Using Mixed Effects Models for the Analysis of Longitudinal Data

Jack Lloyd, supervised by Dr. Reza Drikvandi

Department of Mathematical Sciences, Durham University

Scan for the R Code

1. Introduction

Longitudinal data is data which is collected from the same individuals sequentially over multiple points in time. We are interested in analysing data of this form to attempt to identify trends and changes in the data over time and to see how the individuals are correlated. This project will introduce a class of models that aims to model longitudinal data whilst taking into account variability between each individual.

2. Building a Linear Mixed Effects Model

One can introduce the structure of a **linear mixed effects model (LMEM)** in two stages [3]. Let Y_{ij} be the response of individual i at time $j = 1, ..., n_i$, where n_i is the total number of observations for the i^{th} individual. Stage 1 assumes that $\boldsymbol{Y_i}$ follows a linear regression model, given by

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \ , \tag{1}$$

where $\boldsymbol{Z_i}$ is an $n_i \times q$ dimensional design matrix, $\boldsymbol{\beta_i}$ is a q-dimensional vector of unknown subject specific regression coefficients. We also have $\epsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$. In Stage 2, we aim to explain the variability between subjects by including random effects in $\boldsymbol{\beta}_i$ such as

$$\boldsymbol{\beta_i} = \boldsymbol{M}_i\boldsymbol{\beta} + \boldsymbol{u}_i \ , \tag{2}$$

where $\boldsymbol{M}_i \in \mathbb{R}^{q\times p}$ is a matrix of known covariates and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown fixed regression coefficients. Here $\boldsymbol{u}_i$ is a vector of subject specific random effects which are assumed to be independent and $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \tilde{\boldsymbol{Q}})$. Substituting (2) into (1) gives

$$\begin{aligned}\boldsymbol{Y}_i &= \boldsymbol{Z}_i\boldsymbol{M}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i\end{aligned}$$

where we have written $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{M}_i$. We have the general framework for the entire model as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{3}$$

Note that $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n)$ and $\boldsymbol{u} \sim N(0, \boldsymbol{Q})$ where $\boldsymbol{Q}$ is a block diagonal matrix with the $i-th$ diagonal block being equal to the covariance matrix of $\boldsymbol{u}_i$.

3. References

[1] Verbeke Fitzmaurice Davidian and Molenberghs. *Longitudinal Data Analysis*. Chapman and Hall/CRC, 2009.

[2] Lynn Roy LaMotte. "A Direct Derivation of the REML Likelihood function". In: *Statistical Papers* (2005).

[3] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, 2000.

[4] Geert Verbeke and Geert Molenberghs. *Models for Discrete Longitudinal Data*. Springer.

4. Distribution of a Linear Mixed Effects Model

The distribution of $\boldsymbol{Y}$ is as follows:

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Q}\boldsymbol{Z}^T + \sigma^2\boldsymbol{I}_N$, with $N = \sum_i n_i$, the total number of observations across the whole dataset. This is due to

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbb{E}(\boldsymbol{u}) + \mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{X}\boldsymbol{\beta}$$

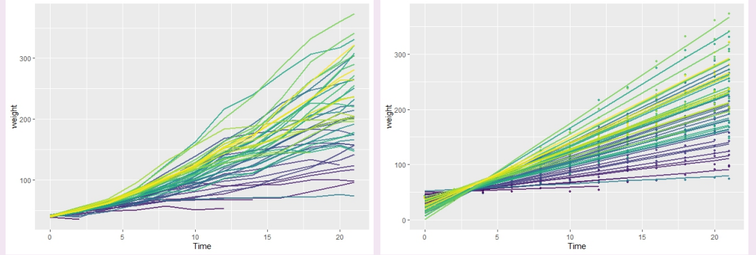
$$\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{Z}\mathrm{Var}(\boldsymbol{u})\boldsymbol{Z}^T + \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{Z}\boldsymbol{Q}\boldsymbol{Z}^T + \sigma^2\boldsymbol{I}_N$$

5. Applying a Linear Mixed Effects Model

We will apply an LMEM with a random intercept and slope to the `ChickWeight` dataset. The data consists of 50 chicks which have been weighed at 12 time points. The chicks were fed one of four different diets, which we denote Diet A, B, C or D. Using `R`, we fit a model with the equation

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 B_i + \beta_3 C_i + \beta_4 D_i + u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}$$

where B_i, C_i and D_i are diet indicator variables.



