Contents

Chapter 1

Review of MCS Methodologies

1 Introduction to Method Comparison Studies

- Accuracy and Precision
- Repeatability (Bland Altman 1999)
- Remarks in Barnhart's Paper
- Regression Techniques (i.e. Orthonormal Regression and Deming Regression)
- Bradley and Blackwood's, and Bartko's Techniques
- Coefficient of Repeatability

2 Bland and Altman Plot

- Bland and Altman 1983 and 1986
- Limits of Agreement
- Variants of the Bland-Altman Technique
- Prevalence and Usage of BA's approach
- Discussion of ShinyMCS appendix

3 Introduction to LME Models

- Model Specification of LME Models

1.1 Bland-Altman methodology

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of paired sample $t-$test, correlation coefficients or simple linear regression. Simple linear regression is unsuitable for method comparison studies because of the required assumption that one variable is measured without error. In comparing two methods, both methods are assume to have attendant random error.

Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyzes and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be 'explained to non-statisticians'.

Notwithstanding previous remarks about linear regression, the first step recommended, which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality should also be shown, as it is necessary to give the correct interpretation of how both methods compare. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

Dewitte et al. (2002) notes that scatter plots were very seldom presented in the Annals of Clinical Biochemistry. This apparently results from the fact that the 'Instructions for Authors' dissuade the use of regression analysis, which conventionally is accompanied by a scatter plot.

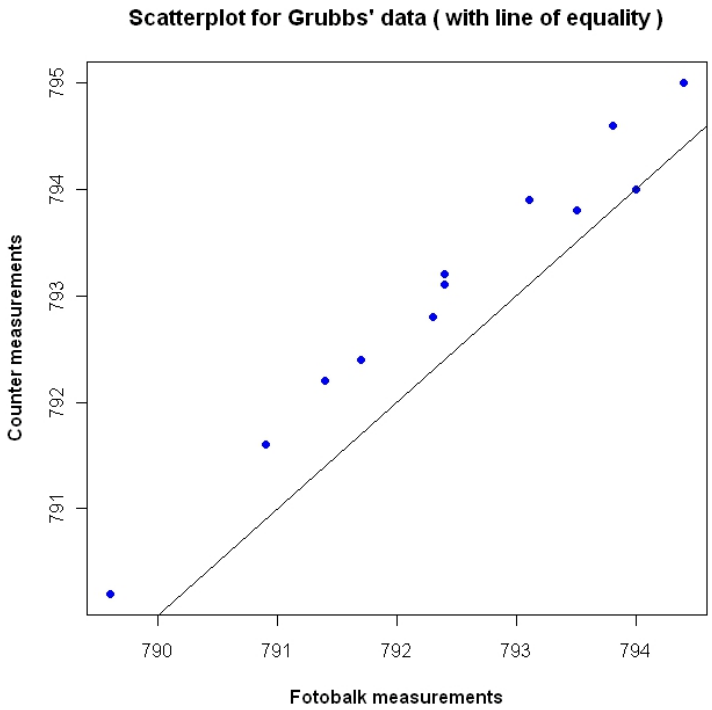

Figure 1.1.1: Scatter plot For Fotobalk and Counter Methods.

1.1.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, \ldots, n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \ldots, n$).

Altman and Bland (1983) proposes a scatterplot of the case-wise averages and differences of two methods of measurement. This scatterplot has since become widely known as the Bland-Altman plot. Altman and Bland (1983) express the motivation for this plot thusly:

> "From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study."

The case wise-averages capture several aspects of the data, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. Bland and Altman (1986) cautions that it would be the difference against either measurement value instead of their average, as the difference relates to both value. This methodology has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical methodology for making a visual assessment of the data.

The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$. This inter-method bias is represented with a line on the Bland-Altman plot. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are also particularly relevant. The variances around this bias is estimated by the standard deviation of these differences S_d.

1.1.2 Bland-Altman plots for the Grubbs data

In the case of the Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

The Bland-Altman plot for comparing the 'Fotobalk' and 'Counter' methods, which shall henceforth be referred to as the 'F vs C' comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.1.1: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.8	793.2	0.6	793.5
2	793.1	793.3	-0.2	793.2
3	792.4	792.6	-0.2	792.5
4	794.0	793.8	0.2	793.9
5	791.4	791.6	-0.2	791.5
6	792.4	791.6	0.8	792.0
7	791.7	791.6	0.1	791.6
8	792.3	792.4	-0.1	792.3
9	789.6	788.5	1.1	789.0
10	794.4	794.7	-0.3	794.5
11	790.9	791.3	-0.4	791.1
12	793.5	793.5	0.0	793.5

Table 1.1.2: Fotobalk and Terma methods: differences and averages.

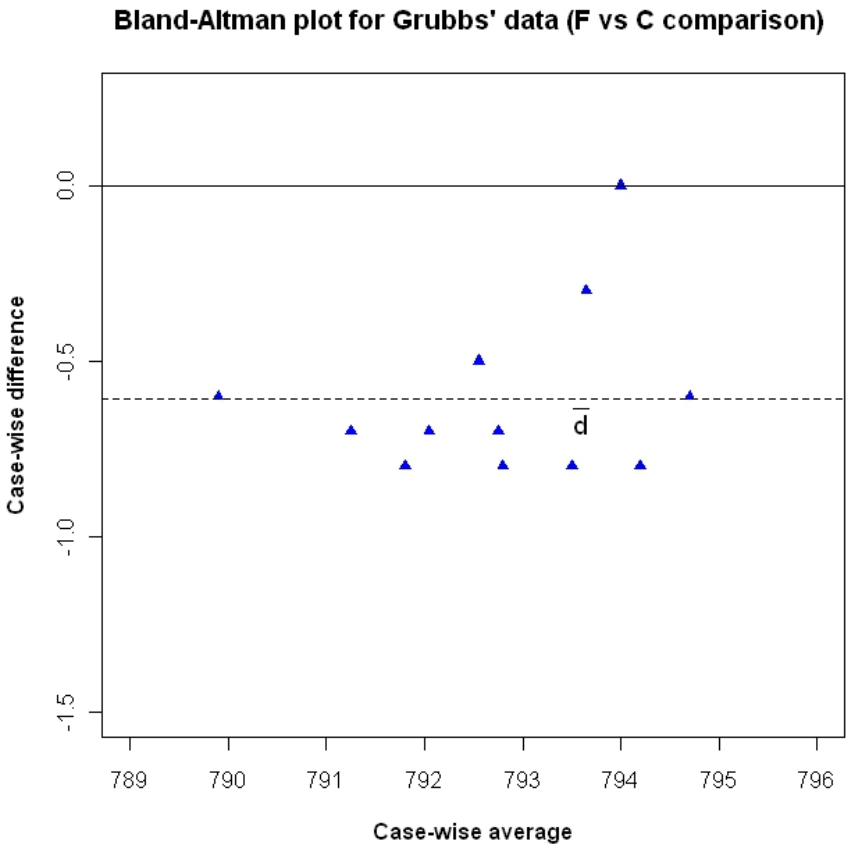

Figure 1.1.2: Bland-Altman plot For Fotobalk and Counter methods.

In Figure 1.3 Bland-Altman plots for the 'F vs C' and 'F vs T' comparisons are shown, where 'F vs T' refers to the comparison of the 'Fotobalk' and 'Terma' methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the 'F vs C' comparison than in the 'F vs T' comparison. Conversely there appears to be less precision in 'F vs T' comparison, as indicated by the greater dispersion of covariates.

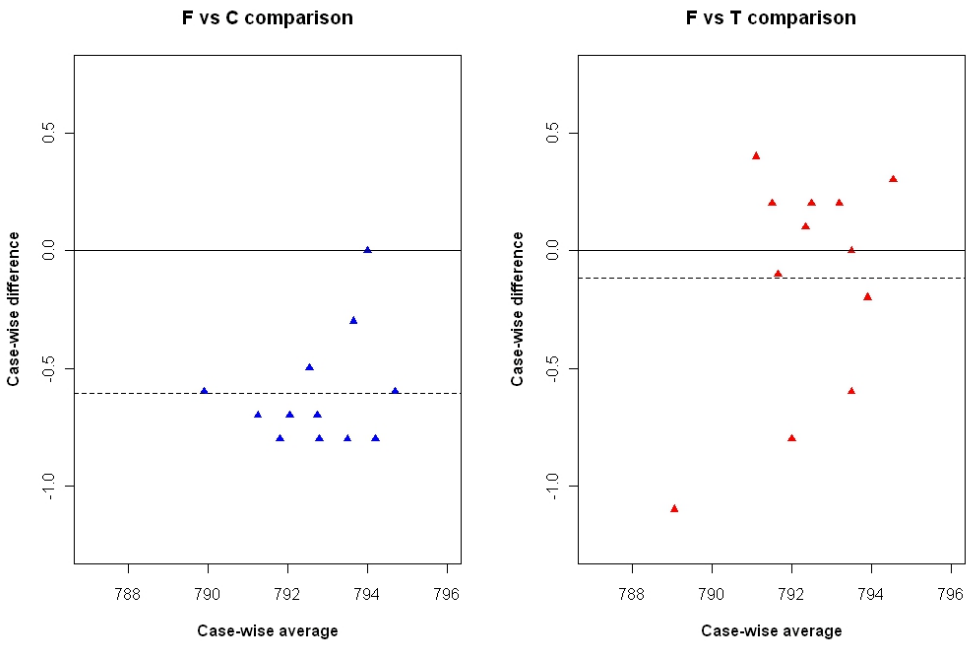

Figure 1.1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

1.1.3 Adverse features

Estimates for inter-method bias and variance of differences are only meaningful if there is uniform inter-bias and variability throughout the range of measurements. Fulfilment of these assumptions can be checked by visual inspection of the plot.The prototype Bland-Altman plots depicted in Figures 1.4, 1.5 and 1.6 are derived from simulated data, for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias, and is defined by Ludbrook (1997) as meaning that 'one method gives values that are higher (or lower) than those from the other by an amount that is proportional to the level of the measured variable'. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for

the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (Benjamini and Hochberg, 1995) test, should be also be used.

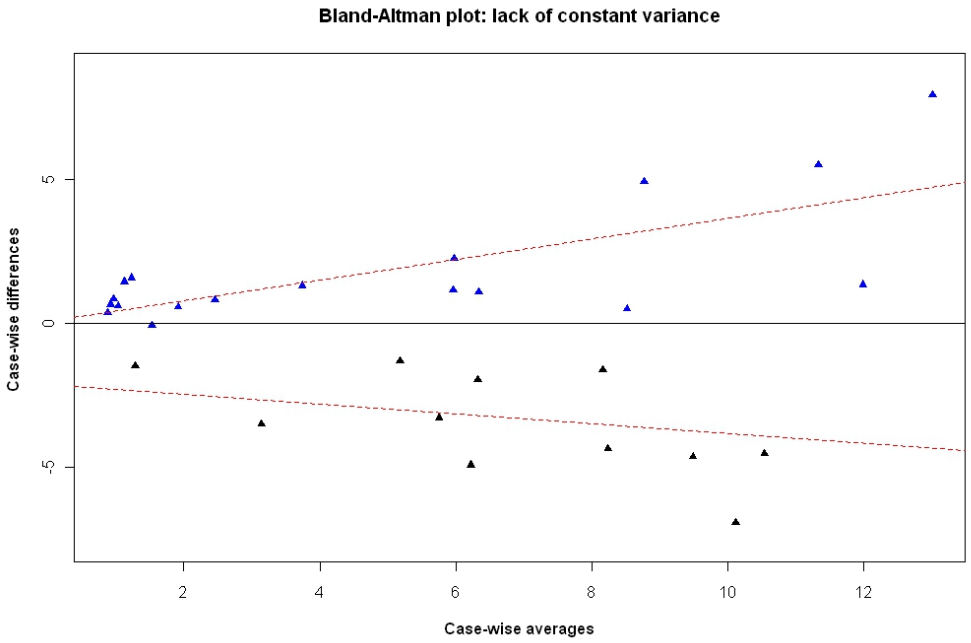

Figure 1.1.4: Bland-Altman plot demonstrating the increase of variance over the range.

1.1.4 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as 'replicate measurements'. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity.

Bland and Altman (1986) address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The

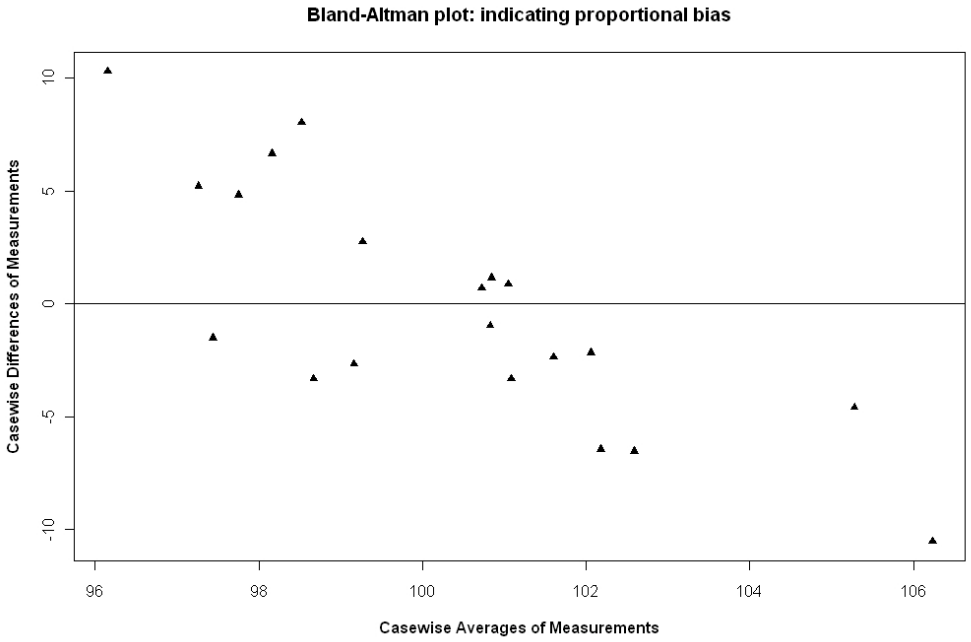

Figure 1.1.5: Bland-Altman plot indicating the presence of proportional bias.

second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. Bland and Altman (1986) propose a correction for this.

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 using the mean of replicates are 'much too narrow as prediction limits for differences between future single measurements'. 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.

1.1.5 Identifiability

Dunn (2002) highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example in literature the variance ratio $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$ must often be assumed to be equal to 1 (Linnet, 1998).Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

1.2 MCS Research Notes

The problem of comparing two methods of measurement is ubiquitous in scientific literature. The use of well-established methodologies, such as the paired t-test, correlation and regression approaches is criticised in Altman and Bland(1983). In the Bland-Altman papers, the British Standards Institute emerge as the key authority on the definition of the Limits of agreement. It is assumed that, in the absence of a specified probability, that the level is 95%.

Bland and Altman proposed a simple graphical technique, plotting the case-wise differences against the case-wise means of the respective measurements. The benefit of such an approach is the plot makes it easier to assess the magnitude of the disagreement (both error and bias), spot outliers, and see whether there is any trend.

1.2.1 Success of Bland-Altmans plot

The success of the Bland-Altman approach is perhaps due to the fact that only a visual inspection of the plot is required. Bland and Altmans paper was later reported to be the sixth most widely cited statistical paper ever (Hollis 1996, for example). Hollis, S (1996), Annals of clinical biochemistry (Annals of Biochemistry 33,1-4) Ryan, T and Woodall W (2005). The most cited statistical papers

Journal of applied Statistics 32(5), 461-474. Bland and Altman emphasis the clinical importance of the range of between the limits of agreement, and use this range as a basis for evaluating agreement. The question arises as to whether or not it is statistically valid to arrive at a decision about the population probability from an observed coverage range in a sample.

Altman and Bland (1983) show that their graphical approach can be supplemented by a test of significance on the Pearson product correlation coefficient of the plotted quantities. This test is equivalent to the test of the hypothesis that the method variances are equal (Pitman 1939) Bland and Altman recommend a test of significance of Spearmans rank correlation coefficient of the absolute differences and the case-wise means. Hayes et al (2006) examines the pitfalls that arises when an outlier is assesses using an informal criterion based on a fixed number of standard deviations rather than a more formal standard approach.

1.2.2 Underlying Model

The model underlying the Bland-Altman approach can be expressed as an LME model with heterogeneous variances.

$$y_{ij} = \beta_j + b_i + \varepsilon_{ij}$$

The case-wise differences and case-wise means follow a bivariate normal distribution, with expected values and variances specified as [input equations].

1.2.3 Outlier detection

Additionally, there is no clear guidance in any of the Bland-Altman papers on the treatment of outliers that may arise in a plot. An example used in Bland-Altman 1986 identifies a clear outlier, where it is advised by the authors that in practice, one could omit this subject. Bland and Altman 1999 recommend the computationally intensive approach of calculating the limits of agreement with, and then without, suspected outliers, in order to assess the impact on the results. However, they are clear that they do not recommend excluding outliers from analyses.

1.3 Westgard et Al

Westgard et al. (1)(2)(3) outlined the basic principles for method comparison in a clear, easy to follow manual. They also introduced the concept of allowable analytical error and gave an overview of published performance criteria. They recommended that the estimated analytical imprecision and bias be compared with these performance criteria in method evaluation as well as in method comparison. Their approach made use of a scatter-plot and calculations based on regression lines, but with confidence limits and judgment of acceptability based on the criteria for allowable analytical error.

These principles of comparing analytical performance with performance criteria, however, have not been universally accepted, and recent publications have criticized the misuse of correlation coefficients (4) and overinterpretation of regression lines in method comparison (5)(6)(7). Bland and Altman (4) recommended the difference plot (or bias plot or residual plot) as an alternative approach for method comparison. On the abscissa they used the mean value of the methods to be compared, to avoid regression towards the mean, and on the ordinate they plotted the calculated difference between measurements by the two methods. They further estimated the mean and standard deviation of differences and displayed horizontal lines for the mean and for 2 the standard deviation. However, they missed the concept of a more objective criterion for acceptability. Recently, Hollis (5) has recommended difference plots as the only acceptable method for method comparison studies for publication in Annals of Clinical Biochemistry, but without specifying criteria for acceptability.

