from the population intercept β_0, while $\epsilon_{ij} \sim N(0, \sigma^2)$ represents the residual error for observation j on subject i.

The independence assumption requires that u_i and ϵ_{ij} are mutually independent and independent across subjects and observations. This assumption is crucial for the validity of the model and the correctness of statistical inference.

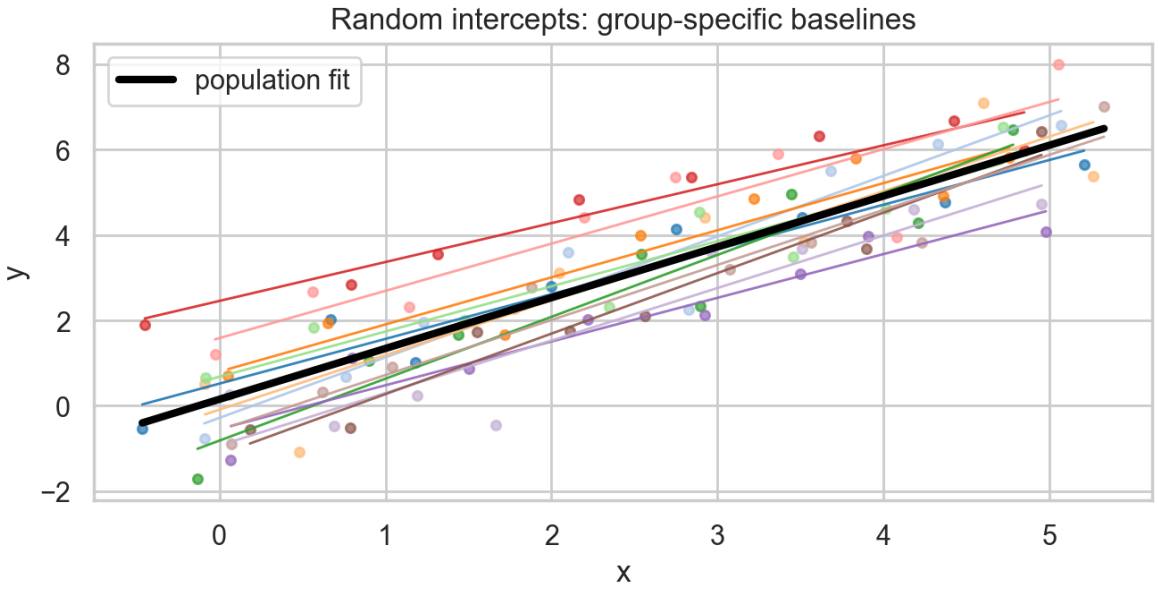

Figure 2: Random intercepts: group-specific baselines with a common population trend.

3.2 Covariance Structure and Intraclass Correlation

The random intercept model induces a specific covariance structure that reflects the clustering of observations within subjects. For any two observations Y_{ij} and Y_{ik} on the same subject i, the covariance is:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_u^2$$

while observations on different subjects are uncorrelated:

$$\text{Cov}(Y_{ij}, Y_{\ell k}) = 0 \quad \text{for } i \neq \ell$$

This covariance structure leads to the concept of intraclass correlation, which quantifies the proportion of total variance attributable to between-subject differences.

Definition 3.1 (Intraclass Correlation Coefficient)**.** *The intraclass correlation coefficient (ICC) is defined as:*

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

where σ_u^2 is the between-subject variance and σ^2 is the within-subject variance.

The ICC has a natural interpretation as the correlation between any two observations from the same subject. Values of ρ close to 1 indicate strong clustering, where observations within subjects are highly similar, while values close to 0 suggest weak clustering, where within-subject correlation is minimal. The ICC also represents the proportion of total variance explained by subject-level differences, making it a key measure for understanding the importance of the clustering structure.

