Contents

Chapter 1

Method Comparison Studies

1.1 What is a method comparison study?

The question of properly assessing "agreement" between two or more methods of measurement is ubiquitous, 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).

Historically comparison of two methods of measurement was carried out by use of paired sample $t-$test, correlation coefficients or simple linear regression. Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these approaches for comparing two methods of measurement, and proposed their own framework with this application in mind. Although the authors acknowledge the opportunity to apply other, more complex methodologies, they argue that simpler approaches are preferable, especially when the results must be 'explained to non-statisticians'.

A method of measurement should ideally be both accurate and precise. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method

may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate 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 smaller the systematic bias, the greater the accuracy of the method. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown 'true value'.

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.

1.1.1 Purpose of Method Comparison Studies

Different authors focus on different aspects of comparison problem. Carstensen (2010) provides a review of many descriptions of the purpose of method comparison studies, several of which are reproduced here.

> "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 interchangeably" (Bland and Altman, 1999).
>
> "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 isotope dilution mass spectroscopy. The question arises as to which method is better" (Mandel, 1991).

"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).

While several major commonalities are present in each definitions, there is a different emphasis for each, which will inevitably give rise to confusion. In the view of Dunn (2002), a question relevant to many practitioners is which of the two methods is more precise. Carstensen (2010) seems to endorse a simple phrasing of the research question that is proposed by Altman and Bland (1983), i.e. "*do the two methods of measurement agree sufficiently closely?*" with Carstensen (2010) expressing the view that other considerations (for example, the "equivalence" of two methods) to be treated as separate research questions. As such, we will revert to other research questions, such as "equivalence of methods" later, focussing on agreement and repeatability of methods.

In many cases the purpose of the study is to calibrate a new method of measurement against a "Gold Standard" method, a known method that is considered most precise in its measurement. For example, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and either a traditional reference or gold standard must be evaluated before the new one is put into practice. Various approaches have been proposed for this purpose in recent years. It must be noted that absence of measurement error should

not be assumed for gold standard methods.

1.1.2 Grubbs' Artillery Round Data

To illustrate the characteristics of a typical method comparison study consider the data in Table 1.1 (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm artillery piece and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels '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 1.1.1: Velocity measurement from 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 1.1 can not be assumed to be 'true values' in any absolute sense.

A simple estimate of the inter-method bias is given by the differences between pairs of measurements, for example, in Table 1.2 shows possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method.

The absence of inter-method bias is, by itself, not sufficient to establish that two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than those of the other, they can not be considered to be in agreement. Hene, method comparison studies are required to take account of both inter-method bias and difference in the precision of measurements.

Round	Fotobalk (F)	Counter (C)	Difference (F-C)
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.1.2: Difference between Fotobalk and Counter measurements.

1.1.3 Agreement

Bland and Altman (1986) defined perfect agreement as the case where all of the pairs

of measurement data, when plotted on a conventional scatter-plot, 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 line $X = Y$ on the Cartesian plane).

Agreement is the extent to which the measure of the variable of interest, under a constant set of experimental conditions, yields the same result on repeated trials (Sanchez and Binkowitz, 1999). The more consistent the results, the more reliable the measuring procedure.

Altman and Bland (1983) define bias (referred to hereafter as inter-method bias) as a *a consistent tendency for one method to exceed the other* and propose estimating its value by determining the mean of the case-wise differences. The variation about this mean shall be estimated by the standard deviation of the case-wise differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

1.2 Improper Method Comparison Techniques

The issue of whether two measurement methods are 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, simple linear regression, or correlation coefficients.

Paired sample *t*-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_d = 0$. The test statistic is distributed as a t random variable with $n - 1$ degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{1.1}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. This method can be potentially misused for method comparison studies. Paired $t-$tests test only whether

the mean responses are the same, and so provides a useful test for inter-method bias. However, no insight can be obtained about the variability of the case-wise differences by the paired $t-$test, critically undermining it as a stand-alone procedure. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

Regression Methods

On account of the fact that one set of measurements are linearly related to another, one could surmise that simple linear Regression is the most suitable approach to analyzing comparisons. However simple linear regression is considered by many authors to be wholly unsuitable for method comparison studies (Altman and Bland, 1983; Cornbleet and Cochrane, 1979; Ludbrook, 1997). Simple linear regression is defined as such with the name 'Model I regression' by Cornbleet and Cochrane (1979), in contrast to 'Model II regression' models, which shall be discussed later on.

A key assumptions of simple linear regression is that the independent variable values are without random error. For method comparison studies, both sets of measurement must be assumed to be measured with imprecision and neither case can be taken to be a reference method. Arbitrarily selecting either method as the reference (i.e. the independent variable) will yield conflicting outcomes: a regression of X on Y would yield an entirely different model from fitting Y on X.

Further criticisms of linear regression exist. Firstly regression methods are uninformative about the variability of the differences. Secondly regression models are unduly influenced by outliers. Lastly, regression models can not be used to effectively analyze repeated measurements.

The Identity Plot

Altman and Bland (1983) states that regression analysis can offer useful insights, and recommending an 'Identity Plot', a simple graphical approach that yields a cursory ex-

amination of how well the measurement methods agree. In the case of good agreement, the co-variates of the Identity plot accord closely with the $X = Y$ line. This plot is not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation. An identity plot shall complement demonstrations of commonly used approaches in the next chapter.

Decomposition of Inter-Method Bias

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing inter-method bias to be decomposed into constant bias and proportional bias. Regression methods can determine the presence of inter-method bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002).

Constant bias describes the case where one method gives values that are consistently different to the other across the whole range. Using a naive estimation of bias, such as the mean of differences, it may incorrectly indicate absence of bias, by yielding a mean difference close to zero. This would be caused by positive differences in the measurements at one end of the range of measurements being canceled out by negative differences at the other end of the scale. Proportional Bias exists when two methods agree on average, but exhibit differences over a range of measurements, i.e. the differences are proportional to the scale of the measurement. A measurement method may be subject to any combination of fixed bias or proportional bias, or both (Ludbrook, 2002).

Constant or proportional bias 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. If there is no constant bias, the intercept is equal to zero and, similarly, if there is no proportional bias, the slope is equal to one. Thus, carrying out hypothesis tests on these coefficients (where the null hypotheses are $\beta_0 = 0$ and $\beta_1 = 1$) allow us to test for the presence of both types of bias.

If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined.

The Correlation Coefficient

Correlation is inadequate to assess agreement because it only evaluates only the linear association of two sets of observations. Nonetheless linear association is not the same as agreement. It is possible for two methods to be highly correlated, yet have poor agreement due to any combination of constant and proportional bias. Arguments against its usage have been made repeatedly in the relevant literature, with Altman and Bland (1983), Bland and Altman (1986), Bland and Altman (2003) and Giavarina (2015) as examples.

1.3 Replicate Measurements and Repeatability

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Repeated measurements on several subjects can be used to quantify measurement error; the variation between measurements of the same quantity on the same individual. Measurements taken in quick succession, so that no real systemic changes can take place on each item, by the same observer using the same instrument on the same item can be considered true replicates (Bland and Altman, 1999). Roy (2009) accords with Bland and Altman's definition, but notes that some measurements may not be 'true' replicates. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity posed by replicate measurements . Bland and Altman (1986) address the additional complexity by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The 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 properly estimate the inter-method bias. However,

Carstensen et al. (2008) is critical of both approaches, offering an alternative approach that shall be introduced later.

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.

1.3.1 Exchangeable and 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.

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

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

1.3.2 Repeatability

Repeatability describes the variation in measurements taken by a single method of measurement on the same item and under the same conditions. A measurement method

can be said to have a good level of repeatability if there is consistency in repeated measurements on the same subject using that method. Conversely, a method has poor repeatability if there is considerable variation in repeated measurements.

Repeatability is defined by the IUPAC (2009) as '*the closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time)*' and is determined by taking multiple measurements on a series of subjects. A similar set of criteria is described in the *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*.

Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study, a view endorsed by Carstensen et al. (2008). The repeatability of two methods influence the amount of agreement which is possible between those methods. Before there can be good agreement between two methods, a method must have good agreement with itself. If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Bland and Altman, 1999; Roy, 2009). Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. However Roy (2009) notes the lack of convenience in such calculations.

Barnhart et al. (2007) remarks that it is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors, while further remarking '*curiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked*.

Statistical procedures on within-item variances of two methods are equivalent to tests on their respective repeatability coefficients. A formal test is introduced by Roy (2009), which will discussed in chapter three.

If replicate measurements by a method are available, it is simple to estimate the measurement error for a method, using a model with fixed effects for item, then taking the residual standard deviation as measurement error standard deviation. However, if

replicates are linked, this may produce an estimate that biased upwards.

1.3.3 Types of Comparison Studies

Lewis et al. (1991) categorize method comparison studies into three different types, namely: calibration, comparison and conversion. 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 as 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 framework 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 for which Bland and Altman's Methodology is intended, and therefore it is the most relevant of the three for this thesis.

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 thesis, it is the least relevant of the three cases.

Roy et al. (2015) discusses the importance of gold Standards in the context of method comparison studies. Currently the phrase 'gold standard' describes the most

accurate method of measurement available. No other criteria are set out. Further to Dunn (2002), various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer (i.e. a blood pressure measurement cuff), which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Dunn (2002) recognizes 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.

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free and that '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'. Pizzi (1999) similarly addresses the issue of gold 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 tests based upon the angiogram are reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8% (ACR, 2008).

