

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	LME models in method comparison studies . . . . .	8
1.2	Introduction to LME Models, Fitting LME Models to MCS Data . . .	10
1.3	Definition of Replicate Measurements (Move to Chapter 1) . . . . .	11
1.4	Definition of Replicate measurements . . . . .	11
1.5	Model for replicate measurements . . . . .	11
1.6	Carstensen's Model . . . . .	12
1.7	Two Way ANOVA . . . . .	13
1.8	Statistical Model For Replicate Measurements . . . . .	14
1.9	Exchangeable and Linked measurements . . . . .	14
1.10	Sampling Scheme : Linked and Unlinked Replicates . . . . .	14
1.11	Replicate measurements . . . . .	15
1.12	Linear Mixed effects Models . . . . .	17
1.13	Laird Ware . . . . .	18
1.14	Statement of the LME model . . . . .	19
1.15	The Linear Mixed Effects Model . . . . .	19
1.16	Likelihood and estimation . . . . .	22
1.17	Estimation for LME Models . . . . .	23
1.18	Henderson's equations . . . . .	23
1.19	Repeated measurements in LME models . . . . .	27
1.19.1	Formulation of the response vector . . . . .	28

1.20	Decomposition of the response covariance matrix . . . . .	28
1.20.1	Correlation terms . . . . .	30
<b>2</b>	<b>LME Model Specification</b>	<b>33</b>
2.1	Model Formula . . . . .	34
<b>3</b>	<b>Introduction to Roy's Procedure</b>	<b>36</b>
3.1	Introduction to Roy's methodology . . . . .	37
3.2	Replicate measurements in Roy's paper . . . . .	38
3.3	Model Set Up . . . . .	38
3.4	Agreement Criteria . . . . .	39
3.5	Test for inter-method bias . . . . .	41
3.6	Variability Tests . . . . .	42
3.7	Variance Covariance Matrices . . . . .	43
3.7.1	Variance-Covariance Structures . . . . .	44
3.8	Roy's Candidate Models : Testing Procedures . . . . .	45
3.9	Hypothesis Testing . . . . .	45
3.10	Roy's hypothesis tests . . . . .	46
3.11	Roy's variability tests . . . . .	47
3.12	Correlation coefficient . . . . .	48
3.13	Roy's variability tests . . . . .	49
3.14	Variability test 1 . . . . .	50
3.15	Variability test 2 . . . . .	50
3.16	Variability test 3 . . . . .	52
3.17	Formal testing for covariances . . . . .	53
3.18	VC structures . . . . .	53
<b>4</b>	<b>Model Specification</b>	<b>55</b>
4.1	Model Specification for Roy's Hypotheses Tests . . . . .	55
4.2	G Component . . . . .	58

4.3	R Component . . . . .	58
4.4	Hamlett . . . . .	61
4.5	For Expository Purposes . . . . .	62
4.6	Kroneckor . . . . .	63
4.7	Overall Variability . . . . .	64
4.8	Off-Diagonal Components in Roy's Model . . . . .	65
4.9	Formal Testing . . . . .	65
4.10	Using LMEs for Method Comparison . . . . .	66
<b>5</b>	<b>Likelihood Ratio Tests</b>	<b>67</b>
5.1	Likelihood . . . . .	67
5.2	Likelihood Ratio Tests . . . . .	68
5.3	Likelihood Ratio Tests . . . . .	68
5.3.1	Nested and Reference Models . . . . .	69
5.4	Likelihood Ratio Tests . . . . .	70
5.5	Likelihood Ratio Tests . . . . .	70
5.6	Nesting: Model Selection Using Likelihood Ratio Tests . . . . .	71
5.7	Other material . . . . .	71
5.8	LRTs for covariance parameters . . . . .	72
5.9	Test Statistic for Likelihood Ratio Tests . . . . .	72
5.10	Pinheiro Bates . . . . .	74
5.11	Empirical p-values of LRT tests . . . . .	74
5.12	Relevance of Estimation Methods . . . . .	75
5.13	Information Criteria . . . . .	75
5.14	Likelihood Ratio Tests in Roy's Analysis . . . . .	77
<b>6</b>	<b>LOAS</b>	<b>78</b>
6.1	Calculation of limits of agreement . . . . .	79
6.2	00-Limits of Agreement in LME models . . . . .	80
6.3	Carstensen et al . . . . .	81

6.4	Computing LoAs from LME models . . . . .	82
6.5	Carstensen's Limits of Agreement . . . . .	82
6.6	Carstensen's Model . . . . .	83
6.7	BXC2008 presents - Carstensen's Limits of agreement . . . . .	86
6.8	Limits of Agreement in LME models . . . . .	87
6.9	BXC2008 uses - Limits of Agreement in LME models . . . . .	88
6.10	Carstensen's LOAs . . . . .	89
6.11	Interaction Terms in Model . . . . .	89
6.12	Computation of limits of agreement . . . . .	90
6.13	BXC - Model Terms . . . . .	91
6.14	Computation of limits of agreement under Roy's model . . . . .	91
6.15	LOAs with Roy . . . . .	92
6.16	Difference Between Approaches . . . . .	93
6.17	Differences Between Models . . . . .	94
6.18	Differences . . . . .	94
6.19	Relevance of Roy's Methodology . . . . .	96
6.20	Difference Variance further to Carstensen . . . . .	96
6.21	Assumptions on Variability . . . . .	97
6.22	Carstensen's Mixed Models . . . . .	99
<b>7</b>	<b>Other Material</b>	<b>102</b>
7.1	Extension of Roy's methodology . . . . .	102
7.2	Roy's methodology for single measurements . . . . .	103
7.3	Correlation . . . . .	104
7.4	Correlation terms . . . . .	104
7.5	Hamlett and Lam . . . . .	106
7.6	LaiShiao . . . . .	106
7.7	Lai Shiao . . . . .	107
7.8	Limits of agreement in LME models . . . . .	109

7.8.1	Variance Ratios . . . . .	110
7.9	Worked Eamples . . . . .	111
7.9.1	LikelihoodRatio Tests . . . . .	111
7.10	Testing Procedures . . . . .	114
7.10.1	Roy's Reference Model . . . . .	115
7.10.2	Nested Model (Between-Item Variability) . . . . .	115
7.11	Worked Eamples . . . . .	117
7.12	Examples: LoAs for Carstensen's data . . . . .	117
7.12.1	Daibetes Example . . . . .	118
7.12.2	Oximetry Data . . . . .	118
7.12.3	Linked replicates . . . . .	122
7.12.4	Linked replicates . . . . .	124
7.12.5	Limits of agreement for Carstensen's data . . . . .	126
7.12.6	Fat Data Examples: LoAs for Carstensen's data . . . . .	126
7.12.7	RV-IV . . . . .	127
7.12.8	Limits of agreement for Oximetry . . . . .	127
7.12.9	Classical Model . . . . .	128
7.13	Classical model for single measurements . . . . .	128
7.13.1	Sampling . . . . .	131
7.13.2	Remarks on the Multivariate Normal Distribution . . . . .	132
7.14	Demonstration of Roy's testing . . . . .	133
7.15	Demonstration of Roy's testing . . . . .	135
7.15.1	Matrix structures . . . . .	135
7.15.2	Variability test 1 . . . . .	136
7.15.3	Variability test 2 . . . . .	137
7.15.4	Variability test 3 . . . . .	138
7.15.5	Test for inter-method bias . . . . .	138
7.15.6	Correlation Test . . . . .	139
7.15.7	Conclusion of procedure . . . . .	139

7.16	Basic Models Fits . . . . .	140
7.16.1	Implementing the Mixed Models Fits . . . . .	140
7.16.2	Model Fit 1 . . . . .	142
7.16.3	Model Fit 1 . . . . .	144
7.16.4	Model Fit 2 . . . . .	146
7.16.5	Model Fit 3 . . . . .	148
7.17	Limits of agreement in LME models . . . . .	151
7.17.1	Linked replicates . . . . .	152
7.18	Implementation in R . . . . .	155
7.19	Conclusion . . . . .	158
<b>8</b>	<b>Influence Diagnostics</b>	<b>159</b>
8.1	Measures of Influence . . . . .	159
8.1.1	DFBETA . . . . .	161
8.1.2	DFFITS . . . . .	162
8.2	DFBETAs . . . . .	162
8.3	DFBETAs . . . . .	162
8.4	Case Deletion Diagnostics . . . . .	163
8.5	Deletion Diagnostics . . . . .	163
8.5.1	Influential Observations : DFBeta and DFBetas . . . . .	173
8.6	Measures of Influence . . . . .	173
8.7	Overall Influence . . . . .	173
8.8	Effects on fitted and predicted values . . . . .	174
8.9	Case Deletion Diagnostics for Mixed Models . . . . .	174
8.10	Terminology for Case Deletion diagnostics . . . . .	174
8.11	Case Deletion Diagnostics . . . . .	174
8.12	Case Deletion Diagnostics . . . . .	176
8.13	Terminology for Case Deletion diagnostics . . . . .	176
8.14	influence.ME . . . . .	176

8.15	Influence() command . . . . .	178
8.16	Cooks's Distance . . . . .	178
<b>9</b>	<b>Appendices 1</b>	<b>179</b>
9.1	LME - Pankaj Choudhury . . . . .	179
9.2	Model Terms (Roy 2009) . . . . .	181
9.3	Algorithms . . . . .	182
9.4	ML v REML . . . . .	182
9.5	ML procedures for LME . . . . .	183
9.6	Estimation of random effects . . . . .	184

# Chapter 1

## Introduction

### 1.1 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.

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. Lai and Shiao (2005) view the LME Models approach as an natural expansion to the Bland Altman method for comparing two measurement methods. 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) is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodology that can used to make such questions tractable.

Lai and Shiao (2005) extends the usual method comparison study question. It correctly identifies LME models as a methodology that can used to make such questions tractable. The Data Set used in 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. 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. Additionally a great understanding of residual analysis and influence analysis for LME models has been achieved 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.

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.

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

## **1.2 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.

In this section, we introduce the LME model, discuss 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 simultaneously. However simple models only can only indicate agreement or 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.

### 1.3 Definition of Replicate Measurements (Move to Chapter 1)

### 1.4 Definition of Replicate measurements

Further to Bland and Altman (1999), a formal definition is required of what exactly replicate measurements are

*By replicates we mean two or more measurements on the same individual taken in identical conditions. In general this requirement means that the measurements are taken in quick succession.*

Roy accords with Bland and Altman's definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

### 1.5 Model for replicate measurements

We generalize the single measurement model for the replicate measurement case, by additionally specifying replicate values. Let  $y_{mir}$  be the  $r$ -th replicate measurement

for subject “i” made by method “m”. Further to ? fixed effect can be expressed with a single term  $\alpha_{mi}$ , which incorporate the true value  $\mu_i$ .

$$y_{mir} = \mu_i + \alpha_m + e_{mir}$$

Combining fixed effects (?), we write,

$$y_{mir} = \alpha_{mi} + e_{mir}.$$

The following assumptions are required

- $e_{mir}$  is independent of the fixed effects with mean  $E(e_{mir}) = 0$ .
- Further to ? between-item and within-item variances  $\text{Var}(\alpha_{mi}) = \sigma_{Bm}^2$  and  $\text{Var}(e_{mir}) = \sigma_{Wm}^2$

## 1.6 Carstensen’s Model

Carstensen (2004) presents a model 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (1.1)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (1.2)$$

Of particular importance is terms of the model, a true value for item  $i$  ( $\mu_i$ ). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

## 1.7 Two Way ANOVA

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model describing  $y_{mir}$ , again the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method ( $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n$ ), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \quad (1.3)$$

The fixed effects  $\alpha_m$  and  $\mu_i$  represent the intercept for method  $m$  and the 'true value' for item  $i$  respectively. The random-effect terms comprise an item-by-replicate interaction term  $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$ , a method-by-item interaction term  $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$ , and model error terms  $\epsilon \sim \mathcal{N}(0, \varphi_m^2)$ . All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item  $i$ ,  $a_{ir}$  can be removed.

The model expressed in (2) describes measurements by  $m$  methods, where  $m = \{1, 2, 3 \dots\}$ . Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}$$

Carstensen et al. (2008) notes that, for  $m = 2$ , separate estimates of  $\tau_m^2$  can not be obtained. To overcome this, the assumption of equality, i.e.  $\tau_1^2 = \tau_2^2$  is required.