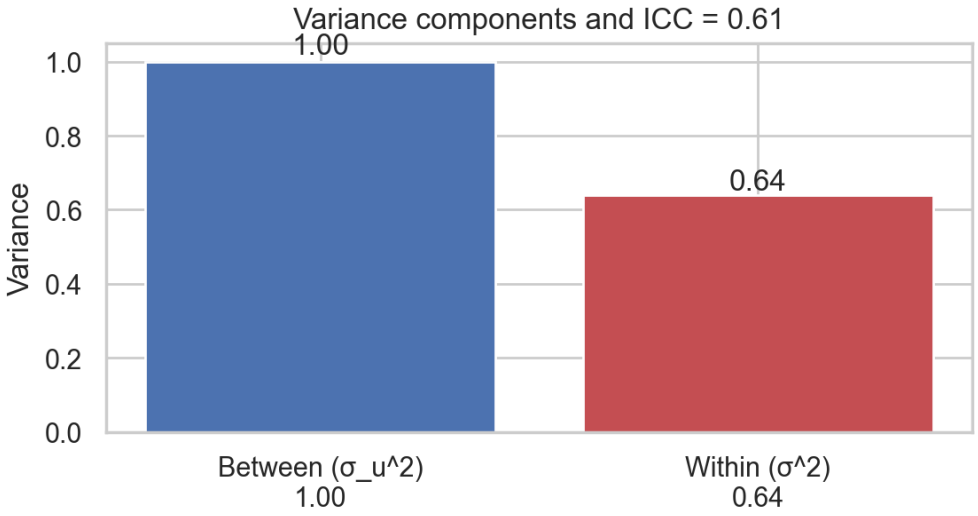

Figure 3: Variance components (between vs within) with the corresponding ICC value.

4 Random Slope Models

4.1 Extending to Random Slopes

While random intercept models capture between-subject differences in baseline levels, they assume that all subjects respond identically to covariates. Random slope models relax this assumption, allowing the effects of covariates to vary across subjects. This extension is particularly important in longitudinal studies where subjects may have different rates of change over time.

The random intercept and slope model for a single covariate is specified as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0i} + u_{1i} X_{ij} + \epsilon_{ij}$$

where u_{0i} is the random intercept and u_{1i} is the random slope for subject i. The random effects follow a bivariate normal distribution:

$$\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix} \right)$$

The covariance parameter σ_{u01} captures the correlation between random intercepts and slopes, providing insight into whether subjects with higher baseline values tend to have steeper or flatter slopes.