However, a few difference plots with evaluation of acceptability according to defined criteria have been published, e.g., in evaluation of estimated biological variation compared with analytical imprecision (8), and in external quality assessment of plasma proteins for the possibilities of sharing common reference intervals (9).

Maybe the scarcity of such publications is more a question of interpretation of the data by plotting than a strict choice between scatter-plot and difference plot, as discussed by Stckl (10) recently. Investigators seem to rely too much on regression lines and r-values, without doing the equally important interpretation of the data points of the plot. This is becoming more and more disadvantageous with the increasing number of Reference Methods available for comparison with field methods, because in these cases, it is not a question of finding some relationships, but simply of judging the field method to

be acceptable or not.

NCCLS has recently published guidelines for method comparison and bias estimation by using patients samples (11), where both scatter-plots and bias plots are advised. The document also recommends plotting of single determinations as mean values and stresses the need of visual inspection of data. Further, comparison with performance criteria is recommended, but these criteria are not specified and they are not used in the graphical interpretation. Recently, Houbouyan et al. (12) used ratio plots in their validation protocol of analytical hemostasis systems, where they used a preset, but arbitrarily chosen, acceptance limit of inaccuracy of 15

In the following, we will use the difference plot (or bias plot) in combination with simple statistics for the principal judgment of the identity or acceptability of a field method. The difference plot makes it easier to apply the concept; in principle, however, the same evaluations could be performed for a scatter-plot in relation to the line of identity (y = x).

The aim of this contribution is to pay attention to the hypothesis of identity and the concept of acceptable analytical quality in method comparison, especially when one of the methods is a Reference Method.

1.4 Other Types of Studies

Lewis et al. (1991) categorize method comparison studies into three different types. The key difference between the first two is whether or not a 'gold standard' method is used. In situations where one instrument or method is known to be 'accurate and precise', it is considered as the'gold standard' (Lewis et al., 1991). A method that is not considered to be a gold standard is referred to as an 'approximate method'. In calibration studies they are referred to a criterion methods and test methods respectively.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). (In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively.) Altman and Bland (1983) make clear that

their methodology is not intended for calibration problems.

2. Comparison problems. When two approximate methods, that use the same units of measurement, are to be compared. This is the case which the Bland-Altman methodology is specfically intended for, and therefore it is the most relevant of the three.

3. Conversion problems. When two approximate methods, that use different units of measurement, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement. Lewis et al. (1991) deals specifically with this issue. In the context of this study, it is the least relevant of the three.

1.5 Gold and Bronze Standards

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free. *It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard.* The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer*leaves considerable room for improvement* (Dunn, 2002). Pizzi (1999) similarly addresses the issue of glod standards: *well-established gold standard may itself be imprecise or even unreliable.*

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years. (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram ,used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the Angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. (This is reported as sensitivity of 95% and a specificity of 92%) (ACR, 2008)

In literature they are, perhaps more accurately, referred to as 'bronze standards'. Consequently

when one of the methods is essentially a bronze standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

1.5.1 Fuzzy Gold Standards

The Gold Standard is considered to be the most accurate measurement of a particular parameter. But even gold standard raters must be assumed to have some level of measurement error. Fuzzy gold standard are considered by Phelps and Hutson (1994)

1.6 Fuzzball Agreement

Fuzzball agreement is a case where the correlation coefficient is close to zero. The sample values is restricted to a narrow range. but an examination of a relevant scatter-plot would indicate that there is agreement between the two methods.
Agreement - a numerical measure Hutson et al define a numerical measure for agreement.
For example, suppose the pairs of rater measurements are (1, 1), (1.1, 1), (1, 1.1), and (1.1, 1.1) then the sample Pearson correlation r = .0, yet the two raters or devices are considered to be in good agreement. We will refer to the instance where r is close to 0, yet there may be good agreement as "fuzzball agreement."
Fuzzball agreement occurs quite often in practice when the sample values have very narrow or restricted ranges. Fuzzball agreement is just one instance where the correlation coefficient is a poor measure of agreement.
Furthermore, note that the ICC is also a poor measure of agreement when there is fuzzball agreement. At the other extreme suppose the same raters given in the previous example had pairs of measurements (1, 101), (2, 102), (3, 103), and (4, 104) on the same relative scale as before. In this instance, r = 1.0, yet there is large disagreement between rater.

Dunn (2002) makes two important points in relation to these categories. Firstly he remarks that there isn't clear cut differences between each category.

Secondly he comments on the clinician gold standard, the sphygmomanometer, *leaves considerable room for improvement.* Pizzi (1999) also attends to this issue: *well-established gold standard may itself be imprecise or even unreliable.* The Magnetic resonance angiogram is considered to the gold standard for measuring aortic dissection, with a sensitivity of 95% and a specificity of 92% . (ACR, 2008) In literature they are, perhaps more accurately, referred to as 'bronze standards'.

Consequently when one of the methods is essentially a bronze standard, as opposed to a true gold standard, the comparison procedure should be considered as being of the second category.

1.6.1 Repeatability and gold standards

Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a 'bronze standard' exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a 'gold standard'. For example, by determining the ratio of CR to the sample mean $\bar{X}$. Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value of $\lambda = CR/\bar{X}$ with $\lambda < 0.1\%$. Similarly, a silver standard may be defined as the method with the lowest value of λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.

1.7 The Conversion Problem

In this section, we will reconsider the conversion problem, where by the methods of measurements are denominated in different units. Conversion problems arise when the comparison is between two

approximate methods of measurement each of which measures the quantity in different units.

This situation can arise when the methods in question proceed by measuring different proxies for the underlying quantity of interest. (lewis 1991)

For the single measurement case, the author can not foresee any scope for insights that are not already offered by using a structural relation model, as proposed by lewis et 1991, or error-in-variables regression. In the case of orthonormal regression, it is not reasonable to assume that both methods have equal measurement variance, when they are denominated in different units. The analyst may attempt to mitigate the problem by scaling the variance of one method, but even still problems remain. Similarly for Deming regression, no further insights on how to properly estimate the variance ratio can be offered.

For the case of conversion problem with replicate measurements, a framework that incorporates the ideas offered by Roy (2009) can be proposed. Estimates for between-subject and within-subject variances may be sought. However Roy's tests on variability are no longer applicable, as one would not expect the method to have similar estimates. An estimate for the scaling factor β may be sought, where $Y_i \approx \beta X$.

$$X_i = \tau_i + \delta_i$$

$$Y_i = \alpha + \beta X \tau_i + \epsilon_i$$

We will simulate a data set based in lewis conversion problems, provide three replicates values for both measurements. To acheive this we add "jitter noise" to three copies of each original measurement.

1.8 Other Types of Studies

Dunn (2002, p.47) cautions that'gold standards' should not be assumed to be error free. 'It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement' (Dunn, 2002). Pizzi (1999) similarly addresses the issue of glod standards, 'well-established gold standard may itself be imprecise or even unreliable'.

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well-being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram, used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. This is reported as sensitivity of 95% and a specificity of 92% (ACR, 2008).

In literature they are, perhaps more accurately, referred to as 'fuzzy gold standards' (Phelps and Hutson, 1995). Consequently when one of the methods is essentially a fuzzy gold standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

Chapter 2

Introduction to Method Comparison Studies

Abstract

The first chapter will consider the topic of Method Comparison Studies, and discuss the impact of the Bland-Altman Methodology. A detailed discussion of the Bland-Altman Methodology will be covered in chapter two.

2.1 Agreement

- The FDA define precision as the *closeness of agreement* (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under prescribed conditions.

- **Barnhart** describes precision as being further subdivided as either within-run, intra-batch precision or repeatability (which assesses precision during a single analytical run), or between-run, inter-batch precision or repeatability (which measures precision over time).

2.2 Outline of Thesis

Thus the study of method comparison is introduced. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models. Chapter two shall describe linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the `R` programming language.

In the first chapter the study of method comparison is introduced, while the second chapter provides a review of current methodologies. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models.

Chapter three shall describes linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the `R` programming language.

Model diagnostics are an integral component of a complete statistical analysis. In chapter three model diagnostics shall be described in depth, with particular emphasis on linear mixed effects models, further to chapter two.

For the fourth chapter, important linear mixed effects model diagnostic methods shall be extended to method comparison studies, and proposed methods shall be demonstrated on data sets that have become well known in literature on method comparison. The purpose is to both calibrate these methods and to demonstrate applications for them. The last chapter shall focus on robust measures of important

parameters such as agreement.

2.3 Purposes of MCS

The question being answered is not always clear, but is usually epxressed as an attempt to quantify the agreement between two methods (Bland and Altman 1995)

Some lack of agreement between different methods of measurement is inevitable. What matters is the amount by which they disagree. we want to know by how much the new method is likely to differ from the old, so that it is not enough to cause problems in the mathematical interpretation we can preplace the old method by the new, or even use the two interchangably.

It often happens that the same physical and chemical property can be measured in different ways. For example, one can determine For example, one can determine sodium in serum by flame atomic emission spectroscopy or by isotops dilution mass spectroscopy. The question arises as to whcih methd is better (Mandel 1991)

In areas of inter-laboratory quality control, method comparisons, assay validations and individual bio-equivalence, etc, the agree between observations and target (reference) value is of interest (lin 2002)

The purpose of comparing two methods of measurement of a continuous biological variable is to uncover systematic differences, not to point to similarities. (ludbrook 1997)

In the pharmaceutical industry, measurement methods that measure the quantity of prdocuts are regulated. The FDA (U.S. Food and Drug Administration) requires that the manufacturer show equivalency prior to approving the new or alternatice method in quality control (Tan & Inglewicz ,1999)

2.4 Method Comparison Studies

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. The 95% limits of agreement, estimated by mean difference +/- 1.96 standard deviation of the differences, provide an interval within which 95% of differences between measurements by the two methods are expected to lie.

2.5 Discussion on Method Comparison Studies

The need to compare the results of two different measurement techniques is common in medical statistics.

In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years.

Indications on how to deal with outliers in Bland Altman plots
We wish to determine how outliers should be treated in a Bland Altman Plot
In their 1983 paper they merely state that the plot can be used to 'spot outliers'.
In their 1986 paper, Bland and Altman give an example of an outlier. They state that it could be omitted in practice, but make no further comments on the matter.
In Bland and Altmans 1999 paper, we get the clearest indication of what Bland and Altman suggest on how to react to the presence of outliers. Their recommendation is to recalculate the limits without them, in order to test the difference with the calculation where outliers are retained.

The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large reduction.
However, they do not recommend removing outliers. Furthermore, they say:
We usually find that this method of analysis is not too sensitive to one or two large outlying differences.
We ask if this would be so in all cases. Given that the limits of agreement may or may not be disregarded, depending on their perceived suitability, we examine whether it would possible that the deletion of an outlier may lead to a calculation of limits of agreement that are usable in all cases?
Should an Outlying Observation be omitted from a data set? In general, this is not considered prudent.
Also, it may be required that the outliers are worthy of particular attention themselves.
Classifying outliers and recalculating We opted to examine this matter in more detail. The following points have to be considered
how to suitably identify an outlier (in a generalized sense)
Would a recalculation of the limits of agreement generally results in a compacted range between the

upper and lower limits of agreement?

2.5.1 Agreement

Bland and Altman (1986) define Perfect agreement as 'The case where all of the pairs of rater data lie along the line of equality'. The Line of Equality is defined as the 45 degree line passing through the origin, or X=Y on a XY plane.

2.5.2 Lack Of Agreement

1. Constant Bias
2. Proportional Bias

Constant Bias

This is a form of systematic deviations estimated as the average difference between the test and the reference method

Proportional Bias

Two methods may agree on average, but they may exhibit differences over a range of

2.6 Methods of assessing agreement

1. Pearson's Correlation Coefficient
2. Intraclass correlation coefficient
3. Bland Altman Plot
4. Bartko's Ellipse (1994)
5. Blackwood Bradley Test

6. Lin's Reproducibility Index

7. Luiz Step function

Bland and Altman attend to the issue of repeated measures in 1996.
Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual.
Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the 't' limits of agreement (the outer pair of dashed lines) centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

2.6.1 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

If the limits of agreement are not clinically important, which is to say that the differences tend not to be substantial, the two methods may be used interchangeably. Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

2.7 Introductory Definitions

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Published examples of method comparison studies can be found in disciplines as diverse as Pharmacology (Ludbrook, 1997), Anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000). Method

Comparison Studies is a branch of statistics used to compare the results of two different method of measurement, measuring the same subject samples. Consider a set of n samples. Measurements are taken on each of the n samples using both methods. This will enable comparison of the method used. In many cases the purpose of the study is to calibrate a new method of measurement against a Gold Standard method. A Gold Standard method is the known method that is considered most precise in its measurement. It should not be assumed that there is no error present in its measurements. The Gold Standard may not be financially feasible for general use, and therefore more economical methods, of suitable levels of precisions, must be devised. Method Comparison studies is used to ascertain the levels of precision of such methods.

To illustrate the characteristics of a typical method comparison study consider the data in Table I, taken from Grubbs (1973). In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured simultaneously (and independently) by three chronographs devices, referred to here as 'Fotobalk', 'Counter' and 'Terma'.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 2.7.1: Measurement of the three chronographs (Grubbs 1973)

An important aspect of the these data is that all three methods of measurement are assumed to have an attended measurement error, and the velocities reported in Table I can not be assumed to be 'true values' in any absolute sense. For expository purposes only the first two methods 'Fotobalk' and 'Counter' will enter in the immediate discussion.

While lack of agreement between two methods is inevitable, the question , as posed by Altman and Bland (1983), is 'do the two methods of measurement agree sufficiently closely?'

A method of measurement should ideally be both accurate and precise.An accurate measurement methods shall give a result close to the 'true value'. Precision of a method is indicated by how tightly clustered its measurements are around their mean measurement value.

A precise and accurate method should yield results consistently close to the true value. However a method may be accurate, but not precise. The average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely an inaccurate method may be quite precise , as it consistently indicates the same level of inaccuracy.

The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The lesser the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero.

A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently there is lack of agreement between the two methods.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.80	794.60	-0.80
2	793.10	793.90	-0.80
3	792.40	793.20	-0.80
4	794.00	794.00	0.00
5	791.40	792.20	-0.80
6	792.40	793.10	-0.70
7	791.70	792.40	-0.70
8	792.30	792.80	-0.50
9	789.60	790.20	-0.60
10	794.40	795.00	-0.60
11	790.90	791.60	-0.70
12	793.50	793.80	-0.30

Table 2.7.2: Difference between Fotobalk and Counter measurements

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree or not. These methods must also have equivalent levels of precision. Should one method yield results considerably more variable than that of the other, they can not be considered to be in agreement.

Therefore a methodology must be introduced that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

2.8 Bland Altman Plots In Literature

Mantha et al. (2000) contains a study the use of Bland Altman plots of 44 articles in several named journals over a two year period. 42 articles used Bland Altman's limits of agreement, wit the other two used correlation and regression analyses. Mantha et al. (2000) remarks that 3 papers, from 42 mention predefined maximum width for limits of agreement which would not impair medical care.

The conclusion of Mantha et al. (2000) is that there are several inadequacies and inconsistencies in the reporting of results ,and that more standardization in the use of Bland Altman plots is required. The authors recommend the prior determination of limits of agreement before the study is carried out. This contention is endorsed by Lin et al. (1991), which makes a similar recommendation for the sample size, noting that*sample sizes required either was not mentioned or no rationale for its choice was given.*

> In order to avoid the appearance of "data dredging", both the sample size and the (limits of agreement) should be specified and justified before the actual conduct of the trial. (Lin et al., 1991)

Dewitte et al. (2002) remarks that the limits of agreement should be compared to a clinically acceptable difference in measurements.

2.8.1 Gold Standard

This is considered to be the most accurate measurement of a particular parameter.

2.9 Bland Altman Plots

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983).

Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge that there are other valid, but complex, methodologies, and argue that a simple approach is preferable to this complex approaches, *especially when the results must be explained to non-statisticians* (Altman and Bland, 1983).

Notwithstanding previous remarks about regression, the first step recommended ,which the authors argue should be mandatory,is construction of a simple scatter plot of the data. The line of equality ($X = Y$) should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 2.1). These differences and averages are then plotted (Figure 2.2).

The dashed line in Figure 2.2 alludes to the inter method bias between the two methods, as mentioned previously. Bland and Altman recommend the estimation of inter method bias by calculating the average of the differences. In the case of Grubbs data the inter method bias is -0.6083 metres per second.

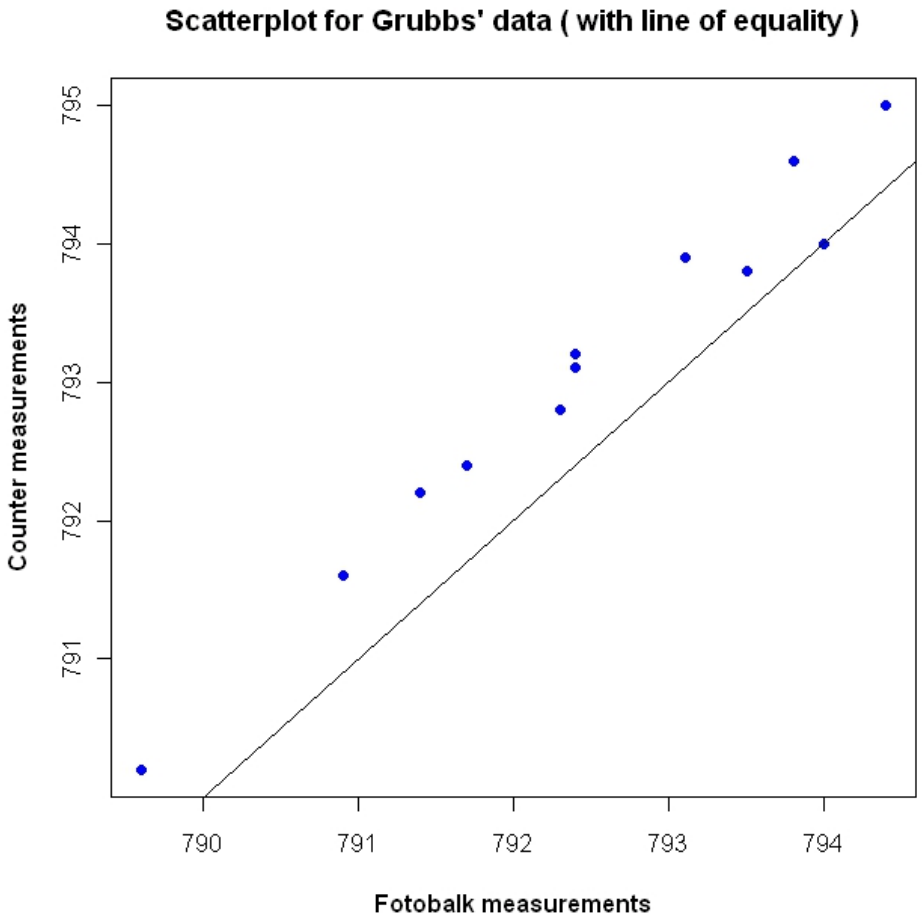

Figure 2.9.1: Scatter plot For Fotobalk and Counter Methods

By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

2.9.1 Inspecting the Data

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. Altman and Bland (1983) express the motivation for this plot thusly:

> "From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 2.9.3: Fotobalk and Counter Methods: Differences and Averages

in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study."

