Contents

Chapter 1

Linear Mixed Effects Models

1.1 Distinction between Classical Models and Mixed Models

Demidenko (2004) discusses the inadequacy of 'classical models' in analysing such data types, with particular reference to the simple linear model. The simple linear model is a well known statistical methodology that describes the relationship between dependent variables Y and an independent predictor variable X. Where $Y = y_1, y_2, ..y_k..y_n$ and $X = x_1, x_2, ...x_k..x_n$, an intercept α and slope β are estimated such that the error terms associated with each observation y_i is minimized.

$$y_k = \alpha + \beta x_k + \epsilon_k \tag{1.1}$$

In classical statistics a typical assumption is that observations are drawn from the same general population, are independent and identically distributed (Demidenko, 2004). Consequently there is no way to account for the grouped nature of data sets described previously, and so there lies the possibility of observations being treated as independent measurements. Demidenko (2004, pg.3) gives a very informative example wherein a classical approach is compared to an approach that does account for grouping. The conclusion to be drawn from Demidenko's example is that failure to account for group-

ing leads to an incorrect conclusion about the data. The approach recommended is known as 'mixed models' and shall be introduced presently.

1.2 Fixed Effects and Random Effects Models

Before proceeding to a description of mixed models, an introduction to fixed effects and random effects models is required. This section follows on from the discussion of measurement error models in the last chapter.

1.2.1 Advantages of Mixed Models

Brown and Prescott (1999) 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.

Demidenko (2004) 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.

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.

1.2.2 Fixed Effects

McCullough and Searle (2001) gives an example of a study where the observations , occurrences of skin tumours called basal cell epithelioma, were classified according to 'factors', i.e. the gender, age and exposure to sunshine of the patients. Levels are the individual classes of each of these factors (e.g. 'Male' and 'Female' would be the levels of the factor 'Gender'). The scientific interest lies in examining the extent to which different factor levels affect the variable of interest. The effects of a level of a factor are one of two types ; fixed effects and random effects. Fixed effects describe effects due to a finite set of levels for a factor (i.e multichotomous factors). The factors described in the skin tumour example are all fixed effects factors. Fixed effects models are the cases where only fixed effects are present, with the exception of random error terms.

To demonstrate fixed effects model Searle (1997) describes a study wherein 24 plants are divided into four groups of six, and each group is subjected to its own treatment regime. Three different fertilizers are used with three of the groups (treatments N,P,K), while no fertilizer is used on the fourth group, (i.e. it is a control group denoted as C). Searle (1997) constructs a model to describe the crop yield resultant from the experiment.

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{1.2}$$

where y_{ij} is the jth plant (i.e. crop yield) on the ith treatment, with μ as the mean yield, α_i is the effect of each fertilizer treatments (i.e. fixed effects) and e_{ij} is the error term. The fixed effect for each observation is an unknown constant that is to be determined from computing the data.

The μ term would not necessarily be present in all formulations. Some authors, such as Demidenko (2004), may use a single term as equivalent to the μ and α terms. (It is customary to centre data prior to using mixed effects methodologies. Centering means subtracting the mean of the observations from each observed value, so mean of the resultant values become zero.)

1.2.3 Variance Components

Each random effect has an associated 'variance component' term. This is a model parameter which quantifies random variation due to that effect only. Therefore for every observation there are two sources of variation, random variation and residual variation, and can be expressed as follows $var(y_{ij}) = \sigma_p^2 + \sigma^2$. These variations are known as the variance components. This is an important difference with fixed effects models, which is subject to residual variation (i.e σ^2) only (Brown and Prescott, 1999).

In fixed effects models there is no covariance between any pair of observations. Brown and Prescott (1999) shows that, while there is no covariance between observations from different subjects (i.e the mice in Searle's example), there exists correlation between observations from the same subject (i.e. litter weights from the same mouse are correlated). In this case the covariance is the subjects variance component (i.e. σ_p^2)

1.2.4 Random Effects

The random effects model describes the case where there is an infinite number of levels in a factor. In other words the factor is a random variable. Searle (1997) demonstrates this with a second example; a study of the maternal ability of mice. In this example 4 female mice , all of the same breed, have 6 litters each over a period of time. The weights w_{ij} of each litter (j) from each mouse (i) were taken to be the proxy for maternal ability and is formulated as follows;

$$w_{ij} = \mu + \delta_i + e_{ij} \tag{1.3}$$

As in the previous exmaple, α is the mean , e_{ij} is the error term and δ_i is a random effect due to each mouse. Notably these four mice are considered as a sample of the overall population of female mice of that breed, consequently an important characteristic of random effects models is that the δ_i values are a random sample of all δ terms. Therefore these random terms can be used for making inferences about populations. (McCullough and Searle, 2001).

1.2.5 Fixed or Random?

In the examples discussed so far, it is clear when effects are fixed or random. In general, however, the difference is not as clear. Searle (1997) discusses the decision whether an effect should be treated as random or fixed, stating that it depends upon the context of the study, and how the data was gathered. *The situation to which the model applies is the deciding factor in determining whether effects are fixed or random.*

Referring to the Grubbs data, the shells fired are a random sample of shells therefore α_i components should be considered random effects. Conversely β_j are fixed effects components because the three measurement devices are the only instruments of interest (Searle, 1997).

1.2.6 Matrix Formulation

There are matrix (i.e multivariate) formulations of both fixed effects models and random effects models. Brown and Prescott (1999) 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;

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{1.4}$$

$\mathbf{Y}$ 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 $q \times 1$.

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{1.5}$$

Again, once the component fixed effects and random effects components are considered, progression to a mixed model formulation is a simple step. Further to Laird and Ware (1982), it is conventional to formulate a mixed effects model in matrix form as follows:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{1.6}$$

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

1.2.7 Advantages of Mixed Models

Brown and Prescott (1999) 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.

Demidenko (2004) 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.

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.

1.2.8 Variance Components

Each random effect has an associated 'variance component' term. This is a model parameter which quantifies random variation due to that effect only. Therefore for every observation there are two sources of variation, random variation and residual variation, and can be expressed as follows $var(y_{ij}) = \sigma_p^2 + \sigma^2$. These variations are known as the variance components. This is an important difference with fixed effects models, which is subject to residual variation (i.e σ^2) only (Brown and Prescott, 1999).

In fixed effects models there is no covariance between any pair of observations. Brown and Prescott (1999) shows that, while there is no covariance between observations from different subjects (i.e the mice in Searle's example), there exists correlation between observations from the same subject (i.e. litter weights from the same mouse are correlated). In this case the covariance is the subjects variance component (i.e. σ_p^2)

1.2.9 More complex examples

Searle (1997) offers elaborations on both examples used so far. In the case of the fixed effects model, the model can be amended to take account for different varieties of each plants being studied. (The groups of six plants are subdivided into three variety types.) y_{ijk} is the yield of the kth plant of the jth variety in the ith treatment, and is described as follows;

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \tag{1.7}$$

This new formulation includes a fixed effect β_j to account for the variety type, and an 'interaction effect' γ_{ij}. An interaction effect describes the combined effects of two or more variables on the observation. Similarly the random effects example is elaborated to account for the effects of three different technicians. Again there is a random effect component τ to account for these technicians, and an interaction effect θ to account for the combined effect of the mice and the technicians.

$$w_{ijk} = \mu + \delta_i + \tau_j + \theta_{ij} + e_{ijk} \tag{1.8}$$

1.3 Using Linear Mixed Effects Models

It is shown that classical models can give different results to linear mixed effects models , based on the same data. Demidenko (2004) illustrates this with a comparison of simple regression model of prices against sales, with a mixed model, that takes groups of data into account.

1.4 Grouped Data Sets

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 (Faraway, 2006). Observations in different groups may be independent, but any assumption that these observations within the same group are independent is inappropriate . Consequently Demidenko (2004) states that there is two sources of variations to be considered, 'within groups' and 'between groups'.

1.5 Classical Models

Demidenko (2004) discusses the inadequacy of 'classical models' in analysing such data types, with particular reference to the simple linear model . The simple linear model is a well known statistical methodology that describes the relationship between dependent variables Y and an independent predictor variable X. Where $Y = y_1, y_2, ..y_k..y_n$ and $X = x_1, x_2, ..x_k..x_n$, an intercept α and slope β are estimated such that the error terms associated with each observation y_i is minimised.

$$y_k = \alpha + \beta x_k + \epsilon_k \tag{1.9}$$

In classical statistics a typical assumption is that observations are drawn from the same general population, are independent and identically distributed (Demidenko, 2004). Consequently there is no way to account for the grouped nature of data sets described previously, and so there lies the possibility of observations being treated as independent measurements. Demidenko (2004, pg.3) gives a very informative example wherein a classical approach is compared to an approach that does account for grouping. The conclusion to be drawn from Demidenko's example is that failure to account for grouping leads to an incorrect conclusion about the data. The approach recommended is known as 'mixed models' and shall be introduced presently.

Chapter 2

Linear Mixed effects Models

2.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (random effects are also known as variance components). LME models are a generalization of the classical linear model, which contain fixed effects only. When the levels of factors are considered to be sampled from a population, and each level is not of particular interest, they are considered random quantities with associated variances. The effects of the levels, as described, are known as random effects. Random effects are represented by unobservable normally distributed random variables. Conversely fixed effects are considered non-random and the levels of each factor are of specific interest.

Fisher (1918) introduced variance components models for use in genetical studies. Whereas an estimate for variance must take an non-negative value, an individual variance component, i.e. a component of the overall variance, may be negative.

The framework has developed since, including contributions from Tippett (1931), who extend the use of variance components into linear models, and Eisenhart (1947), who introduced the 'mixed model' terminology and formally distinguished between mixed and random effects models. Henderson (1950) devised a framework for deriving

estimates for both the fixed effects and the random effects, using a set of equations that would become known as 'mixed model equations' or 'Henderson's equations'. LME methodology is further enhanced by Henderson's later works (Henderson, 1953; Henderson et al., 1959a, 1963, 1973, 1984). The key features of Henderson's work provide the basis for the estimation techniques.

Hartley and Rao (1967) demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. However these estimates are known to be biased 'downwards' (i.e. underestimated) , because of the assumption that the fixed estimates are known, rather than being estimated from the data. Patterson and Thompson (1971) produced an alternative set of estimates, known as the restricted maximum likelihood (REML) estimates, that do not require the fixed effects to be known. Thusly there is a distinction the REML estimates and the original estimates, now commonly referred to as ML estimates.

? provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (Singer, 1998). Pinheiro and Bates (1994) described how to compute LME models in the `S-plus` environment.

Using Laird-Ware form, the LME model is commonly described in matrix form,

$$y = X\beta + Zb + \epsilon \tag{2.1}$$

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and

$$\mathrm{var} \begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}$$

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of

observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

2.2 Statement of the LME model

A linear mixed effects model is a linear mdoel that combined fixed and random effect terms formulated by **?** as follows;

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

- Y_i is the $n \times 1$ response vector
- X_i is the $n \times p$ Model matrix for fixed effects
- β is the $p \times 1$ vector of fixed effects coefficients
- Z_i is the $n \times q$ Model matrix for random effects
- b_i is the $q \times 1$ vector of random effects coefficients, sometimes denoted as u_i
- ϵ is the $n \times 1$ vector of observation errors

The linear mixed effects model is given by

$$Y = X\beta + Zu + \epsilon \tag{2.2}$$

2.2.1 Statement of the LME model

These models are used when there are both fixed and random effects that need to be incorporated into a model.

Fixed effects usually correspond to experimental treatments for which one has data for the entire population of samples corresponding to that treatment.

Random effects,on the other hand, are assigned in the case where we have measurements on a group of samples, and those samples are taken from some larger sample pool, and are presumed to be representative.

As such, linear mixed effects models treat the error for fixed effects differently than the error for random effects.