In literature gold standards are, perhaps more accurately, can be 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 both in the context of a comparison study and a calibration study.

According to Bland and Altman, one should use the methodology previous outlined, even when one of the methods is a gold standard.

1.4 Outline of Thesis

Thus, the basic concepts of, and need for method comparison are introduced. The intention of this thesis is to develop the theory of method comparison studies using Linear Mixed Effects models. Chapter two will provide a review of the prevalent methods, highlighting particular flaws where relevant. Chapter three shall describe Linear Mixed effects models, and how the use of the linear mixed effects models can be extended to method comparison studies. Implementations of important existing work is presented using the `R` programming language.

Chapter three 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 again, using the `R` programming language.

Model diagnostics are an integral component of a complete statistical analysis. In chapter four model diagnostics are described in depth, with particular emphasis on linear mixed effects models.

In the fifth chapter, important linear mixed effects model diagnostic methods shall be extended to method comparison studies, and proposed methods are 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 deals with robust measures of important parameters such as agreement.

Chapter 2

Bland-Altman Methodology

2.1 Anatomy of the Bland-Altman Plot

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. Altman and Bland (1983) recognized the inadequacies of several analyses and articulated quite thoroughly the basis on which they are unsuitable for comparing two methods of measurement. Instead they recommended the use of graphical techniques to assess agreement. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 2.1.1). These differences and averages are then plotted (Figure 2.1.2).

In 1983 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes (Altman and Bland, 1983). 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:

> "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."

Principally their method is calculating, for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference d_i and mean a_i: case-wise differences of measurements of two methods $d_i = x_i - y_i$, for $i = 1, 2, \ldots, n$, on the same subject should be calculated, and then the average of those measurements, $a_i = (x_i + y_i)/2$ for $i = 1, 2, \ldots, n$. An important requirement is that the two measurement methods use the same scale of measurement. Following a technique known as the Tukey mean-difference plot, as noted by Kozak and Wnuk (2014), Altman and Bland (1983) proposed that a_i should be plotted against d_i, a plot now widely known as the Bland-Altman plot, and motivated this plot as follows:

> "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 approach has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical tool for making a visual assessment of the data.

As the objective of the Bland-Altman plot is to advise on the agreement of two methods, the individual case-wise differences are also particularly relevant. The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$, and is represented with a line on the Bland-Altman plot. 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 (1986) do, however, state that the absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

Furthermore they propose 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).

2.1.1 Identity Plot

The first step recommended, which the authors argue should be mandatory, is construction of an identity plot, introduced in the last chapter as a simple scatter-plot approach of measurements for both methods on either axis, with the line of equality (the $X = Y$ line, i.e. the 45 degree line through the origin).

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. However, they are not useful for a thorough examination of the data. This plot can gives the analyst a cursory examination of how well the measurement methods agree. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation. A scatter plot of the Grubbs data is shown in Figure 2.1.1. Visual inspection confirms the previous conclusion that inter-method bias is present, i.e. the Fotobalk device has a tendency to record a lower velocity.

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 2.1.2, using data from Table 2.1.1. The dashed line in Figure 2.1.2 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

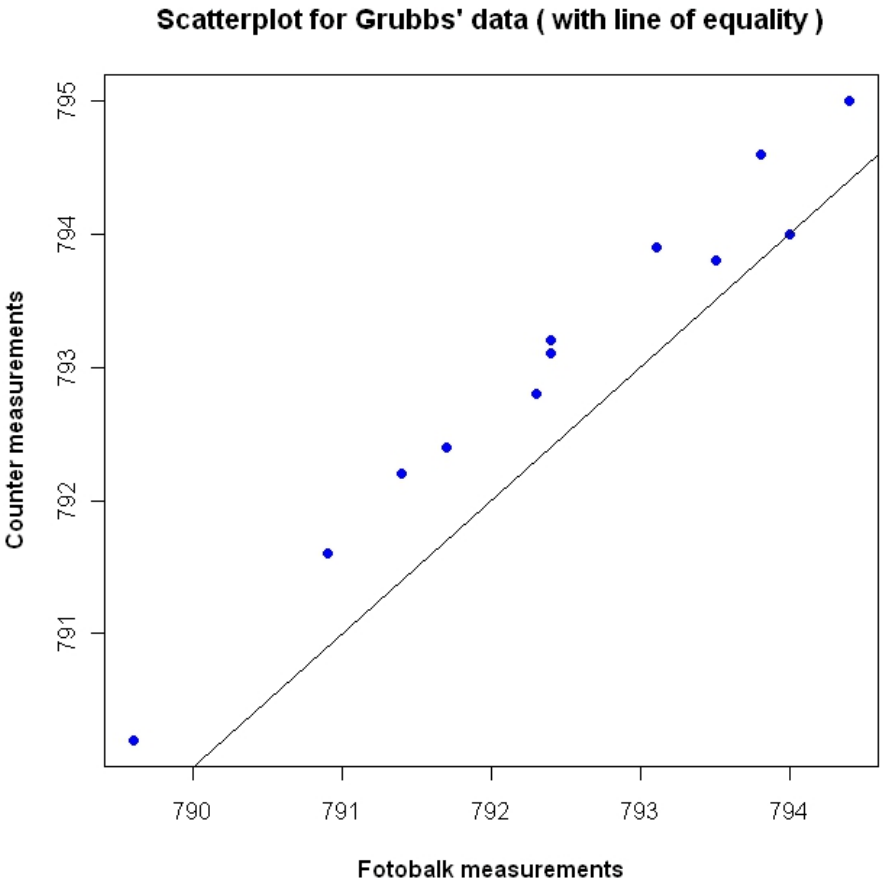


Figure 2.1.1: Scatter plot for Fotobalk and Counter methods.

the case of Grubbs data the inter-method bias is -0.6083 metres per second.

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.

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 2.1.1: 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 2.1.2: Fotobalk and Terma methods: Differences and Averages.

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'

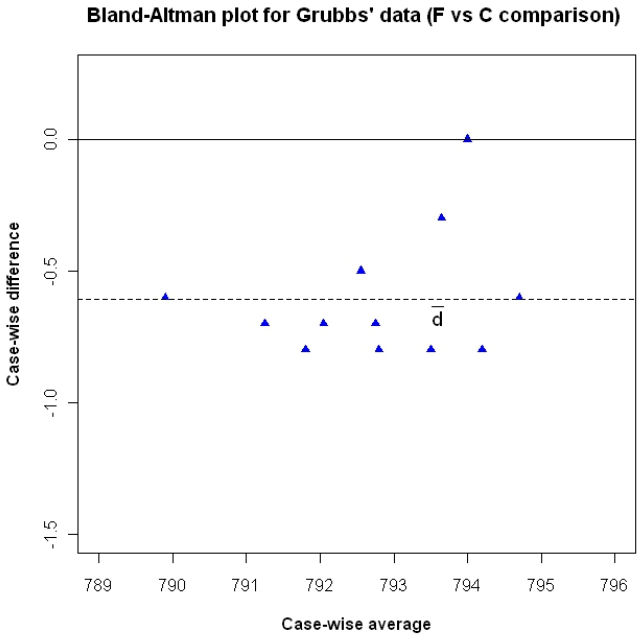


Figure 2.1.2: Bland-Altman plot For Fotobalk and Counter methods.

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.

2.1.2 Inspecting the Data

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.

Figures 2.1.4, 2.1.5 and 2.2.7 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 2.1.4 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, have been added to indicate the trend. Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant

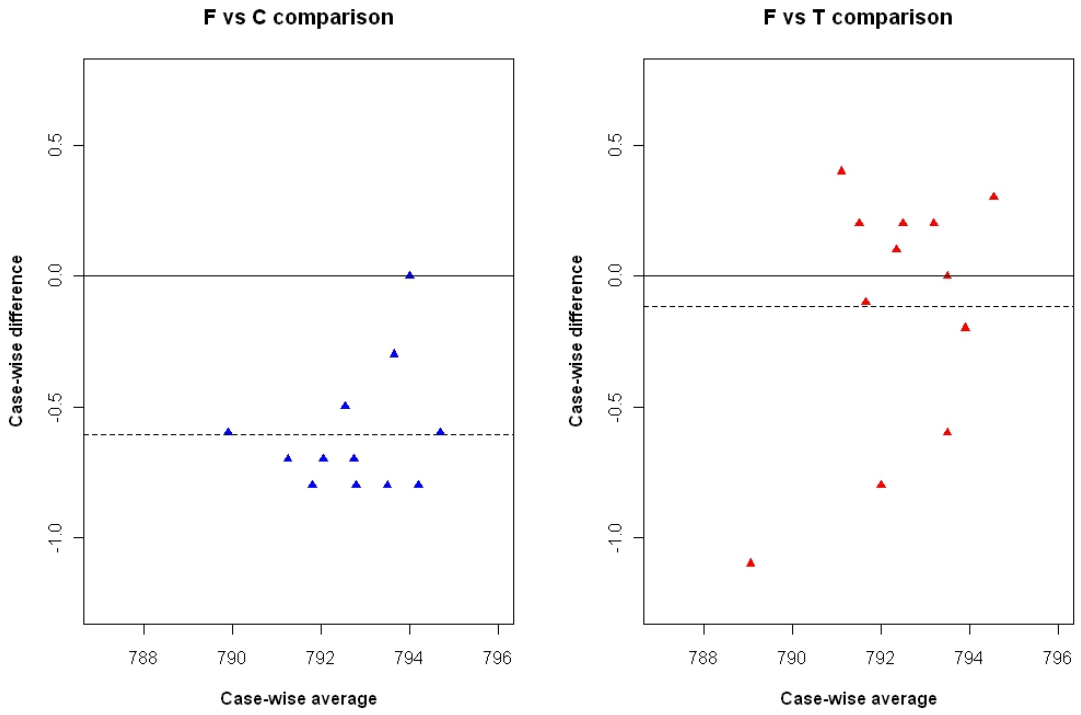


Figure 2.1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

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 could be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (Benjamini and Hochberg, 1995) test, are advisable.