Figures 1.3 1.4 and 1.5 are three Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of trends that would adversely affect use of the recommended methodology. Figure 1.3 demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range. Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (Ludbrook, 1997).

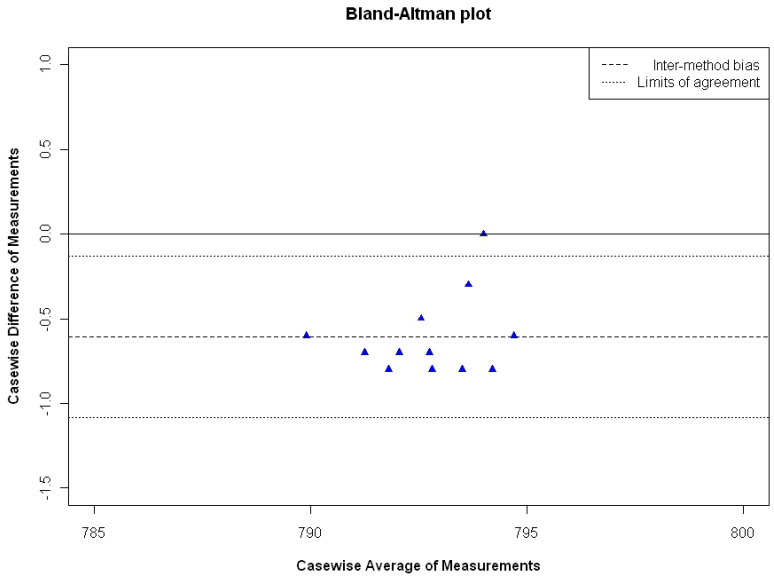

Figure 2.9.2: Bland Altman Plot For Fotobalk and Counter Methods

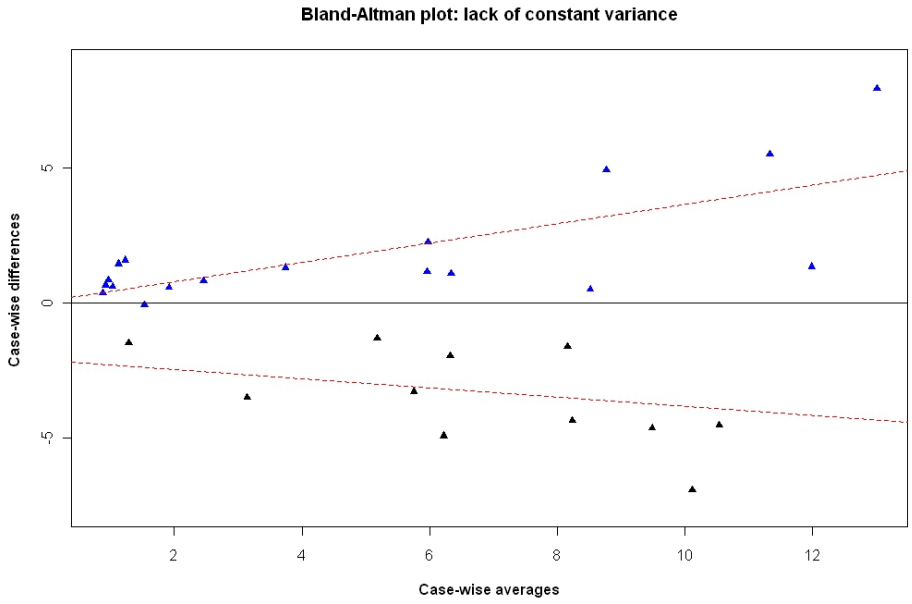

Figure 2.9.3: Bland-Altman Plot demonstrating the increase of variance over the range

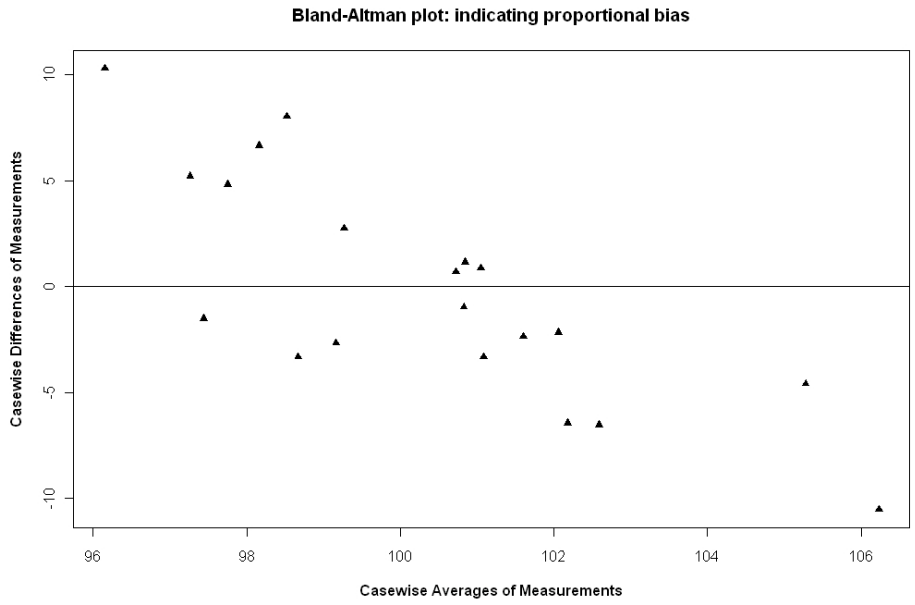

Figure 2.9.4: Bland-Altman Plot indicating the presence of proportional bias

Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as *proportional bias* (Ludbrook, 1997). Both of these cases violate the assumptions necessary for further analysis using limits of agreement ,which shall be discussed later. The plot also can be used to identify outliers. An outlier is an observation that is numerically distant from the rest of the data. Classification thereof is a subjective decision in any analysis, but must be informed by the logic of the formulation. Figure 1.5 is a Bland Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively.

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Hence any observation , such as the one on the extreme right of figure 1.5, should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster. The one on the extreme left should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

Bland and Altman (1999) do not recommend excluding outliers from analyses. However recalculation of the inter-method bias estimate , and further calculations based upon that estimate, are useful for assessing the influence of outliers.(Bland and Altman, 1999) states that *"We usually find that this*

method of analysis is not too sensitive to one or two large outlying differences."

2.9.2 Limits of Agreement

Bland and Altman (1986) introduces an elaboration of the plot, adding to the plot 'limits of agreement' to the plot. These limits are based upon the standard deviation of the differences. The discussion shall be reverted to these limits of agreement in due course.

2.9.3 Variations of the Bland Altman Plot

Bland and Altman (1999) remarks that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman (1999) acknowledge that this is not easy to interpret, and that it is not suitable in all cases.

Bland and Altman (1999) offers two variations of the Bland -Altman plot that are intended to overcome potential problems that the conventional plot would inappropriate for.

The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages.

2.9.4 Agreement

Bland and Altman (1986) defined perfect agreement as the case where all of the pairs of rater data lie along the line of equality, where the line of equality is defined as the 45 degree line passing through the origin(i.e. the $X = Y$ line).

Bland and Altman (1986)expressed this in the terms *we want to know by how much the new method is likely to differ from the old; if this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably. How far apart measurements can be*

without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparisonand to choose the sample size .

2.9.5 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

2.9.6 Inappropriate assessment of Agreement

Paired T tests

This method can be applied to test for statistically significant deviations in bias. This method can be potentially misused for method comparison studies.
It is a poor measure of agreement when the rater's measurements are perpendicular to the line of equality[Hutson et al]. In this context, an average difference of zero between the two raters, yet the scatter plot displays strong negative correlation.

Inappropriate Methodologies

Use of the Pearson Correlation Coefficient , although seemingly intuitive, is not appropriate approach to assessing agreement of two methods. Arguments against its usage have been made repeatedly in the relevant literature. It is possible for two analytical methods to be highly correlated, yet have a poor level of agreement.

Pearson's Correlation Coefficient

It is well known that Pearson's correlation coefficient is a measure of the linear association between two variables, not the agreement between two variables (e.g., see Bland and Altman 1986)..This is a well known as a measure of linear association between two variables.Nonetheless this is not necessarily the

same as Agreement. This method is considered wholly inadequate to assess agreement because it only evaluates only the association of two sets of observations.

2.9.7 Inappropriate use of the Correlation Coefficient

It is intuitive when dealing with two sets of related data, i.e the results of the two raters, to calculate the correlation coefficient (r). Bland and Altman attend to this in their 1999 paper.

They present a data set from two sets of meters, and an accompanying scatterplot. An hypothesis test on the data set leads us to conclude that there is a relationship between both sets of meter measurements. The correlation coeffiecient is determined to be r =0.94.However, this high correlation does not mean that the two methods agree. It is possible to determine from the scatterplot that the intercept is not zero, a requirement for stating both methods have high agreement. Essentially, should two methods have highly correlated results, it does not follow that they have high agreement.

2.9.8 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.

Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

2.9.9 The Bland Altman Plot

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. It has proved highly popular ever since. This is a simple, and widely used , plot of the differences of each data pair, and the corresponding average value. An important requirement is that the two measurement methods use the same scale of measurement.

scatter plots

The authors advise the use of scatter plots to identify outliers, and to determine if there is curvilinearity present. In the region of linearity ,simple linear regression may yield results of interest.

2.9.10 Effect of Outliers

Another argument against the use of model I regression is based on outliers. Outliers can adversely influence the fitting of a regression model. Cornbleet and Cochrane compare a regression model influenced by an outlier with a model for the same data set, with the outlier excluded from the data set. A demonstration of the effect of outliers was made in Bland Altman's 1986 paper. However they discourage the exclusion of outliers.

2.9.11 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

The bias is computed as the average of the difference of paired assays.

If one method is sometimes higher, and sometimes the other method is higher, the average of the differences will be close to zero. If it is not close to zero, this indicates that the two assay methods are producing different results systematically.

Precision of Limits of Agreement

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. A different sample would give different limits of agreement. Bland and Altman (1986) advance a formulation for confidence intervals of the inter-method bias and the

limits of agreement. These calculations employ quantiles of the 't' distribution with $n-1$ degrees of freedom.

2.9.12 Appropriate Use of Limits of Agreement

Importantly Bland and Altman (1999) makes the following point:

> These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

Carstensen attends to the issue of repeated data, using the expression replicate to express a repeated measurement on a subject by the same methods. Carstensen formulates the data as follows Repeated measurement - Arrangement of data into groups, based on the series of results of each subject.

2.9.13 The Bland Altman Plot - Variations

Variations of the Bland Altman plot is the use of ratios, in the place of differences.

$$D_i = X_i - Y_i \tag{2.1}$$

Altman and Bland suggest plotting the within subject differences $D = X_1 - X_2$ on the ordinate versus the average of x_1 and x_2 on the abscissa.

measurements

2.10 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying

quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.

Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

2.10.1 Bland Altman plots using 'Gold Standard' raters

According to Bland and Altman, one should use the methodology previous outlined, even when one of the raters is a Gold Standard.

2.10.2 Bias Detection

further to this method, the presence of constant bias may be indicated if the average value differences is not equal to zero. Bland and Altman does, however, indicate the indication of absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

2.10.3 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line. How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis

A third element of the Bland-Altman methodology, an interval known as 'limits of agreement' is introduced in Bland and Altman (1986), (sometimes referred to in literature as 95% limits of agreement). Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. Bland and Altman (1986) refer to this as the 'equivalence' of two measurement methods. It must be established clearly the specific purpose of the limits of agreement. Bland and Altman (1995)

comment that the limits of agreement "how far apart measurements by the two methods were likely to be for most individuals", a definition echoed in their 1999 paper:

> "We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits of agreement. These values define the range within which most differences between measurements by the two methods will lie."

The limits of agreement (LoA) are computed by the following formula:

$$LoA = \bar{d} \pm 1.96 S(d) \tag{2.2}$$

with $\bar{d}$ as the estimate of the inter method bias, $S(d)$ as the standard deviation of the differences and 1.96 is the 95% quantile for the standard normal distribution. (However, in some literature, 2 standard deviations are used instead for simplicity.) For the Grubbs 'F vs C' comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure 1.9 shows the resultant Bland-Altman plot, with the limits of agreement shown in dashed lines.

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As Bland and Altman (1986) point out this may not be the case. Bland and Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion.

> 'How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size.'(Bland and Altman, 1986)

Bland and Altman (1999) note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as "being like a reference interval."

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential

for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the limits, the mean and standard deviation, are not based on any sample used for an analysis, but on the process's historical values, a key difference with Bland-Altman limits of agreement.

Carstensen et al. (2008) regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. Carstensen et al. (2008) offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{2.3}$$

where n is the number of subjects. Only for 61 or more subjects is there a quantile less than 2.

Luiz et al. (2003) describes limits of agreement as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence.

2.10.4 Appropriate Use of Limits of Agreement

Importantly Bland and Altman (1999) makes the following point:

> These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

2.10.5 Problems with Limits of Agreement

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values.

The formulation is also heavily influenced by outliers. An Example in Altman and Bland (1983) demonstrates the effect of recalculating without a particular outlier. Refering to the VCF data set in the same paper, there is more than one outlier.

Chapter 3

Linear Mixed effects Models

3.1 Statement of the LME model

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

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

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

The linear mixed effects model is given by

$$Y = X\beta + Zu + \epsilon \tag{3.1}$$

3.2 Extended LME model

The extended single level LME model relaxes the independence assumption, allowing heteroscedastic and correlated within group errors.

$$\epsilon_i = \mathcal{N}(0, \sigma^2 \Lambda_i) \tag{3.2}$$

Λ_i are positive definite matrices. σ^2 is factored out of the matrix for computational reasons.

3.3 Variance functions

Variance functions are applied to LME models through the 'weights' argument. R supports several variance functions.

'varIdent' cosntructs a model with different variances per stratum.

3.3.1 Diagnostic plots

Diagnostic plots for identifying within-group heteroscedascity and assessing the adequacy of a variance function can also be used with 'nlme' objects.

3.4 Introduction to Mixed Models

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

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

$$y_{ij} = \mu + \delta_i + \phi_j + \gamma_{ij} + \epsilon_{ijk} \tag{3.3}$$

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

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

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

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{3.4}$$

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

3.5 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 ($\ell(\theta)$).

3.6 Linear Mixed effects Models

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

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

The 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. SAS Institute added PROC MIXED to its software suite in 1992 (Singer, 1998). Pinheiro and Bates (1994) described how to compute LME models in the `S-plus` environment.

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

$$y = X\beta + Zb + \epsilon \tag{3.5}$$

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

3.6.1 Estimation

Estimation of LME models involve two complementary estimation issues'; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and $\hat{b}$ and estimating the variance covariance matrices D and Σ. Inference about fixed effects have become known as 'estimates', while inferences about random effects have become known as 'predictions'. The most common approach to obtain estimators are Best Linear Unbiased Estimator (BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (3.5), the BLUE of $\hat{\beta}$ is given by

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

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

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

Estimation of the fixed parameters

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

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{3.6}$$

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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{3.7}$$

Substituting $\hat{\beta}$ from (3.7) into $\ell(\beta, \theta|y)$ from (3.6) returns the *profile* log-likelihood

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

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

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

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

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (Dodge, 2003). Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $DZ'V^{-1}(y - X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (3.7) 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 (**??**).

Algorithms for likelihood function optimization

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

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

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

The Newton Raphson (NR) method is the most common, and recommended technique for ML and REML estimation. The NR algorithm minimizes an objective function defines as -2 times the log likelihood for the covariance parameters θ. At every iteration the NR algorithm requires the calculation of a vector of partial derivatives, known as the gradient, and the second derivative matrix with respect to the covariance parameters. This is known as the observed Hessian matrix. Due to the Hessian matrix, the NR algorithm is more time-consuming, but convergence is reached with fewer iterations

compared to the EM algorithm. The Fisher scoring algorithm is an variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

3.6.2 Formulation of the response vector

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

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

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

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

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

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

3.6.3 Decomposition of the response covariance matrix

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

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

$\boldsymbol{R_i}$ can be shown to be the Kronecker product of a correlation matrix $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$. The correlation matrix $\boldsymbol{V}$ of the repeated measures on a given response variable is assumed to be the same for all

response variables. Both Hamlett et al. (2004) and Lam et al. (1999) use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. Roy (2009) 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. **?** 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 (2009) indicate its use.

