# Mixed effects Models

Stephen Coleman

February 23, 2019

## 1 Introduction

### 1.1 What, when and why

Traditional parametric models incorporate only *fixed effects*. That is, they have a set of parameters describing with the entire population. For example, consider a system of $n \in \mathbb{N}$ observations of dependent variable, $Y = (y_1, \ldots, y_n)$, and the $n \times (p+1)$ matrix of associated independent variable measurements, $X = \{x_{i,j}\}\, i \in [1, n], j \in [1, p]$ for $p \in \mathbb{N}$, and a $(p+1)$-vector of weights, $\beta$, represented:

$$Y = X\beta + \epsilon \tag{1}$$

Here $X$ is assumed to contain a column of 1's in the first position (hence the $+1$ in the description of its dimensionality) and $\epsilon$ is the $n$-vector of associated errors. We assume that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = (1, \ldots, n)$. If we consider the intuitive case of *biological* and *technical replicates*, where we have $n$ biological replicates and for the $i^{th}$ sample $m_i$ associated technical replicates. As each of the $m_i$ measurements is on the same sample we expect there to be a non-negligible correlation between the $m_i$ technical replicates for each $i$. The model in (1) does not allow for the within group effects due to the correlation between technical replicates. Consider the simplest possible model for the fixed effects model including only the intercept:

$$y_{ij} = \beta + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{2}$$

Some of this model's limitations become apparent if one considers the case of unbalanced data. Continuing the previous example of biological and technical replicates, consider the case that $m_i \neq m_j$ for any $i, j \in (1, \ldots, n)$. In this case the model is skewed by the within sample data rather than by the true observations, the biological replicates. A possible solution is the inclusion of a individual intercept for each group of technical replicates, accommodating the within-sample variability, or *random effects*:

$$y_{ij} = \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{3}$$

While this better describes the observed data, it has certain inherent flaws; most notably the number of parameters scales linearly with the number of observations and the

model only describes the measurements included in the sample. Consider as a solution a combination of these models, containing both a sample mean, $\beta$, and a random variable for each group representing the deviation from the population mean, $b_i$, i.e. a *mixed effects* model:

$$y_{ij} = \beta + b_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{4}$$

The linear mixed effects model described in (4) contains information at both a population level (in the fixed effects) but also at the individual level (in the random effects).

For now we assume $b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2)$ for $i = 1, \ldots, n$. This means that the variance of the observations is divided into two parts, $\sigma_b^2$ for the biological variability and $\sigma^2$ for the technical variability:

$$b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2), \quad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{5}$$

The assumption of normality can be modified if deemed inappropriate and it is possible to generalise the model to allow for heteroscedasticity.

The $b_i$ are called *random effects* as they are associated with the experimental unit and selected at random from the population of interest (at least in theory, obviously there are limitations on this particularly in the area of medicine) They represent that the effect of choosing the sample $i$ is to shift the mean expression of $Y$ from $\beta$ to $\beta + b_i$ - i.e. they effect a deviation from an overall mean. Technical replicates share the same random effect $b_i$ and are correlated. The covariance between technical replicates on the same experimental unit is $\sigma_b^2$; this corresponds to a correlation of $\frac{\sigma_b^2}{(\sigma_b^2 + \sigma^2)}$.