# Contents

# 1 Introduction to LMEs

In modern statistical analysis , data sets have very complex structures, such as clustered data, repeated data and hierarchical data (henceforth referred to as grouped data).

Repeated data considers various observations periodically taken from the same subjects. 'Before and after' measurements, as used in paired t tests, are a well known example of repeated measurements. Clustered data is simply the grouping of observations according to common characteristics. For example, an study of pupils of a school would account for the fact that they are grouped according to their classes.

Hierarchical structures organize data into a tree-like structure, i.e. groups within groups. Using the previous example, the pupils would be categorized according to their years (i.e the parent group) and then their classes (i.e the child group). This can be extended again to multiple schools, where each school would be the parent group of each year.

An important feature of such data sets is that there is correlation between observations within each of the groups (**?**). Observations in different groups may be independent, but any assumption that these observations within the same group are independent is inappropriate . Consequently **?** states that there is two sources of variations to be considered, 'within groups' and 'between groups'.

1. Classical models
2. Grouped data sets
3. variance components
   (a) Fixed effects
   (b) Random effects

# 2 Important Definitions

## 2.1 Advantages of Mixed Models

- **cite:BrownPrescott** discusses the following advantages of using mixed effects models. In the case of repeated measurements , it is appropriate to take account of the correlation of each group of observations.
- Mixed models lead to more appropriate estimates and standard errors for fixed effects, particularly in the case of repeated measures. Analysis using a mixed model is more appropriate for inference on a hierarchical data. In the case of unbalanced data, mixed models are more appropriate than other methodologies.
- **cite:Demidenko** comments that mixed models are the correct approach for dealing with grouped data. The use of linear mixed effects models has advanced greatly with increased usage of statistical software.
- This author also notes that mixed models are a hybrid of bayesian and frequentist methodologies and that mixed model approaches are more flexible than Bayesian.

### 2.1.1 Unbalanced Data

- Unbalanced data refers to situations where these groups are of different sizes. Mixed Effects Models are suitable for studying unbalanced data sets.
- The variance components of random effects for these set can not be derived using alternative methods such as ANOVA.

# 3 Matrix Formulation

There are matrix (i.e multivariate) formulations of both fixed effects models and random effects models. **?** remarks that the matrix notation makes the underlying theory of mixed effects models much easier to work with. The fixed effects models can be specified as follows;

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

$\mathbf{Y}$ is the vector of $n$ observations, with dimension $n \times 1$. $\mathbf{b}$ is a vector of fixed $p$ effects, and has dimension $p \times 1$. It is composed of coefficients, with the first element being the population mean. For the skin tumour example, with the three specified fixed effects, $p = 4$. $\mathbf{X}$ is known as the design 'matrix', model matrix for fixed effects, and comprises 0s or 1s, depending on whether the relevant fixed effects have any effect on the observation is question. $\mathbf{X}$ has dimension $n \times p$. $\mathbf{e}$ is the vector of residuals with dimension $n \times 1$.

The random effects models can be specified similarly. $\mathbf{Z}$ is known as the 'model matrix for random effects', and also comprises 0s or 1s. It has dimension $n \times q$. $\mathbf{u}$ is a vector of random $q$ effects, and has dimension $q \times 1$.

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{2}$$

Again, once the component fixed effects and random effects components are considered, progression to a mixed model formulation is a simple step. Further to **?**, it is conventional to formulate a mixed effects model in matrix form as follows:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{3}$$

($E(\mathbf{u}) = 0$, $E(\mathbf{e}) = 0$ and $E(\mathbf{y}) = \mathbf{Xb}$)