Figure 2.1.5 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'. Both of these cases violate the assumptions necessary for further analysis using limits of agreement, which shall be discussed later.

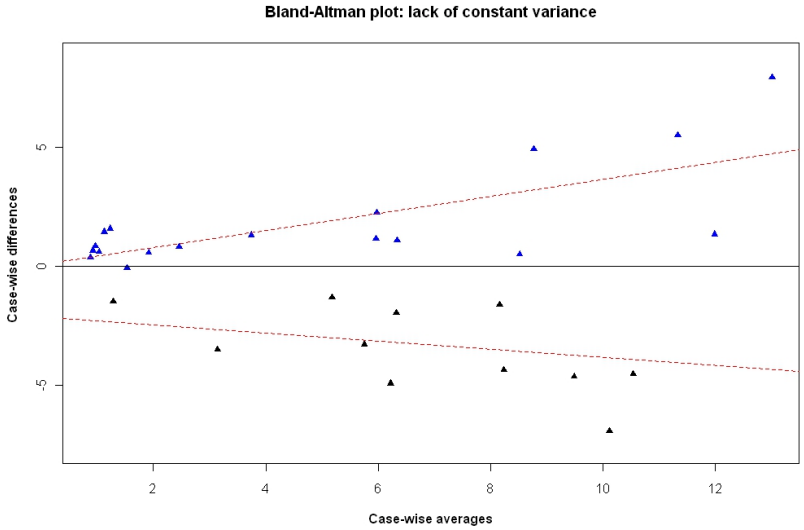


Figure 2.1.4: Bland-Altman Plot demonstrating the increase of variance over the range

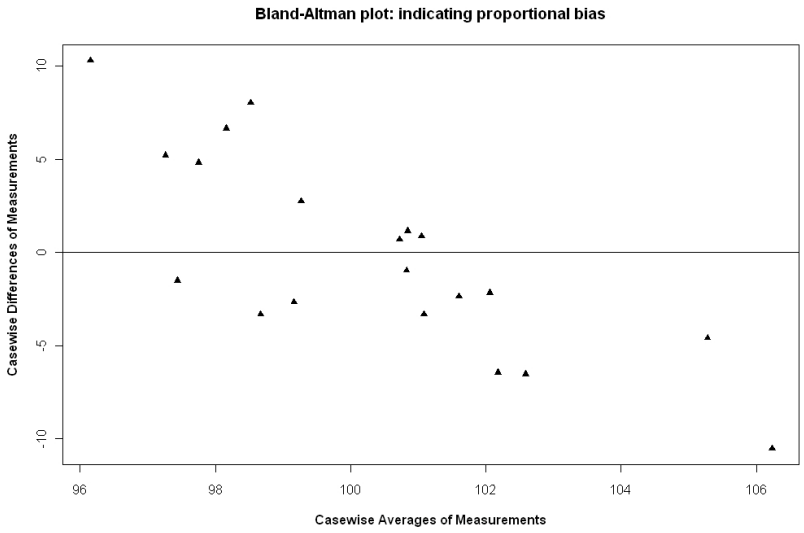


Figure 2.1.5: Bland-Altman Plot indicating the presence of proportional bias

2.1.3 Limits of Agreement

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 liter-

ature as 95% 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.

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 2.1.6 shows the resultant Bland-Altman plot, with the limits of agreement shown in dotted.

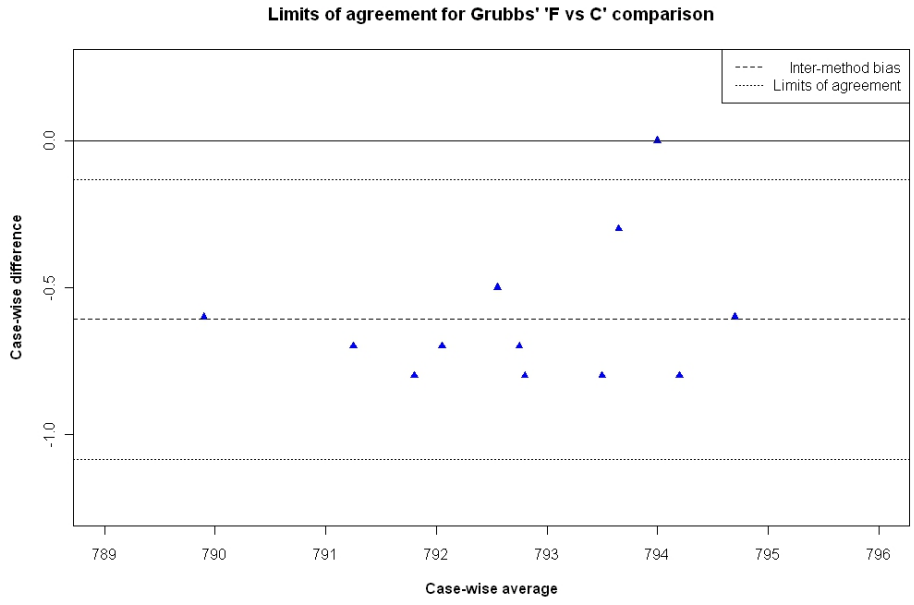


Figure 2.1.6: Bland-Altman plot with limits of agreement

Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably, by demonstrating the range in which 95% of the sample data should lie. Following basic principles of the normal probability distribution, the Limits of Agreement are centred on the average difference line (which indicates the inter-method bias) and are 1.96 times the standard deviation above and below the average difference line. The limits of agreement methodology assumes a constant level of bias throughout the range of measurements.

Bland and Altman (1986) refer to this as the 'equivalence' of two measurement

methods. The specific purpose of the limits of agreement must be established clearly. 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:

$$LOA = \bar{d} \pm 1.96 s_d$$

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.

Importantly the authors recommend prior determination of what would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion. However, Mantha et al. (2000) highlight inadequacies in the correct application of limits of agreement, resulting in contradictory estimates of limits of agreement in various papers.

2.1.4 Normality of Case-wise Differences

The difference are assumed to be normally distributed, although the measurements themselves are not assumed to follow any distribution. Therefore the authors argue that the 95% of differences are expected to lie within these limits.

Calculation of the limits of agreement relies on the assumption that the case-wise differences are normally distributed. The calculation removes a lot of the variation between subjects, leaving measurement error, which is likely to be normally distributed. Bland and Altman (1999) remark that this assumption is easy to check using a normal plot.

This assumption is justified because variation between subjects has been removed, leaving only measurement error (Bland and Altman, 1986). There are formal methodologies to test whether this assumption holds.

2.1.5 Interpretation of Limits Of Agreement

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*', offering no elaboration.

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 they were Shewhart control limits. Importantly the parameters used to determine the limits, the mean and standard deviation, are not based on any randomly ordered sample used for an analysis, but on a statistical process's time ordered 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) offer an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.025,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{2.1}$$

where n is the number of subjects. With consideration of the effect of the sample size on the interval width, Carstensen et al. (2008) remarks that only for 61 or more subjects is the 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. Luiz et al. (2003) offers an alternative description of limits of agreement, this time as tolerance limits.

Barnhart et al. (2007) describes them as a probability interval, and offers a clear description of how they should be used; '*if the absolute limit is less than an acceptable difference d_0, then the agreement between the two methods is deemed satisfactory*'.

Various other interpretations as to how limits of agreement should properly be defined. The prevalence of contradictory definitions of what limits of agreement strictly will inevitably attenuate the poor standard of reporting using limits of agreement, as discussed by Mantha et al. (2000).

2.1.6 Criticism of Limits of Agreement

The Bland-Altman methodology is well noted for its ease of use, and can be easily implemented with most software packages. Also it does not 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.

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) percentage points. An knowledgeable 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.

If one method is sometimes higher, or sometimes lower, 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.

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence they are similar to confidence intervals, they are not constructed as such; they are designed for future values. Ludbrook (1997, 2002) criticizes Bland-Altman plots on the basis that they present 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 the effect they would have on the limits of agreement, but offer no guidance on how to correct for those effects.

2.2 Detection of Outliers

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. 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 Altman (1999), we get the clearest indication of what they 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. 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." Classification thereof is a subjective decision in any analysis, but must be informed by the logic of the formulation. Figure 2.2.7 is a Bland Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively. Importantly, outlier classification must be informed by the logic of the mechanism that produces the data.