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

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

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

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

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

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

3.6.4 Correlation terms

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

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

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

ρ_A describe the correlations of measurements made by the method A at different times. Similarly ρ_B describe the correlation of measurements made by the method B at different times. Correlations among repeated measures within the same method are known as intra-class correlation coefficients. ρ_{AB} describes the correlation of measurements taken at the same same time by both methods. The coefficient δ is added for when the measurements are taken at different times, and is a constant of less than 1 for linked replicates. This is based on the assumption that linked replicates measurements taken at the same time would have greater correlation than those taken at different times. For unlinked replicates δ is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

3.7 Matrix Formulation

There are matrix (i.e multivariate) formulations of both fixed effects models and random effects models. Brown and Prescott (1999) remarks that the matrix notation makes the underlying theory of mixed effects models much easier to work with. The fixed effects models can be specified as follows;

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{3.8}$$

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

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

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{3.9}$$

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

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{3.10}$$

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

3.8 BXC - Model Terms

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

3.9 Other Approaches

3.9.1 Random coefficient growth curve model

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

$$\mathbf{d} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{3.11}$$

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

3.9.2 Marginal Modelling

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

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

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

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

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

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

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 (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet 346*, 1085–87.

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

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

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

Carstensen, B. (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.

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

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

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

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.

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

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

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.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

Lin, S. C., D. M. Whipple, and charles S Ho (1991). Evaluation of statistical equivalence using limits of agreement and associates sample size calculation. *Communications in Statistics - Theory and Methods 27*(6), 1419–1432.

Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry 44*, 1024–1031.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

Luiz, R., A. Costa, P. Kale, and G. Werneck (2003). Assessment of agreement of a quantitative variable: a new graphical approach. *Journal of Clinical Epidemiology 56*, 963–967.

Mantha, S., M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss (2000). Comparing methods of clinical measurement: Reporting standards for bland and altman analysis. *Anaesthesia and Analgesia 90*, 593–602.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

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

Phelps, C. and A. Hutson (1995). Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical decision making 15*, 144–57.

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

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.

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.

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

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

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

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*(2), 153–177.

3.10 LME

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

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{3.12}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{3.13}$$

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

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{3.14}$$

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.

3.11 Other Approaches

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

3.12 Remarks

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.

Chapter 4

LME Likelihood

4.1 One Way ANOVA

4.1.1 Page 448

Computing the variance of $\hat{\beta}$

$$\text{var}(\hat{\beta}) = (X'V^{-1}X)^{-}1 \tag{4.1}$$

It is not necessary to compute V^{-1} explicitly.

$$V^{-1}X = \Sigma^{1}X - Z()Z'\Sigma^{-1}X \tag{4.2}$$

$$= \Sigma^{-1}(X - Zb_x) \tag{4.3}$$

The estimate b_x is the same term obtained from the random effects model; $X = Zb_x + e$, using X as an outcome variable. This formula is convenient in applications where b_x can be easily computed. Since X is a matrix of p columns, b_x can simple be computed column by column. according to the columns of X.

4.1.2 Page 448- simple example

Consider a simple model of the form;

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}.$$

The iterative procedure is as follows Evaluate the individual group mean $\bar{y}_i$ and variance $\hat{Sigma^2}_i$. Then use the variance of the group means as an estimate of the σ_b^2. The average of the the variances of the groups is the initial estimate of the σ_e^2.

Iterative procedure

The iterative procedure comprises two steps, with 0 as the first approximation of b_i.

The first step is to compute λ, the ratio of variabilities,

$$\lambda = \frac{\sigma_b^2}{\sigma_e^2}$$

$$\mu = \frac{1}{N}\sum_{ij}(y_{ij} - b_i)$$

$$b_i = \frac{n(\bar{y}_i - \mu)}{n + \lambda}$$

The second step is to updat $sigma_e^2$

$$\sigma_e^2 = \frac{e'e}{N - df} \tag{4.4}$$

where e is the vector of $e_{ij} = y_{ij} - \mu - b_i$ and $df = qn/n + \lambda$ and

$$\sigma_b^2 = \frac{1}{q}\sum_{i=1}^{q} b_1^2 + (\frac{n}{\sigma_e^2} + \frac{1}{\sigma_b^2})^{-1} \tag{4.5}$$

Worked Example

Further to [pawitan 17.1] the initial estimates for variability are $\sigma_b^2 = 1.7698$ and $\sigma_e^2 = 0.3254$. At convergence the following results are obtained.

n=16, q=5

$$\hat{\mu} = \bar{y} = 14.175$$
$$\hat{\sigma}^2 = 0.325$$
$$\hat{\sigma}_b^2 = 1.395$$
$$\sigma = 0.986$$

At convergene the following estimates are obtained,

$$\hat{\mu} = 14.1751$$
$$\hat{b} = (-0.6211, 0.2683, 1.4389, -1.914, 0.8279)$$
$$\hat{\sigma}_b^2 = 1.3955$$
$$\hat{\sigma}_e^2 = 0.3254$$

4.1.3 Extention to several random effects

[pawitan section 17.7]

4.2 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)

4.3 Conclusion

Carstensen et al. (2008) and **?** highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. **?** 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 ARoy2009's methodology.

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 ARoy2009 (2009) deals with the relevant assumptions regrading 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*".

Chapter 5

General Appendices

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

In the graph above, you can predict non-zero values for the residuals based on the fitted value. For example, a fitted value of 8 has an expected residual that is negative. Conversely, a fitted value of 5 or 11 has an expected residual that is positive.

The non-random pattern in the residuals indicates that the deterministic portion (predictor variables) of the model is not capturing some explanatory information that is leaking into the residuals. The graph could represent several ways in which the model is not explaining all that is possible.

Possibilities include:

- A missing variable
- A missing higher-order term of a variable in the model to explain the curvature
- A missing interction between terms already in the model

Identifying and fixing the problem so that the predictors now explain the information that they missed before should produce a good-looking set of residuals!

In addition to the above, here are two more specific ways that predictive information can sneak into the residuals:

The residuals should not be correlated with another variable. If you can predict the residuals with another variable, that variable should be included in the model. In Minitabs regression, you can plot

the residuals by other variables to look for this problem.

Autocorrelation

Adjacent residuals should not be correlated with each other (**autocorrelation**). If you can use one residual to predict the next residual, there is some predictive information present that is not captured by the predictors. Typically, this situation involves time-ordered observations. For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated. You can include a variable that captures the relevant time-related information, or use a time series analysis.

In Minitabs regression, you can perform the ***Durbin-Watson*** test to test for autocorrelation.

5.1 Gold and Bronze Standards

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free. *It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard.* The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer*leaves considerable room for improvement* (Dunn, 2002). Pizzi (1999) similarly addresses the issue of glod standards: *well-established gold standard may itself be imprecise or even unreliable.*

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years. (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram ,used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the Angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. (This is reported as sensitivity of 95% and a specificity of 92%) (ACR, 2008)

In literature they are, perhaps more accurately, referred to as 'bronze standards'. Consequently when one of the methods is essentially a bronze standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

5.1.1 Fuzzy Gold Standards

The Gold Standard is considered to be the most accurate measurement of a particular parameter. But even gold standard raters must be assumed to have some level of measurement error. Fuzzy gold standard are considered by Phelps and Hutson (1994)

5.2 Fuzzball Agreement

Fuzzball agreement is a case where the correlation coefficient is close to zero. The sample values is restricted to a narrow range. but an examination of a relevant scatter-plot would indicate that there is agreement between the two methods.

Agreement - a numerical measure Hutson et al define a numerical measure for agreement.

For example, suppose the pairs of rater measurements are (1, 1), (1.1, 1), (1, 1.1), and (1.1, 1.1) then the sample Pearson correlation r = .0, yet the two raters or devices are considered to be in good agreement. We will refer to the instance where r is close to 0, yet there may be good agreement as "fuzzball agreement."

Fuzzball agreement occurs quite often in practice when the sample values have very narrow or restricted ranges. Fuzzball agreement is just one instance where the correlation coefficient is a poor measure of agreement.

Furthermore, note that the ICC is also a poor measure of agreement when there is fuzzball agreement. At the other extreme suppose the same raters given in the previous example had pairs of measurements (1, 101), (2, 102), (3, 103), and (4, 104) on the same relative scale as before. In this instance, r = 1.0, yet there is large disagreement between rater.

5.3 Types of Method Comparisons

? categorize method comparison studies into three different types, with the first two being of immediate concern. A method that is not considered to be a gold standard is referred to as an 'approximate method'.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard.

2. Comparison problems - When two approximate methods, that use the same units of measurement, are to be compared.

3. Conversion problems - When two approximate methods, that use different units of measure-

ment, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement.

Dunn (2002) makes two important points in relation to these categories. Firstly he remarks that there isn't clear cut differences between each category.

Secondly he comments on the clinician gold standard, the sphygmomanometer, *leaves considerable room for improvement.* Pizzi (1999) also attends to this issue: *well-established gold standard may itself be imprecise or even unreliable.* The Magnetic resonance angiogram is considered to the gold standard for measuring aortic dissection, with a sensitivity of 95% and a specificity of 92% . (ACR, 2008) In literature they are, perhaps more accurately, referred to as 'bronze standards'.

Consequently when one of the methods is essentially a bronze standard, as opposed to a true gold standard, the comparison procedure should be considered as being of the second category.

5.4 Structural Equation Modelling

This is a statistical technique used for testing and estimating causal relationships using a combination of statistical data and qualitative causal assumptions. This technique was proposed by **?** as a method of assessing the reliability of a new measurement technique. It can indicate the presence of bias. However Bland and Altman (1987) have criticized it on the basis that it offers no insights into the variability about the line of equality.
In this paper, the SEM method is used to assess the linear relationship between the new method and the standard method.
Structural analysis is a generalization of regression analysis.
In Hopkins papers, a critique of the Bland-Altman plot he makes the following remark:
What's needed for a comparison of two or more measures is a generic approach more powerful even than regression to model the relationship and error structure of each measure with a latent variable representing the true value.

Hopkins also adds that he himself is collaborating in research utilising SEM and Mixed Effects modelling.This is a methodology proposed by Kelly (1985).

5.5 ICC, Reproducibility Index and Passing-Bablok

5.5.1 Intraclass Correlation Coefficient

This measure of agreement is estimated using variance components from appropriate analysis of variance models. Measures of agreement are variance dependent, and so the ICC can be misleading. The ICC takes a value between 0 and 1, and is based on Analysis of Variance methodologies.
The ICC is a measure of reliability.

Bartko (1994) considers the ICC as just another measure of agreement.

5.5.2 Passing and Bablok (1983)

Passing & Bablok have described a linear regression model that are without the usual assumptions regarding the distribution of the samples and the measurement errors. The result does not depend on the assignment of the methods (or instruments) to X and Y. The slope and intercept are calculated with their 95% confidence interval.Hypothesis tests on the slope and intercept maybe then carried out. If the hypothesis of the intercept is rejected, then it is concluded that it is significant different from 0 and both raters differ at least by a constant amount.
If the hypothesis of the slope is rejected, then it is concluded that the slope is significant different from 1 and there is at least a proportional difference between the two raters.

5.5.3 Lin's Reproducibility Index

Lin proposes the use of a reproducibility index, called the Concordance Correlation Coefficent (CCC). While it is not strictly a measure of agreement as such, it can form part of an overall method comparision methodology.

5.6 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman 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 too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

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.

5.7 Overview

1. Extending deletion diagnostics to LMEs
2. Christensen et al
3. Haslett hayes
4. Schabenberger
5. Tewomir

1. Residual Diagnostics
    (a) Marginal and Conditional Diagnostics
    (b) Scaled Residuals
2. Influence Diagnostics
    (a) Underlying Concepts

(b) Managing the Covariance Parameters

(c) Predicted Values, PRESS Residual and the PRESS Statistic

(d) Leverage

(e) Internally and Externally Studentized Residuals

(f) DFFITs and MDFFITs

(g) Covariance Ratio and Trace

(h) Likelihood Distance

(i) Non-iterative Update Procedures

5.8 Likelihood ratio test

The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit between two models. A relatively more complex model is compared to a simpler model to see if it fits a particular dataset significantly better. If so, the additional parameters of the more complex model are often used in subsequent analyses. The LRT is only valid if used to compare hierarchically nested models. That is, the more complex model must differ from the simple model only by the addition of one or more parameters. Adding additional parameters will always result in a higher likelihood score. However, there comes a point when adding additional parameters is no longer justified in terms of significant improvement in fit of a model to a particular dataset. The LRT provides one objective criterion for selecting among possible models.

The LRT begins with a comparison of the likelihood scores of the two models:

LR = 2*(lnL1-lnL2) This LRT statistic approximately follows a chi-square distribution. To determine if the difference in likelihood scores among the two models is statistically significant, we next must consider the degrees of freedom. In the LRT, degrees of freedom is equal to the number of additional parameters in the more complex model. Using this information we can then determine the critical value of the test statistic from standard statistical tables.

The LRT is explained in more detail by Felsenstein (1981), Huelsenbeck and Crandall (1997), Huelsenbeck and Rannala (1997), and Swofford et al. (1996). While the focus of this page is using the

LRT to compare two competing models, under some circumstances one can compare two competing trees estimated using the same likelihood model. There are many additional considerations (e.g., see Kishino and Hasegawa 1989, Shimodaira and Hasegawa 1999, and Swofford et al. 1996).

In statistics, a likelihood ratio test is used to compare the fit of two models, one of which is nested within the other. This often occurs when testing whether a simplifying assumption for a model is valid, as when two or more model parameters are assumed to be related.

Both models are fitted to the data and their log-likelihood recorded. The test statistic (usually denoted D) is twice the difference in these log-likelihoods:

The model with more parameters will always fit at least as well (have a greater log-likelihood). Whether it fits significantly better and should thus be preferred can be determined by deriving the probability or p-value of the obtained difference D. In many cases, the probability distribution of the test statistic can be approximated by a chi-square distribution with (df1 df2) degrees of freedom, where df1 and df2 are the degrees of freedom of models 1 and 2 respectively.

The test requires nested models, that is, models in which the more complex one can be transformed into the simpler model by imposing a set of linear constraints on the parameters.

In a concrete case, if model 1 has 1 free parameter and a log-likelihood of 8012 and the alternative model has 3 degrees of freedom and a LL of 8024, then the probability of this difference is that of chi-square of $24 = 2(8024 \;\; 8012)$ under $2 = 3 \;\; 1$ degrees of freedom. Certain assumptions must be met for the statistic to follow a chi-squared distribution and often empirical p-values are computed.

5.9 RSquared for LME models

As a complement to this, one can also consider how to properly employ the R^2 measure, in the context of Methoc Comparison Studies, further to the work by Edwards et al, namely "An R^2 statistic for fixed effects in the linear mixed model".

Abstract for "An R^2 statistic for fixed effects in the linear mixed model"
Statisticians most often use the linear mixed model to analyze Gaussian longitudinal data.

The value and familiarity of the R2 statistic in the linear univariate model naturally creates great interest in extending it to the linear mixed model. We define and describe how to compute a model R2 statistic for the linear mixed model by using only a single model.

The proposed R2 statistic measures multivariate association between the repeated outcomes and the fixed effects in the linear mixed model. The R2 statistic arises as a 11 function of an appropriate F statistic for testing all fixed effects (except typically the intercept) in a full model.

The statistic compares the full model with a null model with all fixed effects deleted (except typically the intercept) while retaining exactly the same covariance structure.

Furthermore, the R2 statistic leads immediately to a natural definition of a partial R2 statistic. A mixed model in which ethnicity gives a very small p-value as a longitudinal predictor of blood pressure (BP) compellingly illustrates the value of the statistic.

In sharp contrast to the extreme p-value, a very small R^2 , a measure of statistical and scientific importance, indicates that ethnicity has an almost negligible association with the repeated BP outcomes for the study.

5.10 Remarks on the Multivariate Normal Distribution

Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution. ARoy2009's model is specified using the bivariate normal distribution. This gives rises to a key difference between the two model, in that a bivariate model accounts for covariance between the variables of interest. The multivariate normal distribution of a k-dimensional random vector $X = [X_1, X_2, \ldots, X_k]$ 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

$$\mu = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]$$

and $k \times k$ covariance matrix

$$\Sigma = [\mathrm{Cov}[X_i, X_j]],\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, k$$

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

   (a)

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

   (b)

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

5.10.1 Lin's Reproducibility Index

Lin proposes the use of a reproducibility index, called the Concordance Correlation Coefficent (CCC).While it is not strictly a measure of agreement as such, it can form part of an overall method comparision methodology.

Chapter 6

Bradley Blackwood

6.1 Bartko's Bradley-Blackwood Test

This is a regression based approach that performs a simultaneous test for the equivalence of means and variances of the respective methods.We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

$$D = (X_1 - X_2) \tag{6.1}$$

$$M = (X_1 + X_2)/2 \tag{6.2}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{6.3}$$

- The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits D on M, where D is the difference and average of a pair of results.
- Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
- We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.

- We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.
- Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept.

6.2 Bartko's Bradley-Blackwood Test

This is a regression based approach that performs a simultaneous test for the equivalence of means and variances of the respective methods.We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

$$D = (X_1 - X_2) \tag{6.4}$$

$$M = (X_1 + X_2)/2 \tag{6.5}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{6.6}$$

- The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits D on M, where D is the difference and average of a pair of results.
- Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
- We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.
- We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.
- Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept.

6.3 Bradley-Blackwood Test (Kevin Hayes Talk)

This work considers the problem of testing $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ using a random sample from a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

The new contribution is a decomposition of the Bradley-Blackwood test statistic (*Bradley and Blackwood, 1989*)for the simultaneous test of $\mu_1 = \mu_2$; $\sigma_1^2 = \sigma_2^2$ as a sum of two statistics.

One is equivalent to the Pitman-Morgan (*Pitman, 1939; Morgan, 1939*) test statistic for $\sigma_1^2 = \sigma_2^2$ and the other one is a new alternative to the standard paired-t test of $\mu_D = \mu_1 = \mu_2 = 0$.

Surprisingly, the classic Student paired-t test makes no assumptions about the equality (or otherwise) of the variance parameters.

The power functions for these tests are quite easy to derive, and show that when $\sigma_1^2 = \sigma_2^2$, the paired t-test has a slight advantage over the new alternative in terms of power, but when $\sigma_1^2 \neq \sigma_2^2$, the new test has substantially higher power than the paired-t test.

While Bradley and Blackwood provide a test on the joint hypothesis of equal means and equal variances their regression based approach does not separate these two issues.

The rejection of the joint hypothesis may be due to two groups with unequal means and unequal variances; unequal means and equal variances, or equal means and unequal variances.

We propose an approach for resolving this (model selection) problem in a manner controlling the magnitudes of the relevant type I error probabilities.