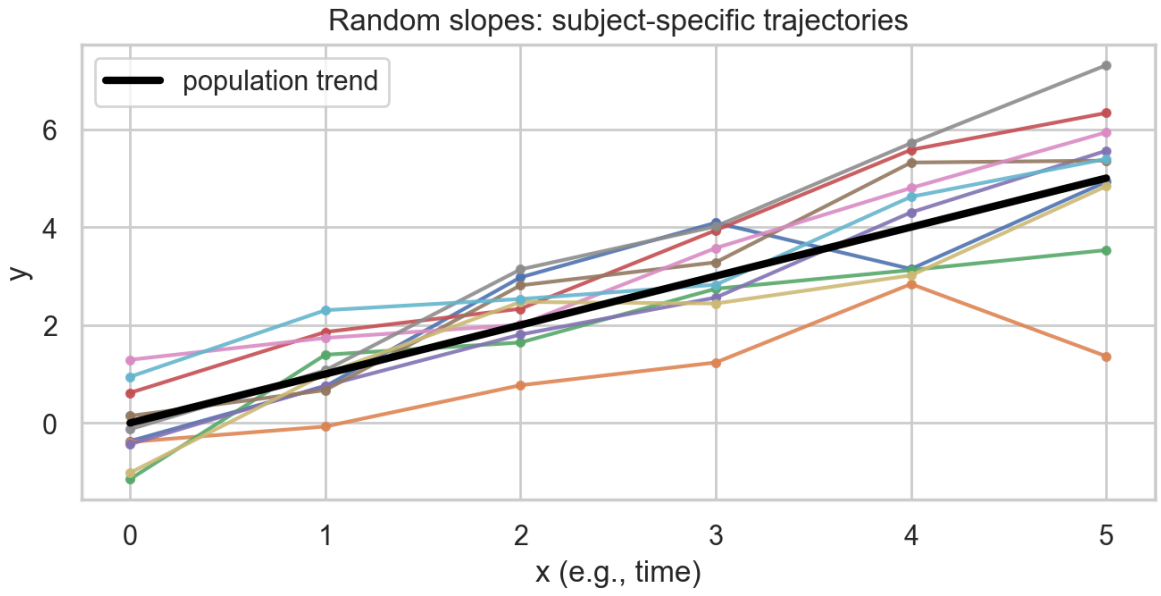

Figure 4: Random slopes: subject-specific trajectories around a population average line.

4.2 Covariance Structure for Random Slopes

The random slope model induces a more complex covariance structure that depends on the values of the covariate. For observations Y_{ij} and Y_{ik} on the same subject with covariate values X_{ij} and X_{ik}, the covariance is:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \sigma^2_{u0} + \sigma_{u01}(X_{ij} + X_{ik}) + \sigma^2_{u1} X_{ij} X_{ik}$$

This structure allows for heteroscedasticity and non-constant correlation, reflecting the reality that observations may be more or less correlated depending on their covariate values. When X represents time, this structure captures the intuitive notion that observations closer in time are more highly correlated than observations farther apart.

5 Maximum Likelihood and REML Estimation

5.1 Maximum Likelihood Estimation

Maximum likelihood estimation for mixed models requires maximizing the marginal likelihood obtained by integrating over the random effects. The marginal likelihood for the linear mixed model is:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

where $\boldsymbol{\theta}$ represents the variance components that determine $\mathbf{V}$.

The computational challenge lies in the evaluation of the determinant and inverse of the potentially large covariance matrix $\mathbf{V}$. Modern algorithms exploit the structure of mixed models to make these computations tractable through techniques such as the sweep operator and Cholesky decomposition.

5.2 Restricted Maximum Likelihood

Restricted Maximum Likelihood (REML) addresses the bias in ML estimation of variance components that arises from not accounting for the loss of degrees of freedom due to fixed effects estimation. REML estimates variance components by maximizing the likelihood of error contrasts that are orthogonal to the fixed effects.

Definition 5.1 (REML Criterion)**.** *The REML log-likelihood is:*

$$\ell_{REML}(\boldsymbol{\theta}) = -\frac{1}{2}\left[\log|\mathbf{V}| + \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right]$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ *is the generalized least squares estimator.*

The additional term $\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|$ in the REML criterion accounts for the uncertainty in fixed effects estimation, leading to less biased estimates of variance components, particularly in small samples.

5.3 Best Linear Unbiased Predictors

The prediction of random effects requires balancing the information from individual subjects against the population average. This balance is achieved through Best Linear Unbiased Predictors (BLUPs), which provide optimal predictions under the mixed model framework.

Theorem 5.1 (BLUPs for Random Effects)**.** *The best linear unbiased predictors of the random effects are:*

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ *is the BLUE of* $\boldsymbol{\beta}$*.*

BLUPs exhibit a shrinkage property, where individual estimates are pulled toward the population mean. The amount of shrinkage depends on the reliability of individual estimates, with less reliable estimates (based on fewer observations or higher variability) being shrunk more heavily toward the population average. This shrinkage property makes BLUPs particularly valuable for prediction in situations where individual-level estimates would be unreliable.

6 Model Selection and Diagnostics

6.1 Information Criteria for Mixed Models

Model selection in mixed models requires careful consideration of the estimation method used. The choice between ML and REML affects the calculation of information criteria and their interpretation.

For comparing models with different fixed effects structures, ML estimation should be used because REML estimates are not directly comparable across different fixed effects specifications. For comparing models with the same fixed effects but different random effects structures, REML is preferred because it provides less biased estimates of variance components.