# 4 Mixed Model Calculations

## 4.1 Formulation of the Variance Matrix V

$\mathbf{V}$ , the variance matrix of $\mathbf{Y}$, can be expressed as follows;

$$\mathbf{V} = var(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \tag{4}$$

$$\mathbf{V} = var(\mathbf{Xb}) + var(\mathbf{Zu}) + var(\mathbf{e})) \tag{5}$$

$var(\mathbf{Xb})$ is known to be zero. $var(\mathbf{Zu})$ can be written as $Zvar(\mathbf{u})Z^T$. $Z$ is a matrix of constants. By letting $var(u) = G$ (i.e $\mathbf{u}$ $N(0, \mathbf{G})$), this becomes $ZGZ^T$. This specifies the covariance due to random effects. The residual covariance matrix $var(e)$ is denoted as $R$, ($\mathbf{e}$ $N(0, \mathbf{R})$). Residual are uncorrelated, hence $\mathbf{R}$ is equivalent to $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The variance matrix $\mathbf{V}$ can therefore be written as;

$$\mathbf{V} = ZGZ^T + \mathbf{R} \tag{6}$$

# 5 Estimation Methods for LMEs

## 5.1 Maximum Likelihood Estimation

Maximum likelihood (ML) estimation is a well known method of obtaining estimates of unknown parameters by optimizing a likelihood function. Models fitted by ML estimation can be compared using the likelihood ratio test. However ML is known to underestimate variance components for finite samples (**?**).

## 5.2 Restricted Likelihood Estimation

A method related to ML is restricted maximum likelihood estimation(REML). REML was developed by **?** and **?** to provide unbiased estimates of variance and covariance parameters. REML obtains estimates of the fixed effects using non-likelihoodlike methods, such as ordinary least squares or generalized least squares, and then using these estimates it maximizes the likelihood of the residuals (subtracting off the fixed effects) to obtain estimates of the variance parameters. In most software packages REML is the default algorithm used to compute coefficients for the predictor variables. REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML.

**?** describes two important outcomes of using REML. Firstly variance components can be estimated without being affected by fixed effects. Secondly in estimating variance components with REML, degrees of freedom for the fixed effects can be taken into account implicitly, whereas with ML they are not. When estimating variance from normally distributed data, the ML estimator for $\sigma^2$ is $\frac{S_{yy}}{n}$ whereas the REML estimator is $\frac{S_{yy}}{n-1}$. ( $S_{yy}$ is the sum of square identity;

$$S_{yy} = \Sigma_{i=i}^{N}(y - \bar{y})^2 \tag{7}$$

# 6 Orthodont Example

- Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14.
- Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head.

```
library(nlme)

data(Orthodont)

summary(Orthodont)
```

```
> summary(Orthodont)
    distance          age          Subject       Sex
 Min.   :16.50   Min.   : 8.0   M16    : 4   Male  :64
 1st Qu.:22.00   1st Qu.: 9.5   M05    : 4   Female:44
 Median :23.75   Median :11.0   M02    : 4
 Mean   :24.02   Mean   :11.0   M11    : 4
 3rd Qu.:26.00   3rd Qu.:12.5   M07    : 4
 Max.   :31.50   Max.   :14.0   M08    : 4
                                (Other):84
```

# 7 "Machines" Data Set - Example

**?** describes an experiment whereby the productivity of six randomly chosen workers are assessed three times on each of three machines, yielding the 54 observations tabulated below.