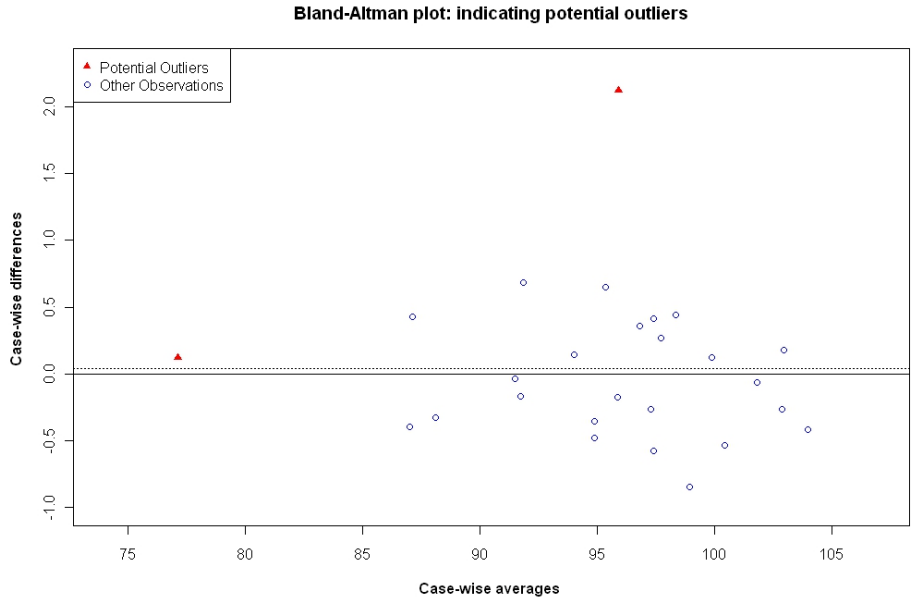


Figure 2.2.7: Bland-Altman plot indicating the presence of outliers

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 2.2.7, 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. Conversely, the fourth observation from the original data set should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

2.2.1 Bartko's Ellipse

As an enhancement on the Bland Altman Plot, Bartko (1994) has expounded a confidence ellipse for the covariates. Bartko (1994) offers a graphical complement to the Bland-Altman plot in the form of a bivariate confidence ellipse as a boundary for dispersion. Altman (1978) provides the relevant calculations for the ellipse. This ellipse is intended as a visual guideline for the scatter plot, for detecting outliers and to assess the within- and between-subject variability. The stated purpose is to 'amplify dispersion', which presumably is for the purposes of outlier detection. The orientation of the the ellipse is key to interpreting the results. Additionally Bartko's ellipse provides a visual aid to determining the relationship between variances.

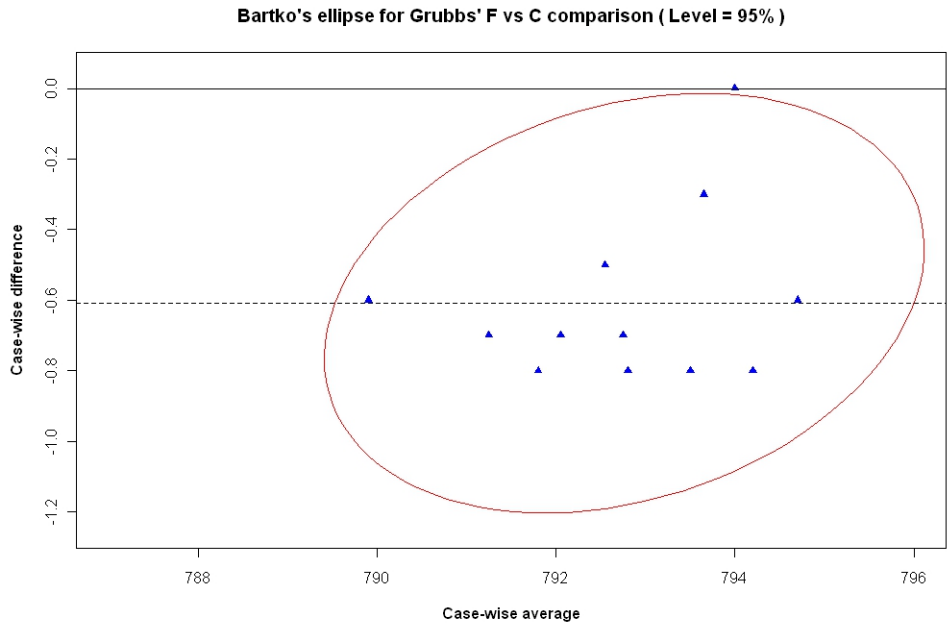


Figure 2.2.8: Bartko's ellipse for Grubbs data

The minor axis relates to the between-subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other.

Furthermore, the ellipse provides a visual aid to determining the relationship be-

tween the variance of the means var(a) and the variance of the differences var(d). If var(a) is greater than var(d), the orientation of the ellipse is horizontal. Conversely if var(a) is less than var(d), the orientation of the ellipse is vertical. The more horizontal the ellipse, the greater the degree of agreement between the two methods being tested.

Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers. The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko's ellipse.

The Bland-Altman plot for the Grubbs data, complemented by Bartko's ellipse, is depicted in figure 2.2.8. The fourth observation is shown to be outside the bounds of the ellipse, indicating that it is a potential outlier.

A covariate is added to the 'F vs C' comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Figure 2.2.9 depicts a 95% confidence ellipse for this manipulated data set. By inspection of the confidence interval, we would conclude that this extra covariate is an outlier, in spite of the fact that this observation is very close to the inter-method bias as determined by this approach.

2.2.2 Grubbs' Test for Outliers

In classifying whether an observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the Grubbs test procedure is the absence of any outliers in the data set.

The test statistic for the Grubbs test (G) is the largest absolute deviation from the sample mean divided by the standard deviation of the differences,

$$G = \max_{i=1,\ldots,n} \frac{\left|d_i - \bar{d}\right|}{S_d}. \tag{2.2}$$

For the 'F vs C' comparison it is the fourth observation that gives rise to the test statistic, $G = 3.64$. The critical value is calculated using Student's t distribution and

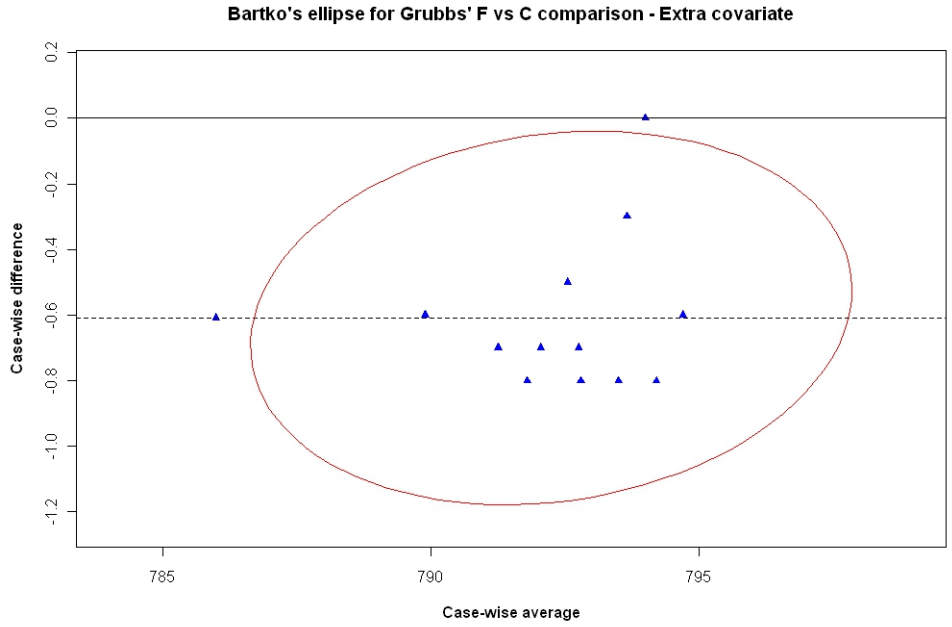


Figure 2.2.9: Bartko's Ellipse for Grubbs' data, with an extra covariate.

the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}.$$

For this test $U = 0.75$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with $p-$value $= 0.003$, in accordance with the previous result of Bartko's ellipse.

2.3 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. Bland and Altman (1986) advance a formulation for confidence intervals of the inter-method bias and the limits of agreement, arguing that it is possible to make such estimates if it is assumed that the case-wise differences approximately follow a normal distribution. However Bland and Altman (1999) caution that such calculations may be 'somewhat optimistic' if the associated

assumptions are not valid. A 95% confidence interval can be determined, by means of the t distribution with $n-1$ degrees of freedom. For the inter-method bias, the confidence interval is simply that of a mean: $\bar{d} \pm t_{(\alpha/2,n-1)} S_d/\sqrt{n}$.

The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

$$\text{Var}(LOA) = \left(\frac{1}{n} + \frac{1.96^2}{2(n-1)}\right) s_d^2.$$

If n is sufficiently large this can be following approximation can be used

$$\text{Var}(LOA) \approx 1.71^2 \frac{s_d^2}{n}.$$

Consequently the standard errors of both limits can be approximated as 1.71 times the standard error of the differences.

2.4 Prevalence of the Bland-Altman plot

Bland and Altman (1986), which further develops the Bland-Altman methodology, was found to be the sixth most cited paper of all time by Ryan and Woodall (2005). Dewitte et al. (2002) reviews the use of Bland-Altman plots by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001, describing the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. This study concludes that use of the Bland-Altman plot increased over the years, from 8% in 1995 to 14% in 1996, and 31-36% in 2002.

The Bland-Altman plot has since become the expected, and often the obligatory, approach for presenting method comparison studies in many scientific journals (Hollis, 1996). Furthermore O'Brien et al. (1990) recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

Mantha et al. (2000) contains a study on 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, while the other two used correlation and regression analyses. Mantha

et al. (2000) remark that 3 papers, from 42 mention predefined maximum width for limits of agreement that 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) remark that the limits of agreement should be compared to a clinically acceptable difference in measurements.