Deming Regression

- Informative analysis for the purposes of method comparison, Deming Regression is a regression technique taking into account uncertainty in both the independent and dependent variables.
- Demings method always results in one regression fit, regardless of which variable takes the place of the predictor variables.
- The measurement error (lambda or λ) is specified with measurement error variance related as

$$\lambda = \sigma_y^2 / \sigma_x^2$$

(where σ_x^2 and σ_y^2 is the measurement error variance of the x and y variables, respectively).

- In the case where λ is equal to one, (i.e. equal error variances), the methodology is equivalent to ***orthogonal regression***.
- Deming approaches the matter by simultaneously minimizing the sum of the square of the residuals of both variables. This derivation results in the best fit to minimize the sum of the squares of the perpendicular distances from the data points.
- To compute the slope by Demings formula, normally distributed error of both variables is assumed, as well as a constant level of imprecision throughout the range of measurements.

6.4 Simple Linear Regression

Simple linear regression is defined as such with the name 'Model I regression' by Cornbleet Gochman (1979), in contrast to 'Model II regression'.

On account of the fact that one set of measurements are linearly related to another, one could surmise that Linear Regression is the most suitable approach to analyzing comparisons. This approach is unsuitable on two counts. Firstly one of the assumptions of Regression analysis is that the independent variable values are without error. In method comparison studies one must assume the opposite; that there is error present in the measurements. Secondly a regression of X on Y would yield and entirely different result from Y on X.

Simple linear regression calculates a line of best fit for two sets of data, n which the independent variable, X, is measured without error, with y as the dependent variable.

SLR (Model I) regression is considered by many Altman and Bland (1983); Cornbleet and Cochrane (1979); Ludbrook (1997) to be wholly unsuitable for method comparison studies, although recommended for use in calibration studies [Corncoch]. Even in the case where one method is a gold standard , it is disputed as to whether it is a valid approach. Model II regression is more suitable for method comparison studies, but it is more difficult to execute. Both Model I and II regression models are unduly influenced by outliers. Regression Models can not be used to analyze repeated measurements

Regression Analysis

Another inappropriate approach is the regressing one set of measurements against the other. According to this methodology the measurement methods could considered equivalent if the confidence interval for the regression coefficient included 1. Analysts sometimes use least squares (referred to by Ludbrook as Model I) regression analysis to calibrate one method of measurement against another. In this technique, the sum of the squares of the vertical deviations of y values from the line is minimized. This approach is invalid, because both y and x values are attended by random error.

The Identity Plot

This is a simple graphical approach, advocated by Bland and Altman (1986), that yields a cursory examination of how well the measurement methods agree. In the case of good agreement, the covariates of the plot accord closely with the $X = Y$ line.

Advantages of Regression Approaches for MCS

- These methods can be employed in conversion problems.
- Bland and Altman have stated that regression analysis offers insights into MCS problems.

Disadvantages

- Regression methods are uninformative about the variability of the differences.
- Regression methods can determine the presence of bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002).

6.5 Constant and Proportional Bias

Linear Regression is a commonly used technique for comparing paired assays. The Intercept and Slope can provide estimates for the constant bias and proportional bias occurring between both methods. If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined. Outliers are a source of error in regression estimates.

Constant or proportional bias in method comparison studies using linear regression can be detected by an individual test on the intercept or the slope of the line regressed from the results of the two methods to be compared.

Bartko's Discussion of BB

Let $y = X_1 - X_2$ and $x = (X_1 - X_2)/2$. The Bradley-Blackwood procedure fits y on x, such that

$$y = \beta_0 + \beta_1 x$$

The slope and intercepte are given bu

$$beta_1 = \frac{(\sigma_1^2 = \sigma_2^2)}{2\sigma_x^2}$$

Pitman's Test on Correlated variances

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

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

Pitman's test is identical to the slope equal to zero in the regression of y on x.

6.6 Bradley-Blackwood Test (Kevin Hayes Talk)

This work considers the problem of testing $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ using a random sample from a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

The new contribution is a decomposition of the Bradley-Blackwood test statistic (*Bradley and Blackwood, 1989*)for the simultaneous test of $\mu_1 = \mu_2$; $\sigma_1^2 = \sigma_2^2$ as a sum of two statistics.

One is equivalent to the Pitman-Morgan (*Pitman, 1939; Morgan, 1939*) test statistic for $\sigma_1^2 = \sigma_2^2$ and the other one is a new alternative to the standard paired-t test of $\mu_D = \mu_1 = \mu_2 = 0$.

Surprisingly, the classic Student paired-t test makes no assumptions about the equality (or otherwise) of the variance parameters.

The power functions for these tests are quite easy to derive, and show that when $\sigma_1^2 = \sigma_2^2$, the paired t-test has a slight advantage over the new alternative in terms of power, but when $\sigma_1^2 \neq \sigma_2^2$, the new test has substantially higher power than the paired-t test.

While Bradley and Blackwood provide a test on the joint hypothesis of equal means and equal variances their regression based approach does not separate these two issues.

The rejection of the joint hypothesis may be due to two groups with unequal means and unequal variances; unequal means and equal variances, or equal means and unequal variances.

We propose an approach for resolving this (model selection) problem in a manner controlling the magnitudes of the relevant type I error probabilities.

6.7 Conclusions about Existing Methodologies

The Bland Altman methodology is well noted for its ease of use, and can be easily implemented with most software packages. Also it doesn't require the practitioner to have more than basic statistical training. The plot is quite informative about the variability of the differences over the range of measurements. For example, an inspection of the plot will indicate the 'fan effect'. They also can be used to detect the presence of an outlier.

Ludbrook (1997, 2002) criticizes these plots on the basis that they presents no information on effect of constant bias or proportional bias. These plots are only practicable when both methods measure in the same units. Hence they are totally unsuitable for conversion problems. The limits of agreement are somewhat arbitrarily constructed. They may or may not be suitable for the data in question. It has been found that the limits given are too wide to be acceptable. There is no guidance on how to deal with outliers. Bland and Altman recognize effect they would have on the limits of agreeement, but offer no guidance on how to correct for those effects.

There is no formal testing procedure provided. Rather, it is upon the practitioner opinion to judge the outcome of the methodology.

Introduction Quotes from Bartko's Paper

"Can two methods of measurement be used interchangeably?"

Bartko's Ellipse

$$\frac{x-\bar{x}}{\sigma_x^2} - \frac{2\rho(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{y-\bar{y}}{\sigma_y^2} = \chi^2(2df_(1-\rho^2)$$

section*Remarks

- Pearson's Correlation of (x,y) is the same as Pitman's correlation of sums and differences.
- Techniques for plotting an ellipse can be found in Douglas Altman's book.

6.8 A regression based approach based on Bland Altman Analysis

Bland and Altman have stated that regression analysis offers insights into method comparison studies. Regression methods can determine the presence of bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002). While they are informative about inter-method bias, Regression methods offer the analyst no insights into the relative precision of both methods. These methods can be employed in conversion problems, however errors are attended. ***Lu et al*** used such a technique in their comparison of DXA scanners. They also used the Blackwood Bradley test. However it was shown that, for particular comparisons, agreement between methods was indicated according to one test, but lack of agreement was indicated by the other.

Remarks

- Pearson's Correlation of (x,y) is the same as Pitman's correlation of sums and differences.
- Techniques for plotting an ellipse can be found in Douglas Altman's book.

6.9 The MCR R pacakge - Regression Techniques for MCS

The ***mcr*** packages provides a set of regression techniques to quantify the relation between two measurement methods.

In particular, it address regression problems with errors in both variables, but without repeated measurements. The ***mcr*** package follows the CLSI EP09-A3 recommendations for analytical method comparison and estimation of bias using patient samples.

*Methods featured in the **mcr** package*

- Deming Regression
- Weighted Deming Regression

- Passing-Bablock Regression

The *creatinine* gives the blood and serum preoperative creatinine measurements in 110 heart surgery patients.

```
library("mcr")
data("creatinine", package="mcr")
tail(creatinine)



fit.lr <- mcreg(as.matrix(creatinine), method.reg="LinReg", na.rm=TRUE)
fit.wlr <- mcreg(as.matrix(creatinine), method.reg="WLinReg", na.rm=TRUE)
compareFit( fit.lr, fit.wlr )
```

6.10 Implementation of Deming Regression with Rs

Thus far, one of the few R implementations of Deming regression is contained in the 'MethComp' package. (Carstensen et al., 2008).

Unless specified otherwise, the variance ratio λ has a default value of one. A means of computing likelihood functions would potentially allow for an algorithm for estimating the true variance ratio.

6.11 Linnet - References

The statistical procedures are described in: Linnet K. Necessary sample size for method comparison studies based on regression analysis. Clin Chem 1999; 45: 882-94. Linnet K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clin Chem 1998; 44: 1024-1031. Linnet K. Evaluation of regression procedures for methods comparison studies.

Clin Chem 1993; 39. 424-432. Linnet K. Estimation of the linear relationship between measurements of two methods with proportional errors. Stat Med 1990; 9: 1463-1473.

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

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

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

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

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

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 (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet 346*, 1085–87.

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

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

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

Carstensen, B. (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.

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

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

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

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.

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

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

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.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

Lin, S. C., D. M. Whipple, and charles S Ho (1991). Evaluation of statistical equivalence using limits of agreement and associates sample size calculation. *Communications in Statistics - Theory and Methods 27*(6), 1419–1432.

Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry 44*, 1024–1031.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

Luiz, R., A. Costa, P. Kale, and G. Werneck (2003). Assessment of agreement of a quantitative variable: a new graphical approach. *Journal of Clinical Epidemiology 56*, 963–967.

Mantha, S., M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss (2000). Comparing methods of clinical measurement: Reporting standards for bland and altman analysis. *Anaesthesia and Analgesia 90*, 593–602.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

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

Phelps, C. and A. Hutson (1995). Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical decision making 15*, 144–57.

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

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.

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.

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

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

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

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*(2), 153–177.

Chapter 7

Residual Diagnostics

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 repeated measures data. However, as a nave analysis, it may be used to explore the data because of the simplicity of the method. Myles states that such misuse of the standards Bland Altman method is widespread in Anaesthetic and critical care literature.

Bland and Altman have provided a modification for analysing repeated measures under stable or chaninging conditions, where repeated data is collected over a period of time. Myers proposes an alternative Random effects model for this purpose.

with repeated measures data, we can calculate the mean of the repeated measurements by each method on each individuals. *The pairs of means can then be used to compare the two methods based on the 95% limits of agreement for the difference of means. The bias between the two methods will not be affected by averaging the repeated measurements.*.However the variation of the differences will be underestimated by this practice because the measurement error is, to some extent, removed. Some advanced statistical calculations are needed to take into account these measurement errors. *Random effects models can be used to estimate the within-subject variation after accounting for other observed and unobserved variations, in which each subject has a different intercept and slope over the observation period .On the basis of the within-subject variance estimated by the random effects model, we can then create an appropriate*

*Bland Altman Plot.*The sequence or the time of the measurement over the observation period can be taken as a random effect.

7.1 Random effects Model

Myles (2007) proposes the use of Random effects models to address the issue of repeated measurement. Myles proposes a formulation of the BlandAltman plot, using the within-subject variance estimated by the random effects model, with the time of the measurement taken as a random effect. He states that *random effects models account for the dependent nature of the data, and additional explanatory variables, to provide reliable estimates of agreement in this setting.*
Agreement between methods is reflected by the between-subject variation.The Random Effects Model takes this into account before calculating the within-subject standard deviation.

7.1.1 Myers Random Effects Model

The presentation of the 95% limits of agreement is for visual judgement of how well two methods of measurement agree. The smaller the range between the two, the better the agreement is The question of small is small is a question of clinical judgement

Repeated measurements for each subjects are often used in clinical research.

7.1.2 Random Effects Modelling

Random effects models are used to examine the within-subject variation after adjusting for known and unknown variables, in which each subject has a different intercept and slope over a time period period.

Myles (2007) remarks that the random effects model is an extension of the analysis of variance method, accounting for more covariates.

A random effect (in Myles's case, time of measurement) is chosen to reflect the different intercept

and slope for each subject with respect to their change of measurements over the time period.

In Myles's methodology, the standard deviation of difference between the means of the repeated measurements can be calculated based on the within-subject standard deviation estimates.

A random effects model (also variance components model)is a type of hierarchical linear model. Hierarchical linear modelling (HLM) is a more advanced form of simple linear regression and multiple linear regression. HLM is appropriate for use with nested data.
Faraway comments that the random effects approach is *more ambitious than the LME model in that it attempts to say something about the wider population beyond the particular sample.*

7.2 Residual

A residual (or fitting error), on the other hand, is an observable estimate of the unobservable statistical error. Residual (or error) represents unexplained (or residual) variation after fitting a regression model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model. Consider the previous example with men's heights and suppose we have a random sample of n people. The sample mean could serve as a good estimator of the population mean. Then we have:

The difference between the observed value of the dependent variable (y) and the predicted value () is called the residual (e). Each data point has one residual.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. .

The difference between the height of each man in the sample and the unobservable population mean is a statistical error, whereas The difference between the height of each man in the sample and the observable sample mean is a residual. Note that the sum of the residuals within a random sample is necessarily zero, and thus the residuals are necessarily not independent. The statistical errors on the other hand are independent, and their sum within the random sample is almost surely not zero.

Other uses of the word "error" in statistics:

The use of the term "error" as discussed in the sections above is in the sense of a deviation of a value from a hypothetical unobserved value. At least two other uses also occur in statistics, both referring to observable prediction errors:

- Mean square error or mean squared error (abbreviated MSE) and root mean square error (RMSE) refer to the amount by which the values predicted by an estimator differ from the quantities being estimated (typically outside the sample from which the model was estimated).

- Sum of squared errors, typically abbreviated SSE or SSe, refers to the residual sum of squares (the sum of squared residuals) of a regression; this is the sum of the squares of the deviations of the actual values from the predicted values, within the sample used for estimation. Likewise, the sum of absolute errors (SAE) refers to the sum of the absolute values of the residuals, which is minimized in the least absolute deviations approach to regression.

Cox and Snell (1968, JRSS-B): general definition of residuals for models with single source of variability Hilden-Minton (1995, PhD thesis UCLA), Verbeke and Lesaffre (1997, CSDA) or Pinheiro and Bates (2000, Springer): extension to define three types of residuals that accommodate the extra source of variability present in linear mixed models, namely:

i) Marginal residuals,

predictors of marginal errors,

ii) Conditional residuals,

$$be = yX\hat{\beta}Zbb = \hat{\sigma}Q\hat{y}$$

, predictors of conditional errors

$$e = yE[y|b] = yX\beta Zb$$

iii) BLUP, Zbb, predictors of random effects,

$$Zb = E[y|b]E[y]$$

7.2.1 Residual Plots

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

In the next lesson, we will work on a problem, where the residual plot shows a non-random pattern. And we will show how to "transform" the data to use a linear model with nonlinear data.

7.3 Studentization

In statistics, a studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation. Typically the standard deviations of residuals in a sample vary greatly from one data point to another even when the errors all have the same standard deviation, particularly in regression analysis; thus it does not make sense to compare residuals at different data points without first studentizing. It is a form of a Student's t-statistic, with the estimate of error varying between points.

This is an important technique in the detection of outliers. It is named in honor of William Sealey Gosset, who wrote under the pseudonym Student, and dividing by an estimate of scale is called studentizing, in analogy with standardizing and normalizing: see Studentization.

7.4 Cooks's Distance - Implementation with R

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.

7.5 Influence measures using R

R provides the following influence measures of each observation.

	dfb.1_	dfb.A	dffit	cov.r	cook.d	hat
1	0.42	-0.42	-0.56	1.13	0.15	0.18
2	0.17	-0.17	-0.34	1.14	0.06	0.11
3	0.01	-0.01	-0.24	1.17	0.03	0.08
4	-1.08	1.08	1.57	0.24	0.56	0.16
5	-0.14	0.14	-0.24	1.30	0.03	0.13
6	-0.00	0.00	-0.11	1.31	0.01	0.08
7	-0.04	0.04	-0.08	1.37	0.00	0.11
8	0.02	-0.02	0.15	1.28	0.01	0.09
9	0.69	-0.68	0.75	2.08	0.29	0.48
10	0.18	-0.18	-0.22	1.63	0.03	0.27
11	-0.03	0.03	-0.04	1.53	0.00	0.19
12	-0.25	0.25	0.44	1.05	0.09	0.12

7.6 LME diagnostic measures

7.6.1 Andrews-Pregibon statistic

- For fixed effect parameters β.

The Andrews-Pregibon statistic AP_i is a measure of influence based on the volume of the confidence ellipsoid. The larger this statistic is for observation i, the stronger the influence that observation will have on the model fit.

7.6.2 Cook's Distance

- For variance components γ

Diagnostic tool for variance components

$$C_{\theta i} = (\hat{(\theta)}_{[i]} - \hat{(\theta)})^T \text{cov}(\hat{(\theta)})^{-1} (\hat{(\theta)}_{[i]} - \hat{(\theta)})$$

7.6.3 Variance Ratio

- For fixed effect parameters β.

7.6.4 Cook-Weisberg statistic

- For fixed effect parameters β.

7.6.5 Andrews-Pregibon statistic

- For fixed effect parameters β.

The Andrews-Pregibon statistic AP_i is a measure of influence based on the volume of the confidence ellipsoid. The larger this statistic is for observation i, the stronger the influence that observation will have on the model fit.

7.7 Residual Diagnostics

Consider a residual vector of the form $\hat{e} = \boldsymbol{PY}$, where $\boldsymbol{P}$ is a projection matrix, possibly an oblique projector. External studentization uses an estimate of Var that does not involve the ith observation. Externally studentized residuals are often preferred over studentized residuals because they have well known distributional properties in the standard linear models for independent data. Residuals that are scaled by the estimated variances of the responses are referred to as Pearson-type residuals. Standardization:

$$\frac{\hat{e}_i}{\sqrt{v_i}}$$

Studentization

$$\frac{\hat{e}_i}{\sqrt{\hat{v}_i}}$$

7.8 Why use 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). 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.

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

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

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

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

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

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.