## 1.8 Statistical Model For Replicate Measurements

Let  $y_{Aij}$  and  $y_{Bij}$  be the  $j$ th repeated observations of the variables of interest  $A$  and  $B$  taken on the  $i$ th 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.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad (1.4)$$

The matrix  $\boldsymbol{\Sigma}$  represents the variance component matrix between response variables at a given time point  $j$ .

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \quad (1.5)$$

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

## 1.9 Exchangeable and Linked measurements

### 1.10 Sampling Scheme : Linked and Unlinked Replicates

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009b) notes that some measurements may not be ‘true’ replicates.

Roy’s methodology assumes the use of ‘true replicates’. However data may not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one  $AR(1)$  structure. However

determining MLEs with such a structure would be computational intense, if possible at all.

*One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this )*

## 1.11 Replicate measurements

Roy (2009b) accords with Bland and Altmans definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

In this model , the variances of the random effects must depend on  $m$ , since the different methods do not necessarily measure on the same scale, and different methods naturally must be assumed to have different variances. Carstensen (2004) attends to the issue of comparative variances.

Bland and Altman (1999) also remark that an important feature of replicate observations is that they should be independent of each other. This issue is addressed by Carstensen (2010), in terms of exchangeability and linkage. Carstensen advises that repeated measurements come in two *substantially different* forms, depending on the circumstances of their measurement: exchangeable and linked.

Repeated measurements are said to be exchangeable if no relationship exists between successive measurements across measurements. If the condition of exchangeability exists, a group of measurement of the same item determined by the same method can be re-arranged in any permutation without prejudice to proper analysis. There is no reason to believe that the true value of the underlying variable has changed over the course of the measurements.

Exchangeable repeated measurements can be treated as true replicates. For the purposes of method comparison studies the following remarks can be made. The  $r$ -th measurement made by method 1 has no special correspondence to the  $r$ -th measurement made by method 2, and consequently any pairing of repeated measurements are as good as each other.

Repeated measurements are said to be linked if a direct correspondence exists between successive measurements across measurements, i.e. pairing. Such measurements are commonly made with a time interval between them, but simultaneously for both methods. Paired measurements are exchangeable, but individual measurements are not.

If the paired measurements are taken in a short period of time so that no real systemic changes can take place on each item, they can be considered true replicates. Should enough time elapse for systemic changes, linked repeated measurements can not be treated as true replicates.



## 1.12 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 a non-negative value, an individual variance component, i.e. a component of the overall variance, may be negative.

The methodology 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 methodology 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., 1959, 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.

Laird and Ware (1982) 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.

## 1.13 Laird Ware

Laird and Ware (1982) provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations.

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

$$y = X\beta + Zb + \epsilon \quad (1.6)$$

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

```
\[
\mathrm{var}
\pmatrix{
b \cr
\epsilon } =
\pmatrix{
D \cr
0 } & 0 \cr
0 & \Sigma }
\]
```

where  $D$  and  $\Sigma$  are positive definite matrices parameterized by an unknown variance component parameter vector  $\theta$ . The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

## 1.14 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. A linear mixed effects model is a linear model that combined fixed and random effect terms formulated by Laird and Ware (1982) as follows;

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i$$

- $Y_i$  is the  $n \times 1$  response vector
- $X_i$  is the  $n \times p$  Model matrix for fixed effects
- $\beta$  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  $u_i$
- $\epsilon$  is the  $n \times 1$  vector of observation errors

## 1.15 The Linear Mixed Effects Model

The linear mixed effects model is given by

$$Y = X\beta + Zu + \epsilon \tag{1.7}$$

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

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

$$\mathbf{V} = \text{Var}(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \quad (1.8)$$

$$\mathbf{V} = \text{Var}(\mathbf{Xb}) + \text{Var}(\mathbf{Zu}) + \text{var}(\mathbf{e}) \quad (1.9)$$

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

By letting  $\text{var}(u) = G$  (i.e  $\mathbf{u} \sim N(0, \mathbf{G})$ ), this becomes  $ZGZ^T$ . This specifies the covariance due to random effects. The residual covariance matrix  $\text{var}(e)$  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;

$$\mathbf{V} = ZGZ^T + \mathbf{R} \quad (1.10)$$

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. (1959), which provides the derivations of both. Inferences about fixed effects have come to be called ‘estimates’, whereas inferences about random effects have come to 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 BLUE of  $\mathbf{b}$ , and the BLUP of  $\mathbf{u}$  can be shown to be;

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \quad (1.11)$$

$$\hat{u} = G Z^T V^{-1} (y - X \hat{b}) \quad (1.12)$$

The practical application of both expressions requires that the variance components be known. An estimate for the variance components must be derived to either maximum likelihood (ML) or more commonly restricted maximum likelihood (REML).

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.

## 1.16 Likelihood and estimation

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 ( $L(\theta)$ ) 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. For computational ease, it is common to use the logarithm of the likelihood function, known simply as the log-likelihood ( $\ell(\theta)$ ).

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.

### Likelihood estimation techniques

Maximum likelihood and restricted maximum likelihood have become the most common strategies for estimating the variance component parameter  $\theta$ . 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).

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  $\beta$ . 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.

## 1.17 Estimation for LME Models

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  $\Sigma$ . 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 Laird and Ware (1982), 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

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

## 1.18 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. (1959, 1963, 1973, 1984) offered a more simpler approach of jointly estimating  $\hat{\beta}$  and  $\hat{b}$ . Henderson (1950) made the (ad-hoc) distributional assumptions  $y|b \sim N(X\beta + Zb, \Sigma)$  and

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

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (1.13)$$

with respect to  $\beta$  and  $b$ , which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)' \Sigma^{-1} (y - X\beta - Zb) + b' D^{-1} b. \quad (1.14)$$

This leads to the mixed model equations

$$\begin{pmatrix} X' \Sigma^{-1} X & X' \Sigma^{-1} Z \\ Z' \Sigma^{-1} X & X' \Sigma^{-1} X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X' \Sigma^{-1} y \\ Z' \Sigma^{-1} y \end{pmatrix}. \quad (1.15)$$

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  $\Sigma$  are known.

Robinson (1991) points out that although Henderson (1950) initially referred to the estimates  $\hat{\beta}$  and  $\hat{b}$  from (1.15) 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 (1.14) 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.

### Estimation of the fixed parameters

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

$$\ell(\beta, \theta | y) = -\frac{1}{2} \log |V| - \frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta), \quad (1.16)$$



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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \quad (1.17)$$

Substituting  $\hat{\beta}$  from (1.17) into  $\ell(\beta, \theta|y)$  from (1.16) returns the *profile* log-likelihood

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

of the variance parameter  $\theta$ . Estimates of the parameters  $\theta$  specifying  $V$  can be found by maximizing  $\ell_P(\theta | y)$  over  $\theta$ . These are the ML estimates. For REML estimation the *restricted* log-likelihood is defined as

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

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  $\beta$ . 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  $\beta$  equals  $DZ'V^{-1}(y - X\beta)$ . In practice  $\beta$  is replaced by an estimator such as  $\hat{\beta}$  from (1.17) 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 (1.15).

## Algorithms for likelihood function optimization

Iterative numerical techniques are used to optimize the log-likelihood function and estimate the covariance parameters  $\theta$ . 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 alternates 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 defined as  $-2$  times the log likelihood for the covariance parameters  $\theta$ . 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 a variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

## The extended likelihood

The desire to have an entirely likelihood-based justification for estimates of random effects, in contrast to Henderson’s equation, has motivated Pawitan (2001, page 429) to define the *extended likelihood*. He remarks “In mixed effects modelling the extended likelihood has been called *h-likelihood* (for hierarchical likelihood) by Lee and Nelder (1996), while in smoothing literature it is known as the *penalized likelihood* (e.g. Green and Silverman 1994).” The extended likelihood can be written  $L(\beta, \theta, b|y) = p(y|b; \beta, \theta)p(b; \theta)$  and adopting the same distributional assumptions used by Henderson (1950) yields the log-likelihood function

$$\begin{aligned}\ell_h(\beta, \theta, b|y) = & -\frac{1}{2} \{ \log |\Sigma| + (y - X\beta - Zb)' \Sigma^{-1} (y - X\beta - Zb) \\ & + \log |D| + b' D^{-1} b \} .\end{aligned}$$

Given  $\theta$ , differentiating with respect to  $\beta$  and  $b$  returns Henderson’s equations in (1.15).

## 1.19 Repeated measurements in LME models

Lam et al. (1999) 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  $\rho_{xy}$  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.

### 1.19.1 Formulation of the response vector

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

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

$$\mathbf{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})' \quad (1.18)$$

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

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (1.19)$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \quad (1.20)$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i) \quad (1.21)$$

$\boldsymbol{\beta}$  is a three dimensional vector containing the fixed effects.  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ .  $\beta_2$  is usually set to zero. Consequently  $\boldsymbol{\beta}$  is the solutions of the means of the two methods, i.e.  $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ . The variance covariance matrix  $\mathbf{D}$  is a general  $2 \times 2$  matrix, while  $\mathbf{R}_i$  is a  $2n_i \times 2n_i$  matrix.

## 1.20 Decomposition of the response covariance matrix

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

$$\text{Cov}(\mathbf{y}_i) = \mathbf{\Omega}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i.$$

$\mathbf{\Omega}_i$  can be expressed as

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

The notation  $\text{dim}_{n_i}$  means an  $n_i \times n_i$  diagonal block.

$\mathbf{R}_i$  can be shown to be the Kronecker product of a correlation matrix  $\mathbf{V}$  and  $\mathbf{\Lambda}$ . The correlation matrix  $\mathbf{V}$  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 ? use the identity matrix, with dimensions  $n_i \times n_i$  as the formulation for  $\mathbf{V}$ . 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  $6 \times 6$  matrix composed of two types of  $2 \times 2$  blocks. Each block represents one separate time of measurement.

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

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

$\mathbf{\Omega}_i$  can be expressed as

$$\mathbf{\Omega}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}). \quad (1.22)$$

The notation  $\text{dim}_{n_i}$  means an  $n_i \times n_i$  diagonal block.

### 1.20.1 Correlation terms

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

$$\mathbf{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_B \rho_{AB} \delta \\ \sigma_A \sigma_B \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix} \quad (1.23)$$

$$\mathbf{\Lambda} = \begin{pmatrix} \sigma_A^2(1 - \rho_A) & \sigma_{AB}(1 - \delta) \\ \sigma_{AB}(1 - \delta) & \sigma_B^2(1 - \rho_B) \end{pmatrix}. \quad (1.24)$$

$\rho_A$  describe the correlations of measurements made by the method  $A$  at different times. Similarly  $\rho_B$  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.  $\rho_{AB}$  describes the correlation of measurements taken at the same same time by both methods. The coefficient  $\delta$  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  $\delta$  is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

Lam et al. (1999) 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  $\rho_{xy}$  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 (2.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ , whereas the model in (1.3) requires  $N + 2$  fixed effects.

Allocating fixed effects to each item  $i$  by (1.3) 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.

subsectionBXC2004 Model Carstensen (2004) presents a model 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (1.25)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (1.26)$$

Of particular importance is terms of the model, a true value for item  $i$  ( $\mu_i$ ). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.



# Chapter 2

## 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  $\mathbf{X}_i$  will be reduced to a  $2 \times 1$  matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.
- $\mathbf{Z}_i$  is the  $2n_i \times 2$  model matrix for the random effects for measurement methods on item  $i$ .
- $\mathbf{b}_i$  is the  $2 \times 1$  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  $i$ .
- $\mathbf{G}$  is the  $2 \times 2$  covariance matrix for the random effects.
- $\mathbf{R}_i$  is the  $2n_i \times 2n_i$  covariance matrix for the residuals on item  $i$ .
- The expected value is given as  $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ . (Hamlett et al., 2004)

- The variance of the response vector is given by  $\text{Var}(\mathbf{y}_i) = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i$  (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  $\Omega_i$ ’ is the addition of  $\hat{D}$  and  $\hat{\Sigma}$ .

