Advanced Methods in Biometry

- Mixed Models-

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Outline

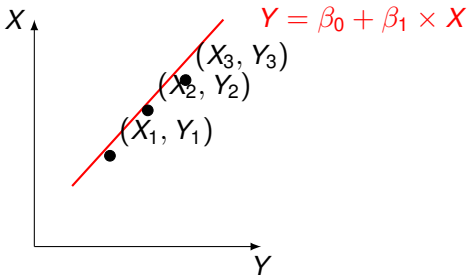
Introduction

- Mixed models, also known as hierarchical or multilevel models, extend linear models by incorporating both fixed and random effects.
- They are particularly useful when dealing with clustered or nested data, where observations are not independent.

Motivation

Linear Regression

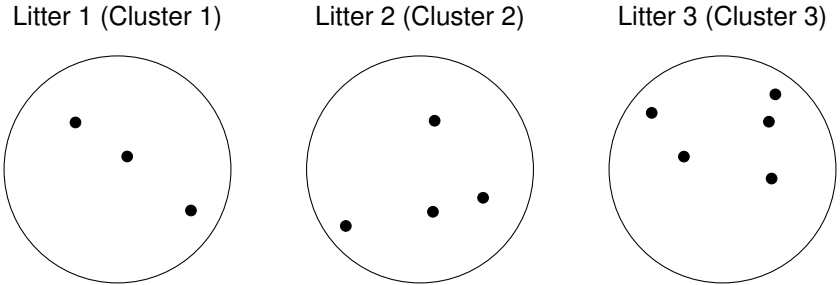
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $\epsilon_i \sim N(0, \sigma^2)$
- **Response values (errors) are independent**
- Are data always independent?



Motivation

- **Example 1**
  - Suppose we are studying the growth of plants over time.
  - Each plant has its own growth trajectory, but there may also be similarities among plants due to shared environmental conditions.
- **Example 2**
  - In education research, students within the same school may have similar academic performance due to shared school-level factors.
  - However, students may also have unique characteristics that influence their achievement.
- **Example 3**
  - Birthweights of rats grouped by litter from different mothers

Data Structure Illustration: Clustered Data



- Linear regression does not account for the data structure.
- Ignoring ends in biased estimates
- **Data within a cluster may be dependent**
- Often we deal with mixtures of independent and dependent data

Mixed = Fixed Effects and Random Effects

Fixed Effects

- Fixed model parameter
- Allow statements on **associations in general** between independent and dependent variables $\Rightarrow$ regression coefficients
- Interpretation as in “normal” regression models

Random Effects

- Random variable
- Account for dependency
- Account for heterogeneity between clusters (independency on higher levels)
- **Variance estimation on each level+residual variance**
- Random intercept / random slope

Fixed Effects versus Random Effects

Fixed Effects

- Effects we are interested in (research question)
- Would be chosen again for another study
- **Examples**
  - Treatments (placebo, verum)
  - Dosages

Random Effects

- Not directly of interest
- Randomly chosen
- Different levels would be chosen for another study
- **Examples**
  - Different mother animals would be chosen for another study
  - Differences between mother animals are often not of interest
  - Different schools would be chosen for another study

QA: Where do we need a mixed model?

- A clinical trial investigates the efficacy of a new drug in reducing blood pressure. Patients from different clinics are recruited and randomly assigned to either the drug or a placebo group. Blood pressure measurements are taken at regular intervals over the study period.
- A study examines the effect of different fertilizers on crop yield across multiple farms. Data is collected on crop yield from various farms over several years, with each farm using different fertilizer types
- A study evaluates the effectiveness of a dental caries prevention program in an elementary school. The aim of the study is to analyse the association of different students characteristics and dental health score with accounting for age and sex differences.

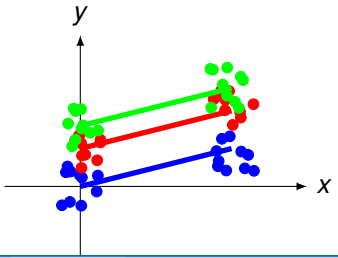
Real Example: Bodyweight from Rats

- In preclinical research, rats are a major animal model. Their body weight is regularly monitored for animal welfare and other characteristics
- Their body weight depends on a few variables, such as age, food intake, activity level, and litter
- We re-simulate a real data set involving the variables
  - LitterID, Age, BodyWeight, FoodIntake, ActivityLevel
  - Explain the data set in detail. Which data are independent, which may be dependent?
  - Model the body weight as a mixed model and estimate the model parameters