2.2.2 Laird Ware Model

? provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (Singer, 1998). Pinheiro and Bates (1994) described how to compute LME models in the `S-plus` environment.

Linear mixed effects models (LME) differs from the conventional linear model in that it has both fixed effects and random effects regressors, and coefficients thereof. Further to a paper published by Laird and Ware in 1982, it is conventional to formulate an LME in matrix form as follows:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

- Y_i is the $n \times 1$ response vector
- X_i is the $n \times p$ Model matrix for fixed effects
- β is the $p \times 1$ vector of fixed effects coefficients
- Z_i is the $n \times q$ Model matrix for random effects
- b_i is the $q \times 1$ vector of random effects coefficients, sometimes denoted as u_i
- ϵ is the $n \times 1$ vector of observation errors

2.3 The Linear Mixed Effects Model

Using Laird-Ware form, the LME model is commonly described in matrix form,

$$y = X\beta + Zb + \epsilon \tag{2.3}$$

$\mathbf{Y}$ 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. $\mathbf{X}$ 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 $q \times 1$.

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and

$$\mathrm{var}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}$$

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ.

The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$.

var($\mathbf{Xb}$) is known to be zero. The variance of the random effects var($\mathbf{Zu}$) can be written as $Z\mathrm{var}(\mathbf{u})Z^T$.

By letting $\mathrm{var}(u) = G$ (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;

$$\mathbf{V} = ZGZ^T + \mathbf{R} \tag{2.4}$$

It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

2.3.1 Formulation of the response vector

Information of individual i is recorded in a response vector $\boldsymbol{y}_i$. The response vector is constructed by stacking the response of the 2 responses at the first instance, then the 2 responses at the second instance, and so on. Therefore the response vector is a $2n_i \times 1$ column vector. The covariance matrix of $\boldsymbol{y_i}$ is a $2n_i \times 2n_i$ positive definite matrix $\boldsymbol{\Omega}_i$.

Consider the case where three measurements are taken by both methods A and B, $\boldsymbol{y}_i$ is a 6×1 random vector describing the ith subject.

$$\boldsymbol{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})\boldsymbol{\prime}$$

The response vector $\boldsymbol{y_i}$ can be formulated as an LME model according to Laird-Ware form.

$$\begin{aligned} \boldsymbol{y_i} &= \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i} \\ \boldsymbol{b_i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}) \\ \boldsymbol{\epsilon_i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{R_i}) \end{aligned}$$

Information on the fixed effects are contained in a three dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)\boldsymbol{\prime}$. For computational purposes β_2 is conventionally set to zero. Consequently $\boldsymbol{\beta}$ is the solutions of the means of the two methods, i.e. $E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. The variance covariance matrix $\boldsymbol{D}$ is a general 2×2 matrix, while $\boldsymbol{R}_i$ is a $2n_i \times 2n_i$ matrix.

2.3.2 Formulation of the Variance Matrix V

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

$$\mathbf{V} = \mathrm{Var}(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \tag{2.5}$$

$$\mathbf{V} = \mathrm{Var}(\mathbf{Xb}) + \mathrm{Var}(\mathbf{Zu}) + var(\mathbf{e})) \tag{2.6}$$

$\mathrm{Var}(\mathbf{Xb})$ is known to be zero. The variance of the random effects $\mathrm{Var}(\mathbf{Zu})$ can be written as $Z\mathrm{Var}(\mathbf{u})Z^T$.

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, ($\mathbf{e} \sim N(0, \mathbf{R})$). 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;

$$\mathbf{V} = ZGZ^T + \mathbf{R} \tag{2.7}$$

2.4 Likelihood and estimation

The likelihood function ($L(\theta)$)is a fundamental concept in statistical inference. It indicates how likely a particular population is t o produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters. For computational ease, it is common to use the logarithm of the likelihood function, known simply as the log-likelihood ($\ell(\theta)$).

2.4.1 Likelihood-based tools

Likelihood functions provide the basis for two important statistical concepts that shall be further referred to; the likelihood ratio test and the Akaike information criterion.

Maximum likelihood (ML) estimation is a method of obtaining parameter estimates by optimizing the likelihood function. The likelihood function is constructed as a function of the parameters in the specified model.

Restricted maximum likelihood (REML) is an alternative methods of computing parameter estimated. REML is often preferred to ML because it produces unbiased estimates of covariance parameters by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$.

REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML. The problem with REML for model building is that the "likelihoods" obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed

effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Assuming a statistical model $f_\theta(y)$ parameterized by a fixed and unknown set of parameters θ, the likelihood $L(\theta)$ is the probability of the observed data y considered as a function of θ (**?**).

The log likelihood $l(\theta)$

2.5 Likelihood estimation techniques

Maximum likelihood and restricted maximum likelihood have become the most common strategies for estimating the variance component parameter θ. Maximum likelihood estimation obtains parameter estimates by optimizing the likelihood function. To obtain ML estimate the likelihood is constructed as a function of the parameters in the specified LME model. The maximum likelihood estimates (MLEs) of the parameters are the values of the arguments that maximize the likelihood function. The REML approach is a variant of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (Dodge, 2003).

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 (Demidenko, 2004).

2.5.1 Restricted Likelihood Estimation

A method related to ML is restricted maximum likelihood estimation(REML). REML was developed by Paterson and Thompson (1971) and Harville (1977) 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.

McCullough and Searle (2001) 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 σ^2 is $\frac{S_{yy}}{n}$ whereas the REML estimator is $\frac{S_{yy}}{n-1}$. (S_{yy} is the sum of square identity;

$$S_{yy} = \Sigma_{i=i}^{N}(y - \bar{y})^2 \tag{2.8}$$

Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Chapter 3

Linear Mixed effects Models

3.1 Computation with R

When tackling linear mixed effects models using the R language, a statistician can call upon the *lme* command found in the *nlme* package.This command fits a LME model to the data set using either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML).

The first two arguments for *lme* are *fixed* and *data*, which give the model for the expected responses (i.e. the fixed part of the model), and the data that the model should be fitted from. The next argument is *random*, a one-sided formula which describes the random effects, and the grouping structure for the model. The *method* argument can specify whether to use 'REML', the default setting, or 'ML'.

3.2 Linear Mixed Effects Models

Applications

So-called mixed-effect models (or just mixed models) include additional random-effect terms, and are often appropriate for representing clustered, and therefore dependent, data arising, for example, when data are collected hierarchically, when observations

are taken on related individuals (such as siblings), or when data are gathered over time on the same individuals.

3.2.1 Restricted Maximum Likelihood

restricted (or residual) maximum likelihood (REML) is a method for fitting linear mixed models. In contrast to conventional maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters.

3.3 Model Selection

The previous section on estimation assumes the specification of a mixed model in terms of X, Z, G, and R. Even though X and Z have known elements, there is some flexibility is specifying the form and construction is flexible, and for a particular data set, there are numerous possibilities that can be considered. Similarly, various potential covariance structures for G and R may be considered.

First, subject matter considerations and objectives are of great importance when selecting a model; refer to Diggle (1988) and Lindsey (1993).

Second, when the data themselves are looked to for guidance, many of the graphical methods and diagnostics appropriate for the general linear model extend to the mixed model setting as well (Christensen, Pearson, and Johnson 1992).

Likelihood-based approachs to the mixed model allow the comparison of candidate models. The most common of these are the likelihood ratio test and Akaike's and Schwarz's information criteria (Bozdogan 1987; Wolfinger 1993).

Chapter 4

Henderson

4.1 Repeated measurements in LME models

In many statistical analyzes, the need to determine parameter estimates where multiple measurements are available on each of a set of variables often arises. Further to Lam et al. (1999), Hamlett et al. (2004) performs an analysis of the correlation of replicate measurements, for two variables of interest, using LME models.

Let y_{Aij} and y_{Bij} be the jth repeated observations of the variables of interest A and B taken on the ith subject. The number of repeated measurements for each variable may differ for each individual. Both variables are measured on each time points. Let n_i be the number of observations for each variable, hence $2 \times n_i$ observations in total.

It is assumed that the pair y_{Aij} and y_{Bij} follow a bivariate normal distribution.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

The matrix Σ represents the variance component matrix between response variables at a given time point j.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

σ_A^2 is the variance of variable A, σ_B^2 is the variance of variable B and σ_{AB} is the covariance of the two variable. It is assumed that $\mathbf{\Sigma}$ does not depend on a particular time point, and is the same over all time points.

? used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

The true correlation ρ_{xy} is repeated measurements can be considered as having two components: between subject and within-subject correlation. The usefulness of estimating repeated measure correlation coefficients is the calculation of between-method and within-method variabilities are produced as by-products.

4.1.1 Decomposition of the response covariance matrix

The variance covariance structure can be re-expressed in the following form,

$$\text{Cov}(\text{y}_i) = \boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R_i}.$$

The variance covariance structure can be re-expressed in the following form,

$$\text{Cov}(\text{y}_i) = \boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R_i}.$$

$\boldsymbol{\Omega_i}$ can be expressed as

$$\boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + (\boldsymbol{I}_{\boldsymbol{n_i}} \otimes \boldsymbol{\Lambda}).$$

The notation $\dim_{n_i}$ means an $n_i \times n_i$ diagonal block. $\boldsymbol{R_i}$ can be shown to be the Kronecker product of a correlation matrix $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$. The correlation matrix $\boldsymbol{V}$ of the repeated measures on a given response variable is assumed to be the same for all response variables. Both Hamlett et al. (2004) and Lam et al. (1999) use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. Roy (2009a) remarks that, with repeated measures, the response for each subject is correlated for each variable, and that such correlation must be taken into account in order to produce a valid inference on correlation estimates. Roy (2006) proposes various correlation structures may be assumed for repeated measure correlations, such as the compound symmetry and autoregressive structures, as alternative to the identity matrix.

However, for the purposes of method comparison studies, the necessary estimates are currently only determinable when the identity matrix is specified, and the results in Roy (2009a) indicate its use.

For the response vector described, Hamlett et al. (2004) presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a 6×6 matrix composed of two types of 2×2 blocks. Each block represents one separate time of measurement.

$$\boldsymbol{\Omega}_i = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{D} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{\Sigma} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{D} & \boldsymbol{\Sigma} \end{pmatrix}$$

The diagonal blocks are Σ, as described previously. The 2×2 block diagonal matrix in $\boldsymbol{\Omega}$ gives $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is the sum of the between-subject variability $\boldsymbol{D}$ and the within subject variability $\boldsymbol{\Lambda}$.

$\boldsymbol{\Omega_i}$ can be expressed as

$$\boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + (\boldsymbol{I}_{\boldsymbol{n_i}} \otimes \boldsymbol{\Lambda}).$$

The notation $\dim_{n_i}$ means an $n_i \times n_i$ diagonal block.

4.1.2 Correlation terms

Hamlett et al. (2004) demonstrated how the between-subject and within subject variabilities can be expressed in terms of correlation terms.

$$\boldsymbol{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_b \rho_{AB} \delta \\ \sigma_A \sigma_b \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_A^2 (1 - \rho_A) & \sigma_{AB} (1 - \delta) \\ \sigma_{AB} (1 - \delta) & \sigma_B^2 (1 - \rho_B) \end{pmatrix}.$$

ρ_A describe the correlations of measurements made by the method A at different times. Similarly ρ_B describe the correlation of measurements made by the method B at different times. Correlations among repeated measures within the same method are known as intra-class correlation coefficients. ρ_{AB} describes the correlation of measurements taken at the same same time by both methods. The coefficient δ is added for when the measurements are taken at different times, and is a constant of less than 1 for linked replicates. This is based on the assumption that linked replicates measurements taken at the same time would have greater correlation than those taken at different

times. For unlinked replicates δ is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

? used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

The true correlation ρ_{xy} is repeated measurements can be considered as having two components: between subject and within-subject correlation. The usefulness of estimating repeated measure correlation coefficients is the calculation of between-method and within-method variabilities are produced as by-products.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (7.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model in (**??**) requires $N + 2$ fixed effects.