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

7.9 Two-tailed testing

A test for equality of variances, based on the likelihood Ratio test, is very simple to implement using existing methodologies. All that is required it to specify the reference model and the relevant nested mode as arguments to the command `anova()`. The output can be interpreted in the usual way.

7.10 One Tailed Testing

The approach proposed by Roy deals with the question of agreement, and indeed interchangeability, as developed by Bland and Altman's 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.

7.11 Enabling One Tailed Testing

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 (or alternatively, the ratio of the variances). 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. However , to facilitate one tailed testing, What is required is the computation of the variance ratios of within-item and between-item standard deviations.

A nave approach would be to compute the variance ratios by relevant F distribution quantiles. However, the question arises as to the appropriate degrees of freedom. However, Douglas Bates has stated that an alternative approach is required (i.e. Profile Likelihoods)

"The omission of standard errors on variance components is intentional. The distribution

of an estimator of a variance component is highly skewed and obtaining an estimate of the standard deviation of a skewed distribution is not very useful. A much better approach is based on profiling the objective function." (Douglas Bates May 2012)

7.12 Profile Likelihood

Normal-based confidence intervals for a parameter of interest are inaccurate when the sampling distribution of the estimate is skewed. The technique known as profile likelihood can produce confidence intervals with better coverage. It may be used when the model includes only the variable of interest or several other variables in addition. Profile-likelihood confidence intervals are particularly useful in nonlinear models.

Profile likelihood confidence intervals are based on the log-likelihood function.

7.13 Implementation of PL Confidence Intervals

The suitable calculation of confidence limits for this variance ratio are to be computed using the profile likelihood approach. The `R` package `profilelikelihood` will be assessed for feasibility, particularly the command `profilelikelihood.lme()`

Normal-based con

dence intervals for a parameter of interest are inaccurate when the sampling distribution of the estimate is skewed. The technique known as profile likelihood can produce confidence intervals with better coverage. It may be used when the model includes only the variable of interest or several other variables in addition. Profile-likelihood confidence intervals are particularly useful in nonlinear models. Profile likelihood confidence intervals are based on the log-likelihood function.

7.14 residuals.lme nlme- Extract lme Residuals

The residuals at level i are obtained by subtracting the fitted levels at that level from the response vector (and dividing by the estimated within-group standard error, if `type="pearson"`).

The fitted values at level i are obtained by adding together the population fitted values (based only on the fixed effects estimates) and the estimated contributions of the random effects to the fitted values at grouping levels less or equal to i.

```
fm1 <- lme(distance ~ age + Sex,
data = Orthodont, random = ~ 1)
head(residuals(fm1, level = 0:1))
summary(residuals(fm1) /
residuals(fm1, type = "p"))

# constant scaling factor 1.432
```

7.15 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.

`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 returns the fixed parameters of these iteratively modified models. These are used to compute measures of influential data.

7.16 Computing DFBETAs with R

- This function computes the DFBETAS based on the information returned by the estex() function.
- 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.
- The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the th observation:

  Missing Formula

  where is the th element of . 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 as a size-adjusted cutoff.

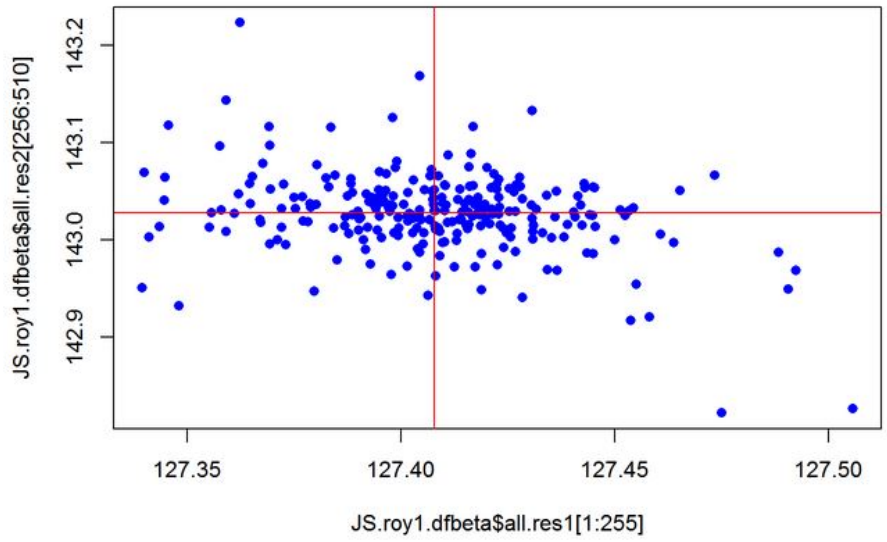

Figure 7.17.1:

7.17 DFbetas for Blood Data

```
plot(JS.ARoy20091.dfbeta$all.res1[1:255],JS.ARoy20091.dfbeta$all.res2[256:510],
pch=16,col="blue")
abline(v=JS.ARoy20091.dfbeta$all.res1[256],col="red")
abline(h=JS.ARoy20091.dfbeta$all.res2[1],col="red")
```

7.18 Diagnostic Tools for the nlme package

With the nlme package, the generic function `lme()` fits a linear mixed-effects model in the formulation described in Laird and Ware (1982) but allowing for nested random effects.

The within-group errors are allowed to be correlated and/or have unequal variances, which is very important in fitting the models for Roy's Tests

The nlme package has a limited set of diagnostic tools that can be used to assess the model fit. A review of the package manual is sufficient to get a sense of the package's capability in that regard.

7.19 The `logLik` Function

`logLik.lme` returns the log-likelihood value of the linear mixed-effects model represented by object evaluated at the estimated coefficients. It is also possible to determine the restricted log-likelihood, if relevant, using this function. For the Blood Data Example, the loglikelihood of the JS.roy1 model can be computed as follows.

```
> logLik(JS.roy1)
'log Lik.' -2030.736 (df=8)
```

7.20 Influence() - Description

`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 returns the fixed parameters of these iteratively modified models.

These are used to compute measures of influential data.

7.21 Leave-One-Out Diagnostics with `lmeU`

Galecki et al discuss the matter of LME influence diagnostics in their book, although not into great detail.

The command `lmeU` fits a model with a particular subject removed. The identifier of the subject to be removed is passed as the only argument

A plot ofthe per-observation diagnostics individual subject log-likelihood contributions can be rendered.

7.22 Partitioning Matrices

Without loss of generality, matrices can be partitioned as if the $i-$th omitted observation is the first row; i.e. $i = 1$.

7.23 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 regrading 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*".

7.24 Zewotir: Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components σ_j^2, compute an estimate for $\boldsymbol{V}$ and then compute the projector matrix A, $\boldsymbol{X\hat{\beta}} = \boldsymbol{AY}$.

Zewotir remarks that $\boldsymbol{D}$ is a block diagonal with the $i-$th block being $u\boldsymbol{I}$

7.25 Haslett Hayes

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

A general theory is presented for residuals from the general linear model with correlated errors. It is demonstrated that there are two fundamental types of residual associated with this model, referred

to here as the marginal and the conditional residual. These measure respectively the distance to the global aspects of the model as represented by the expected value and the local aspects as represented by the conditional expected value. These residuals may be multivariate.

In contrast to classical linear models, diagnostics for LME are difficult to perform and interpret, because of the increased complexity of the model

7.26 Confounded Residuals

Hilden-Minton (1995, PhD thesis, UCLA): residual is pure for a specific type of error if it depends only on the fixed components and on the error that it is supposed to predict Residuals that depend on other types of errors are called ***confounded residuals***

Chapter 8

Fitting LME Models

Further to previous material, an appraisal of the current state of development for statistical software for fitting for LME models, particularly for `nlme` and `lme4` fitted models.

The **lme4** pacakge is used to fit linear and generalized linear mixed-effects models in the R environment. The **lme4** package is also under active development, under the leadership of Ben Bolker (McMaster Uni., Canada).

Crucially, a review of internet resources indicates that almost all of the progress in this regard has been done for `lme4` fitted models, specifically the *Influence.ME* `R` package. (Nieuwenhuis et al 2014) Conversely there is very little for `nlme` models. One would immediately look at the current development workflow for both packages.

As an aside, Douglas Bates was arguably the most prominent `R` developer working in the LME area. However Bates has now prioritised the development of LME models in another computing environment , i.e Julia.

With regards to `nlme`, the package is now maintained by the `R` core development team. The most recent major text is by Galecki & Burzykowski, who have published *Linear Mixed Effects Models using R*. Also, the accompanying `R` package, nlmeU package is under current development, with a version being released $0.70 - 3$.

8.1 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.

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. Carstenen advises that repeated measurements come in two *substantially different* forms, depending on the circumstances of their measurement: exchangable and linked.

8.1.1 Exchangeable measurements

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.

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.

Exchangeable repeated measurements can be treated as true replicates.

8.1.2 Linked measurements

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.

8.1.3 Replicate measurements in ARoy2009's paper

Roy (2009) takes its definition of replicate measurement: two or more measurements on the same item taken under identical conditions. ARoy2009 also assumes linked measurements, but it is can be used for the non-linked case.

8.1.4 Random effects

Further to Barnhart et al. (2007), 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}. Carstensen et al. (2008) also addresses this issue by the addition of an interaction term (i.e. a random effect) $u_m i$, yielding

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

The additional interaction term is characterized as $u_{mi} \sim \mathcal{N}(0, \tau_m^2)$ (Carstensen et al., 2008).

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

Carstensen et al. (2008) 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.

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.

Roy (2009) also assigns a random effect u_{mi} for each response y_{mir}. Importantly ARoy2009's model assumes linkage.

8.2 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 Barnhart et al. (2007) fixed effect can be expressed with a single term α_{mi}, which incorporate the true value μ_i.

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

Combining fixed effects (Barnhart et al., 2007), we write,

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

The following assumptions are required

- e_{mir} is independent of the fixed effects with mean $\mathrm{E}(e_{mir}) = 0$.
- Further to Barnhart et al. (2007) between-item and within-item variances $\mathrm{Var}(\alpha_{mi}) = \sigma^2_{Bm}$ and $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$
- In keeping with Roy (2009), these variance shall be considered as part of the between-item variance covariance matrix $\boldsymbol{D}$ and the within-item variance covariance matrix $\boldsymbol{\Sigma}$ respectively, and will be denoted accordingly (i.e. d^2_m and σ^2_m).
- Additionally, the total variability of method "m", denoted ω^2_m is the sum of the within-item and between-item variabilities.

$$\omega^2_m = d^2_m + \sigma^2_m$$

8.3 Lai Shiao

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

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

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

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

Single fixed effect

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

$$\text{fixed effects}: 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \tag{8.1}$$
$$(\text{p-values}: = 0.0054, < 0.0001, < 0.0001)$$

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

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;

$$\text{fixed effects}: -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl}$$
$$(\text{p-values}: = 0.8113, < 0.0001, < 0.0001)$$

$$\text{random effects}: u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577)$$
$$(\text{p-values}: = 0.0446, < 0.0001, < 0.0001) \tag{8.2}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically signif-

icant. 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.

Chapter 9

BXC

9.1 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{9.1}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.2}$$

9.2 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 ARoy2009'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 ARoy2009'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 ARoy2009'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 ARoy2009'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 rigourous approach to computing Limits of Agreement. The respective papers also discuss several shortcoming for techniques for dealing with replicate measurements, as proposed by Bland-Altman 1999.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{9.3}$$

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

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

m;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{9.4}$$

. Under the assumption that the μs are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr} \sim N(0, \omega_m^2)$to account for this.

Carstensen (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 (μ_i). 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. A distinction can be made between the two models: ARoy2009's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

Let y_{mir} denote the rth replicate measurement on the ith item by the mth method, where $m = 1, 2$; $i = 1, \ldots, N$; and $r = 1, \ldots, n_i$. When the design is balanced and there is no ambiguity we can set $n_i = n$. The LME model underpinning ARoy2009's approach can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{9.5}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009) 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, ARoy2009 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

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \tag{9.6}$$

The fixed effects α_m and μ_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 $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. 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 $m = \{1, 2, 3 \ldots\}$. 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 τ_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 (9.5) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model in (9.6) requires $N + 2$ fixed effects.

Allocating fixed effects to each item i by (9.6) 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.

9.3 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.

9.4 Computing LoAs with LMEs

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_m.

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

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

9.5 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 rth replicate measurement on subject i with method m.

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

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

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The

import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that 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 rth replicate measurement on the ith item by the mth method, where $m = 1, 2, \ldots, M$ $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + a_{ir} + \epsilon_{mir}, \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), a_{ir} \sim \mathcal{N}(0, \varsigma^2), \varepsilon_{mi} \sim \mathcal{N}(0, \varphi_m^2). \tag{9.8}$$

Here the terms α_m and μ_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 ε_{mir}. The c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$.

The random error term for each response is denoted ε_{mir} having $\mathrm{E}(\varepsilon_{mir}) = 0$, $\mathrm{Var}(\varepsilon_{mir}) = \varphi_m^2$. 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 τ_m^2 can not be obtained. Instead the average value τ^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 ARoy2009'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 μ_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, ARoy2009 considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: ARoy2009's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

9.6 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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.9}$$

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

$$var(y_{1j} - y_{2j}) \tag{9.10}$$

9.6.1 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 for all m

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

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

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

Carstensen page 58

var(y10-y20) =12+22+12+22

1-2222+12+22

ARoy2009 further to Carstensen

ymir=m+i+cmi+emir

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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{9.11}$$

Carstensen et al [5] 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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.12}$$

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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{9.13}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \tag{9.14}$$

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.15}$$

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \tag{9.16}$$

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

The intercept term α 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 μ_i. c_{mi} is a interaction term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis.

There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. ***Exchangeability*** means that future samples from a population behaves like earlier samples).

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.

9.6.2 Tau Identifibaility

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_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 τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{9.17}$$

9.6.3 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 μs 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.

9.6.4 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 rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{9.18}$$

The intercept term α and the $\beta_m \mu_i$ term follow from Dunn[7], 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 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.

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

9.6.5 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.

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 rth replicate measurement on the ith item by the mth method, where $m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{9.19}$$

Here the terms α_m and μ_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 ϵ_{mir}. The c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$. Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$. 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 μ_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, ARoy2009 considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: ARoy2009's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

9.7 Carstensen 2004 's Mixed Models

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

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{9.20}$$

. Under the assumption that the μs are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr}d_{mr} \sim N(0, \omega_m^2)$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 10

BXC Limits of Agreement

10.1 Intervals

10.1.1 Purpose of Limits of Agreement

It must be established clearly the specific purpose of the limits of agreement. Bland and Altman (1995) comment that the limits of agreement *how far apart measurements by the two methods were likely to be for most individuals.*, a definition echoed in their 1999 paper:

> We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits of agreement. These values define the range within which most differences between measurements by the two methods will lie(Bland and Altman, 1999).

? offers an alternative, more specific, definition of the limits of agreement *"a prediction interval for the difference between future measurements with the two methods on a new individual."* Luiz et al. (2003) describes them as tolerance limits.

Importantly they have the same construction as Shewhart Control limits.

Chapter 11

BXC materials

11.1 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 simulataneously for all patients.

11.1.1 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 $(-12.18, 17.12)$.

11.1.2 Using LME models to create Prediction Intervals

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{11.1}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{11.2}$$

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{11.3}$$

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \tag{11.4}$$

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

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).

11.1.3 Carstensen's LOAs

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_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.

11.2 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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{11.5}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{11.6}$$

Roy (2009) has demonstrated a methodology whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. Using ARoy2009's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{11.7}$$

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 ARoy2009's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

For Carstensen's 'fat' data, the limits of agreement computed using ARoy2009's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

11.2.1 Limits of agreement for Carstensen's data

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 $(-0.23, 0.32)$

11.3 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 ARoy2009al 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) 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) 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 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 $(-12.18, 17.12)$.

Limits of agreement are determined using ARoy2009's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. ARoy2009's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of ARoy2009'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 ARoy2009's model and the modified model (denoted 1 and 2 respectively);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{11.8}$$

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

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

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

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

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

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. Therefore the test's proposed by Roy (2009) can not be implemented. Conversely, accurate limits of agreement as determined by Carstensen's model may also be found using ARoy2009's method. Addition of the interaction term erodes the capability of ARoy2009'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 $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

11.4 RV-IV

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

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \tag{11.9}$$

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

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \tag{11.10}$$

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

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \tag{11.11}$$

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.

Chapter 12

Repeatability

12.1 Coefficient of Repeatability

The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999). Once the within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

Chapter 13

Alternative agreement indices

13.1 Coverage Probability and Tolerance Deviation Index

Individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI) as proposed by Lin (2000) and Lin et al. (2002).

If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{13.1}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. his boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

The CP is the most intuitively clear approach; it mirrors the information provided by the TDI. Both TDI and CP depend on the normality assumption and offer better power for inference than the CCC. The CP would have difficulty discriminating among instruments or assays that have excellent agreement, all because the CP values would be very close to 1. In this case, the TDI can be used to discriminate among these. When a meaningful clinical range is known and the study is conducted over that range, the CCC offers a meaningful geo- metric interpretation and is unit free. Furthermore, the

accuracy and precision components of the CCC offer more insight. Therefore, the CCC, accuracy, and precision remain very useful tools. Note that when Y and X are not linearly related, the CCC will capture the total deviation. However, it will treat the nonlinear deviation as imprecision rather than inaccuracy. The CCC, ICC, and Pearson correlation coef cient depend largely on the analytical range and the intrasample variation.

13.2 Mean Square Deviation

Mean Square deviation is defined as the expectation of the squared difference of two readings. The MSD is usually used for the case of two methods, each making a single reading.

13.3 Probability Based Approachs to MCS

Coverage Probability and Total Deviation Index

As elaborated by Lin and colleagues (Lin, 2000; Lin et al., 2002), an intuitive measure of agreement is a measure that captures a large proportion of data within a boundary for allowed observers differences.

The proportion and boundary are two quantities that correspond to each other. If we set d0 as the predetermined boundary; i.e., the maximum acceptable absolute difference between two observers readings, we can compute the probability of absolute difference between any two observers readings less than d0.

This probability is called coverage probability (CP). On the other hand, if we set $SYMBOL$ as the predetermined coverage probability, we can find the boundary so that the probability of absolute difference less than this boundary is ?.

