

# Chapter 1

**BXC2010**

# Chapter 2

## BXC

### 2.1 Carstensen

Let  $y_{mir}$  be 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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (2.1)$$

Here  $\beta_0$  and  $\beta_m$  are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ . The  $\beta$  terms can be gathered 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  $\text{Cov}(b_{mi}, b_{m'i}) = g_{12}$ . 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  $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$ . Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \beta_1 = \beta_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  hold simultaneously. Roy (2009a) proposes 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. Let  $\omega_m^2 = \sigma_m^2 + g_m^2$  represent the overall variability of method  $m$ . Roy also integrates  $H_2$  and  $H_3$  into a single testable hypothesis  $H_4: \omega_1^2 = \omega_2^2$ . CONCERNS?

? demonstrates how to implement a method comparison study further to model (1) using the SAS proc mixed package. Carstensen et al. (2008) demonstrates how to construct limits of agreement using SAS, STATA and R. In the case of SAS, the PROC MIXED procedure is used. Implementation in R is performed using the nlme package (?).

Carstensen et al. (2008) remarks that the implementation using R is quite “arcane”.

As R is freely available, this paper demonstrates an implementation of Roy’s model using R.

The R statistical software package is freely available.

The LME model is very easy to implement using PROC MIXED of SAS and the results are also easy to interpret. The SAS proc mixed procedure has very simple syntax.

As the required code to fit the models is complex, R code necessary to fit the models is provided.

A demonstration is provided on how to use the output to perform the tests, and to compute limits of agreement.

We assume the data are formatted as a dataset with four columns named:

meth, method of measurement, the number of methods being M, item, items (persons, samples) measured by each method, of which there are I, repl, replicate indicating repeated measurement of the same item by the same method, and y, the measurement.

### 2.1.1 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)

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

## 2.2 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.

## 2.3 2004 Model

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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (2.2)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.3)$$

## 2.4 Carstensen's Model

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.

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.

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.

Bendix Carstensen et al. proposed the use of LME models to allow for a more statistically rigorous approach to computing Limits of Agreement. The respective papers also discuss several shortcomings for techniques for dealing with replicate measurements, as proposed by Bland-Altman 1999.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (2.4)$$

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  $m$ ;

$$BLUP_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + c_{mi} \quad (2.5)$$

. 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 (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.

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.

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (2.6)$$

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  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$ . Additionally these parameter are assumed to have Gaussian distribution. 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. Additionally, Roy combines  $H_2$  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  $m$ .

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 can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \quad (2.7)$$

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  $\epsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$ . 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.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (2.6) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ , whereas the model in (2.7) requires  $N + 2$  fixed effects.

Allocating fixed effects to each item  $i$  by (2.7) accords with earlier work on comparing methods of measurement, such as ?. 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 (2009b) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

## 2.5 Using Interaction Terms

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.

## 2.6 Computing LoAs with LMEs

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

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

## 2.7 Carstensen's Model

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  $m$ .

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.8)$$

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 mobservations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

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  $m = 1, 2, \dots, M$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$  is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + a_{ir} + \epsilon_{mir}, \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2), a_{ir} \sim \mathcal{N}(0, \varsigma^2), \epsilon_{mi} \sim \mathcal{N}(0, \varphi_m^2). \quad (2.9)$$

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  $\text{Var}(c_{mi}) = \tau_m^2$ .

The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \varphi_m^2$ . All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

When only two methods are to be compared, separate estimates of  $\tau_m^2$  can not be obtained. Instead the average value  $\tau^2$  is obtained and used.

Carstensen's approach is that of a standard two-way mixed effects ANOVA with replicate measurements. With regards to the specification of 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, 2010).

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:  $\tau_1^2 = \tau_2^2 = \tau^2$ . 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.

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

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

The presence of the true value term  $\mu_i$  gives rise to an important difference between Carstensen’s and ARoy2009’s models. The fixed effect of ARoy2009’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: ARoy’s model is a standard LME model, whereas Carstensen’s model is a more complex additive model.

## 2.8 Carstensen’s Mixed Models

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.10)$$

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:

$$var(y_{1j} - y_{2j}) \tag{2.11}$$

### 2.8.1 Carstensen Methods

Components

Section 5.3 Models for replicate measurements

Section 5 Replicate measurements.

Carstensen page 56

%-----%

air extra random effect that does not depend on method.

It is treated as an extension of i.

The variance of air represents the variation between replication condition (common

$$ymir = m + i + cmi + emir$$

$$cmi = N(0, m2)$$

$$emir = N(0, m2)$$

Carstensen page 58

$$\text{var}(y_{10}-y_{20}) = 12+22+12+22$$

$$1-2222+12+22$$

ARoy2009 further to Carstensen

$$y_{mir}=m+i+c_{mi}+e_{mir}$$

Section 7 A general model for method comparisons.

Carstensen discusses the model and its use as if all parameter estimates are available.