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation in 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:

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

Barnhart et al. (2007) 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 easily determinable, thus creating a drawback to this methodology.

Alternative indices, proposed by Barnhart et al. (2007), are the square root of the MSD and the expected absolute difference (EAD).

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

Both of these indices can be interpreted intuitively, since their units are the same as that of the original measurements. They can also be compared to the maximum acceptable absolute difference between two methods of measurement d_0. For the sake of brevity, the EAD will be considered solely.

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. A consequence of using absolute differences is that high variances would result in a higher EAD value.

	U	V	$U - V$	$\lvert U - V \rvert$
1	98.05	99.53	-1.49	1.49
2	99.17	96.53	2.64	2.64
3	100.31	97.55	2.75	2.75
4	100.35	96.03	4.32	4.32
5	99.51	99.00	0.51	0.51
6	98.50	100.76	-2.26	2.26
7	100.66	99.37	1.29	1.29
8	99.66	108.87	-9.21	9.21
9	99.70	105.16	-5.45	5.45
10	101.55	94.31	7.24	7.24

Table 2.4.3: Example data set

To illustrate the use of EAD, consider Table 2.4.3. The inter-method bias of 0.03, which is desirably close to zero in the context of agreement. However, an identity plot would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, which is a wide interval for this data. As with the identity plot, this would indicate lack of agreement. As with inter-method bias, an EAD value close to zero is desirable. However, from Table 2.4.3, the EAD can be computed as 3.71. The Bland-Altman plot remains a useful part of the analysis. In

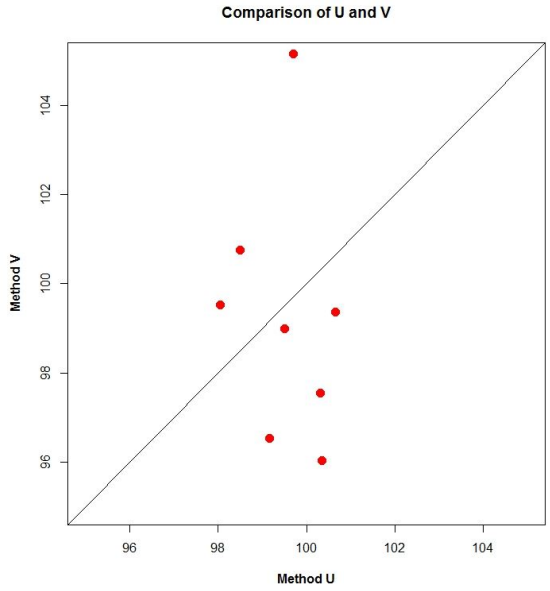


Figure 2.4.10: Identity Plot for example data

Figure 2.4.11, it is clear there is a systematic decrease in differences across the range of measurements.

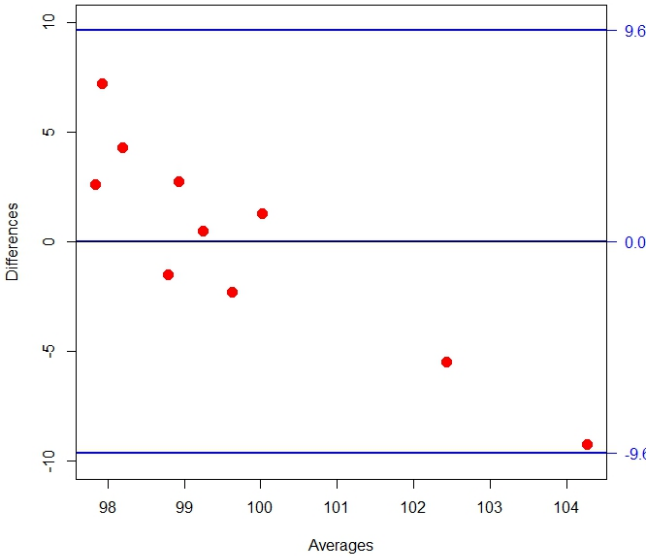


Figure 2.4.11: Bland-Altman Plot for UV comparison

Barnhart et al. (2007) 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. The EAD values for both comparisons are therefore much closer.

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

Table 2.4.4: Agreement indices for Grubbs' data comparisons.

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).

$$CP = P(|x_i - y_i| \leq d_0) \tag{2.3}$$

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. 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.

2.5 Variations of the Bland-Altman Plot

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. 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 importance 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.

Due to limitations of the conventional difference plot, a series of alternative formulations for the Bland-Altman approach have been proposed. Referring to the assumption that bias and variability are constant across the range of measurements, Bland and Altman (1999) address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits of agreement would be 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 acknowledge that this is not easy to interpret, and may not be 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 be inappropriate for. The first variation is a plot of case-wise 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 case-wise ratios as percentage of averages, removing the need for logarithmic transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. Eksborg (1981) proposed such a ratio plot, independently of Bland and Altman. Dewitte et al. (2002) commented on the reception of this article by saying '*Strange to say, this report has been overlooked*'.

2.6 Limits of Agreement for Replicate Measurements

Computing limits of agreement features prominently in many method comparison studies since the publication of Bland and Altman (1986). Bland and Altman (1999) addresses the issue of computing LOAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods. However, the original Bland-Altman method was developed for two sets of measurements done on one occasion, and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method. In addition to Bland and Altman (1999), Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models. This approach will be discussed in due course.

2.7 Coefficient of Repeatability

Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement by collecting replicated data. Roy (2009) notes the lack of convenience in such calculations. The coefficient of repeatability is a measure of how

well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999).

As mentioned previously, Barnhart et al. (2007) emphasize the importance of repeatability as part of an overall method comparison study. The coefficient of repeatability was proposed by Bland and Altman (1999), and is referenced in subsequent papers, such as Carstensen et al. (2008). BSI (1975) define a coefficient of repeatability as *the value below which the difference between two single test results....may be expected to lie within a specified probability.* Bland and Altman (1999) defines the repeatability coefficient as the upper limits of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances.

Once the within-item variability for both methods has been estimated, the relevant calculations for the coefficients of repeatability are straightforward. The coefficient is calculated from the within-item variability σ_m^2 as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$. For 95% of subjects, two replicated measurement by the same method will be within this repeatability coefficient.

The coefficient of repeatability may provide the basis for the formulation a formal definition of a 'gold standard'. For example, by determining the ratio of the repeatability coefficient (CR) to the sample mean $\bar{X}$. Advisably the sample size should 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.

Chapter 3

Linear Mixed effects Models

In cases where there are repeated measurements by each of the two methods on the same subjects, Bland and Altman (1999) suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods. The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be incorrect (Carstensen, 2004). Carstensen (2004) recommends that replicate measurements for each method, but recognizes that resulting data are more difficult to analyze. To this end, Carstensen (2004) and Carstensen et al. (2008) recommend the use of LME models as a suitable framework for method comparison in the case of repeated measurements. Roy (2009) uses an LME model approach to provide a set of formal tests for method comparison studies.

Several authors, such as Kelly (1985) and Voelkel and Siskowski (2005), recommend the use of Structural Equation Model from Method Comparison. Structural Equation Model provides a statistically rigourous analysis, but the approach is undermined in several ways. LME models have greater flexibility and can be adapted to any variant of the method comparison research question, whereas SEM is suitable for some specific cases only. Highly complex models can be developed using SEM, but to overcome the problem of identifiability, a large quantity of data must be gathered. Often this is beyond what is practical in the main applications of method comparison studies, namely the medical sciences. Once simplifications are applied, there is little functional

difference between SEM and LMEs.

3.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (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.

3.1.1 Laird Ware Model

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.

Linear mixed effects models (LME) differs from the conventional linear model in that it has both fixed effects and random effects regressors, and coefficients thereof. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course. Using Laird-Ware form, the LME model is commonly described in matrix form,

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

$\mathbf{Y}$ is the $n \times 1$ response vector, where n is the number of observations. β is a $p \times 1$ vector of fixed p effects, with the first element being the population mean. X and Z are $n \times p$ and $n \times q$ "model matrices" for fixed effects and random effects respectively, comprising 0s or 1s, depending on the observation is question. The vector of residuals, v$\mathbf{e}$ has dimension $n \times 1$. The random effects are contained in the $q \times 1$ vector $\mathbf{b}$.

3.1.2 LME Model 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 (3.1), 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

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

Henderson's equations

Because of the dimensionality of V (i.e. $n \times n$) computing the inverse of V can be difficult. As a way around the this problem Henderson (1953); Henderson et al. (1959, 1963, 1973, 1984) offered a more simpler approach of jointly estimating $\hat{\beta}$ and $\hat{b}$. Henderson (1950) made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and $b \sim \mathrm{N}(0, D)$, and proceeded to maximize the joint density of y and b

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (3.2)$$

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

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (3.3)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (3.4)$$

Using these equations, obtaining the estimates requires the inversion of a matrix of dimension $p + q \times p + q$, considerably smaller in size than V. Henderson et al. (1963) shows that these mixed model equations do not depend on normality and that $\hat{\beta}$ and $\hat{b}$ are the BLUE and BLUP under general conditions, provided D and Σ are known.

Robinson (1991) points out that although Henderson (1950) initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (3.4) as "joint maximum likelihood estimates", Henderson (1973) later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (3.3) is a joint density rather than a likelihood function. Lee et al. (2006) remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

Estimation of the fixed parameters

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

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{3.5}$$

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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{3.6}$$

Substituting $\hat{\beta}$ from (3.6) into $\ell(\beta, \theta|y)$ from (3.5) 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 (3.6) 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 (3.4).