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (2.1)$$

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

$$\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \quad (2.2)$$

- $\mathbf{V}$  represents the correlation matrix of the replicated measurements on a given method.  $\mathbf{\Sigma}$  is the within-subject VC matrix.
- $\mathbf{V}$  and  $\mathbf{\Sigma}$  are positive definite matrices. The dimensions of  $\mathbf{V}$  and  $\mathbf{\Sigma}$  are  $3 \times 3 (= p \times p)$  and  $2 \times 2 (= k \times k)$ .
- It is assumed that  $\mathbf{V}$  is the same for both methods and  $\mathbf{\Sigma}$  is the same for all replications.
- $\mathbf{V} \otimes \mathbf{\Sigma}$  creates a  $6 \times 6 (= kp \times kp)$  matrix.  $\mathbf{R}_i$  is a sub-matrix of this.

## 2.1 Model Formula

Let  $y_{mir}$  denote the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method, where  $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$ . 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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (2.3)$$

Here  $\beta_0$  and  $\beta_m$  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  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{mi}, b_{m'i}) = g_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$ . When two methods of measurement are in agreement, there is no significant differences between  $\beta_1$  and  $\beta_2$ ,  $g_1^2$  and  $g_2^2$ , and  $\sigma_1^2$  and  $\sigma_2^2$ . Here  $\beta_0$  and  $\beta_m$

are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ . The model can be reparameterized by gathering the  $\beta$  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  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$ . Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  hold simultaneously. Roy (2009a) 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. 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.

## Chapter 3

# Introduction to Roy's Procedure

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  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  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. 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.

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).

### 3.1 Introduction to Roy’s methodology

For the purposes of comparing two methods of measurement, Roy (2009b) 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.

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. 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.

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.

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.

## 3.2 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.

## 3.3 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 (?).

Roy proposes an LME model with Kronecker product covariance structure in a doubly multivariate setup. Response for  $i$ th subject can be written as

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

- $\beta_1$  and  $\beta_2$  are fixed effects corresponding to both methods. ( $\beta_0$  is the intercept.)
- $b_{1i}$  and  $b_{2i}$  are random effects corresponding to both methods.

Overall variability between the two methods ( $\Omega$ ) is sum of between-subject ( $D$ ) and within-subject variability ( $\Sigma$ ),

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

For the purposes of comparing two methods of measurement, Roy (2009b) 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.

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).

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 ?, it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

### 3.4 Agreement Criteria

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.

(Work this in) 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.

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

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.

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 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 that 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 that 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.

### **3.5 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

$$H_1 : \mu_1 = \mu_2,$$

$$K_1 : \mu_1 \neq \mu_2.$$

## 3.6 Variability Tests

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  $D$  and  $\Lambda$  matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix  $A$ ,

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

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  $a_{11} = a_{22}$ .

### 3.7 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  $2 \times 2$  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- $\mathbf{\Omega}_i$ , but equally applicable to the component variabilities  $\mathbf{G}$  and  $\mathbf{\Sigma}$ );

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

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.

### 3.7.1 Variance-Covariance Structures

#### 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  $\psi^2$ ) for all 3 of the time points and a single covariance (represented by  $\psi_{ij}$ ) 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.

### 3.8 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.

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.

### 3.9 Hypothesis Testing

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.

### **3.10 Roy's hypothesis tests**

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,

$$H_2 : g_1^2 = g_2^2$$

$$K_2 : g_1^2 \neq g_2^2$$

and a formal test for the equality of within-item variances.

$$H_3 : \sigma_1^2 = \sigma_2^2$$

$$K_3 : \sigma_1^2 \neq \sigma_2^2$$

A formal test for the equality of overall variances is also presented.

$$H_4 : \omega_1^2 = \omega_2^2$$

$$K_4 : \omega_1^2 \neq \omega_2^2$$

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.

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.

### 3.11 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 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 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 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  $D$  and  $\Lambda$  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.

### 3.12 Correlation coefficient

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.



### 3.13 Roy's variability tests

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.

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

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

$$H_0 : g_1^2 = g_2^2$$

$$H_1 : g_1^2 \neq g_2^2$$

a formal test for the equality of within-item variances,

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

and finally, a formal test for the equality of overall variances.

$$H_0 : \omega_1^2 = \omega_2^2$$

$$H_1 : \omega_1^2 \neq \omega_2^2$$

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.

### 3.14 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  $\hat{\mathbf{A}}$  has a symmetric form for both models, and will be the same for both.

The first test allows of the comparison the begin-subject variability of two methods.

### 3.15 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.

$$H_0 : \lambda_A = \lambda_B$$

$$H_A : \lambda_A \neq \lambda_B$$

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

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.

### 3.16 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.

$$H_0 : \sigma_A = \sigma_B$$

$$H_A : \sigma_A \neq \sigma_B$$

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

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \quad (3.1)$$

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

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (3.2)$$

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \quad (3.3)$$

### 3.17 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:

$$H_5 : \sigma_{12} = 0$$

$$K_5 : \sigma_{12} \neq 0$$

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.

### 3.18 VC structures

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

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

$\Psi$  is the variance-covariance matrix of the random effects , with  $2 \times 2$  dimensions.

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{3.4}$$

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

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

$\Psi$  is the variance-covariance matrix of the random effects , with  $2 \times 2$  dimensions.

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{3.5}$$

# Chapter 4

## Model Specification

### 4.1 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  $\mathbf{y}_i = (y_{1i1}, y_{2i1}, y_{1i2}, \dots, y_{mir}, \dots, y_{1in_i}, y_{2in_i})'$ . The LME model can be written

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$  is a vector of fixed effects, and  $\mathbf{X}_i$  is a corresponding  $2n_i \times 3$  design matrix for the fixed effects. The random effects are expressed in the vector  $\mathbf{b} = (b_1, b_2)'$ , with  $\mathbf{Z}_i$  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  $\mathbf{b}_i \sim N(0, \mathbf{G})$ ,  $\boldsymbol{\epsilon}_i$  is a matrix of random errors distributed as  $N(0, \mathbf{R}_i)$  and that the random effects and residuals are independent of each other.

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

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

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

It is assumed that  $\mathbf{b}_i \sim N(0, \mathbf{G})$ ,  $\boldsymbol{\epsilon}_i$  is a matrix of random errors distributed as  $N(0, \mathbf{R}_i)$  and that the random effects and residuals are independent of each other. Assumptions made on the structures of  $\mathbf{G}$  and  $\mathbf{R}_i$  will be discussed in due course.

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

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

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

The matrix of random errors  $\boldsymbol{\epsilon}_i$  is distributed as  $\mathcal{N}_2(0, \mathbf{R}_i)$ . 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.  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}$ .

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

where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  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  $\mathbf{G}$  and  $\mathbf{R}_i$ .

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



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

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

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

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

## 4.2 G Component

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

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

It is assumed that  $\mathbf{b}_i \sim N(0, \mathbf{G})$ ,  $\boldsymbol{\epsilon}_i$  is a matrix of random errors distributed as  $N(0, \mathbf{R}_i)$  and that the random effects and residuals are independent of each other. Assumptions made on the structures of  $\mathbf{G}$  and  $\mathbf{R}_i$  will be discussed in due course.

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

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

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

## 4.3 R Component

$\mathbf{R}_i$  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  $\mathcal{N}_2(0, \mathbf{R}_i)$ .

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.  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}$ .

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

where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. The within-item variance covariance matrix  $\mathbf{\Sigma}$  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  $\mathbf{G}$  and  $\mathbf{R}_i$ .

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

Computational analysis of linear mixed effects models allow for the explicit analysis of both  $\mathbf{G}$  and  $\mathbf{R}_i$ . The above terms can be used to express the variance covariance matrix  $\mathbf{\Omega}_i$  for the responses on item  $i$ ,

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

The partial within-item variance covariance matrix of two methods at any replicate is denoted  $\mathbf{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of both methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. The within-item variance covariance matrix  $\mathbf{\Sigma}$  is assumed to be the same for all replications.

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

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 the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

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  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  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 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  $i$ ,

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

## 4.4 Hamlett

$\boldsymbol{R}_i$  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  $\boldsymbol{R}_i$ .

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  $2 \times 2$  block diagonal Block- $\boldsymbol{\Omega}_i$  represents the covariance matrix between two methods, and is the sum of  $\boldsymbol{G}$  and  $\boldsymbol{\Sigma}$ .

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

Hamlett et al. (2004) shows that  $\mathbf{R}_i$  can be expressed as  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ . The partial within-item variance?covariance matrix of two methods at any replicate is denoted  $\mathbf{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix  $\mathbf{\Sigma}$  is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (4.1)$$

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

## 4.5 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

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

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

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

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

## 4.6 Kroneckor

The between-item variance covariance matrix  $\mathbf{G}$  is constructed as follows:

Both covariance matrices can be written as follows;

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

and

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

Hamlett et al. (2004) shows that  $\mathbf{R}_i$  can be expressed as  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ . The partial within-item variance?covariance matrix of two methods at any replicate is denoted  $\mathbf{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and

$\sigma_{12}$  is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix  $\Sigma$  is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (4.2)$$

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

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

## 4.7 Overall Variability

The overall variability between the two methods is the sum of between-item variability  $\mathbf{G}$  and within-item variability  $\Sigma$ . Roy (2009b) denotes the overall variability as Block -  $\Omega_i$ . 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- $\Omega_i$ . 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 -  $\Omega_i$  matrix. Lack of agreement can arise if there is a disagreement in overall



variabilities. This may be due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

## 4.8 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.

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

$\sigma_x^2$  and  $\sigma_y^2$  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.

## 4.9 Formal Testing

A formal test can be performed to test the hypothesis that the off-diagonal terms are zero.

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \text{ vs } \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

## 4.10 Using LMEs for Method Comparison

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)

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

# Chapter 5

## Likelihood Ratio Tests

### 5.1 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.

- 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  $\beta$ .

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  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively.

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.

## 5.2 Likelihood Ratio Tests

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.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986).

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

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

## 5.3 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) 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.

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.

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

### **5.3.1 Nested and Reference Models**

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship [CITE: West et al].

LRTs are a class of tests used to compare the value of likelihood functions for two models defining a hypothesis to be tested (i.e. the nested and reference model).

The relationship between the respective models presented by Roy (2009a) 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.

## 5.4 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) 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 (West et al., 2007). Conversely, Pinheiro and Bates (1994) 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.

## 5.5 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  $\chi^2$  distribution with  $k_2 - k_1$  degrees of freedom (Pinheiro and Bates, 1994, 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  $\chi^2$  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, Pinheiro and Bates (1994) 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.

## 5.6 Nesting: Model Selection Using Likelihood Ratio Tests

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).

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.

## 5.7 Other material

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 likelihood), 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.

## 5.8 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]

## 5.9 Test Statistic for Likelihood Ratio Tests

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  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively.

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.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \quad (5.1)$$

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



$$\nu = [\text{LRT df under } H_0 \text{ model}] - [\text{LRT df under } H_A \text{ model}] \quad (5.2)$$

The probability distribution of the test statistic can be approximated by a chi-square distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively.

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

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

The probability distribution of the test statistic is approximated by the  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  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  $\theta$ ,

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

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

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

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

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 ( $\mathcal{I}(\theta)$ ).

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.

## 5.10 Pinheiro Bates

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  $\chi^2$  distribution with  $k_2 - k_1$  degrees of freedom (Pinheiro and Bates, 1994, 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.

## 5.11 Empirical p-values of LRT tests

For both REML and ML estimates, the nominal  $p$ -values for the LRT statistics under a  $\chi^2$  distribution with 2 degrees of freedom are much greater than empirical values. A number of ways of dealing with this issues are discussed (Pinheiro and Bates, 1994, 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.(Pinheiro and Bates, 1994, 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.

## 5.12 Relevance of Estimation Methods

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, Pinheiro and Bates (1994) 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.

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.

## 5.13 Information Criteria

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 variance 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. Treatment of items as fixed effects [Pinheiro Bates 2000] 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 Bates 2000] advises the specification of random effects to correspond to items; treating the item effects as random deviations from the population mean.

## 5.14 Likelihood Ratio Tests in Roy's Analysis

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.

## Chapter 6

### LOAS

## 6.1 Calculation of limits of agreement

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

However, the original BlandAltman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

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.