In this model, intermethod bias is assumed to be constant at all measurement levels.

$i$  : True value for item  $i$

The parameter  $i$  can be thought of as the underlying, but unobtainable, true measurement for item  $i$ .

$m$ : Fixed effect for method  $m$

### **Carstensen et al - Mixed Models**

Carstensen et al [4] 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (2.12)$$

Carstensen *et al* 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.

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.13)$$

Carstensen *et al* 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (2.14)$$

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.



$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (2.15)$$

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  $m$ .

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.16)$$

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (2.17)$$

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

The intercept term  $\alpha$  and the  $\beta_m \mu_i$  term follow from *Dunn* 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).

*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.*

### 2.8.2 Tau Identifiability

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

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

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:

$$var(y_{1j} - y_{2j}) \tag{2.18}$$

### 2.8.3 Computation

Modern software packages can be used to fit models accordingly. The best linear unbiased predictor (BLUP) for a specific subject  $i$  measured with method  $m$  has the form  $BLUP_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + c_{mi}$ , under the assumption that the  $\mu$ s are the true item values.

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

## 2.8.4 Carstensen’s Mixed Models

Carstensen *et al*[4] 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.

Carstensen *et al*[4] 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  $m$ .

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (2.19)$$

This model includes a method by item interaction term.

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 ARoy2009’s LoAs are lower than those of Carstensen.

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

### 2.8.5 Computing LoAs from LME models

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

Carstensen et al. (2008) formulates an LME model, both in the absence and the presence of an interaction term. Carstensen et al. (2008) 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.

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;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (2.20)$$

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.

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, 2010).

The presence of the true value term  $\mu_i$  gives rise to an important difference between

Carstensen’s and ARoy2009’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.

## 2.9 Carstensen 2004’s Mixed Models

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

$$BLUP_{mir} = \alpha_m + \hat{\beta}_m \mu_i + c_{mi} \quad (2.21)$$

. 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 ( $d_{mr}d_{mr} \sim N(0, \omega_m^2)$ ) 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.

# Chapter 3

## BXC Limits of Agreement

### 3.1 Carstensen Model (mir model)

A measurement  $y_{mi}$  by method  $m$  on individual  $i$  is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (3.1)$$

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.

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  $m$ .

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;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (3.2)$$

Let  $y_{mir}$  be 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 can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (3.3)$$

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  $\text{Var}(b_{mi}) = d_m^2$  and  $\text{Cov}(b_{1i}, b_{2i}) = d_{12}$ .

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  $\beta_2$ ,  $d_1^2$  and  $d_2^2$ , and  $\sigma_1^2$  and  $\sigma_2^2$ .

Additionally these parameter are assumed to have Gaussian distribution. 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  $H_3: d_1^2 = d_2^2$  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. Additionally, Roy combines  $H_2$  and  $H_3$  into a single testable hypothesis  $H_4: \omega_1^2 = \omega_2^2$ , where  $\omega_m^2 = \sigma_m^2 + d_m^2$  represent the overall variability of method  $m$ .

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 nor-

mally 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.

The presence of the true value term  $\mu_i$  gives rise to an important difference between Carstensen's and Roy's models. 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. 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.

## 3.2 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  $m$ .

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (3.4)$$

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  $m$ ;

$$BLUP_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + c_{mi} \quad (3.5)$$

. 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.

### 3.2.1 Using LME models to create Prediction Intervals

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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (3.6)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (3.7)$$

## 3.3 Linked replicates

Carstensen et al. (2008) proposes the addition of a 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.

### 3.4 Bendix Carstensen’s data sets

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.

Carstensen 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 simultaneously for all patients.

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

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  $(-12.18, 17.12)$ .

### 3.4.2 Using LME models to create Prediction Intervals

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (3.8)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (3.9)$$

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  $m$ .

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (3.10)$$

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (3.11)$$

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

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).

### 3.4.3 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.

BXC2008 formulates an LME model, both in the absence and the presence of an interaction term. BXC2008 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.

## 3.5 The Fat Data Set

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (3.12)$$

All of these terms are given or determinable in computer output. The limits of

agreement can therefore be evaluated using

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (3.13)$$

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 ARoy2009's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (3.14)$$

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

## 3.6 Oxymetry 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 replicates 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);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \quad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \quad (3.15)$$

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, indi-

cating 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 2).

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 (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.

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



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  $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$ .

### 3.7 Oxymetry Data

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 Children’s 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.

Carstensen et al. (2008) demonstrates the use of the interaction term when com-

puting 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  $(-12.18, 17.12)$ .

Limits of agreement are determined using ARoy2009’s methodology, without adding any additional terms, are found to be consistent with the ‘interaction’ model;  $(-9.562, 14.504)$ .

### 3.8 RV-IV

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

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \quad (3.16)$$

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

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \quad (3.17)$$

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

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \quad (3.18)$$

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.

# Bibliography

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723.
- 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).
- Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.
- 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.