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.

The extended Likelihood

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

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

Given θ, differentiating with respect to β and b returns Henderson's equations in (3.4).

The LME model as a general linear model

Henderson's equations in (3.4) can be rewritten $(T'W^{-1}T)\delta = T'W^{-1}y_a$ using

$$\delta = \begin{pmatrix} \beta \\ b \end{pmatrix},\ y_a = \begin{pmatrix} y \\ \psi \end{pmatrix},\ T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix},\ \text{and } W = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix},$$

where **?** describe $\psi = 0$ as quasi-data with mean $\mathrm{E}(\psi) = b$. Their formulation suggests that the joint estimation of the coefficients β and b of the linear mixed effects model can be derived via a classical augmented general linear model $y_a = T\delta + \varepsilon$ where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = W$, with *both* β and b appearing as fixed parameters. The usefulness of this reformulation of an LME as a general linear model will be revisited.

3.2 Repeated measurements in LME models

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

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

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

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

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

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

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

3.2.1 Formulation of the Response Vector

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

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

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

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

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

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

3.2.2 Correlation Terms

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

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

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

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.

Lam et al. (1999) used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

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

The Variance Covariance Matrix

The LME model can be written

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a vector of fixed effects, and $\boldsymbol{X}_i$ is a corresponding $2n_i \times 3$ design matrix for the fixed effects. The random effects are expressed in the vector $\boldsymbol{b} = (b_1, b_2)'$, with $\boldsymbol{Z}_i$ the corresponding $2n_i \times 2$ design matrix. The vector $\boldsymbol{\epsilon}_i$ is a $2n_i \times 1$ vector of residual terms. Random effects and residuals are assumed to be independent of each other. The variance matrix of $\mathbf{Y}$, denoted $\mathbf{V}$, is an $n \times n$ matrix that can be expressed as follows;

$$\mathbf{V} = \text{Var}(\mathbf{Xb} + \mathbf{Zb} + \mathbf{e}) \tag{3.7}$$

$$\mathbf{V} = \text{Var}(\mathbf{Xb}) + \text{Var}(\mathbf{Zb}) + \text{Var}(\mathbf{e}) \tag{3.8}$$

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

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

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$.

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

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

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

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

$$\boldsymbol{D} = \text{Var}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix}$$

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$. The random effects and residuals are assumed to be independent. The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$.

By letting $\text{var}(b) = D$ (i.e $\mathbf{b}$ $N(0, \mathbf{D})$), this becomes ZDZ^T. This specifies the covariance due to random effects. The residual covariance matrix $\text{var}(e)$ is denoted as R, ($\mathbf{e}$ $N(0, \mathbf{R})$). Residual are uncorrelated, hence $\mathbf{R}$ is equivalent to $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The variance matrix $\mathbf{V}$ can therefore be written as;

$$\mathbf{V} = \mathbf{ZDZ}^T + \mathbf{R} \tag{3.9}$$

3.3 LME models in Method Comparison Studies

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. Consequently LME approaches have 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 a great understanding of residual analysis and influence analysis for LME models has been adchieved thanks to authors such as Schabenberger (2004), Christensen et al. (1992), Cook (1986) West et al. (2007), amongst others.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. These authors remark that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' approaches, as advocated in Bland and Altman (1999), describing them as tedious, unnecessary and 'outdated'. Rather than using the 'by hand' methods, estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes associated constraints, such as the need for the design to be perfectly balanced.

Barnhart et al. (2007) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Varying degrees of importances should be attached to each the three agreement criteria listed by Barnhart et al. (2007). Between-item variance d_i^2 is fundamentally a measure of the variability of the item-wise means, as measured by method i, but it does contain limited information on the precision of that method.

For conventional method comparison problems, both methods measures the same set of items using the same unit of measurement. Convergence to equality of between-item variance inevitable as the number of items n increases. Significantly different estimates for d_1^2 and d_2^2 should not be expected for any practical problem.

Therefore a violation of third criterium (i.e. different between-item variances) criterium is contingent upon, and a possible consequence of, the violation of the other two agreement criteria. However, a violation of the third criterium will not occur in isolation. As noted elsewhere, the matter of inter-method bias can be easily accounted

for, once detected. Both between-items and within-items variances must be calculated such that sources of variances are properly assigned, and to compute limits of agreement. However, testing the within-item criterium is the most informative analysis and therefore requires the most attention.

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement.

Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem, which extends beyond the conventional method comparison study question. The data used for their examples is unavailable for independent use.

3.3.1 Roy's Methodology

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

Roy (2009) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

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 $\mathrm{H}_3 \colon d_1^2 = d_2^2$ 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.

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

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

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

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

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

3.3.2 Replicate measurements in Roy's paper

Roy (2009) uses the same definition of replicate measurement as Bland and Altman (1999); measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. under identical conditions. Roy (2009) notes that some measurements may not be 'true' replicates, as data can not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one $AR(1)$ structure. However determining MLEs with such a structure would be computational intense, if possible at all.

3.3.3 Test for inter-method bias

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

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

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

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

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

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

3.3.4 Roy's hypothesis tests : Roy's variability tests

Lack of agreement can also arise if there is a disagreement in overall variabilities. This lack of agreement may be due to differing between-item variabilities, differing within-item variabilities, or both. The formulation previously presented by Roy usefully facilitates a series of significance tests that assess if and where such differences arise. These tests are comprised of a formal test for the equality of between-item variances. The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models. The models are compared using the likelihood ratio test, a general method for comparing nested models fitted by ML (Lehmann and Romano, 2006).

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

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

$$\mathrm{H}_3 : \sigma_1^2 = \sigma_2^2$$
$$\mathrm{K}_3 : \sigma_1^2 \neq \sigma_2^2$$

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

$$\mathrm{H}_4 : \omega_1^2 = \omega_2^2$$
$$\mathrm{K}_4 : \omega_1^2 \neq \omega_2^2$$

Two methods can be considered to be in agreement if criteria based upon these

methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

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

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

3.3.5 Model Specification for Roy's Hypotheses Tests

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.

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

3.3.6 Specifying the Models

Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed.

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

The difference in the models are specifically in how the the D and Σ matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively.

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

3.3.7 Variability Tests

Variability tests proposed by Roy (2009) affords the opportunity to expand upon Carstensen's approach. Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

Variability test 1

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

$$H_0 : \ d_1 = d_2$$
$$H_A : \ d_1 \neq d_2$$

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

Variability test 2

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

$$H_0 : \ \sigma_1 = \sigma_2$$

$$H_A : \ \sigma_1 \neq \sigma_2$$

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

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

Variability test 3

Roy also integrates H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + d_m^2$ represent the overall variability of method m. Roy (2009) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. An examination of this topic is useful because a method for computing Limits of Agreement follows from here.

Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

The estimated overall variance covariance matrix 'Block Ω_i' is the addition of estimate of the between-subject variance covariance matrix $\hat{D}$ and the within-subject variance covariance matrix $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{3.10}$$

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability (Σ), Roy (2009) denotes the overall variability as Block - $\mathbf{\Omega}_i$. The overall variation for methods 1 and 2 are given by

$$\text{Block } \mathbf{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

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

$$H_0 : \ \omega_1 = \omega_2$$

$$H_A : \ \omega_1 \neq \omega_2$$

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

3.4 Correlation terms

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

In addition to the variability tests, Roy (2009) advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked, and demonstrates that placing undue importance to it can lead to incorrect conclusions.

Roy (2009) remarks that PROC MIXED only gives overall correlation coefficients, but not their variances. Similarly variance are not determinable in R as yet either. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

The methodology proposed by Roy (2009) is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999). Hamlett re-analyses the data of Lam et al. (1999) to generalize their model to cover other settings not covered by the Lam method. In many cases, repeated observation are collected from each subject in sequence and/or longitudinally.

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

Also, implementation requires that the between-item variances are estimated as the same value: $d_1^2 = d_2^2 = d^2$. Necessarily Carstensen's method does not allow for a formal test of the between-item variability.

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

In cases where the off-diagonal terms in the overall variability matrix are close to zero, the limits of agreement due to Carstensen et al. (2008) are very similar to the limits of agreement that follow from the general model.

3.4.1 Correlation

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

In addition to the variability tests, Roy advises that it is preferable that a correlation

of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions.

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

3.4.2 Formal Testing for Covariances

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

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

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

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

As it is pertinent to the difference between the two described methodologies, the facilitation of a formal test would be useful. Extending the approach proposed by ARoy2009, the test for overall covariance can be formulated:

$$\mathrm{H}_5 : \sigma_{12} = 0$$

$$\mathrm{K}_5 : \sigma_{12} \neq 0$$

As with the tests for variability, this test is performed by comparing a pair of model fits corresponding to the null and alternative hypothesis. In addition to testing the

overall covariance, similar tests can be formulated for both the component variabilities if necessary.

3.5 Extension of Roy's methodology

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

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

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

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

3.5.1 Roy's methodology for single measurements

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

3.6 Conclusion

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

Chapter 4

Limits of Agreement

4.1 Introduction to LME Methods for Computing Limits of Agreement

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

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

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the

effect of repeated measurement error. Bland and Altman (1999) propose a correction for this. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Carstensen et al. (2008) takes issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation.

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

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

Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements, by computing an appropriate estimate for the standard deviation of case-wise differences, so as to determine the limits of agreement. This approach is similar to Deming's regression, and for estimating variance components for measurements

by different methods.