Computing limits of agreement features prominently in many method comparison studies, further to Bland and Altman (1986, 1999).

Bland and Altman (1999) addresses the issue of computing LoAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods.

However, the original Bland-Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method. Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models.

Further to Bland and Altman (1986), the computation of the limits of agreement follows from the intermethod bias, and the variance of the difference of measurements.

The computation of the inter-method bias is a straightforward subtraction calculation. The variance of differences is easily computable from the variance estimates in the Block -  $\Omega_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, Carstensen et al. (2008) use a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations. Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

## 6.2 00-Limits of Agreement in LME models

Carstensen et al. (2008) computes the limits of agreement to the case with repeated



measurements by using LME models.

Roy (2009b) formulates a very powerful method of assessing whether two methods of measurement, with replicate measurements, also using LME models. Roy's approach is based on the construction of variance-covariance matrices.

Importantly, Roy's approach does not address the issue of limits of agreement (though another related analysis, the coefficient of repeatability, is mentioned).

This paper seeks to use Roy's approach to estimate the limits of agreement. These estimates will be compared to estimates computed under Carstensen's formulation.

In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus hereon is solely on the variance of differences.

### **6.3 Carstensen et al**

Carstensen et al. (2008) also use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, they recommend a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

## 6.4 Computing LoAs from LME models

*One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice.*

## 6.5 Carstensen's Limits of Agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. The method of computation is the same as Roy's model, but with the covariance estimates set to zero.

Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required.

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required. The method of computation is the same as Roy's model, but with the covariance estimates set to zero.

Carstensen et al. (2008) uses LME models to determine the limits of agreement. In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus is solely on the standard deviation.

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example,

it is shown that Roy's LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ . Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

Further to Bland and Altman (1986), the computation of the limits of agreement follows from the intermethod bias, and the variance of the difference of measurements. The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\mathbf{\Omega}_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

## 6.6 Carstensen's Model

Using Carstensen's notation, a measurement  $y_{mi}$  by method  $m$  on individual  $i$  the measurement  $y_{mir}$  is the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method, where  $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$  is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (6.1)$$

Of particular importance is terms of the model, a true value for item  $i$  ( $\mu_i$ ). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

The classical model is based on measurements  $y_{mi}$  by method  $m = 1, 2$  on item  $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Here the terms  $\alpha_m$  and  $\mu_i$  represent the fixed effect for method  $m$  and a true value for item  $i$  respectively. The random effect terms comprise an interaction term  $c_{mi}$  and the residuals  $\epsilon_{mir}$ . The  $c_{mi}$  term represent random effect parameters corresponding to the two methods, having  $E(c_{mi}) = 0$  with  $\text{Var}(c_{mi}) = \tau_m^2$ . Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such,  $\text{Cov}(c_{mi}, c_{m'i}) = 0$ . All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate  $\alpha$  can not be estimated, only their difference can be estimated as  $\bar{D}$

The presence of the true value term  $\mu_i$  gives rise to an important difference between Carstensen's and Roy's models. The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. In other words, Roy considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

With regards to specifying the variance terms, Carstensen remarks that using his approach is common, remarking that *The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods (?)*.

Carstensen et al. (2008) makes some interesting remarks in this regard.

The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

As the difference between methods is of interest, the item term can be disregarded.

We assume that the variance of the measurements is different for both methods, but it does not mean that the separate variances can be estimated with the data available.

## 6.7 BXC2008 presents - Carstensen's Limits of agreement

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen et al. (2008) use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available, instead proposing a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest lies specifically in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy's model, but

for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by (1.3). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (1.3), when covariance between methods is present.

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required.

## 6.8 Limits of Agreement in LME models

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  $\lambda_m^2$ . 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  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (6.2)$$

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

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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.3)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.4)$$

## 6.9 BXC2008 uses - Limits of Agreement in LME models

Carstensen et al. (2008) uses an approach based on linear mixed effects (LME) models for the purpose of computing the limits of agreement for two methods of measurement, where replicate measurements are taken on items. As the emphasis of this methodology lies on the inter-method bias and the limits of agreement, the two key elements of the Bland-Altman methodology, other formal tests are not described.

Between-subject variation for method  $m$  is given by  $d_m^2$  and within-subject variation is given by  $\lambda_m^2$ . 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  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (6.5)$$



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

## 6.10 Carstensen's LOAs

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

## 6.11 Interaction Terms in Model

Further to ?, if the measurements by a method on an item are not necessarily true replications, e.g., repeated measures over time, then additional terms may be needed for  $e_{mir}$ . ? also addresses this issue by the addition of an interaction term (i.e. a random effect)  $u_{mi}$ , yielding

$$y_{mir} = \alpha_{mi} + u_{mi} + e_{mi}.$$

The additional interaction term is characterized as  $u_{mi} \sim \mathcal{N}(0, \tau_m^2)$  (?).

This extra interaction term provides a source of extra variability, but this variance is not relevant to computing the case-wise differences.

? advises that the formulation of the model should take the exchangeability (in other words, whether or not the measurements are ‘true replicates’) into account. If there is a linkage between measurements (therefore not ‘true’ replicates), the ‘item by replicate’ should be included in the model. If there is no linkage, and the replicates are indeed true replicates, the interaction term should be omitted.

? demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. As a surplus source of variability is excluded from the computation, the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates.

? also assigns a random effect  $u_{mi}$  for each response  $y_{mir}$ . Importantly Roy's model assumes linkage.

Carstensen et al. (2008) formulates an LME model, both in the absence and the presence of an interaction term.? uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in over-estimation of the limits of agreement. For the Carstensen estimates below, an interaction term was included when computed.

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

## 6.12 Computation of limits of agreement

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\Omega_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy's model, but for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by (1.3). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using

models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (1.3), when covariance between methods is present.

### 6.13 BXC - Model Terms

- Let  $y_{mir}$  be the response of method  $m$  on the  $i$ th subject at the  $r$ -th replicate.
- Let  $\mathbf{y}_{ir}$  be the  $2 \times 1$  vector of measurements corresponding to the  $i$ -th subject at the  $r$ -th replicate.
- Let  $\mathbf{y}_i$  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  $\alpha_{mi}$  be the fixed effect parameter for method for subject  $i$ .
- Formally Roy uses a separate fixed effect parameter to describe the true value  $\mu_i$ , 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  $i$ .
- $\epsilon_i$  is a  $n_i$ -dimensional vector comprised of residual components. For the blood pressure data  $n_i = 85$ .
- $\beta$  is the solutions of the means of the two methods. In the LME output, the bias and corresponding t-value and p-values are presented. This is relevant to Roy's first test.

### 6.14 Computation of limits of agreement under Roy's model

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject

VC matrix and the within-subject VC matrix. The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\mathbf{\Omega}_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

The standard deviation of the differences of methods  $x$  and  $y$  is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

## 6.15 LOAs with Roy

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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.6)$$

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject VC matrix.

The standard deviation of the differences of methods  $x$  and  $y$  is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.7)$$

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\Omega_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

## 6.16 Difference Between Approaches

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen.

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

## 6.17 Differences Between Models

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations. Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

There is a substantial difference in the number of fixed parameters used by the respective models. For the model in (2.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ . In contrast, the model described by (1.3) requires  $N + 2$  fixed effects for  $N$  items. The inclusion of fixed effects to account for the 'true value' of each item greatly increases the level of model complexity.

When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of  $\tau_m^2$  can not be obtained due to the model over-specification. To overcome this, the assumption of equality, i.e.  $\tau_1^2 = \tau_2^2$ , is required.

## 6.18 Differences

Roy (2009b) has demonstrated a methodology whereby  $d_A^2$  and  $d_B^2$  can be estimated

separately. Also covariance terms are present in both  $\mathbf{D}$  and  $\mathbf{\Lambda}$ . Using Roy's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.8)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.9)$$

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand. Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

In contrast to Roy's model, Carstensen's model requires that commonly used assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off-diagonal elements are also zero. Also, implementation requires that the between-item variances are estimated as the same value:  $g_1^2 = g_2^2 = g^2$ . As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

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

## 6.19 Relevance of Roy's Methodology

The relevance of Roy's methodology is that estimates for the between-item variances for both methods  $\hat{d}_m^2$  are computed. Also the VC matrices are constructed with covariance terms and, so the difference variance must be formulated accordingly.

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{\hat{d}_1^2 + \hat{d}_1^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{d}_{12} - 2\hat{\sigma}_1\hat{\sigma}_2}$$

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  $\Omega_i$ ' is the addition of  $\hat{D}$  and  $\hat{\Sigma}$ .

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{6.10}$$

## 6.20 Difference Variance further to Carstensen

Carstensen et al. (2008) states a model where the variation between items for method  $m$  is captured by  $\tau_m$  (our notation  $d_m^2$ ) and the within-item variation by  $\sigma_m$ .

*The formulation of this model is general and refers to comparison of any number of methods however, if only two methods are compared, separate values of  $\tau_1^2$  and  $\tau_2^2$*



cannot be estimated, only their average value  $\tau$ , so in the case of only two methods we are forced to assume that  $\tau_1 = \tau_2 = \tau$  (Carstensen et al., 2008).

Another important point is that there is no covariance terms, so further to Carstensen et al. (2008) the variance covariance matrices for between-item and within-item variability are respectively.

$$\mathbf{D} = \begin{pmatrix} d_2^1 & 0 \\ 0 & d_2^2 \end{pmatrix}$$

and  $\mathbf{\Sigma}$  is constructed as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Under this model the limits of agreement should be computed based on the standard deviation of the difference between a pair of measurements by the two methods on a new individual,  $j$ , say:

$$\text{var}(y_{1j} - y_{2j}) = 2d^2 + \sigma_1^2 + \sigma_2^2$$

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{d}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

## 6.21 Assumptions on Variability

Aside from the fixed effects, another important difference is that Carstensen's model requires that particular assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off diagonal elements are also zero.

Also, implementation requires that the between-item variances are estimated as the same value:  $g_1^2 = g_2^2 = g^2$ . Necessarily Carstensen's method does not allow for a formal test of the between-item variability.

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

In cases where the off-diagonal terms in the overall variability matrix are close to zero, the limits of agreement due to ? are very similar to the limits of agreement that follow from the general model.

## 6.22 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  $r$ th replicate measurement on subject  $i$  with method  $m$ .

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (6.11)$$

The intercept term  $\alpha$  and the  $\beta_m \mu_i$  term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value  $\mu_i$ .  $c_{mi}$  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  $m$ ;

$$BLUP_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + c_{mi} \quad (6.12)$$

. Under the assumption that the  $\mu$ s 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (6.13)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (6.14)$$

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  $\tau_1^2$  and  $\tau_2^2$  separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \quad (6.15)$$

# Chapter 7

## Other Material

### 7.1 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.

## 7.2 Roy's methodology for single measurements

Roy's methodology 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, simple existing methodologies would be the correct approach where there only one measurements by each method. 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.

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.

### 7.3 Correlation

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.

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 (2009b) remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

### 7.4 Correlation terms

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

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



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

$\rho_A$  describe the correlations of measurements made by the method  $A$  at different times. Similarly  $\rho_B$  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.  $\rho_{AB}$  describes the correlation of measurements taken at the same same time by both methods. The coefficient  $\delta$  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  $\delta$  is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

## 7.5 Hamlett and Lam

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

Hamlett re-analyses the data of ? to generalize their model to cover other settings not covered by the Lam method.

In many cases, repeated observation are collected from each subject in sequence and/or longitudinally.

$$y_i = \alpha + \mu_i + \epsilon$$

## 7.6 LaiShiao

Lai and Shiao (2005) advocates the use of LME models to study method comparison problems. The authors analyse a data set typical of method comparison studies using SAS software, with particular use of the ‘*Proc Mixed*’ package. The stated goal of this study is to determine which factor from a specified group of factors is the key contributor to the difference in the two methods.

The study relates to oxygen saturation, the most investigated variable in clinical nursing studies (Lai and Shiao, 2005). The two method compared are functional saturation (SO<sub>2</sub>, percent functional oxy-hemoglobin) and fractional saturation (HbO<sub>2</sub>, percent fractional oxy-hemoglobin), which is considered to be the ‘gold standard’ method of measurement.