Allocating fixed effects to each item i by (**??**) accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, Roy (2009a) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

4.2 Henderson's equations

Because of the dimensionality of V (i.e. $n \times n$) computing the inverse of V can be difficult. As a way around the this problem Henderson (1953); Henderson et al. (1959a, 1963, 1973, 1984) offered a more simpler approach of jointly estimating $\hat{\beta}$ and $\hat{b}$. Henderson (1950) made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and

$b \sim \mathrm{N}(0, D)$, and proceeded to maximize the joint density of y and b

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (4.1)$$

with respect to β and b, which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (4.2)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (4.3)$$

Using these equations, obtaining the estimates requires the inversion of a matrix of dimension $p+q \times p+q$, considerably smaller in size than V. **?** shows that these mixed model equations do not depend on normality and that $\hat{\beta}$ and $\hat{b}$ are the BLUE and BLUP under general conditions, provided D and Σ are known.

Robinson (1991) points out that although Henderson (1950) initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (4.3) as "joint maximum likelihood estimates", Henderson (1973) later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (4.5) is a joint density rather than a likelihood function. Lee et al. (2006) remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

4.3 Computation on a mixed effects model

? 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.

(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 (β_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 (τ_{ij}) describes whether the effect of changing from one machine to another is different for each worker. The productivity score y_{ijk} is the kth observation taken on worker i on machine j, and is formulated as follows;

$$y_{ijk} = \beta_j + u_i + \tau_{ij} + \epsilon_{ijk} \tag{4.4}$$

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 $\sigma_\epsilon^2 = (0.96)^2$.

In simple examples V^{-1} is a straightforward calculation, but with higher dimensions it becomes a very complex calculation.

4.3.1 Henderson's equations

Henderson (1950) made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and $b \sim \mathrm{N}(0, D)$, and proceeded to maximize the joint density of y and b with respect to β and b, which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \tag{4.5}$$

This leads to the solutions Robinson (1991) points out that although Henderson (1950) initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (4.3) as "joint maximum likelihood estimates" Henderson (1973) later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (4.5) is a joint density rather than a likelihood function.

4.3.2 Henderson's Mixed Model Equations

Henderson et al. (1959b, 1963, 1973, 1984) derived the 'mixed model equations (MME)' to provide estimates for **b** and **u**without the need to calculate the inverse of **V**.

When **R** and **G** are diagonal, determining the inverses thereof are trivial calculations, and therefore the above matrices are much simpler to solve , and overcomes the problem posed by the inverse of **V**.

Each of the elements of the above matrices are submatrices. $X^T R^{-1} X$ is a $p \times p$ matrix, $Z^T R^{-1} Z + G^{-1}$ is a $q \times q$ matrix. The remaining elements, which are transposes of each other, are of dimensions $p \times q$ and $q \times p$ respectively. Therefore the overall matrix is of dimension $(p+q) \times (p+q)$. These dimensions are notably smaller than $n \times n$, which would have been the case if V^{-1}, and therefore the inversion is easier to compute.

Henderson et al. (1959b, 1963, 1973, 1984) derived the 'Mixed Model Equations (MME)' to provide estimates for β and u without the need to calculate V^{-1}.

Rearranging the equation CC, the BLUE of β, and the BLUP of u can be shown to be;

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{4.6}$$

$$\hat{u} = G Z^T V^{-1} (y - X\hat{\beta}) \tag{4.7}$$

4.3.3 Estimation

Estimation of LME models involve two complementary estimation issues'; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and $\hat{b}$ and estimating the variance covariance matrices D and Σ.

Inference about fixed effects have become known as 'estimates', while inferences about random effects have become known as 'predictions'.

The most common approach to obtain estimators are Best Linear Unbiased Estimator (BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (2.3), the BLUE of $\hat{\beta}$ is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

whereas the BLUP of $\hat{b}$ is given by

$$\hat{b} = DZ'V^{-1}(y - X\hat{\beta}).$$

Estimation of the fixed parameters

The vector y has marginal density $y \sim \mathrm{N}(X\beta, V)$, where $V = \Sigma + ZDZ'$ is specified through the variance component parameters θ. The log-likelihood of the fixed parameters (β, θ) is

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{4.8}$$

and for fixed θ the estimate $\hat{\beta}$ of β is obtained as the solution of

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{4.9}$$

Maximum likelihood and restricted maximum likelihood have become the most common strategies for estimating the variance component parameter θ.

Substituting $\hat{\beta}$ from (4.9) into $\ell(\beta, \theta|y)$ from (4.8) returns the *profile* log-likelihood

$$\begin{aligned} \ell_P(\theta \mid y) &= \ell(\hat{\beta}, \theta \mid y) \\ &= -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) \end{aligned}$$

of the variance parameter θ. Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta \mid y)$ over θ. These are the ML estimates. Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta \mid y)$ over θ.

For REML estimation the *restricted* log-likelihood is defined as

$$\ell_R(\theta \mid y) = \ell_P(\theta \mid y) - \frac{1}{2}\log|X'VX|.$$

In practice the *restricted* log-likelihood is preferred. This approach is based on maximizing the likelihood of linear combinations of y that do not depend on β, and in this way takes into account the estimation of β.

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (Dodge, 2003).

Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account

the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood.

The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $DZ'V^{-1}(y-X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (4.9) so that $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and $\hat{b}$ satisfy the equations in (4.3).

Algorithms for likelihood function optimization

Iterative numerical techniques are used to optimize the log-likelihood function and estimate the covariance parameters θ. The procedure is subject to the constraint that R and D are both positive definite. The most common iterative algorithms for optimizing the likelihood function are the Newton-Raphson method, which is the preferred method, the expectation maximization (EM) algorithm and the Fisher scoring methods.

The EM algorithm, introduced by Dempster et al. (1977), is an iterative technique for maximizing complicated likelihood functions. The algorithm alternates between performing an expectation (E) step and the maximization (M) step. The 'E' step computes the expectation of the log-likelihood evaluated using the current estimate for the variables. In the 'M' step, parameters that maximize the expected log-likelihood, found on the previous 'E' step, are computed. These parameter estimates are then

used to determine the distribution of the variables in the next 'E' step. The algorithm alternatives between these two steps until convergence is reached.

The main drawback of the EM algorithm is its slow rate of convergence. Consequently the EM algorithm is rarely used entirely in LME estimation, instead providing an initial set of values that can be passed to other optimization techniques.

The Newton Raphson (NR) method is the most common, and recommended technique for ML and REML estimation. The NR algorithm minimizes an objective function defines as -2 times the log likelihood for the covariance parameters θ. At every iteration the NR algorithm requires the calculation of a vector of partial derivatives, known as the gradient, and the second derivative matrix with respect to the covariance parameters. This is known as the observed Hessian matrix. Due to the Hessian matrix, the NR algorithm is more time-consuming, but convergence is reached with fewer iterations compared to the EM algorithm. The Fisher scoring algorithm is an variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

4.3.4 Estimators and Predictors

The best linear unbiased predictor (BLUP) is used to estimating random effects, i.e to derive $\mathbf{u}$. The best linear unbiased estimator (BLUE) is used to estimate the fixed effects, $\mathbf{b}$. They were formulated in a paper by Henderson et al. (1959a), which provides the derivations of both. Inferences about fixed effects have come tobe called 'estimates', whereas inferences about random effects have come be called 'predictions'. hence the naming of BLUP is to reinforce distinction between the two , but it is essentially the same principal involved in both cases, (GK, 1991). The procedures are known as the 'best' in the sense that they minimise the sampling variance and unbiased in the sense that E[BLUE($\mathbf{b}$)]= $\mathbf{b}$ and E[BLUP($\mathbf{u}$)] = $\mathbf{u}$. The BLUE of $\mathbf{b}$, and the BLUP of $\mathbf{u}$ can be shown to be;

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{4.10}$$

$$\hat{u} = G Z^T V^{-1} (y - X\hat{b}) \tag{4.11}$$

The practical application of both expressions requires that the variance components be known. Therefore an estimate for the variance components must be derived to analysis by either ANOVA, or REML, a method that shall be introduced shortly. Importantly calculations based on the above formulae require the calculation of the inverse of $\mathbf{V}$. In simple examples V^{-1} is a straightforward calculation, but with higher dimensions it becomes a very complex calculation.

The estimate for the fixed effects are referred to as the best linear unbiased estimates (BLUE). Henderson's estimate for the random effects is known as the best linear unbiased predictor (BLUP).

4.4 Extension of Roy's methodology

Roy's methodology is constructed to compare two methods in the presence of replicate measurements. Necessarily it is worth examining whether this methodology can be adapted for different circumstances.

An implementation of Roy's methodology, whereby three or more methods are used, is not feasible due to computational restrictions. Specifically there is a failure to reach convergence before the iteration limit is reached. This may be due to the presence of additional variables, causing the problem of non-identifiability. In the case of two variables, it is required to estimate two variance terms and four correlation terms, six in all. For the case of three variabilities, three variance terms must be estimated as well as nine correlation terms, twelve in all. In general for n methods has $2 \times T_n$ variance terms, where T_n is the triangular number for n, i.e. the addition analogue of the factorial. Hence the computational complexity quite increases substantially for every increase in n.

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

The tests would not be informative as to whether any two of those three were interchangeable, or equivalently if one method in particular disagreed with the other two. Indeed it is easier to perform three pair-wise comparisons separately and then to combine the results.

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector y_i, as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

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 4.3.1: Machines Data , Pinheiro Bates

Chapter 5

Introduction

In this section, we introduce the LME model, discusss how it can be applied to MCS problems, and how it is desirable in the case of replicate measurements, giving some examples from previous work (i.e. Carstensen et Al, Lai & Shaio, and Roy). Further to that, there will be a demonstration on fitting various types LME models using freely available software.

While the MCS problem is conventionally poised in the context of two methods of measurements, LME models allow for a straightforward analysis whereby several methods of measurement can be measured simulataneously. However simple models only can only indicate agreement of lack thereof, and the presence of inter-method bias. To consider more complex questions, more complex LME models are required. Useful approaches will be introduced in a later section.

5.1 LME models in method comparison studies

Barnhart et al. (2007) describes the sources of disagreement in a method comparison study problem as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods. Further to this, Roy (2009b) states three criteria for

two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Roy (2009b) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

The LME model approach has seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples)

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement.

5.2 LAI-SHIAO

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement.Lai and Shiao (2005) view the LME Models approach as an natural expansion to the Bland ? Altman method for comparing two measurement methods. Lai and Shiao (2005) is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem.

Lai and Shiao (2005) extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. The data used for their examples is unavailable for independent use. Therefore, for the sake of consistency, a data set will be simulated based on the Blood Data that will allow for extra variables, and an exploration shall be provided in the

appendices.

5.3 Lai Shiao

Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

If the parameter **b**, and the variance components are not significantly different from zero, the conclusion that there is no inter-method bias can be drawn. If the fixed effects component contains only the intercept, and a simple correlation coefficient is used, then the estimate of the intercept in the model is the inter-method bias. Conversely the estimates for the fixed effects factors can advise the respective influences each factor has on the differences. It is possible to pre-specify different correlation structures of the variance components **G** and **R**.

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. 'Fractional saturation' (HbO_2) is considered to be the gold standard method of measurement, with 'functional saturation' (SO_2) being an alternative method. The method of examining the causes of differences between these two methods is applied to a clinical study conducted by **?**. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods (described by the variable 'Bloodage') and are categorized according to haemoglobin percentages(i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

Lai and Shiao (2005) fits two models on this data, with the lab technicians and the replicate measurements as the random effects in both models. The first model uses haemoglobin level as a fixed effects component. For the second model, blood age is added as a second fixed factor.