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

4.2 Limits of Agreement in LME models

Carstensen's approach is that of a standard two-way mixed effects ANOVA with replicate measurements. With regards to the specification of the variance terms, Carstensen remarks that using his approach is common, remarking that "The only slightly non-standard feature is the differing residual variances between methods" (Carstensen, 2010).

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{4.1}$$

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

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{4.2}$$

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

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

This formulation doesn't require the data set to be balanced, but does require a sufficient number of replicates and measurements to overcome the problem of identifiability. Consequently more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{4.3}$$

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

4.3 Computation of Limits of agreement in LME models

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.Between-subject variation for method m is given by d_m^2 (in the author's notation τ_m^2) and within-subject variation is given by σ_m^2.

Carstensen et al. (2008) remarks that for two methods A and B, separate values of d_A^2 and d_B^2 cannot be estimated, only their average. Hence the assumption that

$d_x = d_y = d$ is necessary.

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

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

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

The variance for method m is $d_m^2 + \sigma_m^2$. Limits of agreement are determined using the standard deviation of the case-wise differences between the sets of measurements by two methods A and B, given by

$$\text{var}(y_A - y_B) = 2d^2 + \sigma_A^2 + \sigma_B^2. \tag{4.4}$$

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

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

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

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

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{4.5}$$

4.3.1 Computing Limits of Agreement using Roy's Model

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 $\boldsymbol{\Sigma}$. Using Roy's methodology, the variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed. The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\boldsymbol{\Omega}_i$ matrix. The variance of differences is easily computable from the variance estimates in the Block - $\boldsymbol{\Omega}_i$ matrix, i.e.

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

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

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject Variance Covariance matrix. For Carstensen's 'fat' data, the limits of agreement computed using Roy's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

4.3.2 Linked Replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods. Carstensen et al. (2008) demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. As a surplus source of variability is excluded from the computation, the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates.

Carstensen et al. (2008) introduces a second data set; the oximetry study. This

study done at the Royal Children's Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

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

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

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

$$\hat{\mathbf{\Sigma}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Sigma}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{4.6}$$

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

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

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

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

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

4.4 Differences Between Models

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations.

Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008)

formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

In contrast to Roy's model, Carstensen's model requires that commonly used assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off-diagonal elements are also zero. Therefore the variance covariance matrices for between-item and within-item variability are respectively.

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

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

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by Carstensen et al. (2008).

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

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

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

item i using Carstensen's model accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population.

However this approach seems contrary to the purpose of LoAs as a prediction interval for a population of items. Conversely, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

4.5 Carstensen Coefficient of Repeatability

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

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

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

Chapter 5

Residual Analysis and Influence Diagnostics for Method Comparison

Model validation and model appraisal are vital parts of the modelling process, yet are too often overlooked. Using a small handful of simple measures and methods, such as the AIC and R^2 measures, is insufficient to properly assess the usefulness of a fitted model. A full and comprehensive analysis that comprises residual analysis and influence analysis for testing model assumptions, should be carried out. In classical linear models model diagnostics are now considered a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers (Christensen et al., 1992; Schabenberger, 2004) that model diagnostics do not often accompany LME model analyses.

5.1 Residual Analysis

In classical linear models, model diagnostics techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations, and have been become a required part of any statistical analysis. Well established methods are commonly available in statistical packages and standard textbooks on ap-

plied regression. However it has been noted by several papers that model diagnostics do not often accompany LME model analyses. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. As with classical models, there are two key techniques: a residual plot and the normal probability plot. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect. that is to say, if the residuals behave randomly, with no discernible trend, the model has fitted the data well. If some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted.

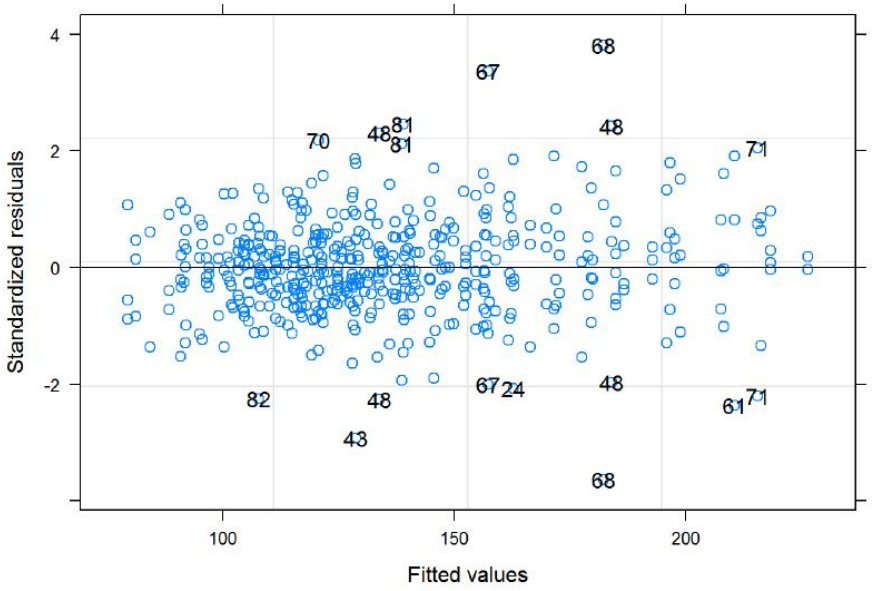
However, for LME models the matter of residual is more complex, both from a theoretical point of view and from the practical matter of implementing a comprehensive analysis using statistical software. Schabenberger (2004) discusses residuals for LME model, providing a useful summary of various techniques. Prominent in literature is the taxonomy of residuals for LME Models, distinguishing between condition residuals, marginal residuals and EBLUPS, including Hilden-Minton (1995); Schabenberger (2004); West et al. (2007); Nobre and Singer (2007). The underlying assumptions for LME models are similar to those of classical linear mdoels.

Statistical software environments, such as the `R` Programming language, provides a suite of tests and graphical procedures for appraising a fitted linear model, with several of these procedures analysing the model residuals. Texts such as Pinheiro and Bates (1994); West et al. (2007); Gałecki and Burzykowski (2013) describe what can be implemented for LME residual analyses with statistical software, such as `R` and `SAS`.

In the context of Method Comparison, a residual analysis would be carried out just as any other LME model would, testing normality. As such there is little scope for adding additional insights, other than to say that it is possible to create plots specific to each method. The figures on the next page depict the residual analysis for the Blood data. This can be used to indicate which methods disagree with the rest, but these would be a confirmation of something detected previously.

Analysis of the residuals could determine if the methods of measurement disagree

systematically, or whether or not erroneous measurements associated with a subset of the cases are the cause of disagreement. The figure depicts residual plot for the Systolic Blood Pressure example. Points are labelled by subjects, with cases 67, 68 and 71 being among the prominent cases. Prominent cases warrant further investigation, but an analyst should procede to influence diagnostics beforehand.



The next figure depicts residual plot for the Systolic Blood Pressure example, panelled by the various measurement methods. It serves to confirm agreement between methods J and R, with lack of agreement between those two methods and method S. However, little insight can be gained as to what actually causes lack of agreement here.

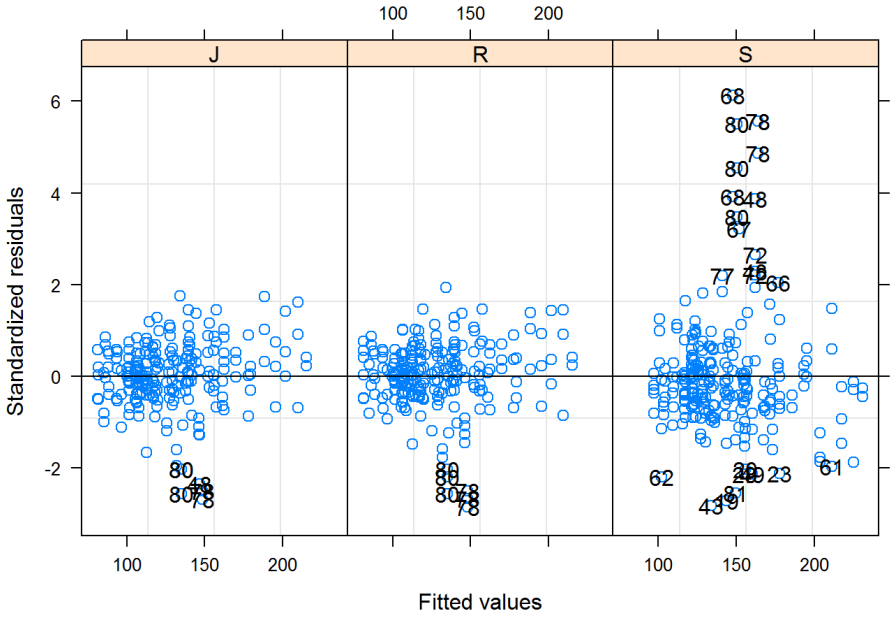


Figure 5.1.1: LME Residuals by Method (Blood Pressure Data)

5.2 Influence Diagnostics

Model diagnostic techniques can determine whether or not the distributional assumptions are satisfied, but also to assess the influence of unusual observations.

Following model specification and estimation, it is of interest to explore the model-data agreement by raising pertinent questions. Pinheiro and Bates provide some insight into how to compute and interpret model diagnostic plots for LME models. Unfortunately this aspect of LME theory is not as expansive as the corresponding body of work for Linear Models. Their particular observations will be reverted to shortly. Further to the analysis of residuals, Schabenberger (2004) recommends the examination of the following questions:

- Does the model-data agreement support the model assumptions?
- Should model components be refined, and if so, which components? For example, should certain explanatory variables be added or removed, and is the covariance of the observations properly specified?