Lai and Shiao (2005) establishes an LME model for analysing the differences  $D_{ijtl}$ , where  $D_{ijtl}$  is the differences of the measurements (i.e.  $SO_{2_{ijtl}} - HbO_{2_{ijtl}}$ ) for the  $i$ th donor at the  $j$ th level of foetal haemoglobin percent (Fhbperct) and the  $t$ th repeated measurement by the  $l$ th practitioner of the experiment.

(Carstensen (2004) also advocates the use of LME models in comparing methods, but with a different emphasis.)

## 7.7 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  $\mathbf{b}$ , 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  $\mathbf{G}$  and  $\mathbf{R}$ .

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’;

$$\begin{aligned} \text{fixed effects : } & 2.5056 - 0.0263\text{Fhbperct}_{ijtl} & (7.1) \\ \text{(p-values : } & = 0.0054, < 0.0001, < 0.0001) \end{aligned}$$

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

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;

$$\begin{aligned} \text{fixed effects : } & - 0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ \text{(p-values : } & = 0.8113, < 0.0001, < 0.0001) \\ \\ \text{random effects : } & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ \text{(p-values : } & = 0.0446, < 0.0001, < 0.0001) & (7.2) \end{aligned}$$

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.

## 7.8 Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessaire 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  $\lambda_m^2$ . 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  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (7.3)$$

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.4)$$

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  $\mathbf{D}$  and  $\mathbf{\Lambda}$ . Using Roy’s methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (7.5)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (7.6)$$

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);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

### 7.8.1 Variance Ratios

Variance Ratios The approach proposed by Roy deals with the question of agreement, and indeed interchangeability, as developed by Bland and Altmans corpus of work.

In the view of Dunn, a question relevant to many practitioners is which of the two methods is more precise.

The relationship between precision and the within-item and between-item variability must be established. Roy establishes the equivalence of repeatability and within-item variability, and hence precision. The method with the smaller within-item variability can be deemed to be the more precise. A useful approach is to compute the confidence intervals for the ratio of within-item standard deviations (equivalent to the ratio of repeatability coefficients), which can be interpreted in the usual manner. In fact, the ratio of within-item standard deviations, with the attendant confidence interval, can be determined using a single R command: `intervals()`. Pinheiro and Bates (pg 93-95) give a description of how confidence intervals for the variance components are computed. Furthermore a complete set of confidence intervals can be computed to complement the variance component estimates. What is required is the computation of the variance ratios of within-item and between-item standard deviations. A naive approach would be to compute the variance ratios by relevant F distribution quantiles. However, the question arises as to the appropriate degrees of freedom. Limits of agreement are easily computable using the LME framework. While we will not be considering this analysis, a demonstration will be provided in the example.

## 7.9 Worked Examples

### 7.9.1 LikelihoodRatio Tests

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat,    #Symm , Symm#  
+   random = list(item=pdSymm(~ meth-1)),  
+   weights=varIdent(form=~1|meth),
```

```
+ correlation = corSymm(form=~1 | item/repl),
+ method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model. Nested Model (Between-Item Variability)

```
> NMB.fit = lme(y ~ meth-1, data = dat, #CS , Symm#
+ random = list(item=pdCompSymm(~ meth-1)),
+ correlation = corSymm(form=~1 | item/repl),
+ method="ML")
```

Nested Model (Within item Variability)

```
> NMW.fit = lme(y ~ meth-1, data = dat, #Symm , CS#
+ random = list(item=pdSymm(~ meth-1)),
+ weights=varIdent(form=~1|meth),
+ correlation = corCompSymm(form=~1 | item/repl),
+ method="ML")
```

Nested Model (Overall Variability) Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.



```

> NMO.fit = lme(y ~ meth-1, data = dat,    #CS , CS#
+   random = list(item=pdCompSymm(~ meth-1)),
+   correlation = corCompSymm(form=~1 | item/repl),
+   method="ML")

```

ANOVAs for Original Fits The likelihood Ratio test is very simple to implement in R. All that is required it to specify the reference model and the relevant nested mode as arguments to the command `anova()`. The figure below displays the three tests described by Roy (2009).

```

> testB    = anova(Ref.Fit,NMB.fit)                # Between-Subject Var
> testW    = anova(Ref.Fit,NMW.fit)                # Within-Subject Variabi
> testO    = anova(Ref.Fit,NMO.fit)                # Overall Variabilitie

```

## 7.10 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.

The probability distribution of the test statistic can be approximated by a chi-square distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  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.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \quad (7.7)$$

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

$$\nu = [\text{LRT df under } H_0 \text{ model}] - [\text{LRT df under } H_A \text{ model}] \quad (7.8)$$

```
> 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
```

### 7.10.1 Roy's Reference Model

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat, #Symm , Symm#
+ random = list(item=pdSymm(~ meth-1)),
+ weights=varIdent(form=~1|meth),
+ correlation = corSymm(form=~1 | item/repl),
+ method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model.

### 7.10.2 Nested Model (Between-Item Variability)

```
> NMB.fit = lme(y ~ meth-1, data = dat, #CS , Symm#
+ random = list(item=pdCompSymm(~ meth-1)),
+ correlation = corSymm(form=~1 | item/repl),
```

```
+     method="ML")
```

## 7.11 Worked Examples

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

? uses the “Blood” data set, which featured in Bland and Altman (1999).

## 7.12 Examples: LoAs for Carstensen’s data

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.9)$$

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);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

? describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are  $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are  $(-9.97, 14.81)$ . Carstensen advises the inclusion of the

interaction term for linked replicates, and hence the limits of agreement are recomputed as  $(-12.18, 17.12)$ .

### 7.12.1 Diabetes Example

Carstensen et al. (2008) 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 et al. (2008) notes that every machine was calibrated every day to the manufacturers guidelines. Measurements are classified by method, individual and replicate. In this case the replicates are clearly not exchangeable, neither within patients nor simulataneously for all patients.

### 7.12.2 Oximetry Data

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 additional 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);

(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 be 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  $\hat{\mathbf{A}}$  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  $\hat{\mathbf{D}}$  and  $\hat{\mathbf{A}}$ . Therefore the test's proposed by Roy (2009b) 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  $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$ .)



Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are  $(-9.97, 14.81)$ . Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as  $(-12.18, 17.12)$ .

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);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

### 7.12.3 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);

(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 be 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  $\hat{\mathbf{A}}$  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  $\hat{\mathbf{D}}$  and  $\hat{\mathbf{A}}$ . Therefore the test's proposed by Roy (2009b) 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  $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$ .

#### 7.12.4 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);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \quad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \quad (7.10)$$

(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  $\hat{\mathbf{D}}$  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  $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$ .

### 7.12.5 Limits of agreement for Carstensen’s data

? describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are  $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are  $(-9.97, 14.81)$ . Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as  $(-12.18, 17.12)$ .

### 7.12.6 Fat Data Examples: LoAs for Carstensen’s data

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.11)$$

Carstensen et al. (2008) describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are  $(-0.23, 0.32)$

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);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

### 7.12.7 RV-IV

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

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \quad (7.12)$$

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

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \quad (7.13)$$

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

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \quad (7.14)$$

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.

### 7.12.8 Limits of agreement for Oximetry

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are  $(-9.97, 14.81)$ . Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as  $(-12.18, 17.12)$ .

### 7.12.9 Classical Model

The classical model is based on measurements  $y_{mi}$  by method  $m = 1, 2$  on item  $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate  $\alpha$  can not be estimated, only their difference can be estimated as  $\bar{D}$

## 7.13 Classical model for single measurements

The classical model is based on measurements  $y_{mi}$  by method  $m = 1, 2$  on item  $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate  $\alpha$  can not be estimated, only their difference can be estimated as  $\bar{D}$

In the first instance, we require a simple model to describe a measurement by method  $m$ . We use the term *item* to denote an individual, subject or sample, to be measured, being randomly sampled from a population. Let  $y_{mi}$  be the measurement for item  $i$  made by method  $m$ .



$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

- $\alpha_m$  is the fixed effect associated with method  $m$ ,
- $\mu_i$  is the true value for subject  $i$  (fixed effect),
- $e_{mi}$  is a random effect term for errors with  $e_{mi} \sim \mathcal{N}(0, \sigma_m^2)$ .

This model implies that the difference between the paired measurements can be expressed as

$$d_i = y_{1i} - y_{2i} \sim \mathcal{N}(\alpha_1 - \alpha_2, \sigma_1^2 - \sigma_2^2).$$

Importantly, this is independent of the item levels  $\mu_i$ . As the case-wise differences are of interest, the parameters of interest are the fixed effects for methods  $\alpha_m$ .

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

Importantly these variance covariance structures are central to Roy methodology.

? proposes a series of hypothesis tests based on these matrices as part of her methodology. These tests shall be reverted to in due course.

The standard deviation of the differences of variables  $a$  and  $b$  is computed as

$$\text{var}(a - b) = \text{var}(a) + \text{var}(b) - 2\text{cov}(a, b)$$

Hence the variance of the difference of two methods, that allows for the calculation of the limits of agreement, can be calculated as

$$\text{var}(d) = \omega_1^2 + \omega_2^2 - 2 \times \omega_1 \omega_2$$

### 7.13.1 Sampling

*One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this )*

### 7.13.2 Remarks on the Multivariate Normal Distribution

Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution. Roy's model is specified using the bivariate normal distribution. This gives rise to a key difference between the two models, in that a bivariate model accounts for covariance between the variables of interest. The multivariate normal distribution of a  $k$ -dimensional random vector  $X = [X_1, X_2, \dots, X_k]$  can be written in the following notation:

$$X \sim \mathcal{N}(\mu, \Sigma),$$

or to make it explicitly known that  $X$  is  $k$ -dimensional,

$$X \sim \mathcal{N}_k(\mu, \Sigma).$$

with  $k$ -dimensional mean vector

$$\mu = [E[X_1], E[X_2], \dots, E[X_k]]$$

and  $k \times k$  covariance matrix

$$\Sigma = [\text{Cov}[X_i, X_j]], \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, k$$

#### 1. Univariate Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

#### 2. Bivariate Normal Distribution

(a)

$$X \sim \mathcal{N}_2(\mu, \Sigma),$$

(b)

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

## 7.14 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;

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

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.

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

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;

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

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  $\hat{\mathbf{r}}_{\Omega_{ii}}$  of the correlation matrix indicate an overall coefficient of 0.7959. This is less than the threshold of 0.82 that Roy recommends.

$$\hat{\mathbf{r}}_{\Omega_{ii}} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix  $\hat{\mathbf{r}}_{\Omega_{ii'}}$  present the correlation

coefficients further to Hamlett et al. (2004).

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

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.

## 7.15 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 his first example, and the 'R' and 'S' methods in his second.

### 7.15.1 Matrix structures

Before discussing the tests, it is useful to point out the difference between symmetric form and compound symmetry form. Consider a generic matrix  $A$ ,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (7.15)$$

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  $a_{11} = a_{22}$ .

### 7.15.2 Variability test 1

This is a test on whether both methods  $A$  and  $B$  have the same between-subject variability or not.

$$H_0 : d_A = d_B \quad (7.16)$$

$$H_A : d_A \neq d_B \quad (7.17)$$

When implemented using **R**, 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  $\hat{\mathbf{A}}$  has a symmetric form for both models, and will be the same for both.