Mixed-Effects Models: Three Main Models

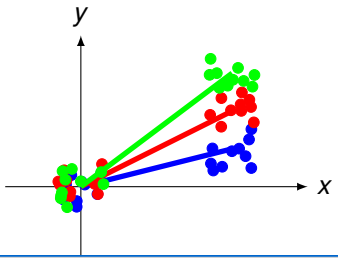
Random Intercept Model

- Different intercepts for each cluster
- association between independent and dependent variable is assumed to be equal across clusters



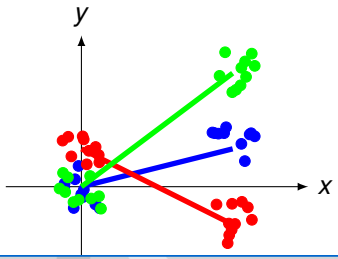
Random Slope Model

- association between independent and dependent variable is not equal across clusters



Random Intercept & Slope

- different intercepts for each cluster
- association between independent and dependent variable is not equal across clusters



Random Intercept Model

Nullmodel

$$Y_{ij} = \beta_0 + \gamma_i + \epsilon_{ij};\ i = 1, \ldots, n; j = 1, \ldots, n_i$$

where:

- i: observation, j: cluster
- **Fixed effect**: β_0
- **Random effects**
  - $\gamma_i \sim N(0, \tau^2)$: cluster effect
  - $\epsilon_{ij} \sim N(0, \sigma^2)$: error term *(Strictly speaking not an effect)*
- **Mixed Effect**: $(\beta_0 + \gamma_i)$: intercept for cluster $i \Rightarrow$ **Different intercepts for clusters**

Random Effect: Models Dependencies

- Moments of the variables

$$E(Y_{ij}) = \mu \quad \text{and} \quad \text{Var}(Y_{ij}) = \tau^2 + \sigma^2$$

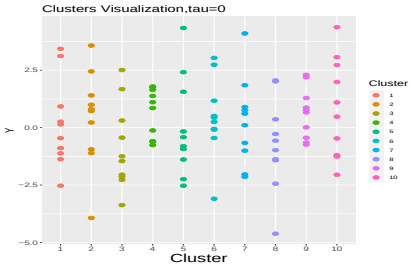
- Animals from the same litter may be dependent
- We compute the covariance between two animals from the same litter and find:

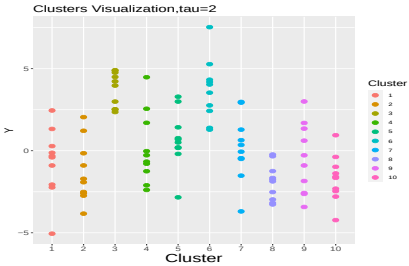
$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Var}(\gamma_i) = \tau^2$$

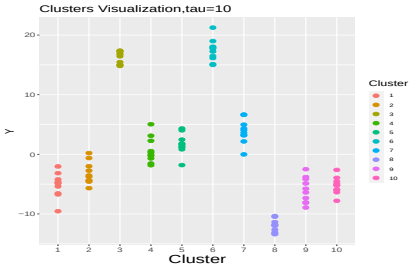
where:

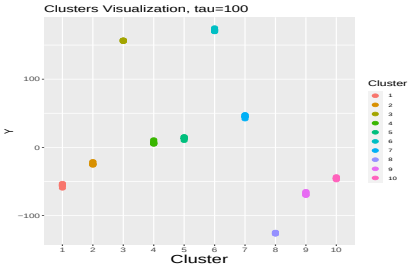
- The degree of linearity is τ^2
- If $\tau^2 = 0$: No effect/impact from the mother

Visualization of the Random Effect









R Code for the Graphics

```
library(lme4);library(ggplot2); set.seed(123)
# Number of clusters
n_clusters <- 10
# Number of observations per cluster
n_per_cluster <- 10
cluster_ids <- rep(1:n_clusters, each = n_per_cluster)
# Generate random effects for clusters
cluster_effects <- rnorm(n_clusters, mean = 0, sd = 100)
# Generate random errors
errors <- rnorm(n_clusters * n_per_cluster, mean = 0, sd = 2)
# Generate response variable with clustered structure
y <-  cluster_effects[cluster_ids] + errors
#Combine data into a data frame
data <- data.frame(y, cluster = factor(cluster_ids))
# Fit a mixed-effects model
model <- lmer(y ~ (1 | cluster), data = data)
# Print summary of the model
summary(model)
# Create a scatter plot with clusters colored differently
cluster_plot <- ggplot(data, aes(x = cluster, y = y, color = factor(cluster))) +
  geom_point(size=3) +
  labs(title = "Clusters Visualization, tau=100",x = "Cluster", y = "Y", color = "Cluster",size=12) +
 theme(axis.text = element_text(size = 12),
axis.title.x=element_text(size=22),
title=element_text(size=15))
print(cluster_plot)
```