The effective number of parameters in mixed models is not always clear, particularly when random effects are involved. The conditional AIC (cAIC) and other modifications have been proposed to address this issue, though the standard AIC and BIC remain widely used in practice.

6.2 Likelihood Ratio Testing

Likelihood ratio tests provide a formal framework for comparing nested mixed models. However, testing variance components presents special challenges because the null hypothesis often places parameters on the boundary of the parameter space.

Theorem 6.1 (LRT for Variance Components)**.** *For testing $H_0 : \sigma_u^2 = 0$ versus $H_1 : \sigma_u^2 > 0$, the likelihood ratio test statistic:*

$$LRT = 2[\ell_1 - \ell_0]$$

has an asymptotic distribution that is a mixture of χ^2 distributions: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, where χ_0^2 represents a point mass at zero.

This non-standard asymptotic distribution arises because the null hypothesis places the variance parameter on the boundary of the parameter space (at zero), violating the regularity conditions for standard likelihood ratio theory.

6.3 Residual Analysis

Mixed models generate multiple types of residuals, each providing different insights into model adequacy. Marginal residuals $\mathbf{r}_m = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ reflect deviations from the population average, while conditional residuals $\mathbf{r}_c = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}$ reflect deviations from subject-specific predictions.

Standardized residuals account for the heteroscedasticity induced by the mixed model structure. The standardization requires careful attention to the covariance structure, as residuals from the same subject are correlated even after fitting the model.

7 Generalized Linear Mixed Models

7.1 Extension to Non-Normal Responses

Generalized Linear Mixed Models (GLMMs) extend the mixed model framework to accommodate non-normal response variables by incorporating a link function and specifying an exponential family distribution for the response.

Definition 7.1 (GLMM)**.** *The generalized linear mixed model is specified as:*

$$g(E[Y_{ij}|\mathbf{u}_i]) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i$$

where $g(\cdot)$ is the link function, Y_{ij} follows an exponential family distribution, and $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$.

Common GLMMs include logistic mixed models for binary responses, Poisson mixed models for count data, and gamma mixed models for positive continuous responses. Each specification requires careful consideration of the appropriate link function and distributional assumptions.

7.2 Estimation Challenges

GLMMs present significant computational challenges because the marginal likelihood involves intractable integrals over the random effects distribution. Various approximation methods have been developed to address this challenge.

The Laplace approximation provides a second-order Taylor expansion around the mode of the integrand, yielding a Gaussian approximation to the integral. Adaptive Gaussian quadrature uses numerical integration with strategically chosen quadrature points. Markov Chain Monte Carlo methods provide exact sampling-based inference but at greater computational cost.

Each method involves trade-offs between computational efficiency and approximation accuracy. The choice of method depends on the complexity of the model, the sample size, and the required precision of the estimates.

8 Advanced Applications in Data Science

8.1 Longitudinal Data Analysis

Mixed models provide the natural framework for analyzing longitudinal data, where repeated measurements are collected on the same subjects over time. The correlation structure induced by repeated measurements violates the independence assumptions of standard regression methods, making mixed models essential for valid inference.

In clinical trials, mixed models accommodate missing data under the missing at random assumption, providing more efficient and less biased estimates than complete case analysis. The ability to include all available data, even from subjects with incomplete follow-up, represents a major advantage over alternative approaches.

Web analytics applications often involve tracking user behavior over time, with natural clustering at the user level. Mixed models can accommodate user-specific baseline levels and time trends, providing insights into both population-level patterns and individual heterogeneity.

8.2 Hierarchical Data Structures

Educational research frequently involves students nested within classrooms within schools, creating multiple levels of clustering. Mixed models can accommodate this hierarchical structure through multiple levels of random effects, allowing for variation at each level of the hierarchy.

In business applications, sales data might be clustered within sales representatives within regions within companies. Mixed models can decompose the total variation into components attributable to each level of the hierarchy, providing insights into where interventions might be most effective.