This boundary is called total deviation index (TDI) and is the 100% percentile of the absolute difference of paired observations. A satisfactory agreement may require a large CP or, equivalently, a small TDI.

Coverage probability (CP)

Another user friendly measure of agreement which is related to the computation of the TDI is the so called coverage probability (CP) [11,12]. The CP describes the proportion captured within a pre-specified boundary of the absolute paired-measurement differences from two devices, i.e., the value of pκ such that P(—D— ¡ κ) = pκ. Therefore one can find pκ for a specified boundary κ using standard methods for computing probability quantities under normal assumptions [11]:

(13) and to obtain a CP estimate, pκ can be computed by replacing μ_D and σ_D by their REML estimate counterparts derived from model (1).

As with the TDI, the CP criterion can also be translated into a hypothesis test specification. In this case the interest is to ensure that a specified boundary of the absolute paired-measurement differences captures at least a predetermined proportion, p0:

The proposed TI method for inference about the TDI can be utilized to perform inferences about the CP estimates. From the TI in (10) it follows that

(14) Now κ is a fixed known boundary, and our interest lies in finding a lower confidence bound for the CP estimate. Thus, one can find a lower confidence bound for a non-central Student-t proportion with confidence level 1 - α by searching the non-centrality parameter, that depends on and hence on pκ, that satisfies

(15) and once the non-centrality parameter is achieved, a lower bound about the proportion pκ is found using equation (5),

However, the non-centrality parameter cannot be found in a closed form, so one may use again a modified version of the binary search algorithm as follows:

1. begin with the interval [low = 0; high = 1], as pκ is bounded by the interval (0,1);
2. calculate the midpoint of the interval *mid = (low + high)/2* and compute the difference ;
3. if d is greater than 0 up to a tolerance bound δ (i.e.,), then recalculate the interval [low = mid + δ; high = 1]; if it is lower than 0 up to a tolerance bound δ (i.e.), then recalculate the interval [low = 0; high = mid - δ];
4. repeat steps 2-3 until convergence, i.e. until d satisfies .

Coverage probability (CP)

Another user friendly measure of agreement which is related to the computation of the TDI is the so called coverage probability (CP) [11,12]. The CP describes the proportion captured within a pre-specified boundary of the absolute paired-measurement differences from two devices, i.e., the value of pκ such that $P(|D| < \kappa) = p_\kappa$. Therefore one can find pκ for a specified boundary κ using standard methods for computing probability quantities under normal assumptions [11]:

(13) and to obtain a CP estimate, pκ can be computed by replacing μ_D and σ_D by their REML estimate counterparts derived from model (1).

As with the TDI, the CP criterion can also be translated into a hypothesis test specification. In this case the interest is to ensure that a specified boundary of the absolute paired-measurement differences captures at least a predetermined proportion, p0:

The proposed TI method for inference about the TDI can be utilized to perform inferences about the CP estimates. From the TI in (10) it follows that

(14) Now κ is a fixed known boundary, and our interest lies in finding a lower confidence bound for the CP estimate. Thus, one can find a lower confidence bound for a non-central Student-t proportion with confidence level 1 - α by searching the non-centrality parameter, that depends on and hence on pκ, that satisfies

(15) and once the non-centrality parameter is achieved, a lower bound about the proportion pκ is found using equation (5),

However, the non-centrality parameter cannot be found in a closed form, so one may use again a modified version of the binary search algorithm as follows:

1. begin with the interval [low = 0; high = 1], as pκ is bounded by the interval (0,1);
2. calculate the midpoint of the interval *mid = (low + high)/2* and compute the difference ;
3. if d is greater than 0 up to a tolerance bound δ (i.e.,), then recalculate the interval [low = mid + δ; high = 1]; if it is lower than 0 up to a tolerance bound δ (i.e.), then recalculate the interval [low = 0; high = mid - δ];
4. repeat steps 2-3 until convergence, i.e. until d satisfies .

13.4 Probability Based Methods

Probability Based Approachs to MCS

Coverage Probability and Total Deviation Index

As elaborated by Lin and colleagues (Lin, 2000; Lin et al., 2002), an intuitive measure of agreement is a measure that captures a large proportion of data within a boundary for allowed observers differences.

- The proportion and boundary are two quantities that correspond to each other.
- If we set d0 as the predetermined boundary; i.e., the maximum acceptable absolute difference between two observers readings, we can compute the probability of absolute difference between any two observers readings less than d0.
- This probability is called coverage probability (CP). On the other hand, if we set as ... the predetermined coverage probability, we can find the boundary so that the probability of absolute difference less than this boundary is ....
- This boundary is called **total deviation index (TDI)** and is the 100th percentile of the absolute difference of paired observations.
- A satisfactory agreement may require a large CP or, equivalently, a small TDI.

13.5 Alternative agreement indices

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation is assessing agreement. The mean square deviation is defined as the expectation of the squared differences of two readings . The MSD is usually used for the case of two measurement methods X and Y , each making one measurement for the same subject, and is given by

$$MSDxy = E[(x-y)^2] = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1-\rho_{xy}).$$

? advises the use of a predetermined upper limit for the MSD value, MSD_{ul}, to define satisfactory agreement. However, a satisfactory upper limit may not be properly determinable, thus creating a drawback to this methodology.

? proposes both the use of the square root of the MSD or the expected absolute difference (EAD) as an alternative agreement indices. Both of these indices can be interpreted intuitively, being denominated in the same units of measurements as the original measurements. Also they can be compare to the maximum acceptable absolute difference between two methods of measurement d_0.

$$EAD = E(|x-y|) = \frac{\sum |x_i - y_i|}{n}$$

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions.

? remarks that a comparison of EAD and MSD , using simulation studies, would be interesting, while further adding that 'It will be of interest to investigate the benefits of these possible new unscaled agreement indices'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement,

	F vs C	F vs T
Inter-method bias	-0.61	0.12 3
Difference variances	0.06	0.22
Limits of agreement	(-1.08, -0.13)	(-0.81,1.04)
EAD	0.61	0.35

Table 13.5.1: Agreement indices for Grubbs' data comparisons.

the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{13.2}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. This boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

13.6 Coverage Probability and Tolerance Deviation Index

Individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI) as proposed by Lin (2000) and Lin et al. (2002).

If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{13.3}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. his boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

13.7 Mean Square Deviation

Mean Square deviation is defined as the expectation of the squared difference of two readings. The MSD is usually used for the case of two methods, each making a single reading.

13.8 Total Deviation Index and Coverage Probability

Lin et al. (2002) proposes a measure called the 'Total Deviation Index'. This assumes that the differences of paired measurements are a random sample from a normal distribution, and consequently the approach is to construct a probability interval, known as a tolerance interval, for these differences. A tolerance interval is a statistical range within which a specified proportion of the population lies. Smaller values of q indicate better agreement. P_0 is specified by the practitioner.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements. These methodologies have been adopted by Mayo Clinic (Research Section).

This measure was coined by Lin as the value $TDI_{1-p} = \kappa$ that a given fraction (1-p) of the differences between two measurement methods will be in a symmetric interval $[-\kappa, kappa]$. This is roughly equivalently to the numerically largest of the 1-p limits of agreement. The measure clearly has its main applicability in equivalence testing.

Lin gives an approximate formula for the calculations.

$$\Theta\left(\frac{TDI - \mu_d}{\sigma_d}\right) - \Theta\left(\frac{-TDI - \mu_d}{\sigma_d}\right) = 1 - p$$

Again, the assumption of the normality of the case-wise differences is relied upon.

The approach is illustrated in a real case example where the agreement between two instruments, a handle mercury sphygmomanometer device and an OMRON 711 automatic device, is assessed in a sample of 384 subjects where measures of systolic blood pressure were taken twice by each device. A simulation study procedure is implemented to evaluate and compare the accuracy of the approach

to two already established methods, showing that the TI approximation produces accurate empirical confidence levels which are reasonably close to the nominal confidence level.

13.9 Unscaled Agreement Indices

- Summary agreement indices based on the absolute difference of readings by observers are grouped here as unscaled agreement indices.
- They are usually defined as the expectation of a function of the difference, or features of the distribution of the absolute difference.
- These indices include mean squared deviation, repeatability coefficient, repeatability variance, reproducibility variance (ISO), limits of agreement (Bland and Altman, 1999), coverage probability (CP) and total deviation index (TDI) (Lin et al., 2002 Choudhary and Nagaraja, 2007; Choudhary, 2007a).

13.10 Information Approach

PURPOSE: Disagreement on the interpretation of diagnostic tests and clinical decisions remains an important problem in medicine. As no strategy to assess agreement seems to be fail-safe to compare the degree of agreement, or disagreement,

13.10.1 Example: Systolic Blood Pressure

Bland and Altman (19) present the example of measurements of systolic blood pressure of 85 individuals, by two observers (observer J and observer R) with sphygmomanometer, and one other measurement, by a semiautomatic device (device S). Luiz et al. (16) re-analyze the data and also observe, with a graphical approach, a greater agreement between the two observers than between the observers and the semiautomatic device. Using our information-based measure of disagreement; we also obtained a significantly more disagreement between each observer and the semiautomatic device than between the two observers (Table 1).

13.10.2 Discussion

- We can look at disagreement between observers as the distance between their ratings, so the metric properties are important. Moreover, the proposed measure of disagreement is scale-invariant, i.e., the degree of disagreement between two observers should be the same if the measurements are analyzed in kilograms or in grams, for example.

- Differential weighting is another property of the proposed information-based measure of disagreement: each comparison between two ratings is divided by a normalizing factor, depending on each pair of ratings alone, before summing. Therefore, the information-based measure of disagreement is appropriate for ratio scale measurements (with a natural 0) and it is not appropriate for interval scale measurements (without a natural 0).

- For example, outside air temperature in Celsius (or Fahrenheit) scale does not have a natural 0. The 0 is arbitrary and it does not make sense to say that 20 is twice as hot as 10. Outside air temperature in Celsius (or Fahrenheit) scale is an interval scale. On the other hand, height has a natural 0 meaning: the absence of height. Therefore, it makes sense to say that 80 inches is twice as large as 40 inches. Height is a ratio scale.

- Suppose the heights of a sample of subjects measured independently by two different observers. A difference between the two observers of 1 inch in a child subject represents a worse observers' error than a disagreement between observers of 1 inch in an adult subject.

- Due to differential weighting property of the information-based measure of disagreement, a difference between the observers of one inch in a child in fact weights less to the estimate of information-based measure of disagreement between observers than a difference between the observers of 1 inch in an adult.

- The usual approaches used to evaluate agreement have the limitation of the comparability of populations. In fact, ICC depends on the variance of the trait in the population; although this characteristic can be considered an advantage it does not permit one to compare the degree of agreement across different populations. Also the interpretation of the limits of agreement depends

on what can be considered clinically relevant or not, which could be subjective and different from reader to reader.

- The comparison of the degree of agreement in different populations is not straightforward. Other approaches 16 and 17 to assess observer agreement have been proposed, however the comparability of populations is still not easy with these approaches.

- The proposed information-based measure of disagreement, used as a complement to current approaches for evaluating agreement, can be useful to compare the degree of disagreement among different populations with different characteristics, namely with different variances.

- Moreover, we believe that information theory can make an important contribution to the relevant problem of measuring agreement in medical research, providing not only better quantification but also better understanding of the complexity of the underlying problems related to the measurement of disagreement.

13.10.3 Coverage probability

This term refers to the probability that a procedure for constructing random regions will produce an interval containing, or covering, the true value. It is a property of the interval producing procedure, and is independent of the particular sample to which such a procedure is applied. We can think of this quantity as the chance that the interval constructed by such a procedure will contain the parameter of interest.

13.11 Coverage Probability

Further to Lin (2000) and Lin et al. (2002), individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{13.4}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. This boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

13.12 Coverage probability

This term refers to the probability that a procedure for constructing random regions will produce an interval containing, or covering, the true value. It is a property of the interval producing procedure, and is independent of the particular sample to which such a procedure is applied. We can think of this quantity as the chance that the interval constructed by such a procedure will contain the parameter of interest.

Coverage probability (CP)

Another user friendly measure of agreement which is related to the computation of the TDI is the so called coverage probability (CP) [11,12]. The CP describes the proportion captured within a pre-specified boundary of the absolute paired-measurement differences from two devices, i.e., the value of pκ such that P(—D— ¡ κ) = pκ. Therefore one can find pκ for a specified boundary κ using standard methods for computing probability quantities under normal assumptions [11]:

(13) and to obtain a CP estimate, pκ can be computed by replacing μ_D and σ_D by their REML estimate counterparts derived from model (1).

As with the TDI, the CP criterion can also be translated into a hypothesis test specification. In this case the interest is to ensure that a specified boundary of the absolute paired-measurement differences captures at least a predetermined proportion, p0:

The proposed TI method for inference about the TDI can be utilized to perform inferences about the CP estimates. From the TI in (10) it follows that

(14) Now κ is a fixed known boundary, and our interest lies in finding a lower confidence bound for the CP estimate. Thus, one can find a lower confidence bound for a non-central Student-t proportion with confidence level 1 - α by searching the non-centrality parameter, that depends on and hence on pκ, that satisfies

(15) and once the non-centrality parameter is achieved, a lower bound about the proportion pκ is found using equation (5),

However, the non-centrality parameter cannot be found in a closed form, so one may use again a modified version of the binary search algorithm as follows:

1. begin with the interval [low = 0; high = 1], as pκ is bounded by the interval (0,1);
2. calculate the midpoint of the interval *mid = (low + high)/2* and compute the difference ;
3. if d is greater than 0 up to a tolerance bound δ (i.e.,), then recalculate the interval [low = mid + δ; high = 1]; if it is lower than 0 up to a tolerance bound δ (i.e.), then recalculate the interval [low = 0; high = mid - δ];
4. repeat steps 2-3 until convergence, i.e. until d satisfies .

13.13 Total Deviation Index and Coverage Probability

Lin et al. (2002) proposes a measure called the 'Total Deviation Index'. This assumes that the differences of paired measurements are a random sample from a normal distribution, and consequently the approach is to construct a probability interval, known as a tolerance interval, for these differences. A tolerance interval is a statistical range within which a specified proportion of the population lies. Smaller values of q indicate better agreement. P_0 is specified by the practitioner.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the

variance of differences as functions of observed values of the average of the paired measurements. These methodologies have been adopted by Mayo Clinic (Research Section).

This measure was coined by Lin as the value

$$TDI_{1-p} = \kappa$$

that a given fraction (1-p) of the differences between two measurement methods will be in a symmetric interval $[-\kappa, \kappa]$. This is roughly equivalently to the numerically largest of the 1-p limits of agreement. The measure clearly has its main applicability in equivalence testing.

Lin gives an approximate formula for the calculations.

$$\Theta\left(\frac{TDI - \mu_d}{\sigma_d}\right) - \Theta\left(\frac{-TDI - \mu_d}{\sigma_d}\right) = 1 - p$$

Again, the assumption of the normality of the case-wise differences is relied upon.

The approach is illustrated in a real case example where the agreement between two instruments, a handle mercury sphygmomanometer device and an OMRON 711 automatic device, is assessed in a sample of 384 subjects where measures of systolic blood pressure were taken twice by each device. A simulation study procedure is implemented to evaluate and compare the accuracy of the approach to two already established methods, showing that the TI approximation produces accurate empirical confidence levels which are reasonably close to the nominal confidence level.

13.14 LME - Pankaj Choudhury

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

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{13.5}$$

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

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

$$v_{ij} \sim N(0, \Sigma), \tag{13.6}$$

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

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{13.7}$$

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.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

Chapter 14

BA99

14.1 Regression-based Limits of Agreement

Assuming that there will be no curvature in the scatter-plot, the methodology regresses the difference of methods (d) on the average of those methods (a) with a simple intercept slope model; $\hat{d} = b_0 + b_1 a$. Should the slope b_1 be found to be negligible, $\hat{d}$ takes the value $\bar{d}$.

The next step to take in calculating the limits is also a regression, this time of the residuals as a function of the scale of the measurements, expressed by the averages a_i; $\hat{R} = c_0 + c_1 a_i$

With reference to absolute values following a half-normal distribution with mean $\sigma\sqrt{\frac{2}{\pi}}$, Bland and Altman (1999) formulate the regression based limits of agreement as follows

$$\hat{d} \pm 1.96\sqrt{\frac{\pi}{2}}\hat{R} = \hat{d} \pm 2.46\hat{R} \tag{14.1}$$

Chapter 15

BXC2010

15.1 1. Introduction

15.2 2. Model for LoA

95% prediction interval

$$\bar{D} \pm 1.96 \times s.d.(D_i)\sqrt{\frac{n+1}{n}}$$

The correct factor is $\sigma_1^2 + \sigma_2^2 \frac{n+1}{n}$

15.3 3.Non constant difference

3.1 Model

$D_i = (\alpha_1 - \alpha_2) + (\beta_1 + \beta_2)\mu_i + (e_{1i} + e_{2i})$

3.2 Regression of differences on averages

$\beta_{2|1} = \frac{1-b/2}{1+b/2} \geq 1 - b$

$Y_{2|1} = -a + (1-b)y_1 \pm 2\tau$

15.4 4. Worked Examples

4.1 Blood Glucose (Plasma and Capillary)

- 46 non diabetic obese people at 120 minutes after a 75g oral glucose challenge
- $D = -2.24 + 0.33A$ with residual standard deviation of 1.08
- Prediction interval for the difference of sizes.
- $Y_{C|P} = 1.92 + 0.71Y_N \pm 1.86$ and $Y_{P|C} = 2.69 + 1.40Y_N \pm 2.60$

4.2 Plasma volume (Nadler Hurley)

15.5 5. Why is it wrong to use the regression of the differences on the averages.

5.1 Substantially wrong

5.2 Statistically wrong It is assumed that the averages are independent of the error terms.

$$\frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} = \frac{\beta_1 - \beta_2}{\beta_1 + \beta_2}$$
$$\therefore \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2} = \frac{\beta_1 - \beta_2}{\beta_1 + \beta_2}$$

5.3 Why are the limits straight lines The prediction limits are straight lines because the estimation variance $\sigma 2, 1$ and $\beta 2, 1$ is ignored.

