Chapter 1

Limits of Agreement

1.1 Introduction to LME Methods for Computing Limits of Agreement

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. 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.

Further to Bland and Altman (1986), the computation of the limits of agreement follows from the inter-method bias, and the variance of the difference of measurements. When repeated measures data are available, it is desirable to use all the data to compare the two methods. However, the classical Bland-Altman method was developed for two sets of measurements done on one occasion, but is inadequate for replicate measurement data. Bland and Altman (1999) addresses this issue by suggesting several computationally simple approaches. One approach suggested by Bland and Altman (1999) is to calculate 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 too small, because of the reduction of the

effect of repeated measurement error. Bland and Altman (1999) propose a correction for this. 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.

Carstensen et al. (2008) takes issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation.

Carstensen et al. (2008) demonstrates how the limits of agreement calculated solely from the mean of replicates are 'much too narrow as prediction limits for differences between future single measurements'. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the 'mean of replicates' approach. Instead, a linear mixed effects model is recommended for appropriate estimates for the variance of the inter-method bias.

Carstensen et al. (2008) proposes the use of LME models to allow for a more statistically rigourous approach to computing Limits of Agreement. This approach is based upon variance component estimates derived using linear mixed effects models. This approach extends the well established Bland-Altman methodology for the case of replicate measurements on each item. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

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, by computing an appropriate estimate for the standard deviation of case-wise differences, so as to determine the limits of agreement. This approach is sim-

ilar to Deming's regression, and for estimating variance components for measurements by different methods.

Roy (2009) formulates a very powerful method of assessing the agreement of two methods of measurement, with replicate measurements, also using LME models. This approach does not directly address the issue of limits of agreement, but does allow for an alternative approach to computing LoAs using LME Models.

1.2 Limits of Agreement in LME models

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 feature is the differing residual variances between methods" (Carstensen, 2010).

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

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

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m. The differences are expressed as $d_i = y_{1i} - y_{2i}$.

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

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

term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

This formulation doesn't require the data set to be balanced, but does require a sufficient number of replicates and measurements to overcome the problem of identifiability. Consequently 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).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.

Carstensen et al. (2008) presents a simplified, but more tractable, model:

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{1.3}$$

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 μs are the true item values.

1.3 Computation of Limits of agreement in LME models

Carstensen et al. (2008) proposes a technique to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman approach in this regard.Between-subject variation for method m is given by d_m^2 (in the author's notation τ_m^2) and within-subject variation is given by σ_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.

When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of τ_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.

Carstensen et al. (2008) states a model where the variation between items for method m is captured by τ_m (our notation d_m^2) and the within-item variation by σ_m. When only two methods are to be compared, separate estimates of τ_m^2 can not be obtained. Instead the average value τ^2 is used. The between-subject variability $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

$$\boldsymbol{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}.$$

The variance for method m is $d_m^2 + \sigma_m^2$. 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

$$\text{var}(y_A - y_B) = 2d^2 + \sigma_A^2 + \sigma_B^2. \tag{1.4}$$

Importantly the covariance terms in both variability matrices are zero, so no covariance components are present. Further to his model, Carstensen computes the limits of agreement as

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

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

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

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{1.5}$$

1.3.1 Computing Limits of Agreement using Roy's Model

Roy (2009) has demonstrated a method whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$. Using Roy's approach, the variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed. The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\boldsymbol{\Omega}_i$ matrix. The variance of differences is easily computable from the variance estimates in the Block - $\boldsymbol{\Omega}_i$ matrix, i.e.

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

Lack of agreement can arise if there is a disagreement in overall variabilities.

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

1.3.2 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) 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.

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 limts 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.

Roy's approach assumes that replicates are linked. Limits of agreement are determined using Roy's method, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. 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);

$$\hat{\mathbf{\Sigma}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Sigma}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{1.6}$$

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

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Sigma}}$. Therefore the test's proposed by Roy (2009) 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 approach 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 $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.

1.4 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.

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. Therefore the variance covariance matrices for between-item and within-item variability are respectively.

$$\boldsymbol{D} = \begin{pmatrix} d_2^1 & 0 \\ 0 & d_2^2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

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 Carstensen et al. (2008).

A consequence of this is that the between-method and within-method covariance are 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.

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

There is a substantial difference in the number of fixed parameters used by the respective models; the model in Roy (2009) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model using the Carstensen Model requires $N + 2$ fixed effects. Allocating fixed effects to each

item i using Carstensen's model 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 (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

1.5 Carstensen Coefficient of Repeatability

The limits of agreement are not always the only issue of interest, the assessment of method specific repeatability and reproducibility are of interest in their own right. Repeatability can only be assessed when replicate measurements by each method are available.

Under the model for linked replicates, there are two possibilities depending on the circumstances. If the variation between replicates within item can be considered a part of the repeatability it will be $2.8\sqrt{\omega^2 + \sigma_m^2}$.

However, if replicates are taken under substantially different circumstances, the variance component ω^2 may be considered irrelevant in the repeatability and one would therefore base the repeatability on the measurement errors alone, i.e. use $2.8\sigma_m$.

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.

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

Roy, A. (2009). 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.