MA50260 Statistical Modelling

Lecture 14: Introduction to Mixed Effects Models

Ilaria Bussoli

March 22, 2024

Reminder : The (normal) linear regression

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumptions on $\epsilon_1, \ldots, \epsilon_n$?

Reminder : The (normal) linear regression

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumptions on $\epsilon_1, \ldots, \epsilon_n$?

- $\epsilon_1, \ldots, \epsilon_n$ are mutually independent;
- $\epsilon_i \sim \text{Normal}(0, \sigma^2), \; i = 1 \ldots, n.$

Reminder : The (normal) linear regression

Recall the linear regression model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n.$$

What are the assumptions on $\epsilon_1, \ldots, \epsilon_n$?

- $\epsilon_1, \ldots, \epsilon_n$ are mutually independent;
- $\epsilon_i \sim \text{Normal}(0, \sigma^2),\ i = 1 \ldots, n.$

Generalized linear models provide us with tools in case the second assumption does not hold.

We will now consider a modelling framework in case the assumption of mutual independence does not hold.

Motivation

In many situations, we will observe data which are grouped in nature, such as

- Observations in medical studies,
- Ecological / agricultural data,
- Student exam scores.

We should take these groupings into account within our model

$\Rightarrow$ **random effects.**

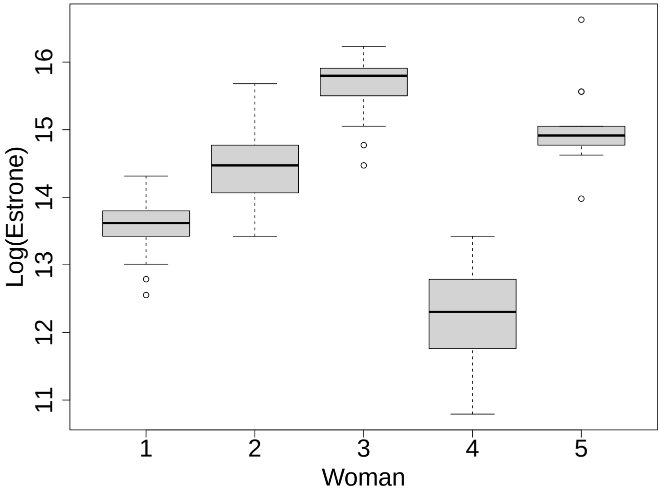
Example (1) - Air Pollution Measurements

Suppose you are recording air pollution daily across several measurement stations. Then, our model has to account for

- Daily variations in the observations at an individual station;
- Variation across measurement stations;
- Variations across space and time are likely to be different.

$\Rightarrow$ We cannot use our standard linear model and this motivates the use of **mixed effect models**.

Example (2) - Female estrone measurements



Mixed Effects Modelling (I)

A **mixed effects model** contains both fixed and random effects.

Let $Y_{i,j}$ denote the raw estrone level. We model

$$Y_{i,j} = \mu + b_i + \epsilon_{i,j} \quad \text{for } i = 1, \ldots, I = 5 \text{ and } j = 1, \ldots, J = 16,$$

where

- μ is the average estrone measurement;
- b_i is the person effect;
- $\epsilon_{i,j}$ is the residual.

We assume $\mathbb{E}(b_i) = \mathbb{E}(\epsilon_{i,j}) = 0$, $\mathrm{Var}(b_i) = \sigma_b^2$ and $\mathrm{Var}(\epsilon_{i,j}) = \sigma_\epsilon^2$.

Mixed Effects Modelling (II)

We are interested in the variability between women, rather than the differences in the individual levels.

Two questions that arise from the data:

1. Is there evidence for variability in estrone **between** women?
2. If so, how large is this variability in relation to the variability of measurements for an individual woman?

Mixed Effects Modelling (III)

The total variance is $\mathrm{Var}(Y_{i,j}) = \sigma^2_{total} = \sigma^2_b + \sigma^2_\epsilon$.

For Question 1, we want to test $H_0 : \sigma^2_b = 0$ vs $H_1 : \sigma^2_b > 0$.

To address Question 2, we use the **intraclass correlation**

$$\rho = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\epsilon}.$$

Connection to ANOVA

The framework is similar to a **one-way ANOVA** with

- m levels (groups) of a factor variable
- n observations per group.

Let $\bar{Y}_i$ denote the mean of the i-th group. Then

$$\bar{Y}_i = \mu + b_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{i,j},$$

where

$$b_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{i,j} \sim \mathrm{Normal}\left(0, \sigma_b^2 + \frac{\sigma_\epsilon^2}{J}\right).$$

General Notation of the Mixed Effects Model (I)

In matrix form, the linear regression model is

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}.$$

In the estrone example, we considered

$$Y_{i,j} = \mu + b_i + \epsilon_{i,j}.$$

How can we define this model in matrix form?

General Notation of the Mixed Effects Model (II)

A mixed effects (normal) linear model can be written as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b} + \underline{\epsilon},$$

where

- $\mathbf{Y}$ is the vector of responses
- $\mathbf{X}$ is the design matrix
- $\underline{\beta}$ are the regression coefficients
- $\mathbf{Z}$ is the matrix of covariates associated to the q (unknown) random effects
- $\mathbf{b}$ is the vector of random effects
- $\underline{\epsilon}$ is the vector of residuals.

General Notation of the Mixed Effects Model (III)

If $\mathbf{b}$ is known and $\epsilon_{i,j} \sim \mathrm{Normal}(0, \sigma_\epsilon^2)$,

$$\mathbf{Y} \sim \mathrm{MVN}_n \left(\mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b}, \sigma_\epsilon^2 \mathbf{I}_n \right),$$

where $n = I \times J$.

General Notation of the Mixed Effects Model (III)

If **b** is known and $\epsilon_{i,j} \sim \mathrm{Normal}(0, \sigma_\epsilon^2)$,

$$\mathbf{Y} \sim \mathrm{MVN}_n\left(\mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b}, \sigma_\epsilon^2 \mathbf{I}_n\right),$$

where $n = I \times J$.

Since the q random effects **b** are in general unknown, we assume

$$\mathbf{b} \sim \mathrm{MVN}_q(0, \mathcal{D}),$$

leading to

$$\mathbf{Y} \sim \mathrm{MVN}_n\left(\mathbf{X}\underline{\beta}, \Sigma + \mathbf{Z}\mathcal{D}\mathbf{Z}^T\right),$$

where $\Sigma = \sigma_\epsilon^2 \mathbf{I}_n$ and $\mathcal{D}$ is parametrized by a vector $\underline{\theta}$, such as $\underline{\theta} = \sigma_b^2$.

Example - Estrone Levels

- $\mathbf{Y} \in \mathcal{M}_{80\times 1}(\mathbb{R})$ (most intuitively with observations grouped according to woman)
- $\mathbf{X} \in \mathcal{M}_{80\times 1}(\mathbb{R})$ with all values equal to 1;
- $\mathbf{Z} \in \mathcal{M}_{80\times 5}(\mathbb{R})$ made up of indicator columns, one for each woman, with 16 ones per column
- $\Sigma = \sigma_\epsilon^2 \mathbf{I}_{80} \in \mathcal{M}_{80\times 80}(\mathbb{R})$
- $\mathcal{D} = \sigma_b^2 \mathbf{I}_5 \in \mathcal{M}_{5\times 5}(\mathbb{R})$
- The vectors involved in the model are $\underline{\beta} = \mu$, $\mathbf{b} = (b_1, \ldots, b_5)^T$ and $\underline{\theta} = (\sigma_b^2, \sigma_\epsilon^2)$.