Single fixed effect

The first model fitted by Lai and Shiao (2005) takes the blood level as the sole fixed effect to be analyzed. The following coefficient estimates are estimated by 'Proc Mixed';

$$\begin{aligned} &\text{fixed effects} : 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \\ &(\text{p-values} : \ = 0.0054, < 0.0001, < 0.0001) \\ \\ \text{random effects} : \ &u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma_e^2 = 0.1525, \rho = 0.6978) \\ &(\text{p-values} : \ = 0.8113, < 0.0001, < 0.0001) \end{aligned} \tag{5.1}$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of SO_2. Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

Two fixed effects

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\begin{aligned} \text{fixed effects}: \; & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ & (\text{p-values}: \; = 0.8113, < 0.0001, < 0.0001) \\ \\ \text{random effects}: \; & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ & (\text{p-values}: \; = 0.0446, < 0.0001, < 0.0001) \end{aligned} \tag{5.2}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidisation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

Carstensen

Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' methods.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Rather than using the 'by hand' methods, estimates for required LME parameters can be read directly from program output.

Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. Rather than using the 'by hand' methods, estimates for required parameters can be gotten directly from output code. Furthermore, using computer approaches removes constraints, such as the need for the design to be perfectly balanced. In part this is due to the increased profile of LME models, and furthermore the availability of capable software.

Additionally a great understanding of residual analysis and influence analysis for LME models has been adchieved thanks to authors such as **?**, **?**, Cook (1986) West et al. (2007), amongst others. In this chapter various LME approaches to method comparison studies shall be examined.

Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

Roys uses and LME model approach to provide a set of formal tests for method comparison studies.

5.4 Introduction to LME Models, Fitting LME Models to MCS Data

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland and Altman (1999) suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods. The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be incorrect, (Carstensen, 2004). Carstensen (2004) recommends that replicate measurements for each method, but recognizes that resulting data are more difficult to analyze. To this end, Carstensen (2004) and Carstensen et al. (2008) recommend the use of LME models as a suitable framework for method comparison in the case of repeated measurements.

Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them.

This approach has seen increased use in method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples). In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

5.5 Model Specification for Roy's Hypotheses Tests

In order to express Roy's LME model in matrix notation we gather all $2n_i$ observations specific to item i into a single vector $\boldsymbol{y}_i = (y_{1i1}, y_{2i1}, y_{1i2}, \ldots, y_{mir}, \ldots, y_{1in_i}, y_{2in_i})'$. The LME model can be written

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a vector of fixed effects, and $\boldsymbol{X}_i$ is a corresponding $2n_i \times 3$ design matrix for the fixed effects. The random effects are expressed in the vector

$\boldsymbol{b} = (b_1, b_2)'$, with $\boldsymbol{Z}_i$ the corresponding $2n_i \times 2$ design matrix. The vector $\boldsymbol{\epsilon}_i$ is a $2n_i \times 1$ vector of residual terms. Random effects and residuals are assumed to be independent of each other.

It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, $\boldsymbol{\epsilon}_i$ is a matrix of random errors distributed as $N(0, \boldsymbol{R}_i)$ and that the random effects and residuals are independent of each other.

The random effects are assumed to be distributed as $\boldsymbol{b}_i \sim \mathcal{N}_2(0, \boldsymbol{G})$. $\boldsymbol{G}$ is the variance covariance matrix for the random effects $\boldsymbol{b}$. i.e. between-item sources of variation. The between-item variance covariance matrix $\boldsymbol{G}$ is constructed as follows:

$$\boldsymbol{G} = \text{Var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$. The random effects and residuals are assumed to be independent.

$$\text{Var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \boldsymbol{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

It is important to note that no special assumptions about the structure of $\boldsymbol{G}$ are made. An example of such an assumption would be that $\boldsymbol{G}$ is the product of a scalar value and the identity matrix. It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, $\boldsymbol{\epsilon}_i$ is a matrix of random errors distributed as $N(0, \boldsymbol{R}_i)$ and that the random effects and residuals are independent of each other. Assumptions made on the structures of $\boldsymbol{G}$ and $\boldsymbol{R}_i$ will be discussed in due course.

The random effects are assumed to be distributed as $\boldsymbol{b}_i \sim \mathcal{N}_2(0, \boldsymbol{G})$. The between-item variance covariance matrix $\boldsymbol{G}$ is constructed as follows:

$$\boldsymbol{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

It is important to note that no special assumptions about the structure of $\boldsymbol{G}$ are made. An example of such an assumption would be that $\boldsymbol{G}$ is the product of a scalar value and the identity matrix.

The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$. Hamlett et al. (2004) shows that the variance covariance matrix for the residuals(i.e. the within-item sources of variation between both methods) can be expressed as the Kroneckor product of an $n_i \times n_i$ identity matrix and the partial within-item variance covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications.Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$. Hamlett et al. (2004) shows that the variance covariance matrix for the residuals(i.e. the within-item sources of variation between both methods) can be expressed as the Kroneckor product of an $n_i \times n_i$ identity matrix and the partial within-item variance covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

The partial within-item variance covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of both methods, and

σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$. The above terms can be used to express the variance covariance matrix $\boldsymbol{\Omega}_i$ for the responses on item i ,

$$\boldsymbol{\Omega}_i = \boldsymbol{Z}_i \boldsymbol{G} \boldsymbol{Z}_i' + \boldsymbol{R}_i.$$

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$.

5.6 Model Set Up

Roy (2009b) proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects. Response for ith subject can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_{1i} z_{i1} + b_{2i} z_{i2} + \epsilon_i$$

- β_1 and β_2 are fixed effects corresponding to both methods. (β_0 is the intercept.)
- b_{1i} and b_{2i} are random effects corresponding to both methods.

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability (Σ),

$$\text{Block } \boldsymbol{\Omega}_i = \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

5.7 Roy's Approach

Roy (2009b) proposes the use of LME models to perform a test on two methods of agreement to comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available, determining whether they can be used interchangeably. This approach uses a Kronecker product covariance structure with doubly multivariate setup to assess the agreement, and is designed such that the data may be unbalanced and with unequal numbers of replications for each subject (Roy, 2009b).

Roy (2009b) uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available. She provides three tests of hypothesis appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

For the purposes of comparing two methods of measurement, Roy (2009b) presents a framework that utilizes linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. Roy (2009b) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. (An item would commonly be a patient). Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2\colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3\colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009b) uses

a Bonferroni correction to control the familywise error rate for tests of $\{H_1, H_2, H_3\}$ and account for difficulties arising due to multiple testing.

The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009b) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. The well-known "Limits of Agreement", as developed by Bland and Altman (1986) are easily computable using the LME framework, proposed by Roy. While we will not be considering this analysis, a demonstration will be provided in the example.

These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

A formal test for inter-method bias can be implemented by examining the fixed effects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means. The inter-method bias and necessary $t-$value and $p-$value are presented in computer output. A decision on whether the first of Roy's criteria is fulfilled can be based on these values.

Importantly Roy (2009b) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Roy's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

A symmetric matrix allows the diagonal terms a_{11} and a_{22} to differ. The compound symmetry structure requires that both of these terms be equal, i.e $a_{11} = a_{22}$.

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models.

Roy also integrates H_2 and H_3 into a single testable hypothesis H_4: $\omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m. Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

5.8 Roy's hypothesis tests : Roy's variability tests

For the purposes of method comparison, Roy presents a methodology utilising linear mixed effects model. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to **?**, it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Lack of agreement can also arise if there is a disagreement in overall variabilities. This lack of agreement may be due to differing between-item variabilities, differing within-item variabilities, or both. The formulation previously presented usefully facilitates a series of significance tests that assess if and where such differences arise. Roy allows for a formal test of each. These tests are comprised of a formal test for the equality of between-item variances, Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed. The difference in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

Conversely, the tests of variability required detailed explanation. Each test is performed by fitting two candidate models, according with the null and alternative hypothesis respectively. The distinction between the models arise in the specification in one, or both, of the variance-covariance matrices.

5.9 Agreement Criteria

Roy sets out three conditions for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Should both the second and third conditions be fulfilled, then the overall variabilities of both methods would be equal. Roy additionally uses the overall correlation coefficient to provide extra information about the comparison, with a minimum of 0.82 being required.

Roy's method considers two methods to be in agreement if three conditions are met.

- no significant bias, i.e. the difference between the two mean readings is not "statistically significant",
- high overall correlation coefficient,
- the agreement between the two methods by testing their repeatability coefficients.

Roy (2009b) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. Further to this, Roy(2009) demonstrates an suite of tests that can be used to determine how well two methods of measurement, in the presence of repeated measures, agree with each other.

- No Significant inter-method bias
- No difference in the between-subject variabilities of the two methods

- No difference in the within-subject variabilities of the two methods

Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects.

Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other.

Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods. Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal.

Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods.

Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

5.10 Using LME for method comparison

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Rather than using the 'by hand' methods, estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes constraints associated with 'by-hand' approaches, such as the need for the design to be perfectly balanced.

5.10.1 Roy's methodology

For the purposes of comparing two methods of measurement, Roy (2009a) presents a methodology utilizing linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009a) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

A formal test for inter-method bias can be implemented by examining the fixed ef-

fects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means. The inter-method bias and necessary $t-$value and $p-$value are presented in computer output. A decision on whether the first of Roy's criteria is fulfilled can be based on these values.

Importantly Roy (2009a) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Roy's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

A symmetric matrix allows the diagonal terms a_{11} and a_{22} to differ. The compound symmetry structure requires that both of these terms be equal, i.e $a_{11} = a_{22}$.

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2. The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in more detail shortly.

5.10.2 Correlation coefficient

These tests are complemented by the ability to the overall correlation coefficient of the two methods, which is determinable from variance estimates. Two methods can be considered to be in agreement if criteria based upon these tests are met. Inference for inter-method bias follows from well-established methods and, as such, will only be noted when describing examples.

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. Roy (2009a) remarks that PROC MIXED only gives overall correlation coefficients, but not their variances. Similarly variance are not determinable in `R` as yet either. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

5.11 Roy's Candidate Models : Testing Procedures

Variability tests proposed by Roy (2009b) affords the opportunity to expand upon Carstensen's approach.

Roy's methodology requires the construction of four candidate models. Using Roy's method, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement.

The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models.

Four candidates models are fitted to the data. These models are similar to one another, but for the imposition of equality constraints.

These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.

5.12 Test for inter-method bias

Firstly, a practitioner would investigate whether a significant inter-method bias is present between the methods. This bias is specified as a fixed effect in the LME model. For a practitioner who has a reasonable level of competency in R and undergraduate statistics (in particular simple linear regression model) this is a straight-forward procedure.

A formal test for inter-method bias can be implemented by examining the fixed effects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means.

The inter-method bias and necessary $t-$value and $p-$value are presented in computer output. A decision on whether the first of Roy's criteria is fulfilled can be based on these values.

Bias is determinable by examination of the 't-table'. Estimate for both methods are given, and the bias is simply the difference between the two. Because the R implementation does not account for an intercept term, a $p-$value is not given. Should a $p-$value be required specifically for the bias, and simple restructuring of the model is required wherein an intercept term is included. Output from a second implementation will yield a $p-$value.

The presence of an inter-method bias is the source of disagreement between two methods of measurement that is most easily identified. As the first in a series of hypothesis tests, Roy (2009a) presents a formal test for inter-method bias. With the null and alternative hypothesis denoted H_1 and K_1 respectively, this test is formulated as

$$\mathrm{H}_1 : \mu_1 = \mu_2,$$

$$\mathrm{K}_1 : \mu_1 \neq \mu_2.$$

5.13 Testing Procedures (LRTs)

Roy's methodology requires the construction of four candidate models. The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models.