5.4 What is the relation to Standard regression The model (2) is not a standard model

Classical regression models are based on the conditional distribution of one method given another.

15.5.1 5.5 What is the relation to Deming Regression

Deming Regression does not solve the prediction problem unless we are willing to assume a known value for the ratio of the variances. In studies without replicates, there is no information about the variance ratio for the two methods. 6. How wrong is it to do it anyway?

Chapter 16

Lesaffre's paper.

16.1 Lesaffre's paper.

Lesaffre considers the case-weight perturbation approach.

(Cook, 1986) Cook's 86 describes a local approach wherein each case is given a weight w_i and the effect on the parameter estimation is measured by perturbing these weights. Choosing weights close to zero or one corresponds to the global case-deletion approach.

Lesaffre describes the displacement in log-likelihood as a useful metric to evaluate local influence

Lesaffre describes a framework to detect outlying observations that matter in an LME model. Detection should be carried out by evaluating diagnostics C_i , $C_i(\alpha)$ and $C_i(D, \sigma^2)$.

Lesaffre defines the total local influence of individual i as

$$C_i = 2|\triangle \prime_i L^{-1} \triangle_i|. \tag{16.1}$$

The influence function of the MLEs evaluated at the ith point IF_i, given by

$$IF_i = -L^{-1} \triangle_i \tag{16.2}$$

can indicate how $\hat{theta}$ changes as the weight of the ith subject changes.

The manner by which influential observations distort the estimation process can be determined by inspecting the interpretable components in the decomposition of the above measures of local influence.

Lesaffre comments that there is no clear way of interpreting the information contained in the angles, but that this doesn't mean the information should be ignored.

16.2 Lai Shiao

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 (SO2, percent functional oxy-hemoglobin) and fractional saturation (HbO2, 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 = $SO2_{ijtl}$ - $HbO2_{ijtl}$) for the ith donor at the jth level of foetal haemoglobin percent (Fhbperct) and the tth repeated measurement by the lth practitioner of the experiment.

(Carstensen (2004) also advocates the use of LME models in comparing methods, but with a different emphasis.) Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

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

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. 'Fractional saturation' (HbO_2) is considered to be the gold standard method of measurement, with 'functional saturation' (SO_2) being an alternative method. The method of examining the causes of differences between these two methods is applied to a clinical study conducted by **?**. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods (described by the variable 'Bloodage')

and are categorized according to haemoglobin percentages(i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

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

Single fixed effect

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

$$\begin{aligned} &\text{fixed effects}: 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \\ &(\text{p-values}: = 0.0054, < 0.0001, < 0.0001) \end{aligned} \tag{16.3}$$

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

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;

$$\begin{aligned} \text{fixed effects}: \; & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ & (\text{p-values}: \; = 0.8113, < 0.0001, < 0.0001) \\ \\ \text{random effects}: \; & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ & (\text{p-values}: \; = 0.0446, < 0.0001, < 0.0001) \end{aligned} \tag{16.4}$$

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.

Chapter 17

Updating Techniques and Cross Validation

17.1 The Hat Matrix

The hat matrix, also known as the projection matrix, is well known in classical linear models. The diagonal elements h_{ii} are known as 'leverages'. The properties of $\boldsymbol{H}$,such as symmetry and idempotency, are well known.

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$$

$$\boldsymbol{H} = \begin{bmatrix} h_{ii} & \boldsymbol{h}_i' \\ \boldsymbol{h}_i & \boldsymbol{H}_{(i)} \end{bmatrix}$$

$\boldsymbol{H}_{(i)}$ is an $(n-1) \times (n-1)$ matrix. It's inversion for each i is computationally expensive.

$$\boldsymbol{C} = \boldsymbol{H}^{-1} = \begin{bmatrix} c_{ii} & \boldsymbol{h}_c' \\ \boldsymbol{c}_i & \boldsymbol{C}_{(i)} \end{bmatrix}$$

17.1.1 The Hat Matrix

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{17.1}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{17.2}$$

The variances of Y and R can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{17.3}$$

Updating techniques allow an economic approach to recalculating the projection matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

17.2 Efficient updating theorem

It is convenient to write partitioned matrices in which the i-th case is isolated. The partitioned matrix is written as $i = 1$, but the results apply in general.

If $\boldsymbol{C'}_i = [c_{ii}, \boldsymbol{c'}_i]$, such that $\boldsymbol{C}_i$ is the i-th column of $\boldsymbol{H}^{-1}$ then

- $m_i = \frac{1}{c_{ii}}$
- $\breve{x}_i = \frac{1}{c_{ii}} \boldsymbol{X'C}_i$
- $\breve{\boldsymbol{z}_{ji}} = \frac{1}{c_{ii}} \boldsymbol{Z'}_j \boldsymbol{C}_i$

- $\breve{y}_i = \frac{1}{c_{ii}} \boldsymbol{y'C}_i$

Once $\boldsymbol{H}^{-1}$ is determined, an efficient updating formula can be applied.

$$\boldsymbol{H}^{-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{D}^{-1} + \boldsymbol{ZZ})^{-1}\boldsymbol{Z'} \tag{17.4}$$

17.2.1 Updating Regression Estimates

Let the observation j be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where (j) denotes that the jth has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^{n}(x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^{n} x_i) - x_j)^2}{n-1}}{n-2} \tag{17.5}$$

$$Syy^{(j)} = \frac{\sum_{i=1}^{n}(y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^{n} y_i) - y_j)^2}{n-1}}{n-2} \tag{17.6}$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^{n}(x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^{n} x_i) - x_j)(\sum_{i=1}^{n} y_i) - y_k)}{n-1}}{n-2} \tag{17.7}$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \tag{17.8}$$

It is necessary to determine the mean for x and y of the remaining $n-1$ terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^{n} x_i) - (x_j)}{n-1}, \tag{17.9}$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^{n} y_i) - (y_j)}{n-1}. \tag{17.10}$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)}\bar{x}^{(j)}. \tag{17.11}$$

17.2.2 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row.

Consider a $p \times p$ matrix X, from which a row x_i^T is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{17.12}$$

This approach allows an economic approach to recalculating the projection matrix, V, by removing the necessity to refit the model each time it is updated.

This approach is known for numerical instability in the case of downdating.

17.2.3 Updating Standard deviation

A simple, but useful, example of updating is the updating of the standard deviation when an observation is omitted, as practised in statistical process control analyzes. From first principles, the variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \tag{17.13}$$

While using bivariate data, the notation Sxx and Syy shall apply hither to the variance of x and of y respectively. The covariance term Sxy is given by

$$Sxy = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1}. \tag{17.14}$$

17.2.4 Inference on intercept and slope

$$\hat{\beta}_1 \pm t_{(\alpha,n-2)}\sqrt{\frac{S^2}{(n-1)S_x^2}} \tag{17.15}$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \tag{17.16}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_0)} \tag{17.17}$$

17.2.5 Inference on correlation coefficient

This test of the slope is coincidentally the equivalent of a test of the correlation of the n observations of X and Y.

$$\begin{aligned} H_0 &: \rho_{XY} = 0 \\ H_A &: \rho_{XY} \neq 0 \end{aligned} \tag{17.18}$$

17.3 Sherman Morrison Woodbury Formula

The 'Sherman Morrison Woodbury' Formula is a well known result in linear algebra;

$$(A + a^T B)^{-1} \quad = \quad A^{-1} - A^{-1}a^T(I - bA^{-1}a^T)^{-1}bA^{-1} \tag{17.19}$$

This result is highly useful for analyzing regression diagnostics, and for matrices inverses in general. Consider a $p \times p$ matrix X, from which a row x_i^T is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{17.20}$$

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{17.21}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{17.22}$$

The variances of Y and R can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{17.23}$$

Updating techniques allow an economic approach to recalculating the projection matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

Chapter 18

Appendices 1

18.1 Model Terms (ARoy2009 2009)

- Let y_{mir} be the response of method m on the ith subject at the $r-$th replicate.
- Let $\boldsymbol{y}_{ir}$ be the 2×1 vector of measurements corresponding to the $i-$th subject at the $r-$th replicate.
- Let $\boldsymbol{y}_i$ be the $R_i \times 1$ vector of measurements corresponding to the $i-$th subject, where R_i is number of replicate measurements taken on item i.
- Let $\alpha_m i$ be the fixed effect parameter for method for subject i.
- Formally ARoy2009 uses a separate fixed effect parameter to describe the true value μ_i, but later combines it with the other fixed effects when implementing the model.
- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i.
- $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.
- $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to ARoy2009's first test.

18.2 Application to MCS

Let $\hat{\beta}$ denote the least square estimate of β based upon the full set of observations, and let $\hat{\beta}^{(k)}$ denoted the estimate with the k^{th} case excluded.

18.3 Grubbs' Data

When considering the regression of case-wise differences and averages, we write $D^{-Q} = \hat{\beta}^{-Q} A^{-Q}$

	F	C	D	A
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.75
7	791.70	792.40	-0.70	792.05
8	792.30	792.80	-0.50	792.55
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.25
12	793.50	793.80	-0.30	793.65

$$Y^{(k)} = \hat{\beta}^{(k)} X^{(k)} \tag{18.1}$$

Consider two sets of measurements , in this case F and C , with the vectors of case-wise averages A and case-wise differences D respectively. A regression model of differences on averages can be fitted with the view to exploring some characteristics of the data.

When considering the regression of case-wise differences and averages, we write

$$D^{-Q} = \hat{\beta}^{-Q} A^{-Q} \tag{18.2}$$

Let $\hat{\beta}$ denote the least square estimate of β based upon the full set of observations, and let $\hat{\beta}^{(k)}$ denoted the estimate with the k^{th} case excluded.

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $k = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

$$Y^{(k)} = \hat{\beta}^{(k)} X^{(k)} \tag{18.3}$$

Consider two sets of measurements , in this case F and C , with the vectors of case-wise averages A and case-wise differences D respectively. A regression model of differences on averages can be fitted with the view to exploring some characteristics of the data.

```
Call: lm(formula = D ~ A)

Coefficients: (Intercept)            A
-37.51896      0.04656
```

When considering the regression of case-wise differences and averages, we write

$$D^{-Q} = \hat{\beta}^{-Q} A^{-Q} \tag{18.4}$$

18.4 Grubbs' data

Let $\hat{\beta}$ denote the least square estimate of β based upon the full set of observations, and let $\hat{\beta}^{(k)}$ denoted the estimate with the k^{th} case excluded.

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $k = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $k = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

$$Y^{-Q} = \hat{\beta}^{-Q} X^{-Q} \tag{18.5}$$

$$Y^{(k)} = \hat{\beta}^{(k)} X^{(k)} \tag{18.6}$$

Consider two sets of measurements , in this case F and C , with the vectors of case-wise averages A and case-wise differences D respectively. A regression model of differences on averages can be fitted with the view to exploring some characteristics of the data.

```
Call: lm(formula = D ~ A)

Coefficients: (Intercept)            A
-37.51896      0.04656
```

	F	C	D	A
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.75
7	791.70	792.40	-0.70	792.05
8	792.30	792.80	-0.50	792.55
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.25
12	793.50	793.80	-0.30	793.65

When considering the regression of case-wise differences and averages, we write

$$D^{-Q} = \hat{\beta}^{-Q} A^{-Q} \tag{18.7}$$

18.5 Grubb's example

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $Q = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

$$Y^{-Q} = \hat{\beta}^{-Q} X^{-Q} \tag{18.8}$$

18.6 Hat Values for MCS regression

With A as the averages and D as the casewise differences.

```
fit = lm(D~A)
```

$$H = A \left(A^\top A\right)^{-1} A^\top,$$

Chapter 19

Augmented GLMs

Generalized linear models are a generalization of classical linear models.

19.1 Augmented GLMs

With the use of h-likihood, a random effected model of the form can be viewed as an 'augmented GLM' with the response varaibkes $(y^t, \phi_m^t)^t$, (with $\mu = E(y)$,$u = E(\phi)$, $var(y) = \theta V(\mu)$. The augmented linear predictor is

$$\eta_{ma} = (\eta^t, \eta_m^t)^t) = T\omega.$$

.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{19.1}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{19.2}$$

$y_a = T\delta + e^*$

Weighted least squares equation

19.1.1 The Augmented Model Matrix

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{19.3}$$

newpage

19.2 Algorithms : ML v REML

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 algoritm, 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.

(***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.

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 $\boldsymbol{y}$ with respect to each element of $\boldsymbol{\beta}$ and $\boldsymbol{b}$.

19.3 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 "controlargument in LME can be used to change the way the model arrives at the optimum.

19.4 Covariance Parameters

The unknown variance elements are referred to as the covariance parameters and collected in the vector θ.

19.4.1 Methods and Measures

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.

Zewotir and Galpin (2005) lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,

- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

19.5 Haslett's Analysis

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

19.6 Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components σ_j^2, compute an estimate for $\boldsymbol{V}$ and then compute the projector matrix A, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$.

Zewotir and Galpin (2005) remarks that $\boldsymbol{D}$ is a block diagonal with the $i-$th block being $u\boldsymbol{I}$

Chapter 20

Generalized linear models

20.1 Generalized Linear model

In statistics, the generalized linear model (GzLM) is a flexible generalization of ordinary least squares regression. The GzLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Mixed Effects Models offer a flexible framework by which to model the sources of variation and correlation that arise from grouped data. This grouping can arise when data collection is undertaken in a hierarchical manner, when a number of observations are taken on the same observational unit over time, or when observational units are in some other way related, violating assumptions of independence.

20.2 Generalized Model(GzLM)

Nelder and Wedderburn (1972) integrated the previously disparate and separate approaches to models for non-normal cases in a framework called "generalized linear models." The key elements of their approach is to describe any given model in terms of it's link function and it's variance function.

20.2.1 What is a GzLM

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{20.1}$$

where $E(Y)$ is the expected value of Y, $X\beta$ is the linear predictor, a linear combination of unknown parameters,β and g is the link function.

$\mathrm{Var}(\mathbf{Y}) = \mathrm{V}(\boldsymbol{\mu}) = \mathrm{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta}))$

20.2.2 GzLM Structure

The GzLM consists of three elements.

1. A probability distribution from the exponential family.
2. A linear predictor $\eta = X\beta$.
3. A link function g such that $E(Y) = \mu = g^{-1}(eta)$.

20.2.3 Link Function

Definition 1 : The link function provides the relationship between the linear predictor and the mean of the distribution function. There are many commonly used link functions, and their choice can be somewhat arbitrary. It can be convenient to match the domain of the link function to the range of the distribution function's mean.

Definition 2 : A link function is the function that links the linear model specified in the design matrix, where columns represent the beta parameters and rows the real parameters.

20.2.4 Canonical parameter

θ, called the dispersion parameter,

20.2.5 Dispersion parameter

τ, called the dispersion parameter, typically is known and is usually related to the variance of the distribution.

20.2.6 Iteratively weighted least square

IWLS is used to find the maximum likelihood estimates of a generalized linear model.
Definition: An iterative algorithm for fitting a linear model in the case where the data may contain outliers that would distort the parameter estimates if other estimation procedures were used. The procedure uses weighted least squares, the influence of an outlier being reduced by giving that observation a small weight. The weights chosen in one iteration are related to the magnitudes of the residuals in the previous iteration with a large residual earning a small weight.

20.2.7 Residual Components

In GzLMS the deviance is the sum of the deviance components

$$D = \sum d_i \tag{20.2}$$

In GzLMS the deviance is the sum of the deviance components

20.3 Generalized linear mixed models

[pawitan section 17.8]

The Generalized linear mixed model (GLMM) extend classical mixed models to non-normal outcome data.

In statistics, a generalized linear mixed model (GLMM) is a particular type of mixed model. It is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. These random effects are usually assumed to have a normal distribution.

Fitting such models by maximum likelihood involves integrating over these random effects.

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

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

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

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

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

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 (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet 346*, 1085–87.

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

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

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

Carstensen, B. (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.

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

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

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

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.

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

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

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.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

Lin, S. C., D. M. Whipple, and charles S Ho (1991). Evaluation of statistical equivalence using limits of agreement and associates sample size calculation. *Communications in Statistics - Theory and Methods 27*(6), 1419–1432.

Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry 44*, 1024–1031.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

Luiz, R., A. Costa, P. Kale, and G. Werneck (2003). Assessment of agreement of a quantitative variable: a new graphical approach. *Journal of Clinical Epidemiology 56*, 963–967.

Mantha, S., M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss (2000). Comparing methods of clinical measurement: Reporting standards for bland and altman analysis. *Anaesthesia and Analgesia 90*, 593–602.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

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

Phelps, C. and A. Hutson (1995). Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical decision making 15*, 144–57.

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

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.

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.

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

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

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

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*(2), 153–177.

20.4 Assessment of Agreements in Linear and Generalized Linear Mixed Models

- Study of measuring agreement is intend to evaluate whether the readings from one rater/ measurement agree with those from other raters/measurements. In this dissertation, we are going to

present a general method to assess agreement for a large variety of data with repeated measurements using linear and generalized linear mixed models.

- In the first place, a set of agreement statistics, including mean square deviation, concordance correlation coefficient, precision and accuracy coefficients, is presented for evaluating the intra-, inter-, and total-rater agreement in the multiple-rater and multiple-replications cases.

- Secondly, likelihood-based approaches are developed to estimate all the agreement statistics. Asymptotic properties of these estimates are also discussed for different data structures.

- Furthermore, our method has the merit of handling missing values and covariates naturally, and a new set of restricted agreement statistics is proposed in order to capture the true random variations and between-instrument effects adjusted for the covariate effects.

- Simulations for both linear and generalized linear mixed models are conducted to show the accuracy and effectiveness of our approaches. In the end, two industry datasets are evaluated using our approach.

- One is the cardiac function measurements used to determine the agreement between impedance cardiography and radionuclide ventriculography estimates, and the other one is an antihypertensive patch dataset given by FDA for assessing individual bioequivalence.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

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.

20.5 Random Effects and MCS

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

20.5.1 Random coefficient growth curve model

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

$$\mathbf{d} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{20.3}$$

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

20.6 Other Approaches

20.6.1 Random coefficient growth curve model

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

20.6.2 Marginal Modelling

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

20.7 KP

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