#### Bland-Altman's blood data

With the alternative model, the MLE of the between-subject variance covariance matrix is given by

$$\hat{D}_{Symm} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix} \quad (7.18)$$

With the null model the MLE is as follows:

$$\hat{D}_{CS} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix} \quad (7.19)$$

A likelihood ratio test is performed to determine which model is more suitable. The outcome of this test is presented in the following **R** code.

```
> 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 test statistic is the difference of the  $-2 \log$  likelihoods; 0.15291. The  $p$ -value is 0.6958. Therefore we fail to reject the hypothesis that both have the same between-subject variabilities.

### 7.15.3 Variability test 2

This is a test on whether both methods  $A$  and  $B$  have the same within-subject variability or not.

$$H_0 : \lambda_A = \lambda_B \quad (7.20)$$

$$H_A : \lambda_A \neq \lambda_B \quad (7.21)$$

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

#### Bland-Altman's blood data

For the null model the MLE of the within-subject variance covariance matrix is given below.

$$\hat{\mathbf{\Lambda}}_{Symm} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix} \quad (7.22)$$

With the alternative model the MLE is as follows:

$$\hat{\mathbf{\Lambda}}_{CS} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix} \quad (7.23)$$

A likelihood ratio test is perform to determine which model is more suitable. The outcome of this test is that it can be assumed that they have equal The test statistic is the difference of the  $-2 \log$  likelihoods; 28.617. The  $p$ -value is less than 0.0001. In

this case we reject the null hypothesis that both models have the same within-subject variabilities.

### 7.15.4 Variability test 3

This is a test on whether both methods  $A$  and  $B$  have the same overall variability or not.

$$H_0 : \sigma_A = \sigma_B \quad (7.24)$$

$$H_A : \sigma_A \neq \sigma_B \quad (7.25)$$

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

#### Bland-Altman's blood data

With the null model the MLE of the within-subject variance covariance matrix is given below.

$$\hat{\Sigma}_{Symm} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix} \quad (7.26)$$

With the alternative model the MLE is as follows:

$$\hat{\Sigma}_{CS} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix} \quad (7.27)$$

Again a likelihood ratio test is used to determine the most suitable of the two candidate models. The test statistic is the difference of the  $-2 \log$  likelihoods; 28.884. The  $p$ -value is less than 0.0001. We again reject the null hypothesis. Each model has a different overall variability, a foregone conclusion from the second variability test.

### 7.15.5 Test for inter-method bias

The inter-method bias between the two method is found to be 15.62 , with a  $p$ -value of

### 7.15.6 Correlation Test

$$\hat{\mathbf{r}}_{\Omega_{ii}} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix} \quad (7.28)$$

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

The off diagonal blocks of the overall correlation matrix  $\hat{\mathbf{r}}_{\Omega_{ii'}}$  are

$$\hat{\mathbf{r}}_{\Omega_{ii'}} = \begin{pmatrix} 0.9611 & 0.7799 \\ 0.7799 & 0.9212 \end{pmatrix}. \quad (7.29)$$

### 7.15.7 Conclusion of procedure

The overall conclusion of the procedure is that the two methods are not in agreement, specifically due to the within-subject variability, and the inter-method bias. The repeatability coefficients are substantially different, one being 49% larger than the other. Additionally the overall correlation coefficient did not exceed the recommended threshold of 0.82.

## 7.16 Basic Models Fits

Further to ?, several simple LME models are constructed for the blood pressure data. This data set is the subject of a method comparison study in Bland and Altman (1999).

### 7.16.1 Implementing the Mixed Models Fits

They are implemented using the following R code, utilising the ‘nlme’ package. An analysis of variance is used to compare the model fits.

The R script:

```
fit1 = lme( BP ~ method, data = dat, random = ~1 | subject )
fit2 = update(fit1, random = ~1 | subject/method )
fit3 = update(fit1, random = ~method - 1 | subject )
#analysis of variance
anova(fit1,fit2,fit3)
```

1. Simplest workable model, allows differences between methods and incorporates a random intercept for each subject. For subject 1 we have

$$\mathbf{X}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{b}_i = b$$

where  $E(b) = 0$  and  $\text{var}(b) = \psi$ .

2.

$$\mathbf{Z}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \mathbf{b}_i = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

where  $E(b_i) = 0$  and  $\text{var}(\mathbf{b}) = \mathbf{\Psi}$ .

The variance of error terms is a  $6 \times 6$  matrix.

### 7.16.2 Model Fit 1

This is a simple model with no interactions. There is a fixed effect for each method and a random effect for each subject.

$$y_{ijk} = \beta_j + b_i + \epsilon_{ijk}, \quad i = 1, \dots, 2, j = 1, \dots, 85, k = 1, \dots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear mixed-effects model fit by REML

Data: dat

Log-restricted-likelihood: -2155.853

Fixed: BP ~ method

(Intercept)      methodS

127.40784      15.61961

Random effects:

Formula: ~1 | subject

(Intercept) Residual

StdDev:      29.39085 12.44454

Number of Observations: 510

Number of Groups: 85

The following output was obtained.

Linear mixed-effects model fit by REML

Data: dat

Log-restricted-likelihood: -2047.582

Fixed: BP ~ method

(Intercept)      methodS

127.40784      15.61961

Random effects:

Formula: ~method - 1 | subject

Structure: General positive-definite, Log-Cholesky parametrization

StdDev      Corr

methodJ   30.455093   methdJ

methodS   31.477237   0.835

Residual   7.763666

Number of Observations: 510

Number of Groups: 85

### 7.16.3 Model Fit 1

This is a simple model with no interactions. There is a fixed effect for each method and a random effect for each subject.

$$y_{ijk} = \beta_j + b_i + \epsilon_{ijk}, \quad i = 1, \dots, 2, j = 1, \dots, 85, k = 1, \dots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear mixed-effects model fit by REML

Data: dat

Log-restricted-likelihood: -2155.853

Fixed: BP ~ method

(Intercept)      methodS

127.40784      15.61961

Random effects:

Formula: ~1 | subject

(Intercept) Residual

StdDev:      29.39085 12.44454

Number of Observations: 510

Number of Groups: 85

The following output was obtained.

Linear mixed-effects model fit by REML

Data: dat

Log-restricted-likelihood: -2047.582

Fixed: BP ~ method

(Intercept)      methodS



127.40784      15.61961

Random effects:

Formula: ~method - 1 | subject

Structure: General positive-definite, Log-Cholesky parametrization

StdDev      Corr

methodJ   30.455093   methdJ

methodS   31.477237   0.835

Residual   7.763666

Number of Observations: 510

Number of Groups: 85

### 7.16.4 Model Fit 2

This is a simple model, this time with an interaction effect. There is a fixed effect for each method. This model has random effects at two levels  $b_i$  for the subject, and another,  $b_{ij}$ , for the respective method within each subject.

$$y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk}, \quad i = 1, \dots, 2, j = 1, \dots, 85, k = 1, \dots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_1^2), \quad b_{ij} \sim \mathcal{N}(0, \sigma_2^2), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, the random interaction terms all have the same variance  $\sigma_2^2$ . These terms are assumed to be independent of each other, even within the same subject.

Linear mixed-effects model fit by REML

Data: dat

Log-restricted-likelihood: -2047.714

Fixed: BP ~ method

(Intercept)      methodS

127.40784      15.61961

Random effects:

Formula: ~1 | subject

(Intercept)

StdDev:      28.28452

Formula: ~1 | method %in% subject

(Intercept) Residual

StdDev:      12.61562 7.763666

Number of Observations: 510

Number of Groups:

subject method %in% subject

85

170

### 7.16.5 Model Fit 3

This model is a more general model, compared to 'model fit 2'. This model treats the random interactions for each subject as a vector and allows the variance-covariance matrix for that vector to be estimated from the set of all positive-definite matrices.  $\mathbf{y}_i$  is the entire response vector for the  $i$ th subject.  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are the fixed- and random-effects design matrices respectively.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, 85$$

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda})$$

For the first subject the response vector,  $\mathbf{y}_1$ , is:

observation	BP	subject	method	replicate
1	100.00	1	J	1
86	106.00	1	J	2
171	107.00	1	J	3
511	122.00	1	S	1
596	128.00	1	S	2
681	124.00	1	S	3

The fixed effects design matrix  $\mathbf{X}_i$  is given by:

(Intercept)	method S
1	0
1	0
1	0
1	1
1	1
1	1

The random effects design matrix  $\mathbf{Z}_i$  is given by:

method J	method S
1	0
1	0
1	0
0	1
0	1
0	1

## 7.17 Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, due to they're being intuitive and easy to use. Necessarily their prevalence in literature has meant that they are now the best known measurement for agreement, and that 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  $\lambda_m^2$ . 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  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

$$\mathbf{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda_A^2 & 0 \\ 0 & \lambda_B^2 \end{pmatrix}. \quad (7.30)$$

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (7.31)$$

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.32)$$

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  $\mathbf{D}$  and  $\mathbf{\Lambda}$ . Using Roy's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (7.33)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (7.34)$$

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);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

### 7.17.1 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 additional 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 both models (denoted 1 and 2 respectively);

$$\begin{aligned} \hat{\boldsymbol{\Lambda}}_1 &= \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \\ \hat{\boldsymbol{\Lambda}}_2 &= \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix} \end{aligned}$$

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

The Akaike information criterion (AIC) for both of models are  $AIC_1 = 2304.226$  and  $AIC_2 = 2306.226$ . Having a difference of AIC values of 2 is equivalent to both models being equally as good as the other. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively.

The  $\hat{\mathbf{A}}$  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 2).

To conclude, Carstensen’s models provided a rigorous way to determine limits of agreement, but don’t provide for the computation of  $\hat{\mathbf{D}}$  and  $\hat{\mathbf{\Lambda}}$ . Therefore the test’s proposed by Roy (2009a) can not be implemented. Conversely, accurate limits of agreement can 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.

(N.B. To complement the blood pressure ‘J vs S’ analysis, the limits of agreement are  $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$ .)

## 7.18 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 ?, pg. 158.

Similarly a variety of structures for the within-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.

## 7.19 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 8

## Influence Diagnostics

### Residual Analysis for LME, Applications to MCS Data

This short section will look at residual analysis for LME models. The underlying assumptions for LME models are similar to those of classical linear models. There are two key techniques: a residual plot and the normal probability plot. Using the nlme package it is possible to create plots specific to each method. This is useful in determine which methods ‘disagree’ with the rest. Analysis of the residuals would determine if the methods of measurement disagree systematically, or whether or not erroneous measurements associated with a subset of the cases are the cause of disagreement. Erroneous measurements are incorrect measurements that indicate disagreement between methods that would otherwise be in agreement.

### 8.1 Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. DFBETA and DFFITS are well known measures of influence. The measure DFBETA is the studentized value of this difference. DFFITS is a statistical measure designed to show how influential

an observation is in a statistical model. DFFITS is closely related to the studentized residual.

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \quad (8.1)$$

$$= B(Y - Y_{\bar{a}}) \quad (8.2)$$

$$DFFITS = \frac{\widehat{y}_i - \widehat{y}_{i(k)}}{s_{(k)}\sqrt{h_{ii}}} \quad (8.3)$$

The prediction residual sum of squares (PRESS) is a value associated with this calculation. When fitting linear models, PRESS can be used as a criterion for model selection, with smaller values indicating better model fits.

$$PRESS = \sum (y - y^{(k)})^2$$



The dfbeta refers to how much a parameter estimate changes if the observation or case in question is dropped from the data set. Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

### 8.1.1 DFBETA

DFBETAS (standardized difference of the beta) is a measure that standardizes the absolute difference in parameter estimates between a (mixed effects) regression model based on a full set of data, and a model from which a (potentially influential) subset of data is removed. `dfbeta()`

The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the  $i$ th observation:

where  $\hat{\beta}_i$  is the  $i$ th element of  $\hat{\beta}$ . In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch (1980) recommend 2 as a general cutoff value to indicate influential observations and  $2/\sqrt{k}$  as a size-adjusted cutoff.

$$DFBETA_{a_i} = \frac{\hat{\beta}_i - \hat{\beta}_{(i)}}{\sqrt{\text{var}(\hat{\beta}_i)}} \quad (8.4)$$

$$= \frac{B(Y_i - \hat{Y}_i)}{\sqrt{\text{var}(\hat{\beta}_i)}} \quad (8.5)$$

In the case of method comparison studies, there are two covariates, and one can construct scatterplots of the pairs of dfbeta values accordingly, both for LOO and LSO calculations. Furthermore 95% confidence ellipse can be constructed around these scatterplots. Note that with  $k$  covariates, there will be  $k + 1$  dfbetas (the intercept,  $\beta_0$ , and 1  $\beta$  for each covariate). For example there would be 2 sets of dfbeta, 510 values for each in the case of LOO, and 85 for LSO diagnostics.

### 8.1.2 DFFITS

DFFITS is a statistical measure designed to show how influential an observation is in a statistical model.

$$DFFITS = \frac{\widehat{y}_i - \widehat{y}_{i(k)}}{s_{(k)}\sqrt{h_{ii}}}$$

It is closely related to the studentized residual. For the sake of brevity, we will concentrate on the Studentized Residuals.

## 8.2 DFBETAs

DFBETAS (standardized difference of the beta) is a measure that standardizes the absolute difference in parameter estimates between a (mixed effects) regression model based on a full set of data, and a model from which a (potentially influential) subset of data is removed. A value for DFBETAS is calculated for each parameter in the model separately. This function computes the DFBETAS based on the information returned by the `estex()` function.

## 8.3 DFBETAs

The measure that measures how much impact each observation has on a particular predictor is DFBETAs. The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

DFBETA is a measure found for each observation in a dataset. The DFBETA for a particular observation is the difference between the regression coefficient for an included variable calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

The cut-off value for DFBETAs is  $\frac{2}{\sqrt{n}}$ , where  $n$  is the number of observations. However, another cut-off is to look for observations with a value greater than 1.00. Here cutoff means, this observation could be overly influential on the estimated coefficient.

## DFFITS

DFFITS is a diagnostic meant to show how influential a point is in a statistical regression. It was proposed in 1980. It is defined as the change ("DFFIT"), in the predicted value for a point, obtained when that point is left out of the regression, "Studentized" by dividing by the estimated standard deviation of the fit at that point:

$$\text{DFFITS} = \frac{\hat{y}_i - \widehat{y_{i(i)}}}{s_{(i)}\sqrt{h_{ii}}}$$

## DFbetas for Blood Data