The probability distribution of the test statistic can be approximated by a chi-square distribution with (ν_1 - ν_2) degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

Likelihood ratio tests are very simple to implement in `R`, simply use the '`anova()`' commands. Sample output will be given for each variability test.

The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{5.3}$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{5.4}$$

5.14 Variability Tests

Variability tests proposed by Roy (2009b) affords the opportunity to expand upon Carstensen's approach. Roy (2009b) considers four independent hypothesis tests. The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

- Testing of hypotheses of differences between the means of two methods
- Testing of hypotheses in between subject variabilities in two methods,
- Testing of hypotheses of differences in within-subject variability of the two methods,
- Testing of hypotheses in differences in overall variability of the two methods.

The formulation presented above usefully facilitates a series of significance tests that advise as to how well the two methods agree. These tests are as follows:

- A formal test for the equality of between-item variances,
- A formal test for the equality of within-item variances,
- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

Importantly Roy (2009b) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Barnhart's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed,

using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

A symmetric matrix allows the diagonal terms a_{11} and a_{22} to differ. The compound symmetry structure requires that both of these terms be equal, i.e $a_{11} = a_{22}$.

The tests are implemented by fitting a four variants of a specific LME model to the data. For the purpose of comparing models, one of the models acts as a reference model while the three other variant are nested models that introduce equality constraints to serves as null hypothesis cases. The methodology uses a linear mixed effects regression fit using a combination of symmetric and compound symmetry (CS) correlation structure the variance covariance matrices.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the reference model.

The methodology uses a linear mixed effects regression fit using compound symmetry (CS) correlation structure on $\mathbf{V}$.

$\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L}$

5.14.1 Variability test 1

The first test determines whether or not both methods A and B have the same between-subject variability, further to the second of Roy's criteria.

$$H_0 : \ d_A = d_B$$
$$H_A : \ d_A \neq d_B$$

This test is facilitated by constructing a model specifying a symmetric form for D (i.e. the alternative model) and comparing it with a model that has compound symmetric form for D (i.e. the null model). For this test $\hat{\mathbf{\Lambda}}$ has a symmetric form for both models, and will be the same for both.

5.14.2 Variability test 2

This test determines whether or not both methods A and B have the same within-subject variability, thus enabling a decision on the third of Roy's criteria.

$$H_0 : \ \lambda_A = \lambda_B$$
$$H_A : \ \lambda_A = \lambda_B$$

This model is performed in the same manner as the first test, only reversing the roles of $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Lambda}}$. The null model is constructed a symmetric form for $\hat{\boldsymbol{\Lambda}}$ while the alternative model uses a compound symmetry form. This time $\hat{\boldsymbol{D}}$ has a symmetric form for both models, and will be the same for both.

As the within-subject variabilities are fundamental to the coefficient of repeatability, this variability test likelihood ratio test is equivalent to testing the equality of two coefficients of repeatability of two methods. In presenting the results of this test, Roy (2009a) includes the coefficients of repeatability for both methods.

5.14.3 Variability test 3

The last of the variability test examines whether or not methods A and B have the same overall variability. This enables the joint consideration of second and third criteria.

$$H_0 : \ \sigma_A = \sigma_B$$
$$H_A : \ \sigma_A = \sigma_B$$

The null model is constructed a symmetric form for both $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Lambda}}$ while the alternative model uses a compound symmetry form for both.

The first test allows of the comparison the begin-subject variability of two methods. As the within-subject variabilities are fundamental to the coefficient of repeatability, this variability test likelihood ratio test is equivalent to testing the equality of two

coefficients of repeatability of two methods. In presenting the results of this test, Roy (2009a) includes the coefficients of repeatability for both methods.

Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

5.14.4 Variability test 3 - Omnibus Test

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as D. The estimate for the within-subject variance covariance matrix is $\hat{\Sigma}$. The estimated overall variance covariance matrix 'Block Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{5.5}$$

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\mathbf{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{5.6}$$

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{5.7}$$

5.14.5 Demonstration of Roy's testing

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples

used are from the 'blood pressure' data set introduced by Bland and Altman (1999). The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted 'J' and 'R') using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted 'S'). Three sets of readings were made in quick succession. Roy compares the 'J' and 'S' methods in the first of her examples.

The inter-method bias between the two method is found to be 15.62 , with a $t-$value of -7.64, with a $p-$value of less than 0.0001. Consequently there is a significant inter-method bias present between methods J and S, and the first of the Roy's three agreement criteria is unfulfilled.

Next, the first variability test is carried out, yielding maximum likelihood estimates of the between-subject variance covariance matrix, for both the null model, in compound symmetry (CS) form, and the alternative model in symmetric (symm) form. These matrices are determined to be as follows;

$$\hat{\boldsymbol{D}}_{CS} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix}, \qquad \hat{\boldsymbol{D}}_{Symm} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix}.$$

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is -2030.7, and for the alternative model -2030.8. The test statistic, presented with greater precision than the log-likelihoods, is 0.1592. The $p-$value is 0.6958. Consequently we fail to reject the null model, and by extension, conclude that the hypothesis that methods J and S have the same between-subject variability. Thus the second of the criteria is fulfilled.

The second variability test determines maximum likelihood estimates of the within-subject variance covariance matrix, for both the null model, in CS form, and the alternative model in symmetric form.

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{CS}} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix}, \qquad \hat{\boldsymbol{\Lambda}}_{Symm} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix}.$$

Again, A likelihood ratio test is perform to compare both candidate models. The

log-likelihood of the alternative model model is -2045.0. As before, the null model has a log-likelihood of -2030.7. The test statistic is computed as 28.617, again presented with greater precision. The $p-$value is less than 0.0001. In this case we reject the null hypothesis of equal within-subject variability. Consequently the third of Roy's criteria is unfulfilled. The coefficient of repeatability for methods J and S are found to be 16.95 mmHg and 25.28 mmHg respectively.

The last of the three variability tests is carried out to compare the overall variabilities of both methods. With the null model the MLE of the within-subject variance covariance matrix is given below. The overall variabilities for the null and alternative models, respectively, are determined to be as follows;

$$\hat{\boldsymbol{\Sigma}}_{CS} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix}, \qquad \hat{\boldsymbol{\Sigma}}_{Symm} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix},$$

The log-likelihood of the alternative model model is -2045.2, and again, the null model has a log-likelihood of -2030.7. The test statistic is 28.884, and the $p-$value is less than 0.0001. The null hypothesis, that both methods have equal overall variability, is rejected. Further to the second variability test, it is known that this difference is specifically due to the difference of within-subject variabilities.

Lastly, Roy considers the overall correlation coefficient. The diagonal blocks $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii}$ of the correlation matrix indicate an overall coefficient of 0.7959. This is less than the threshold of 0.82 that Roy recommends.

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'}$ present the correlation coefficients further to Hamlett et al. (2004).

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'} = \begin{pmatrix} 0.9611 & 0.7799 \\ 0.7799 & 0.9212 \end{pmatrix}.$$

The overall conclusion of the procedure is that method J and S are not in agreement, specifically due to the within-subject variability, and the inter-method bias. The

repeatability coefficients are substantially different, with the coefficient for method S being 49% larger than for method J. Additionally the overall correlation coefficient did not exceed the recommended threshold of 0.82.

5.15 Formal testing for covariances

As it is pertinent to the difference between the two described methodologies, the facilitation of a formal test would be useful. Extending the approach proposed by Roy, the test for overall covariance can be formulated: As with the tests for variability, this test is performed by comparing a pair of model fits corresponding to the null and alternative hypothesis. In addition to testing the overall covariance, similar tests can be formulated for both the component variabilities if necessary.

5.16 Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessarily their prevalence in literature has meant that they are now the best known measurement for agreement, and therefore any newer methodology would benefit by making reference to them.

Carstensen et al. (2008) uses LME models to determine the limits of agreement. Between-subject variation for method m is given by d_m^2 and within-subject variation is given by λ_m^2. Carstensen et al. (2008) remarks that for two methods A and B, separate values of d_A^2 and d_B^2 cannot be estimated, only their average. Hence the assumption that $d_x = d_y = d$ is necessary. The between-subject variability $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

$$\boldsymbol{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_A^2 & 0 \\ 0 & \lambda_B^2 \end{pmatrix}.$$

The variance for method m is $d_m^2 + \lambda_m^2$. Limits of agreement are determined using the standard deviation of the case-wise differences between the sets of measurements

by two methods A and B, given by

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \tag{5.8}$$

Importantly the covariance terms in both variability matrices are zero, and no covariance component is present.

Carstensen et al. (2008) presents a data set 'fat', which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen. Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be 'true' replicates.

A linear mixed effects model is formulated, and implementation through several software packages is demonstrated. All of the necessary terms are presented in the computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{5.9}$$

Roy (2009a) has demonstrated a methodology whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. Using Roy's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{5.10}$$

All of these terms are given or determinable in computer output. The limits of agreement can therefore be evaluated using

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{5.11}$$

For Carstensen's 'fat' data, the limits of agreement computed using Roy's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

5.16.1 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by

replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Childrens Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is unbalanced. Carstensen et al. (2008) describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

Carstensen et al. (2008) demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the 'oximetry' data set using a model with the additional term, and one without. When the interaction is accounted for the limits of agreement are (-9.62,14.56). When the interaction is not accounted for, the limits of agreement are (-11.88,16.83). It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy's model and the modified model (denoted 1 and 2 respectively);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{5.12}$$

(The variance of the additional random effect in model 2 is 3.01.)

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked

according to their AIC values, with the model having the lowest AIC being considered the best fit.Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

The Akaike information criterion (AIC) for both models are $AIC_1 = 2304.226$ and $AIC_2 = 2306.226$, indicating little difference in models. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

The $\hat{\mathbf{\Lambda}}$ matrices are informative as to the difference between Carstensen's unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively) are of similar magnitudes to the variance terms. Conversely for the 'fat' data the covariance term (-0.00032) is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the 'fat' data, the difference in AIC values is also approximately 2).

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. Therefore the test's proposed by Roy (2009a) can not be implemented. Conversely, accurate limits of agreement as determined by Carstensen's model may also be found using Roy's method. Addition of the interaction term erodes the capability of Roy's methodology to compare candidate models, and therefore shall not be adopted.

Finally, to complement the blood pressure (i.e.'J vs S') method comparison from the previous section (i.e.'J vs S'), the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

Chapter 6

Introduction to Roy's Procedure

6.1 Replicate measurements in Roy's paper

Roy (2009b) takes its definition of replicate measurement: two or more measurements on the same item taken under identical conditions. Roy also assumes linked measurements, but it is can be used for the non-linked case.

6.2 Variance Covariance Matrices

Under Roy's model, random effects are defined using a bivariate normal distribution. Consequently, the variance-covariance structures can be described using 2×2 matrices. A discussion of the various structures a variance-covariance matrix can be specified under is required before progressing. The following structures are relevant: the identity structure, the compound symmetric structure and the symmetric structure.

The identity structure is simply an abstraction of the identity matrix. The compound symmetric structure and symmetric structure can be described with reference to the following matrix (here in the context of the overall covariance Block-$\boldsymbol{\Omega}_i$, but equally applicable to the component variabilities $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$);

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}$$

Symmetric structure requires the equality of all the diagonal terms, hence $\omega_1^2 = \omega_2^2$. Conversely compound symmetry make no such constraint on the diagonal elements. Under the identity structure, $\omega_{12} = 0$. A comparison of a model fitted using symmetric structure with that of a model fitted using the compound symmetric structure is equivalent to a test of the equality of variance.