Intraclass Correlation Coefficient (ICC)

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2}$$

ICC = Proportion of total variance due to differences between clusters

- ICC = 0: No variance explained by differences between clusters
- ICC = 1: All variance is explained by differences between clusters, measures within clusters are equal

Example with visual illustration: `http://mfviz.com/hierarchical-models/`

R Package lme4 and lmer function

- **R-package lme4**: Fit linear and generalized linear mixed-effects models
- Function **lmer**: Fit a linear mixed-effects model (LMM) to data, via REML or maximum likelihood.
- Random effects are modeled using $(|)$ operator, see ?*lmer*

Example Evaluation: Random Intercept Model

```
library(lme4)
fit <- lmer(BodyWeight~1+(1|LitterID),data=Bodyweight)
summary(fit)
Linear mixed model fit by REML ['lmerMod']
Formula: BodyWeight ~ 1 + (1 | LitterID)

Random effects:
 Groups   Name        Variance Std.Dev.
 LitterID (Intercept)  0.7351  0.8574
 Residual             19.8851  4.4593
Number of obs: 89, groups:  LitterID, 15

Fixed effects:
            Estimate Std. Error t value
(Intercept) 249.8444     0.5235   477.3
```

Model: $Bodyweight_j = 249.84 + \gamma_j,\ \gamma_j \sim N(0, 0.74)$

Random Intercept with Fixed Slope Model

$$Y_{ij} = \beta_0 + \gamma_i + \beta_1 x_{ij} + \epsilon_{ij};\ i = 1, \ldots, n; j = 1, \ldots, n_i$$

where:

- i: observation, j: cluster
- **Fixed effect**: β_0 (intercept); β_1: fixed effect of X on Y
- **Random effects**
  - $\gamma_i \sim N(0, \tau^2)$: cluster effect
  - $\epsilon_{ij} \sim N(0, \sigma^2)$: error term *(Strictly speaking not an effect)*
- **Mixed Effect**: $(\beta_0 + \gamma_i)$: intercept for cluster i
  $\Rightarrow$ **Different intercepts for clusters**
- Association of X and Y is fixed across all clusters

Example Evaluation: Random Intercept Model with Fixed Slope

```
 fit2 <- lmer(BodyWeight~1 + FoodIntake+(1|LitterID),data=Bodyweight)
summary(fit2)

Random effects:
 Groups   Name        Variance Std.Dev.
 LitterID (Intercept)  0.812   0.9011
 Residual             19.713   4.4399
Number of obs: 89, groups:  LitterID, 15

Fixed effects:
            Estimate Std. Error t value
(Intercept) 249.7616     0.5308 470.561
FoodIntake    2.6145     2.1444   1.219
```

$$Bodyweight_j = 249.76 + 2.61\,FoodIntake + \gamma_j + N(0, 19.71);\ \gamma_j \sim N(0, 0.812)$$

Random Intercept Model with Random Slope

$$Y_{ij} = (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i})x_{ij} + \epsilon_{ij};\ i = 1, \ldots, n; j = 1, \ldots, n_i$$

where:

- i: observation, j: cluster
- **Fixed effect**: β_0 (intercept); β_1: fixed effect of X on Y
- **Random effects**
  - $\gamma_{0i} \sim N(0, \tau^2)$: cluster effect; $\gamma_{1i} \sim N(0, \tau_1^2)$: random slope of cluster i
  - $\epsilon_{ij} \sim N(0, \sigma^2)$: error term *(Strictly speaking not an effect)*
- **Mixed Effect**: $(\beta_0 + \gamma_{0i})$: intercept of cluster i
- $(\beta_1 + \gamma_{1i}) \Rightarrow$ **Individual intercepts AND slopes for each cluster**

Example Evaluation: Random Intercept Model with Random Slope

```
fit3 <- lmer(BodyWeight~ FoodIntake+(1|LitterID) +(0+FoodIntake|LitterID),data=Bodyweight)
summary(fit3)
Random effects:
 Groups      Name         Variance  Std.Dev.
 LitterID    (Intercept) 8.120e-01 0.9011308
 LitterID.1 FoodIntake   1.177e-08 0.0001085
 Residual                1.971e+01 4.4399205

Fixed effects:
            Estimate Std. Error t value
(Intercept) 249.7616     0.5308 470.561
FoodIntake    2.6145     2.1444   1.219
```

$$Bodyweight_j = (249.76 + \gamma_j) + (2.5 + \gamma_{1j})FoodIntake + N(0, 19.67);$$
$$\gamma_j \sim N(0, 0.812),\ \gamma_{1j} \sim N(0, 1.17e-08)$$