```
plot(JS.roy1.dfbeta$all.res1[1:255],JS.roy1.dfbeta$all.res2[256:510],
     pch=16,col="blue")
abline(v=JS.roy1.dfbeta$all.res1[256],col="red")
abline(h=JS.roy1.dfbeta$all.res2[1],col="red")
```

## 8.4 Case Deletion Diagnostics

? develops case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

## 8.5 Deletion Diagnostics

Since the pioneering work of Cook in 1977, deletion measures have been applied to many statistical models for identifying influential observations.

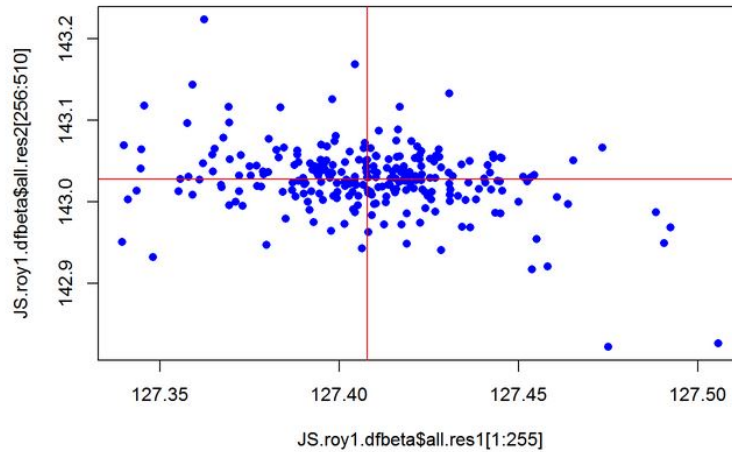


Figure 8.3.1:

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models.

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of  $\beta$  and  $\sigma^2$ , which exclude the  $i$ th observation, can be computed without re-fitting the model. Such update formulas are available in the mixed model only if you assume that the covariance parameters are not affected by the removal of the observation in question. This is rarely a reasonable assumption.

### 3. Case Deletion Diagnostics for LME Data: Cooks Distance, DFBetas

In this section we introduce influence analysis and case deletion diagnostics. A full overview of the topic will be provided although there are specific tools that are particularly useful in the case of MCS problems: specifically the Cook's Distance and the DFBeta.

A discussion of how leave-k-out diagnostics would work in the context of MCS problems is required. There are several scenarios. Suppose we have two methods of measurement X and Y, each with three measurements for a specific case:  $(x_1, x_2, x_3, y_1, y_2, y_3)$

- Leave One Out - one observation is omitted (e.g.  $x_1$ )
- Leave Pair Out - one pair of observation is omitted (e.g.  $x_1$  and  $y_1$ )
- Leave Case (or Subject) Out - All observations associated with a particular case or subject are omitted. (e.g.  $\{x_1, x_2, x_3, y_1, y_2, y_3\}$ )

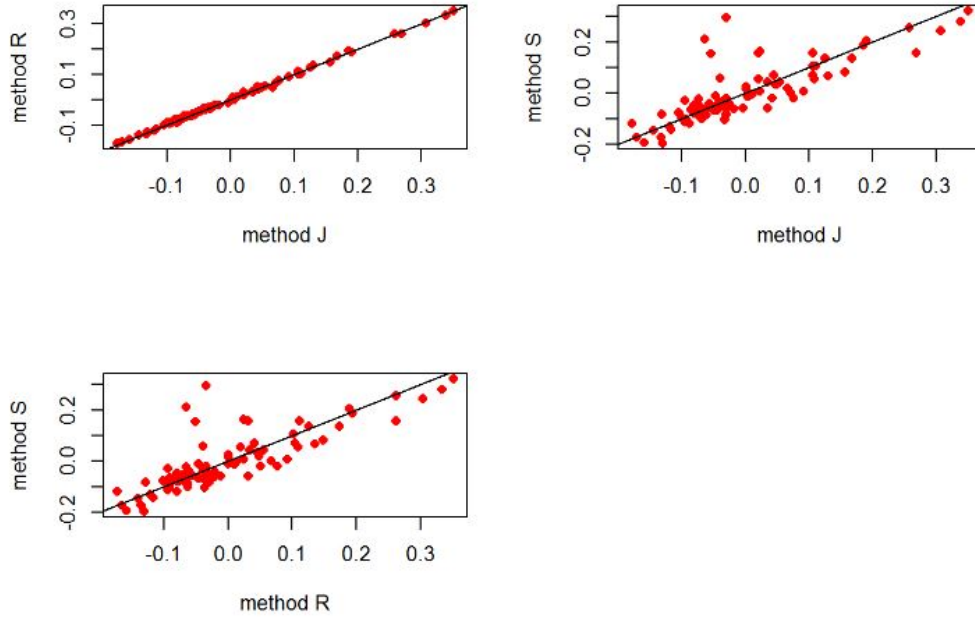
Other metrics, such as the likelihood distance, will also be introduced, and revisited in a later section.

#### 4. Using DFBETAs to Assess Agreement

Suppose an LME model was formulated to model agreement for various (i.e. 2 or more) methods of measurement, with replicate measurements. If the methods are to be agreement, the DFBetas for each case would be the same for both methods. **As such, agreement between any two methods can be determined by a simple scatterplot of the DFBetas. If the points align along the line of equality, then both methods can be said to be in agreement.**

For the model fitted to the blood data with the lme4 R package, the results tabulated below can be produced. All 85 subjects are ranked by Cook's Distance (with only the top 6 being presented here). The remaining columns are the DFBeta for each of the fixed effects, for each of the 85 subject.

Subject	Cook's D	methodJ	methodR	methodS
78	0.61557407	-0.02934556	-0.03387780	0.2954937
80	0.41590973	-0.06305026	-0.06515241	0.2123881
68	0.22536651	-0.05334867	-0.05062375	0.1555187
72	0.09348500	0.02388626	0.02419887	0.1617474
48	0.08706988	0.02147541	0.03145273	0.1581591
30	0.07118415	0.26925807	0.26215970	0.1581569



In the first of the three plots (*Top Right*), strong agreement between method J and method R is indicated. The other plots indicate lack of agreement of methods J and R with method S.

If lack of agreement is indicated, a subsequent analysis using a technique proposed by Roy(2009) can be used to identify the specific cause for this lack of agreement (see next section).

The Pearson Correlation coefficient of the DFBetas can be used in conjunction with this analysis. A high correlation confirms good agreement. No threshold value for agreement is suggested, and analysts are advised to perform model diagnostics regardless of the correlation coefficient.

The Bonferroni Outlier Test and Cook's Distance values can be used to identify unusual cases, when the relationship between sets of dfbeta is modelled as a (classical) linear model. In this model, the covariates should be homoskedastic. A test for non-constant variance may be used to verify this. These diagnostic procedures are implementable using the *car* R package.

Deming Regression can be used to verify the line of equality. Significance test for Deming regression estimates are not available, but 95% bootstrap confidence intervals for the slope estimate and intercept estimates can be computed.

Additionally a mean difference plot can be used to identify outliers. This mean-difference plot differs from the Bland-Altman plot in that the plot is denominated in terms of dfbeta values, and not in measurement units.

If lack of agreement is indicated between methods of measurement, use of Roy's Testing is advised (This is the subject of the next section).



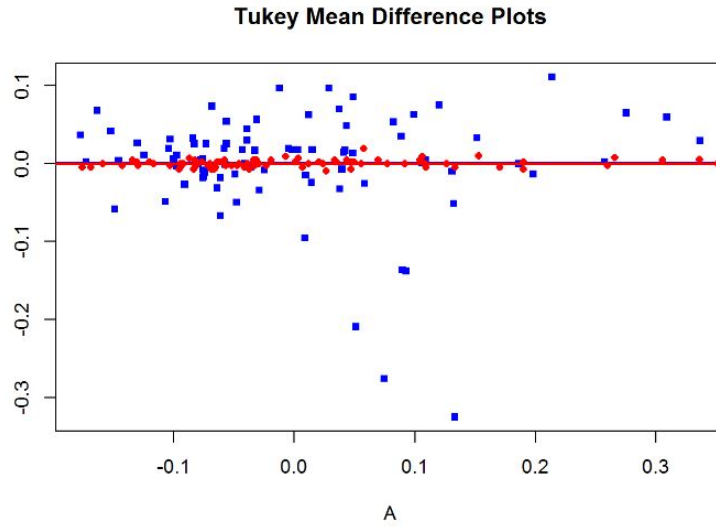


Figure 8.5.2:

## 5. Using Roy's Test to Identify cause of Lack of agreement

Barnhart specifies three conditions for method of measurement that are required for two methods of measurement to be considered in agreement.

- (i) No Significant Inter-method bias
- (ii) No significant Difference in Within-Subject Variance
- (iii) No significant Difference in Within-Subject Variance

Roy(2009) demonstrates a LME model specification, and a series of tests that look at each of these agreement criteria individually. If two methods of measurement lack agreement, the specific reason or reasons for this lack of agreement can be identified.

Roy proposes an LME model with Kronecker product covariance structure in a doubly multivariate setup. Response for  $i$ th subject can be written as

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

- $\beta_1$  and  $\beta_2$  are fixed effects corresponding to both methods. ( $\beta_0$  is the intercept.)

- $b_{1i}$  and  $b_{2i}$  are random effects corresponding to both methods.

Overall variability between the two methods ( $\Omega$ ) is sum of between-subject ( $D$ ) and within-subject variability ( $\Sigma$ ),

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

## 6. Using Roy's Model to Compute LoAs and CR

In this short section, a demonstration of how Roy's technique can be used to compute two common MCS metrics: Limits of Agreement and the Coefficient of Repeatability. While Limits of Agreement are not used in the analysis proposed here, they are ubiquitous in literature, and a demonstration on how to compute them with the Roy Model would assist the adoption of this proposed method.

The coefficient of repeatability is encountered in Gage R & R analysis. *(A future exploration of how LME models can be used in that field would be of interest. This is something to include in the Conclusions Section).*

## 7. Model Diagnostics for Roy's Models

Further to previous work, this section revisits case-deletion and residual diagnostics, and explores how approaches devised by Galecki & Burzykowski (2013) can be used to appraise Roy's model. These authors specifically look at Cook's Distances and Likelihood Distances. For the Roy Model, Cook's Distances may also be generated using the ***predictmeans***

As the model is structurally different from the models discussed in the earlier sections, Residual analysis will be briefly revisited.

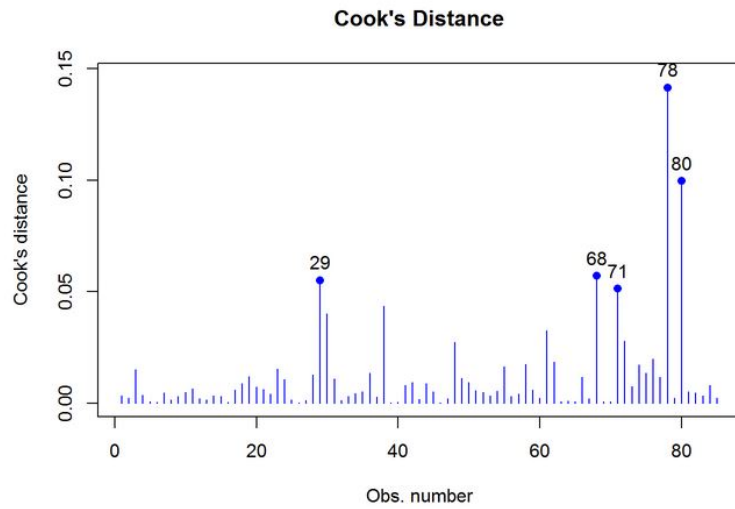


Figure 8.5.3:

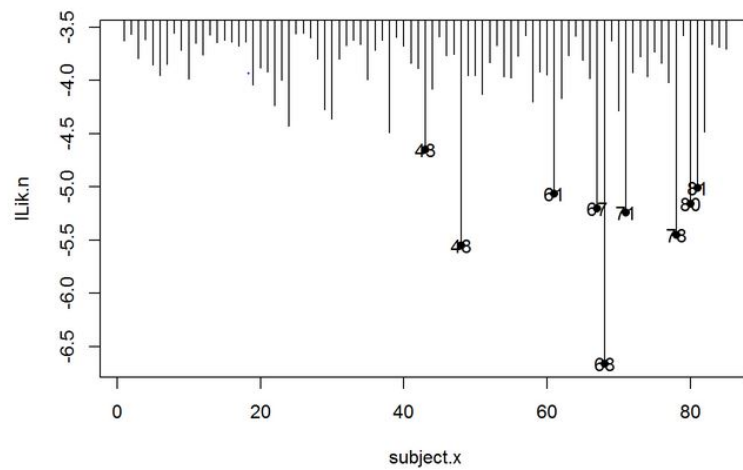


Figure 8.5.4:

## 8. Case Deletion Diagnostics for the Variance Ratios

Schabenberger advises on the use of deletion diagnostics for variance components of an LME model. Taking the core principals of his methods, and applying them to the Method Comparison problem, case deletion diagnostics are used on the variance components of the Roy model., specifically the ratio of between subject variances and the within subject covariances respectively.

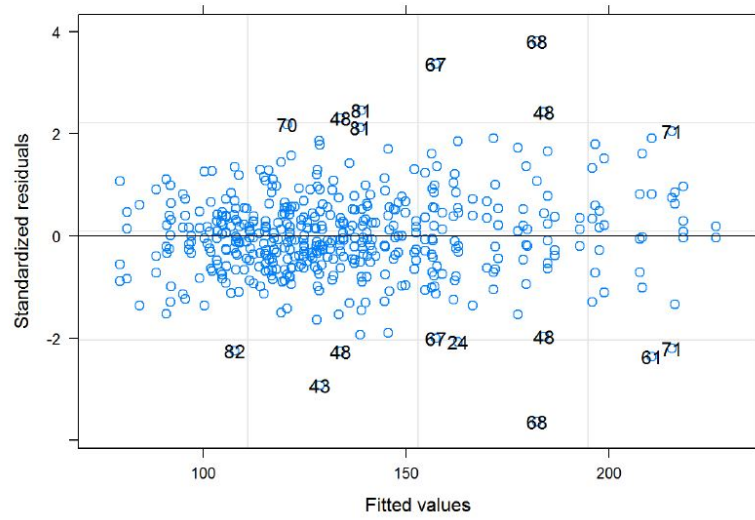


Figure 8.5.5:

$$\text{BSVR} = \frac{\sigma_2^2}{\sigma_2^2} \quad \text{WSVR} = \frac{d_2^2}{d_2^2}$$

These variance ratios are re-computed for each case removed, and may be analysed separately or jointly for outliers.

The Grubbs' Test for Outliers is a commonly used technique for assessing outlier in a univariate data set. As there may be several outliers (i.e. influential cases) present, the Grubbs test is not practical. However outlier detection using to Tukey's specification for boxplots (i.e. greater than  $Q_3 + 1.5IQR$  or less than  $Q_1 - 1.5IQR$ ), will suffice. Ranking the absolute values of the standardized scores can also be used to identify influential cases, even if the data is not normally distributed.

Bivariate Analyses may be applied jointly to the both sets of data sets, e.g Mahalanobis distances. The Mahalanobis distance, while not an intuitive measure in the context of the data, can be used to rank highly influential cases.

## 9. Permutation Test, Power Tests and Missing Data

This section explores topics such as dependent variable simulation and power analysis, introduced by Galecki & Burzykowski (2013), and implementable with their *nlmeU* R package. Using the *predictmeans* R package, it is possible to perform permutation t-tests for coefficients of (fixed) effects and permutation F-tests.

The matter of missing data has not been commonly encountered in either Method Comparison Studies or Linear Mixed Effects Modelling. However Roy (2009) deals with the relevant assumptions regarding missing data. Galecki & Burzykowski (2013) approaches the subject of missing data in LME Modelling. The *nlmeU* package includes the `patMiss` function, which “*allows to compactly present pattern of missing data in a given vector/matrix/data frame or combination of thereof*”.

### 8.5.1 Influential Observations : DFBeta and DFBetas

## 8.6 Measures of Influence

The impact of an observation, or a case with multiple observations, on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