Independence

As though analyzed using between subjects analysis.

$$\begin{pmatrix} \psi^2 & 0 & 0 \\ 0 & \psi^2 & 0 \\ 0 & 0 & \psi^2 \end{pmatrix}$$

Compound Symmetry

Assumes that the variance-covariance structure has a single variance (represented by ψ^2) for all 3 of the time points and a single covariance (represented by ψ_{ij}) for each of the pairs of trials.

$$\begin{pmatrix} \psi^2 & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi^2 & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi^2 \end{pmatrix}$$

Unstructured

Assumes that each variance and covariance is unique. Each trial has its own variance (e.g. s12 is the variance of trial 1) and each pair of trials has its own covariance (e.g. s21 is the covariance of trial 1 and trial2). This structure is illustrated by the half matrix below.

Autoregressive

Another common covariance structure which is frequently observed in repeated measures data is an autoregressive structure, which recognizes that observations which are more proximate are more correlated than measures that are more distant.

6.3 VC structures

$\mathbf{\Psi}$ is the variance-covariance matrix of the random effects , with 2×2 dimensions.

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{6.1}$$

There is three alternative structures for $\mathbf{\Psi}$, the diagonal form, the identity form and the general form.

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix} \quad \text{or} \quad \mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \quad \text{or} \quad \mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$$

Chapter 7

LME Model Specification

Model Terms (Roy 2009)

It is important to note the following characteristics of this model.

Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.

Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.

$\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.

$\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.

$\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.

$\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.

$\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.

The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)

The variance of the response vector is given by $\text{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as D. The estimate for the within-subject variance covariance matrix is $\hat{\Sigma}$. The estimated overall variance covariance matrix 'Block Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{7.1}$$

$\boldsymbol{b}_i$ is a $m-$dimensional vector comprised of the random effects.

$$\boldsymbol{b}_i = \begin{pmatrix} b_{1i} \\ b_{21} \end{pmatrix} \tag{7.2}$$

$\boldsymbol{V}$ represents the correlation matrix of the replicated measurements on a given method. $\boldsymbol{\Sigma}$ is the within-subject VC matrix.

$\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are positive definite matrices. The dimensions of $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are $3 \times 3(= p \times p)$ and $2 \times 2(= k \times k)$.

It is assumed that $\boldsymbol{V}$ is the same for both methods and $\boldsymbol{\Sigma}$ is the same for all replications.

$\boldsymbol{V} \bigotimes \boldsymbol{\Sigma}$ creates a $6 \times 6(= kp \times kp)$ matrix. $\boldsymbol{R}_i$ is a sub-matrix of this.

7.1 Model Formula

Let y_{mir} denote the rth replicate measurement on the ith item by the mth method, where $m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$. When the design is balanced and there is no ambiguity we can set $n_i = n$. The LME model underpinning Roy's approach can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{7.3}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m.

The b_{1i} and b_{2i} terms represent random effect parameters corresponding to the two methods, having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$.

When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, g_1^2 andg_2^2, and σ_1^2 andσ_2^2.
Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and

an overall effect for method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$.

The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009a) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing.

Roy also integrates H_2 and H_3 into a single testable hypothesis H_4: $\omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m.
Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

7.2 G Component

$\boldsymbol{G}$ is the variance covariance matrix for the random effects $\boldsymbol{b}$. i.e. between-item sources of variation.

It is important to note that no special assumptions about the structure of $\boldsymbol{G}$ are made. An example of such an assumption would be that $\boldsymbol{G}$ is the product of a scalar value and the identity matrix.

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$. The random effects and residuals are assumed to be independent.

7.3 R Component

Hamlett et al. (2004) shows that the variance covariance matrix for the residuals(i.e. the within-item sources of variation between both methods) can be expressed as the Kroneckor product of an $n_i \times n_i$ identity matrix and the partial within-item variance covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Again it is important

to note that no special assumptions are made about the structure of the matrix. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & \dots & \dots & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The above terms can be used to express the variance covariance matrix $\boldsymbol{\Omega}_i$ for the responses on item i ,

$$\boldsymbol{\Omega}_i = \boldsymbol{Z}_i \boldsymbol{G} \boldsymbol{Z}_i' + \boldsymbol{R}_i.$$

The variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed.

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\boldsymbol{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$. The random effects and residuals are assumed to be independent. Both covariance matrices can be written as follows;

The above terms can be used to express the variance covariance matrix $\boldsymbol{\Omega}_i$ for the responses on item i ,

$$\boldsymbol{\Omega}_i = \boldsymbol{Z}_i \boldsymbol{G} \boldsymbol{Z}_i' + \boldsymbol{R}_i.$$

7.4 Hamlett

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The partial within-item variance?covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{7.4}$$

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates. The 2×2 block diagonal Block-$\boldsymbol{\Omega}_i$ represents the covariance matrix between two methods, and is the sum of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$.

$$\text{Block-}\boldsymbol{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates. The 2×2 block diagonal Block-$\boldsymbol{\Omega}_i$ represents the covariance matrix between two methods, and is the sum of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$.

$$\text{Block-}\boldsymbol{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed.

7.5 For Expository Purposes

For expository purposes consider the case where each item provides three replicates by each method. Then in matrix notation the model has the structure

$$\boldsymbol{y}_i = \begin{pmatrix} y_{1i1} \\ y_{2i1} \\ y_{1i2} \\ y_{2i2} \\ y_{1i3} \\ y_{2i3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i1} \\ \epsilon_{2i1} \\ \epsilon_{1i2} \\ \epsilon_{2i2} \\ \epsilon_{1i3} \\ \epsilon_{2i3} \end{pmatrix}.$$

The between item variance covariance $\boldsymbol{G}$ is as before, while the within item variance covariance is given as

$$\boldsymbol{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Assumptions made on the structures of $\boldsymbol{G}$ and $\boldsymbol{R}_i$ will be discussed in due course.

7.6 Overall Variability

The overall variability between the two methods is the sum of between-item variability $\boldsymbol{G}$ and within-item variability $\boldsymbol{\Sigma}$. Roy (2009b) denotes the overall variability as Block - $\boldsymbol{\Omega}_i$. The overall variation for methods 1 and 2 are given by

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The variance of case-wise difference in measurements can be determined from Block-$\mathbf{\Omega}_i$. Hence limits of agreement can be computed.

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\mathbf{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

7.7 Off-Diagonal Components in Roy's Model

The Within-item variability is specified as follows, where x and y are the methods of measurement in question.

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

σ_x^2 and σ_y^2 describe the level of measurement error associated with each of the measurement methods for a given item. Attention must be given to the off-diagonal elements of the matrix. It is intuitive to consider the measurement error of the two methods as independent of each other. A formal test can be performed to test the hypothesis that the off-diagonal terms are zero.

$$\begin{pmatrix} \sigma_x^2 & \sigma_x y \\ \sigma_x y & \sigma_y^2 \end{pmatrix} vs \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

Chapter 8

Extending Current Methodologies

8.1 Extension of Roy's Methodology

Roy's methodology is constructed to compare two methods in the presence of replicate measurements. Necessarily it is worth examining whether this methodology can be adapted for different circumstances.

An implementation of Roy's methodology, whereby three or more methods are used, is not feasible due to computational restrictions. Specifically there is a failure to reach convergence before the iteration limit is reached. This may be due to the presence of additional variables, causing the problem of non-identifiability. In the case of two variables, it is required to estimate two variance terms and four correlation terms, six in all. For the case of three variabilities, three variance terms must be estimated as well as nine correlation terms, twelve in all. In general for n methods has $2 \times T_n$ variance terms, where T_n is the triangular number for n, i.e. the addition analogue of the factorial. Hence the computational complexity quite increases substantially for every increase in n.

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null

hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

The tests would not be informative as to whether any two of those three were interchangeable, or equivalently if one method in particular disagreed with the other two. Indeed it is easier to perform three pair-wise comparisons separately and then to combine the results.

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector y_i, as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

8.2 Conclusion

Carstensen et al. (2008) and Roy (2009a) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009a) presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

Chapter 9

LME models for MCS

9.1 Roy's LME methodology for assessing agreement

Roy (2009b) proposes the use of LME models to perform a test on two methods of agreement to determine whether they can be used interchangeably.

Bivariate correlation coefficients have been shown to be of limited use in method comparison studies (Bland and Altman, 1986). However, recently correlation analysis has been developed to cope with repeated measurements, enhancing their potential usefulness. Roy incorporates the use of correlation into his methodology.

Roy's method considers two methods to be in agreement if three conditions are met.

- no significant bias, i.e. the difference between the two mean readings is not "statistically significant",
- high overall correlation coefficient,
- the agreement between the two methods by testing their repeatability coefficients.

The methodology uses a linear mixed effects regression fit using compound symmetry (CS) correlation structure on $\mathbf{V}$.

$\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L}$

Roy (2009b) considers the problem of assessing the agreement between two methods with replicate observations in a doubly multivariate set-up using linear mixed effects models.

Roy (2009b) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

Roy (2009b) proposes a LME based approach with Kronecker product covariance structure with doubly multivariate setup to assess the agreement between two methods. This method is designed such that the data may be unbalanced and with unequal numbers of replications for each subject.

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as D. The estimate for the within-subject variance covariance matrix is $\hat{\Sigma}$. The estimated overall variance covariance matrix 'Block Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{9.1}$$

For the the RV-IC comparison, $\hat{D}$ is given by

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \tag{9.2}$$

The estimate for the within-subject variance covariance matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \tag{9.3}$$

The estimated overall variance covariance matrix for the the 'RV vs IC' comparison is given by

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \tag{9.4}$$

The power of the likelihood ratio test may depends on specific sample size and the specific number of replications, and the author proposes simulation studies to examine this further.

9.2 LME models in method comparison studies

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement. Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement. Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' methods. In this chapter various LME approaches to method comparison studies shall be examined.

9.3 Statement of the LME model

Further to a paper published by Laird and Ware in 1982, a linear mixed effects model is a linear mdoel that combined fixed and random effect terms formulated as follows;

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

- Y_i is the $n \times 1$ response vector
- X_i is the $n \times p$ Model matrix for fixed effects
- β is the $p \times 1$ vector of fixed effects coefficients
- Z_i is the $n \times q$ Model matrix for random effects
- b_i is the $q \times 1$ vector of random effects coefficients, sometimes denoted as u_i
- ϵ is the $n \times 1$ vector of observation errors

9.3.1 Bendix Carstensen's data sets

?describes the sampling method when discussing of a motivating example.Diabetes patients attending an outpatient clinic in Denmark have their HbA_{1c} levels routinely measured at every visit.Venous and Capillary blood samples were obtained from all patients appearing at the clinic over two days.

Samples were measured on four consecutive days on each machines, hence there are five analysis days.Carstensen notes that every machine was calibrated every day to the manufacturers guidelines.

9.4 Hamlett and Lam

The methodology proposed by **?** is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999).

9.4.1 Roy's variability tests

Variability tests proposed by **?** affords the opportunity to expand upon Carstensen's approach.

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

9.5 Carstensen's Mixed Models

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.5}$$

The intercept term α and the $\beta_m \mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value μ_i. c_{mi} is a interaction term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

Carstensen (2004) uses the above formula to predict observations for a specific individual i by method m;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta}_m \mu_i + c_{mi} \tag{9.6}$$

. Under the assumption that the μs are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates.

The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr} \sim N(0, \omega_m^2)$to account for this.

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{9.7}$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$ For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.8}$$

Carstensen et al. (2008) proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{9.9}$$

9.6 Carstensen's Mixed Models

Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr}d_{mr} \sim N(0, \omega_m^2)$to account for this.

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

9.6.1 KP

Most residual covariance structures are design for one within-subject factor. However two or more may be present. For such cases, an approppriate approach would be the residual covariance structure using Kronecker product of the underlying within-subject factor specific covariances structure.