Note: Model is almost singular; variance of random effect γ_{1j} estimated 0

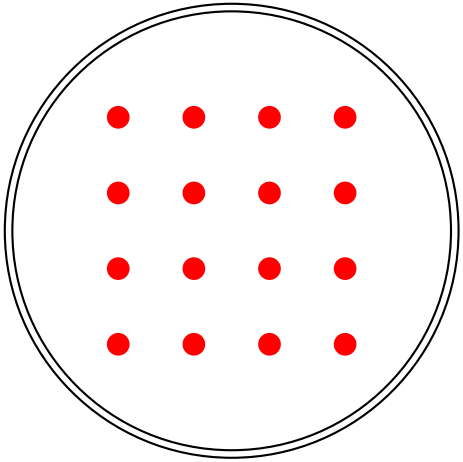
Mixed ANOVA Models

- Regression is not the only framework where random effects are important
- ANOVA models are of interest as well
- Here: Testing whether means of several groups are of interest.
- A common example of an experiment involving measurements in petri dishes across different groups is the growth of bacterial colonies under different treatment conditions. This type of experiment is often used in microbiology to evaluate the effect of various treatments (e.g., antibiotics, environmental conditions) on bacterial growth.

An Example

Evaluate the effect of different antibiotic treatments on bacterial growth.

- Groups: Control, Antibiotic A, Antibiotic B.
- Replicates: 10 petri dishes per group.
- Measurement: Number of bacterial colonies in each petri dish.
- Mixed models are used to account for the variability between petri dishes and evaluate the fixed effects of the treatments.



Petri Dish with Bacterial Colonies

An Example

- Statistical model

$$Y_{ijk} = \mu + \alpha_i + Z_{ij} + \epsilon_{ijk}$$

- $i = 1, \ldots, a$: a groups
- $j = 1, \ldots, n_i$: cluster per group
- $k = 1, \ldots, m_{ij}$: observations within cluster (i, j)
- μ: Overall mean (fixed)
- α_i: Effect of group i (fixed)
- $Z_{ij} \sim N(0, \tau^2)$: Random cluster effect
- $\epsilon_{ijk} \sim N(0, \sigma^2)$: Error term
- **Mixed Effect:** $\mu + \alpha_i + Z_{ij}$

Data EvalationRandom Intercept Model

- Fit a mixed model with a random intercept for petri dishes and fixed effects for the treatment groups:

```
library(lme4)
model1 <- lmer(y ~ group + (1 | dish_id), data = data)
summary(model1)
```

- Interpretation:
  - Fixed effects: Estimates for treatment groups.
  - Random effects: Variance component for the random intercepts.

Statistical Model

The general form of a linear mixed model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{Y}$ is the vector of responses.
- $\mathbf{X}$ is the design matrix for fixed effects.
- $\boldsymbol{\beta}$ is the vector of fixed effect coefficients.
- $\mathbf{Z}$ is the design matrix for random effects.
- $\boldsymbol{\gamma}$ is the vector of random effect coefficients.
- $\boldsymbol{\epsilon}$ is the vector of residuals.

Matrix Notation

In matrix notation, the model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{Z}_1 \\ \mathbf{X}_2 & \mathbf{Z}_2 \\ \vdots & \vdots \\ \mathbf{X}_n & \mathbf{Z}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{bmatrix}$$

Model Assumptions

- Linearity: The relationship between the response variable and predictors is linear.
- Independence: Observations across clusters are independent of each other.
- Normality: The residuals are normally distributed.
- Homoscedasticity: The variance of the residuals is constant across all levels of predictors.
- Random Effects Distribution: The random effects are normally distributed with mean zero and a common variance τ^2.

Interpretation of Model Parameters

- β: Represents the fixed effects. Each element of β corresponds to the change in the response variable for a one-unit change in the corresponding predictor variable, assuming all other variables are held constant.
- γ: Represents the random effects. Each element of γ corresponds to the variability in the intercept or slope across different groups or clusters.

Conclusion

- Mixed models are powerful tools for analyzing data with complex structures.
- They allow for the incorporation of both fixed and random effects, providing insights into both average effects and variability across groups.
- Various estimation methods are available, depending on the assumptions and goals of the analysis.