(a) `ChickWeight` plotted (b) LME fitted to `ChickWeight`

Figure 1: Plots Showing How the LME Models `ChickWeight`

The REML estimation (see **Section 7**) for the coefficients is as follows:

$$Y_{ij} = 26.4 + 8.4t_{ij} + 2.8B_i + 2.0C_i + 9.3D_i + u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}$$

6. Inference for the Random Effects

One can estimate $\boldsymbol{u}_i$ by taking $\hat{\boldsymbol{u}}_i(\theta) = \mathbb{E}(\boldsymbol{u}_i|\boldsymbol{Y}_i = \boldsymbol{y}_i)$, its posterior mean. We know that

$$\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{Y}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ \boldsymbol{X}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\boldsymbol{Q} & \mathrm{Cov}(\boldsymbol{u}, \boldsymbol{Y})\\ \mathrm{Cov}(\boldsymbol{u}, \boldsymbol{Y})^T & \boldsymbol{\Sigma}\end{pmatrix}\right)$$

where $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{Y}) = \boldsymbol{Q}\boldsymbol{Z}^T$. Using a general result for multivariate conditional normal distributions, the posterior mean can be specified as $\hat{\boldsymbol{u}}_i(\theta) = \boldsymbol{Q}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ and hence this is our estimate for $\boldsymbol{u}_i$.

7. REML Estimator for the Likelihood

One can introduce an estimator for the likelihood which is not dependent on $\boldsymbol{\beta}$ [4]. One does this by multiplying the general equation by the transpose of a matrix $\boldsymbol{A}$ which is orthogonal to $\boldsymbol{X}$. One has

$$\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{A}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}^T\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{A}^T\boldsymbol{\epsilon} = \boldsymbol{A}^T\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{A}^T\boldsymbol{\epsilon}$$

$$\boldsymbol{A}^T\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{A}).$$

LaMotte [2] proves that the likelihood of this function can be written in a form which does not depend on $\boldsymbol{A}$, with $\hat{\boldsymbol{\beta}}$ being the generalised least squares (GLS) estimate. This is given by:

$$(2\pi)^{-(n-p)}|\Sigma|^{-\frac{1}{2}}|X^T\Sigma^{-1}X|^{-\frac{1}{2}}|X^TX|^{\frac{1}{2}}\exp\left(-\tfrac{1}{2}(Y - X\hat{\beta})^T\Sigma^{-1}(Y - X\hat{\beta})\right). \tag{5}$$

8. Missing Data in Longitudinal Datasets

Longitudinal datasets often include missing data. The `lmer` function in `R` can still fit an LMEM to a dataset that contains missing values. We randomly set a percentage of the response values equal to `NA` and calculated the residual sum of squares (RSS) of the model fitted using `lmer`.

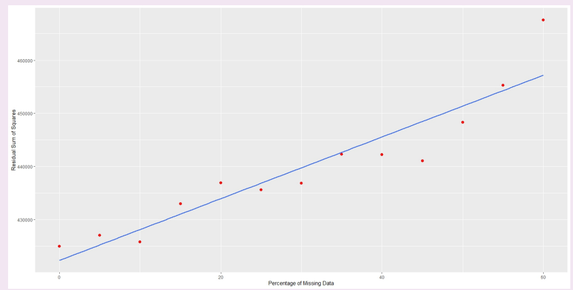


Figure 2: RSS of an LME Fitted to `ChickWeight` with Missing Data

The higher the percentage of missing data, the more `R` struggles to fit the model. There are many ways that one can impute missing data from simple methods such as single mean imputation and last observation carried forward to more advanced methods such as multiple imputation.

9. Conclusions and Future Work

We have seen that we can model longitudinal datasets effectively with LMEMs which capture the trends of the data for each individual.

A future step in the project is to create a function on `R` to fit a selection model following the theory from [1]. This is a hierarchical model which assumes the data follows a multivariate normal distribution and the missingness follows a logistic regression.