9.7 LME

Consistent with the conventions of mixed models, **?** formulates the measurement $y_{ij}$from method i on individual j as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{9.10}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random

effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{9.11}$$

These vectors are assumed to be independent for different is, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{9.12}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate x is the key step in how this methodology assesses agreement.

9.8 Fixed Effects and Random Effects Models

Before proceeding to a description of mixed models, an introduction to fixed effects and random effects models is required. This section follows on from the discussion of measurement error models in the last chapter.

9.8.1 Fixed Effects

McCullough and Searle (2001) gives an example of a study where the observations , occurrences of skin tumours called basal cell epithelioma, were classified according to 'factors', i.e. the gender, age and exposure to sunshine of the patients. Levels are the individual classes of each of these factors (e.g. 'Male' and 'Female' would be the levels

of the factor 'Gender'). The scientific interest lies in examining the extent to which different factor levels affect the variable of interest. The effects of a level of a factor are one of two types ; fixed effects and random effects. Fixed effects describe effects due to a finite set of levels for a factor (i.e multichotomous factors). The factors described in the skin tumour example are all fixed effects factors. Fixed effects models are the cases where only fixed effects are present, with the exception of random error terms.

To demonstrate fixed effects model Searle (1997) describes a study wherein 24 plants are divided into four groups of six, and each group is subjected to its own treatment regime. Three different fertilizers are used with three of the groups (treatments N,P,K), while no fertilizer is used on the fourth group, (i.e. it is a control group denoted as C). Searle (1997) constructs a model to describe the crop yield resultant from the experiment.

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{9.13}$$

where y_{ij} is the jth plant (i.e. crop yield) on the ith treatment, with μ as the mean yield, α_i is the effect of each fertilizer treatments (i.e. fixed effects) and e_{ij} is the error term. The fixed effect for each observation is an unknown constant that is to be determined from computing the data.

The μ term would not necessarily be present in all formulations. Some authors, such as Demidenko (2004), may use a single term as equivalent to the μ and α terms. (It is customary to centre data prior to using mixed effects methodologies. Centering means subtracting the mean of the observations from each observed value, so mean of the resultant values become zero.)

9.8.2 Random Effects

The random effects model describes the case where there is an infinite number of levels in a factor. In other words the factor is a random variable. Searle (1997) demonstrates this with a second example; a study of the maternal ability of mice. In this example

4 female mice , all of the same breed, have 6 litters each over a period of time. The weights w_{ij} of each litter (j) from each mouse (i) were taken to be the proxy for maternal ability and is formulated as follows;

$$w_{ij} = \mu + \delta_i + e_{ij} \tag{9.14}$$

As in the previous exmaple, α is the mean , e_{ij} is the error term and δ_i is a random effect due to each mouse. Notably these four mice are considered as a sample of the overall population of female mice of that breed, consequently an important characteristic of random effects models is that the δ_i values are a random sample of all δ terms. Therefore these random terms can be used for making inferences about populations. (McCullough and Searle, 2001).

9.8.3 Variance Components

Each random effect has an associated 'variance component' term. This is a model parameter which quantifies random variation due to that effect only. Therefore for every observation there are two sources of variation, random variation and residual variation, and can be expressed as follows $var(y_{ij}) = \sigma_p^2 + \sigma^2$. These variations are known as the variance components. This is an important difference with fixed effects models, which is subject to residual variation (i.e σ^2) only (Brown and Prescott, 1999).

In fixed effects models there is no covariance between any pair of observations. Brown and Prescott (1999) shows that, while there is no covariance between observations from different subjects (i.e the mice in Searle's example), there exists correlation between observations from the same subject (i.e. litter weights from the same mouse are correlated). In this case the covariance is the subjects variance component (i.e. σ_p^2)

9.8.4 More complex examples

Searle (1997) offers elaborations on both examples used so far. In the case of the fixed effects model, the model can be amended to take account for different varieties of each

plants being studied. (The groups of six plants are subdivided into three variety types.) y_{ijk} is the yield of the kth plant of the jth variety in the ith treatment, and is described as follows;

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \tag{9.15}$$

This new formulation includes a fixed effect β_j to account for the variety type, and an 'interaction effect' γ_{ij}. An interaction effect describes the combined effects of two or more variables on the observation. Similarly the random effects example is elaborated to account for the effects of three different technicians. Again there is a random effect component τ to account for these technicians, and an interaction effect θ to account for the combined effect of the mice and the technicians.

$$w_{ijk} = \mu + \delta_i + \tau_j + \theta_{ij} + e_{ijk} \tag{9.16}$$

9.9 Mixed Models

All models are characterized by the mean α and the error terms. In addition to these terms, any model described so far will have either random effects terms or fixed effects terms and accordingly are referred to as random or fixed models. Models that have both fixed effects terms and random effects terms are known as 'mixed effects models'. Once the theory underlying fixed and random effects models has been fully understood, the progression to understanding mixed models is very simple.

Elaborating on the original mice litter example, the six litters by each mouse were fed according to three different dietary treatments (Searle, 1997). Therefore a fixed effect ϕ_j has been added to the model, which is now formulated as follows;

$$y_{ij} = \mu + \delta_i + \phi_j + \gamma_{ij} + \epsilon_{ijk} \tag{9.17}$$

As before, an interaction effect γ_{ij} must also be added to the model. In cases where the interaction term describes the combined effect of fixed and random components, it

should be treated as random effect. The variance of the above model is composed of the σ^2_δ, σ^2_γ and σ^2_ϵ .

It may be shown that the interaction factors make no contribution to the outcome, i.e γ_{ij} is consistently calculated as zero. Considering the skin tumour example, a person's age would bear no relation to their gender and hence there would be plausible interaction between the two factors. Indeed , in keeping with the 'Law of Parsimony', factors should be specified such that each would convey separate information. However, interaction terms are extant when the model specifies repeated observations, as there is necessarily a relationship between observations from the same subject. Importantly, interaction effects, being random effects, are attended by variance component terms and therefore also contribute to the overall variance of the model.

Searle (1997) gives a mixed effects model formulation for the Grubbs artillery study. y_{ij} is the muzzle velocity of the ith shell, as measured by the jth chronometer.

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{9.18}$$

In this formulation α_i is the random effect of round i, and the fixed effect component β_j is the bias in chronometer j. (Also, no interaction term is used).

9.10 Mixed Model Calculations

9.11 Estimability of Fixed Effects

Potentially it may be impossible to compute unique BLUE estimates for all the fixed factors in a model. This may be due to linear dependence in the model matrix $\mathbf{X}$. Consider the following example;

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{9.19}$$

HERE

9.12 Random Effects and MCS

The methodology comprises two calculations. The second calculation is for the standard deviation of means Before the modified Bland and Altman method can be applied for repeated measurement data, a check of the assumption that the variance of the repeated measurements for each subject by each method is independent of the mean of the repeated measures. This can be done by plotting the within-subject standard deviation against the mean of each subject by each method. Mean Square deviation measures the total deviation of a

9.13 Conclusion

Carstensen et al. (2008) and Roy (2009a) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009a) presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

Chapter 10

Likelihood Ratio Tests

10.1 Likelihood

Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. Likelihood differs from probability in that probability refers to future occurrences, while likelihood refers to past known outcomes.

The likelihood function is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986). Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. Each of these three test shall be examined in more detail shortly.

Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs.

10.2 Likelihood ratio tests

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs).

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test.

Likelihood ratio tests (LRTs) are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model).

LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs.

The test statistic for the LRT is the difference of the log-likelihood functions, multiplied by -2.

The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

The score function $S(\theta)$ is the derivative of the log likelihood with respect to θ,

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta),$$

and the maximum likelihood estimate is the solution to the score equation

$$S(\theta) = 0.$$

The Fisher information $I(\theta)$, which is defined as

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta),$$

give rise to the observed Fisher information $(I(\hat{\theta}))$ and the expected Fisher information $(\mathcal{I}(\theta))$.

? used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

The true correlation ρ_{xy} is repeated measurements can be considered as having two components: between subject and within-subject correlation. The usefulness of estimating repeated measure correlation coefficients is the calculation of between-method and within-method variabilities are produced as by-products.

10.3 Test Statistic for Likelihood Ratio Tests

The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models.

The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2. The test statistic for the LRT is the difference of the log-likelihood functions, multiplied by -2.

L= - 2ln is approximately distributed as 2 under H_0 for large sample size and under the normality assumption.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{10.1}$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{10.2}$$

The power of the likelihood ratio test may depends on specific sample size and the specific number of replications, and [Roy 2009] proposes simulation studies to examine this further.

10.4 Nesting: Model Selection Using Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) is known as "nesting".

An important step in the process of model selection is to determine, for a given pair of models, if there is a "nesting relationship" between the two.

We define Model A to be "nested" in Model B if Model A is a special case of Model B, i.e. Model B with a specific constraint applied.

One model is said to be *nested* within another model, i.e. the reference model, if it represents a special case of the reference model (Pinheiro and Bates, 1994).

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship West et al. (2007).

10.5 Statistical Assumptions for Likelihood Ratio Tests

If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample", distribution of the LRT statistic, under the null hypothesis that the

restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (**?**, pg.83).

We generally use LRTs to evaluate the significance of terms in the random effects structure, i.e. different nested models are fitted in which the random effects structure is changed.

The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters (West et al., 2007). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

A general method for comparing nested models fit by maximum likelihood is the ***likelihood ratio test***. This test can be used for models fit by REML (restricted maximum liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method="ML" must be employed (ML = maximum likelihood).

- Example of a likelihood ratio test used to compare two models:
  `>anova(modelA, modelB)`

- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.

- Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.

- A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple "anova"

function. Example:

```
>anova(modelA)
```

will give the most reliable test of the fixed effects included in model1.

10.5.1 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009b) is known as "nesting". A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters.

Conversely, Roy (2009b) advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2. The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in more detail shortly.

10.5.2 Testing Procedures

Roy's methodology requires the construction of four candidate models. The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models.

10.6 LRTs for covariance parameters

[cite: West et al] When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters [cite: Morrel98]

10.7 Relevance of Estimation Methods

The problem with REML for model building is that the "likelihoods" obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters (West et al., 2007). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

Nested LME models, fitted by ML estimation, can be compared using the likelihood ratio test [Lehmann (1986)]. Models fitted using REML estimation can also be compared, but only if both were fitted using REML, and both have the same fixed effects specifications.

Likelihood ratio tests are generally used to test the significance of terms in the random effects structure.

REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986). Such a test can also be used for models fitted using

REML, but only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample", distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (**?**, pg.83).

We generally use LRTs to evaluate the significance of terms in the random effects structure, i.e. different nested models are fitted in which the random effects structure is changed.

For both REML and ML estimates, the nominal $p-$values for the LRT statistics under a χ^2 distribution with 2 degrees of freedom are much greater than empirical values. A number of ways of dealing with this issues are discussed (**?**, pg.86).

One should be aware that these p-values may be conservative. That is, the reported p-value may be greater than the true p-value for the test and, in some cases, it may be much greater.(**?**, pg.87).

Pinheiro & Bates (2000; p. 88) argue that Likelihood Ratio Test comparisons of models varying in fixed effects tend to be anticonservative i.e. will see you observe significant differences in model fit more often than you should.

10.8 Information Criteria

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.

Additionally nested models may be compared by using the Akaike Information Criterion,(AIC) and the Bayesian Information Criterion (BIC).

When comparing the respective scores for nested models, the model with the smaller score is considered to be the preferable model. ML / REML [Morrell 1998] The vari-

ance components in the LME model may be estimated by ML or REML. Maximum Likelihood estimates do not take into account the estimation of fixed effects and so are biased downwards. REML estimates accounts for the presence of these nuisance parameters by maximising the linearly independent error contrasts to obtain more unbiased estimates.