| Observation | Worker | Machine | score | Observation | Worker | Machine | score |
|---|---|---|---|---|---|---|---|
| 1 | 1 | A | 52.00 | 28 | 4 | B | 63.20 |
| 2 | 1 | A | 52.80 | 29 | 4 | B | 62.80 |
| 3 | 1 | A | 53.10 | 30 | 4 | B | 62.20 |
| 4 | 2 | A | 51.80 | 31 | 5 | B | 64.80 |
| 5 | 2 | A | 52.80 | 32 | 5 | B | 65.00 |
| 6 | 2 | A | 53.10 | 33 | 5 | B | 65.40 |
| 7 | 3 | A | 60.00 | 34 | 6 | B | 43.70 |
| 8 | 3 | A | 60.20 | 35 | 6 | B | 44.20 |
| 9 | 3 | A | 58.40 | 36 | 6 | B | 43.00 |
| 10 | 4 | A | 51.10 | 37 | 1 | C | 67.50 |
| 11 | 4 | A | 52.30 | 38 | 1 | C | 67.20 |
| 12 | 4 | A | 50.30 | 39 | 1 | C | 66.90 |
| 13 | 5 | A | 50.90 | 40 | 2 | C | 61.50 |
| 14 | 5 | A | 51.80 | 41 | 2 | C | 61.70 |
| 15 | 5 | A | 51.40 | 42 | 2 | C | 62.30 |
| 16 | 6 | A | 46.40 | 43 | 3 | C | 70.80 |
| 17 | 6 | A | 44.80 | 44 | 3 | C | 70.60 |
| 18 | 6 | A | 49.20 | 45 | 3 | C | 71.00 |
| 19 | 1 | B | 62.10 | 46 | 4 | C | 64.10 |
| 20 | 1 | B | 62.60 | 47 | 4 | C | 66.20 |
| 21 | 1 | B | 64.00 | 48 | 4 | C | 64.00 |
| 22 | 2 | B | 59.70 | 49 | 5 | C | 72.10 |
| 23 | 2 | B | 60.00 | 50 | 5 | C | 72.00 |
| 24 | 2 | B | 59.00 | 51 | 5 | C | 71.10 |
| 25 | 3 | B | 68.60 | 52 | 6 | C | 62.00 |
| 26 | 3 | B | 65.80 | 53 | 6 | C | 61.40 |
| 27 | 3 | B | 69.70 | 54 | 6 | C | 60.50 |

Table 1: Machines Data , Pinheiro Bates

(Overall mean score = 59.65, mean on machine A = 52.35 , mean on machine B = 60.32, mean on machine C = 66.27)

The 'worker' factor is modelled with random effects($u_i$), whereas the 'machine' factor is modelled with fixed effects ($\beta_j$). Due to the repeated nature of the data, interaction effects between these factors are assumed to be extant, and shall be examined accordingly. The interaction effect in this case ($\tau_{ij}$) describes whether the effect of changing from one machine to another is different for each worker. The productivity score $y_{ijk}$ is the $k$th observation taken on worker $i$ on machine $j$, and is formulated as follows;

$$y_{ijk} = \beta_j + u_i + \tau_{ij} + \epsilon_{ijk} \tag{8}$$
$$u_i \sim N(0, \sigma_u^2), \epsilon_{ijk} \sim N(0, \sigma^2), \tau_i \sim N(0, \sigma_\tau^2)$$

The 'nlme' package is incorporated into the R programming to perform linear mixed model calculations. For the 'Machines' data, **?** use the following code, with the hierarchical structure specified in the last argument.

```
lme(score~Machine, data=Machines, random=~1|Worker/Machine)
```

The output of the R computation is given below.

```
Linear mixed-effects model fit by REML
  Data: Machines
  Log-restricted-likelihood: -107.8438
  Fixed: score ~ Machine
(Intercept)    MachineB    MachineC
  52.355556    7.966667   13.916667

Random effects:
 Formula: ~1 | Worker
        (Intercept)
StdDev:     4.78105

 Formula: ~1 | Machine %in% Worker
        (Intercept)  Residual
StdDev:    3.729532 0.9615771

Number of Observations: 54 Number of Groups:
             Worker Machine %in% Worker
                  6                  18
```

The crucial pieces of information given in the programme output are the estimates of the intercepts for each of the three machines. Machine A, which is treated as a control case, is estimated to have an intercept of 52.35. The intercept estimates for machines B and C are found to be 60.32 and 66.27 (by adding the values 7.96 and 13.91 to 52.35 respectively). Estimate for the variance components are also given; $\sigma_u^2 = (4.78)^2$ , $\sigma_\tau^2 = (3.73)^2$ and $\sigma_\epsilon^2 = (0.96)^2$.