8.3 Machine Learning Integration

Modern machine learning applications increasingly recognize the value of mixed models for handling structured data. Personalized recommendation systems can incorporate user-specific random effects to capture individual preferences while learning population-level patterns.

Multi-task learning problems, where related tasks are learned simultaneously, can be formulated as mixed models with task-specific random effects. This approach allows for information sharing across tasks while accommodating task-specific differences.

Transfer learning applications can use mixed models to adapt population-level models to new domains through domain-specific random effects, providing a principled approach to domain adaptation.

9 Summary

Mixed models represent a fundamental advancement in statistical methodology that addresses the limitations of classical regression when dealing with correlated data structures. The mathematical framework provides a principled approach to modeling complex correlation patterns while maintaining interpretability and computational tractability.

The distinction between fixed and random effects provides both conceptual clarity and practical flexibility. Fixed effects capture population-level relationships of primary scientific interest, while random effects account for unobserved heterogeneity that creates correlation among observations. This decomposition allows mixed models to accommodate a wide variety of data structures commonly encountered in modern data science applications.

The estimation theory for mixed models, including maximum likelihood and REML approaches, provides the foundation for statistical inference. The development of BLUPs offers optimal prediction of random effects through a principled shrinkage approach that balances individual information against population averages.

Model selection and diagnostic procedures adapted for mixed models ensure that these powerful methods can be applied reliably in practice. The extension to generalized linear mixed models broadens the applicability to non-normal response variables, though at the cost of increased computational complexity.

The applications in longitudinal data analysis, hierarchical data structures, and modern machine learning demonstrate the continued relevance and importance of mixed models in contemporary data science. As data structures become increasingly complex and sample sizes continue to grow, mixed models provide essential tools for extracting meaningful insights while accounting for the correlation structures inherent in real-world data.

10 Exercises

1. **Intraclass Correlation Analysis:** A study examines student test scores across different schools with $\sigma_u^2 = 25$ (between-school variance) and $\sigma^2 = 100$ (within-school variance). Calculate the intraclass correlation coefficient and provide a detailed interpretation of its meaning for educational policy. Discuss how this ICC value would influence decisions about school-level versus student-level interventions.

2. **Random Effects Covariance Structure:** For a random intercept and slope model with time as the covariate, derive the covariance between observations at times t_1 and t_2 for the same subject. Show how this covariance depends on the time points and explain the implications for study design in longitudinal research.

3. **REML versus ML Estimation:** Explain the theoretical basis for why REML provides less biased estimates of variance components compared to ML. Construct a simple example with small sample size to demonstrate this bias numerically, and discuss when you might prefer ML estimation despite this bias.

4. **Model Selection Strategy:** Design a systematic approach for selecting the random effects structure in a longitudinal study with multiple covariates. Include considerations of both statistical criteria (AIC, BIC, likelihood ratio tests) and practical interpretability. Discuss how the choice between ML and REML affects your selection strategy.

5. **BLUP Properties:** Prove that BLUPs exhibit shrinkage toward the population mean and derive the shrinkage factor for a simple random intercept model. Explain how the amount of shrinkage depends on the number of observations per subject and the intraclass correlation coefficient.

6. **Boundary Testing Problem:** For testing $H_0 : \sigma_u^2 = 0$ in a random intercept model, explain why the standard chi-square distribution doesn't apply and derive the correct mixture distribution. Provide practical guidance for interpreting p-values in this context.

7. **GLMM Approximation Methods:** Compare the Laplace approximation and adaptive Gaussian quadrature for a logistic mixed model. Discuss the trade-offs between computational efficiency and accuracy, and provide guidelines for choosing between these methods in practice.

8. **Hierarchical Data Application:** Design a mixed model for analyzing customer satisfaction scores where customers are nested within stores within regions. Specify the complete model including all random effects, discuss the interpretation of variance components at each level, and explain how you would test the significance of each level of clustering.