Pinheiro and Bates (1994) addresses the issue of treating items as fixed effects. Such a specification is useful only for the specific sample of items, rather than the population of items, where the interest would naturally lie.

Pinheiro and Bates (1994) advises the specification of random effects to correspond to items; treating the item effects as random deviations from the population mean.

10.9 Model Selection

10.9.1 Akaike Information Critierion

This is a model selection method, assessing how the goodness of fit of a model. It is computed as follows:

$$AIC = -2l_{max} + 2k$$

with l_{max} as the log-likelihood maximum and k as the number of parameters. The candidate model with the lowest AIC value is considered the best fitting of the candidate models.

Demidenko (2004, p.13) reports that some researchers have noted that there is a bias present in AIC estimation, and have proposed alternative formulations to rectify it. It is also reported that AIC doesn't address the issue of multicollinearity sufficiently. Demidenko (2004) formulate an adaption; the Healthy AIC. It is constructed to overcome the issue of ill posed models. The HAIC will choose the candidate model with the shortest parameter vector length.

Chapter 11

Errata

11.1 Fixed Effects and Random Effects Models

Before proceeding to a description of mixed models, an introduction to fixed effects and random effects models is required. This section follows on from the discussion of measurement error models in the last chapter.

11.2 Mixed Models

All models are characterized by the mean α and the error terms. In addition to these terms, any model described so far will have either random effects terms or fixed effects terms and accordingly are referred to as random or fixed models. Models that have both fixed effects terms and random effects terms are known as 'mixed effects models'. Once the theory underlying fixed and random effects models has been fully understood, the progression to understanding mixed models is very simple.

Elaborating on the original mice litter example, the six litters by each mouse were fed according to three different dietary treatments (Searle, 1997). Therefore a fixed effect ϕ_j has been added to the model, which is now formulated as follows;

$$y_{ij} = \mu + \delta_i + \phi_j + \gamma_{ij} + \epsilon_{ijk} \tag{11.1}$$

As before, an interaction effect γ_{ij} must also be added to the model. In cases where the interaction term describes the combined effect of fixed and random components, it should be treated as random effect. The variance of the above model is composed of the σ_δ^2, σ_γ^2 and σ_ϵ^2 .

It may be shown that the interaction factors make no contribution to the outcome, i.e γ_{ij} is consistently calculated as zero. Considering the skin tumour example, a person's age would bear no relation to their gender and hence there would be plausible interaction between the two factors. Indeed , in keeping with the 'Law of Parsimony', factors should be specified such that each would convey separate information. However, interaction terms are extant when the model specifies repeated observations, as there is necessarily a relationship between observations from the same subject. Importantly, interaction effects, being random effects, are attended by variance component terms and therefore also contribute to the overall variance of the model.

Searle (1997) gives a mixed effects model formulation for the Grubbs artillery study. y_{ij} is the muzzle velocity of the ith shell, as measured by the jth chronometer.

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{11.2}$$

In this formulation α_i is the random effect of round i, and the fixed effect component β_j is the bias in chronometer j. (Also, no interaction term is used).

11.2.1 Matrix Formulation

There are matrix (i.e multivariate) formulations of both fixed effects models and random effects models. Brown and Prescott (1999) 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;

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{11.3}$$

$\mathbf{Y}$ 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 $q \times 1$.

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{11.4}$$

Again, once the component fixed effects and random effects components are considered, progression to a mixed model formulation is a simple step. Further to Laird and Ware (1982), it is conventional to formulate a mixed effects model in matrix form as follows:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{11.5}$$

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

11.3 Linear Mixed effects Models

11.4 Computation of LMEs using R

Pinheiro and Bates (1994) advises how to implement LME models in statistical software (ostensibly for S and S PLUS, but R is very similar). When tackling linear mixed effects models using the R language, a statistician can call upon the *lme* command found in the *nlme* package.This command fits a LME model to the data set using either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). ML may be referred to as 'full maximum likelihood' estimation.

The first two arguments for *lme* are *fixed* and *data*, which give the model for the expected responses (i.e. the fixed part of the model), and the data that themodel should be fitted from. The next argument is *random*, a one-sided formula which describes the random effects, and the grouping structure for the model. The *method* argument can specify whether to use 'REML', the default setting, or 'ML'.

Pinheiro and Bates (1994) describes an experiment whereby the productivity of six randomly chosen workers are assessed three times on each of three machines, yielding the 54 observations in the following table.

(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 (β_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 (τ_{ij}) describes whether the effect of changing from one machine to another is different for each worker. The productivity score y_{ijk} is the kth observation taken on worker i on machine j, and is formulated as follows;

$$y_{ijk} = \beta_j + u_i + \tau_{ij} + \epsilon_{ijk} \tag{11.6}$$

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 hier-

archical 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 $\sigma_\epsilon^2 = (0.96)^2$.

11.5 Implementation in R

To implement an LME model in `R`, the `nlme` package is used. This package is loaded into the `R` environment using the library command, (i.e. `library(nlme)`). The `lme` command is used to fit LME models. The first two arguments to the `lme` function specify the fixed effect component of the model, and the data set to which the model is to be fitted. The first candidate model ('MCS1') fits an LME model on the data set 'dat'. The variable 'method' is assigned as the fixed effect, with the response variable 'BP' (i.e. blood pressure).

The third argument contain the random effects component of the formulation, describing the random effects, and their grouping structure. The `nlme` package provides a set of positive-definite matrices , the `pdMat` class, that can be used to specify a structure for the between-subject variance-covariance matrix for the random effects. For Roy's methodology, we will use the `pdSymm` and `pdCompSymm` to specify a symmetric structure and a compound symmetry structure respectively. A full discussion of these structures can be found in Pinheiro and Bates (1994, pg. 158).

Similarly a variety of structures for the with-subject variance-covariance matrix can be implemented using `nlme`. To implement a particular matrix structure, one must specify both a variance function and correlation structure accordingly. Variance functions are used to model the variance structure of the within-subject errors. `varIdent` is a variance function object used to allow different variances according to the levels of a classification factor in the data. A compound symmetry structure is implemented using the `corCompSymm` class, while the symmetric form is specified by `corSymm` class. Finally, the estimation methods is specified as "ML" or "REML".

The first of Roy's candidate model can be implemented using the following code;

```
MCS1 = lme(BP ~ method-1, data = dat,
random =  list(subject=pdSymm(~ method-1)),
weights=varIdent(form=~1|method),
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

For the blood pressure data used in Roy (2009a), all four candidate models are implemented by slight variations of this piece of code, specifying either `pdSymm` or `pdCompSymm` in the second line, and either `corSymm` or `corCompSymm` in the fourth line. For example, the second candidate model 'MCS2' is implemented with the same code as MCS1, except for the term `pdCompSymm` in the second line, rather than `pdSymm`.

```
MCS2 = lme(BP ~ method-1, data = dat,
random = list(subject=pdCompSymm(~ method-1)),
weights = varIdent(form=~1|method),
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

Using this `R` implementation for other data sets requires that the data set is structured appropriately (i.e. each case of observation records the index, response, method and replicate). Once formatted properly, implementation is simply a case of re-writing the first line of code, and computing the four candidate models accordingly.

To perform a likelihood ratio test for two candidate models, simply use the `anova` command with the names of the candidate models as arguments. The following piece of code implement the first of Roy's variability tests.

```
> anova(MCS1,MCS2)
Model df    AIC    BIC  logLik   Test L.Ratio p-value
MCS1     1  8 4077.5 4111.3 -2030.7
MCS2     2  7 4075.6 4105.3 -2030.8 1 vs 2 0.15291  0.6958
>
```

The fixed effects estimates are the same for all four candidate models. The inter-method bias can be easily determined by inspecting a summary of any model. The summary presents estimates for all of the important parameters, but not the complete variance-covariance matrices (although some simple R functions can be written to overcome this). The variance estimates for the random effects for MCS2 is presented below.

```
Random effects:
Formula: ~method - 1 | subject
Structure: Compound Symmetry
StdDev Corr
methodJ  30.765
methodS  30.765 0.829
Residual  6.115
```

Similarly, for computing the limits of agreement the standard deviation of the differences is not explicitly given. Again, A simple R function can be written to calculate the limits of agreement directly.

11.6 LRTs wtih R

Likelihood ratio tests are very simple to implement in R, simply use the 'anova()' commands. Sample output will be given for each variability test. The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' (M2LL) is computed. The test statistic for each of the three hypothesis tests is the difference of the M2LL for each pair of models. If the p-value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-??2ln\Lambda_d = \text{[M2LL under H0 model]} - \text{[M2LL under HA model]} \quad (11.7)$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

$$\nu_{=}\text{[LRT df under H0 model]} - \text{[LRT df under HA model]} \quad (11.8)$$

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
MCS1	8	4077.5	4111.3	-2030.7			
MCS2	7	4075.6	4105.3	-2030.8	1 vs 2	0.15291	0.6958

11.7 BXC - Model Terms

- Let y_{mir} be the response of method m on the ith subject at the $r-$th replicate.
- Let $\boldsymbol{y}_{ir}$ be the 2×1 vector of measurements corresponding to the $i-$th subject at the $r-$th replicate.
- Let $\boldsymbol{y}_i$ be the $R_i \times 1$ vector of measurements corresponding to the $i-$th subject, where R_i is number of replicate measurements taken on item i.
- Let $\alpha_m i$ be the fixed effect parameter for method for subject i.
- Formally Roy uses a separate fixed effect parameter to describe the true value μ_i, but later combines it with the other fixed effects when implementing the model.
- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i.
- $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.
- $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to Roy's first test.

11.8 Other Approaches

11.8.1 Random coefficient growth curve model

(Chincilli 1996) Random coefficient growth curve model, a special type of mixed model have been proposed a single measure of agreement for repeated measurements.

$$\mathbf{d} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{11.9}$$

The distributional asummptions also require $\mathbf{d}$ to $\mathbf{N}$

11.8.2 Marginal Modelling

(Diggle 2002) proposes the use of marginal models as an alternative to mixed models.m Marginal models are appropriate when interences about the mean response are of specific interest.

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 11.4.1: Machines Data , Pinheiro Bates

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

Brown, H. and R. Prescott (1999). *Applied Mixed Models In Medicine*. John Wiley and Sons.

Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics 5*(3), 399–413.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(2), 133–169.

Demidenko, E. (2004). *Mixed Models: Theory And Application*. Dartmouth College: Wiley Interscience.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithme. *ournal of the Royal Statistical Society. Series B 39*(1), 1–38.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics 3*(1), 1–21.

Faraway, J. J. (2006). *Extending The Linear Model With R: Generalized Linear , Mixed Effects and Nonparametric Regression Models*. Chapman and Hall / CRC.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh 2*, 399–433.

GK, R. (1991). That blups are a good thing: The estimation of random effects. *Statistical Science 6*(1), 15–32.

Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association 43*, 243–264.

Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International 198-229*, 1–7.

Hartley, H. and J. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika 54*(1/2), 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association 72*(358), 320–338.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics 9*(2), 226–252.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1959a). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1959b). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1963). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1973). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1984). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics 21*, 309–310.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and genetics Symposium in Honor of Dr Jay L. Lush*, pp. 10–41. Champaign, Illinois: Americian Society of Animal Science and American Dairy Science Association.

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

Lee, Y., J. A. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall Ltd.

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

Paterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (Disc: P32-51). *Statistical Science 6*, 15–32.

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects models. *Biometric Journal 2*, 286–301.

Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

Searle, S. (1997). *Linear Models*. Wiley classics Library.

Singer, J. D. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics 24*(4), 323–355.

Tippett, L. (1931). *The Methods of Statistics* (1st ed.). London: Williams and Norgate.

West, B., K. Welch, and A. Galecki (2007). *Linear Mixed Models: a Practical Guide Using Statistical Software*. Chapman and Hall CRC.