## 8.7 Overall Influence

An overall influence statistic measures the change in the objective function being minimized. For example, in OLS regression, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance [Cook and Weisberg ].

## 8.8 Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \quad (8.6)$$

## 8.9 Case Deletion Diagnostics for Mixed Models

? notes the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect.

? develops these techniques in the context of REML

## 8.10 Terminology for Case Deletion diagnostics

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called 'observation-diagnostics'. For multiple observations, Preisser describes the diagnostics as 'cluster-deletion' diagnostics.

## 8.11 Case Deletion Diagnostics

Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of  $\beta$  and  $\sigma^2$ , which exclude the  $i$ th observation, can be computed without re-fitting the model. Such update formulas are available in the mixed model only if you assume that the covariance parameters are not affected by the removal of the observation in question. This is rarely a reasonable assumption.

Linear models for uncorrelated data have well established measures to gauge the influence of one or more observations on the analysis. For such models, closed-form update expressions allow efficient computations without refitting the model.

Since the pioneering work of Cook in 1977, deletion measures have been applied to

many statistical models for identifying influential observations. Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of  $\beta$  and  $\sigma^2$ , which exclude the  $i$ -th observation, can be computed without re-fitting the model.

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called ‘*observation-diagnostics*’. For multiple observations, Preisser describes the diagnostics as ‘*cluster-deletion*’ diagnostics. When applied to LME models, such update formulas are available only if one assumes that the covariance parameters are not affected by the removal of the observation in question. However, this is rarely a reasonable assumption.

## 8.12 Case Deletion Diagnostics

Since the pioneering work of Cook in 1977, deletion measures have been applied to many statistical models for identifying influential observations.

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models.

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models. Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of  $\beta$  and  $\sigma^2$ , which exclude the  $i$ th observation, can be computed without re-fitting the model. Such update formulas are available in the mixed model only if you assume that the covariance parameters are not affected by the removal of the observation in question. This is rarely a reasonable assumption.

## 8.13 Terminology for Case Deletion diagnostics

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called 'observation-diagnostics'. For multiple observations, Preisser describes the diagnostics as 'cluster-deletion' diagnostics.

## 8.14 `influence.ME`

*influence.ME* allows you to compute measures of influential data for mixed effects models generated by `lme4`.



*influence.ME* provides a collection of tools for detecting influential cases in generalized mixed effects models. It analyses models that were estimated using lme4. The basic rationale behind identifying influential data is that when iteratively single units are omitted from the data, models based on these data should not produce substantially different estimates.

To standardize the assessment of how influential a (single group of) observation(s) is, several measures of influence are common practice, such as DFBETAS and Cook's Distance. In addition, we provide a measure of percentage change of the fixed point estimates and a simple procedure to detect changing levels of significance.

## 8.15 `Influence()` command

`influence()` is the workhorse function of the `influence.ME` package. Based on a priorly estimated mixed effects regression model (estimated using `lme4`), the `influence()` function iteratively modifies the mixed effects model to neutralize the effect a grouped set of data has on the parameters, and which returns the fixed parameters of these iteratively modified models. These are used to compute measures of influential data.

## 8.16 Cook's Distance

Cook's Distance is a measure indicating to what extent model parameters are influenced by (a set of) influential data on which the model is based. This function computes the Cook's distance based on the information returned by the `estex()` function.

# Chapter 9

## Appendices 1

### 9.1 LME - Pankaj Choudhury

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

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2, \dots, n) \quad (9.1)$$

The design matrix  $P_{ij}$ , with its associated column vector  $\theta$ , specifies the fixed effects common to both methods. The fixed effect specific to the  $j$ th 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  $i$ th subject by the  $j$ th 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 include a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \quad (9.2)$$

These vectors are assumed to be independent for different  $i$ s, 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;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \quad (9.3)$$

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.

## 9.2 Model Terms (Roy 2009)

- Let  $y_{mir}$  be the response of method  $m$  on the  $i$ th subject at the  $r$ —th replicate.
- Let  $\mathbf{y}_{ir}$  be the  $2 \times 1$  vector of measurements corresponding to the  $i$ —th subject at the  $r$ —th replicate.
- Let  $\mathbf{y}_i$  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  $\alpha_{mi}$  be the fixed effect parameter for method for subject  $i$ .
- Formally Roy uses a separate fixed effect parameter to describe the true value  $\mu_i$ , 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 and corresponding t-value and p-values are presented. This is relevant to Roy's first test.

## 9.3 Algorithms

Maximum likelihood estimation is a method of obtaining estimates of unknown parameters by optimizing a likelihood function. The ML parameter estimates are the values of the argument that maximise the likelihood function, i.e. the estimates that make the observed values of the dependent variable most likely, given the distributional assumptions

The most common iterative algorithms used for the optimization problem in the context of LMEs are the EM algorithm, fisher scoring algorithm and NR algorithm, which [cite:West] commends as the preferred method.

A mixed model is an extension of the general linear models that can specify additional random effects terms.

Parameter of the mixed model can be estimated using either ML or REML, while the AIC and the BIC can be used as measures of "goodness of fit" for particular models, where smaller values are considered preferable.

## 9.4 ML v REML

(***Wikipedia***)The restricted (or residual, or reduced) maximum likelihood (REML) approach is a particular form of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data, so that nuisance parameters have no effect.

In contrast to the earlier maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters.

## 9.5 ML procedures for LME

The maximum likelihood procedure of Hartley and Rao yields simultaneous estimates for both the fixed effects and the random effect, by maximising the likelihood of  $\mathbf{y}$  with respect to each element of  $\boldsymbol{\beta}$  and  $\mathbf{b}$ .

## 9.6 Estimation of random effects

Estimation of random effects for LME models in the NLME package is accomplished through use of both EM (Expectation-Maximization) algorithms and Newton-Raphson algorithms.

- EM iterations bring estimates of the parameters into the region of the optimum very quickly, but convergence to the optimum is slow when near the optimum.
- Newton-Raphson iterations are computationally intensive and can be unstable when far from the optimum. However, close to the optimum they converge quickly.
- The LME function implements a hybrid approach, using 25 EM iterations to quickly get near the optimum, then switching to Newton-Raphson iterations to quickly converge to the optimum.
- If convergence problems occur, the “control argument in LME can be used to change the way the model arrives at the optimum.



# 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.
- Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics* 5(3), 399–413.
- Carstensen, B. (2010). *Comparing Clinical Measurement Methods : A Practical Guide*. Wiley.
- 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.

- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal 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.
- 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.
- Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall Ltd.
- 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.
- Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics* 9(2), 226–252.

- Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1959). 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: American 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. and J. A. Nelder (1996). Hierarchical generalized linear models (Disc: P656-678). *Journal of the Royal Statistical Society, Series B: Methodological* 58, 619–656.
- Lee, Y., J. A. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall Ltd.
- Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3), 545–554.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.
- Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.
- Preisser, J. S. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* 83(3), 551–556.
- 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.
- 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.