- Are the results sensitive to model and/or data? Are individual data points or groups of cases particularly influential on the analysis?

The last of these three questions, regarding influential points, is of particular interest in the context of Method Comparison. After fitting an LME model, it is important to carry put model diagnostics to check whether distributional assumptions for the residuals as satisfied and whether the fit the model is sensitive to unusual assumptions. The process of carrying out model diagnostic involves several informal and formal techniques, which will mentioned throughout the chapter.

Influential points have a large influence on the fit of the model. Influential points are a set of one or more observations whose removal would cause a different conclusion in the analysis, e.g. substantially changes the estimate of the regression coefficients. West et al. (2007) remarks that influence diagnostics play an important role in the interpretation of results, because influential data can negatively influence the statistical model and generalizability of the model. Schabenberger (2004) remarks that the concept of critiquing the model-data agreement applies in mixed models in the same way as in linear fixed-effects models. In fact, because of the more complex model structure, you can argue that model and data diagnostics are even more important (West et al., 2007).

5.2.1 A Procedure for Quantifying Influence

Schabenberger (2004) describes a simple procedure for quantifying influence for LME Models. Firstly a model should be fitted to the data, and estimates of the parameters should be obtained. The second step is that either single or multiple data points, specifically outliers, should be omitted from the analysis, with the original parameter estimates being updated. This is known as *leave one out* or *leave k out* analysis. The final step of the procedure is comparing the sets of estimates computed from the entire and reduced data sets to determine whether the absence of observations changed the analysis.

Influence can be thought of as consequence of leverage and outlierness. Outliers are the most noteworthy data points in an analysis, and an objective of influence analysis is how influential they are, and the manner in which they are influential. They can point to a model breakdown and lead to development of a better model. The linear mixed effects model is a useful methodology for fitting a wide range of models. However, linear mixed effects models are known to be sensitive to outliers. Specifically likelihood based estimation techniques, such as ML and REML, are sensitive to outliers. Christensen et al. (1992) advises that identification of outliers is necessary before conclusions may be drawn from the fitted model. The leverage of an observation is a further consideration.

An observation with an extreme, but not unusual, value on a predictor variable is a point with high leverage. High leverage points can have a great amount of effect on the estimate of regression coefficients. In general. a high leverage point means a extreme value for the one or more of the independent variables, and a greater potential of overly influencing the final fitted model. However, if a case has extreme values for the independent variables but is fitted very well by a regression model, this case is not necessarily overly influential.

In classical linear models, leverages are the diagonal elements h_{ii} of the Projection matrix, also known as the Hat Matrix $\boldsymbol{H}$. Schabenberger (2004) describes two analogues of $\boldsymbol{H}$ for LME models. However the practical use for either approach is not made clear.

5.2.2 Analyzing Influence in LME models

Model diagnostic techniques, well established for classical models, have since been adapted for use with linear mixed effects models. Diagnostic techniques for LME models are inevitably more difficult to implement, due to the increased complexity.

Influence diagnostics are formal techniques allowing for the identification of observations that exert substantial influence on the estimates of fixed effects and variance covariance parameters. While linear models and GLMS can be studied with a wide

range of well-established diagnostic technqiues, the choice of methodology is much more restricted for the case of LMEs. However influence diagnostics for LME Models is an area of active research. Research on diagnostic analyses for LME models are presented in Beckman et al. (1987), Christensen et al. (1992), Hilden-Minton (1995), Lesaffre and Verbeke (1998), Banerjee and Frees (1997), Fung et al. (2002), Demidenko (2004), Zewotir and Galpin (2005), Zewotir (2008) and Nobre and Singer (2007, 2011).

Schabenberger (2004) states that goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis.

5.2.3 Measuring of Influence for LME Models

Influence analysis methodologies have been used extensively in classical linear models, and provided the basis for methodologies for use with LME models. Computationally inexpensive diagnostics tools have been developed to examine the issue of influence (Zewotir and Galpin, 2005).

Zewotir and Galpin (2005) remarks the development of efficient computational formulas is crucial making deletion diagnostics useable, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model. A number of approaches to model diagnostics are described, including variance components, dixed effects parameters, prediction of the response variable and of random effects, and the likelihood function. Influence statistics can be grouped by the aspect of estimation that is their primary target:

- **overall measures compare changes in objective functions**: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2)
- **influence on parameter estimates**: Cook's (Cook 1977, 1979), MDFFITS (Belsley, Kuh, and Welsch 1980, p. 32)

- **influence on precision of estimates**: CovRatio and CovTrace
- **influence on fitted and predicted values**: PRESS residual, PRESS statistic (Allen 1974), DFFITS (Belsley, Kuh, and Welsch 1980, p. 15)
- **outlier properties**: internally and externally studentized residuals, leverage

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, and the Andrews-Prebigon statistic.

The subscript (U) is used to denote quantities computed from data with subset of cases U omitted. If the global measure suggests that the points in U are influential, you should next determine the nature of that influence. In particular, the points can affect

- the estimates of fixed effects
- the estimates of the precision of the fixed effects
- the estimates of the covariance parameters
- the estimates of the precision of the covariance parameters
- fitted and predicted values

For example, if observations primarily affect the precision of the covariance parameters without exerting much influence on the fixed effects, then their presence in the data may not distort hypothesis tests or confidence intervals about β. Schabenberger (2004) notes that removing observations or sets of observations affects fixed effects and covariance parameter estimates.

5.2.4 Deletion Diagnostics

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on parameters inferences for a fitted model. For classical linear models, Cook (1977) greatly expands the study of residuals and influence 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. Cook's key observation was the effects of deleting each observation in turn could be calculated with little additional computation. Cook proposed a measure that combines the information of leverage and residual of the observation, now known simply as the Cook's Distance, $D_{(i)}$, which can be calculated without fitting a new regression coefficient each time an observation is deleted. Consequently deletion diagnostics have become an integral part of assessing linear models.

It must be pointed out that the effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's distance, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large.

Christensen et al. (1992) notes the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. Christensen et al. (1992) developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted resgression problem (conditional on the estimated covariance matrix) for fixed effects.

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

undermines the use of many proposed procedures for Method Comparison.

5.2.5 Cook's Distance

As previously described, Cooks Distance (D_i) is a diagnostic technique used in classical linear models, that functions as an overall measure of the influence of an observation that is a measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Cook's Distance as a measure of the influence of observations in subset U on a vector of parameter estimates is given below (Cook, 1977)

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)}.$$

Observations, or sets of observations, that have high Cook's distance usually have high residuals, although this is not necessarily the case.

If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Large values for Cook's distance indicate observations for special attention. Cook's distance can be used in several ways: to indicate data points that are particularly worth checking for validity; to indicate regions of the design space where it would be good to be able to obtain more data points.

Use of threshold values for Cook's Distance is discouraged (Fox, 1997). However, informal heuristics do exist for OLS models; Obervations for which Cook's distance is higher than 1 are usually considered as influential. Another informal threshold of $4/n$ or $4/(n-k-1)$, where n is the number of observations and k the number of explanatory variables. Fox (1997) advises the use of diagnostic plotting and to examine in closer details the points with *"values of D that are substantially larger than the rest"*, and that thresholds should feature only to enhance graphical displays.

The effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's distance, for example, can still greatly

affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large.

Christensen et al. (1992) develops case deletion diagnostics, in particular the equivalent of Cook's distance for diagnosing influential observations when estimating the fixed effect parameters and variance components, adapting the Cook's Distance measure for the analysis of LME models. For LME models, two formulations exist; a Cook's distance that examines the change in fixed fixed parameter estimates, and another that examines the change in random effects parameter estimates. The outcome of either Cook's distance is a scaled change in either β or θ. Zewotir and Galpin (2005) gives a detailed discussion of the various formulation for Cook's distances for LME Models.

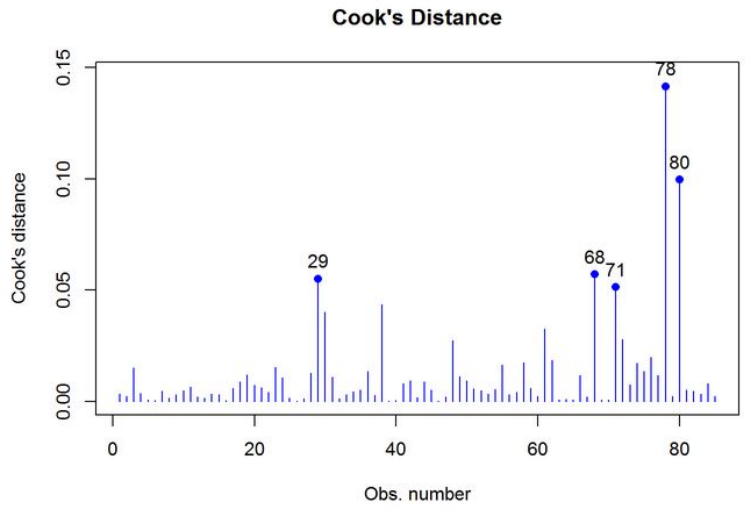


Figure 5.2.2:

Consideration of how leave-U-out diagnostics would work in the context of Method Comparison problems is required. There are several scenarios. Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called '*observation-diagnostics*'. For multiple observations, Preisser describes the diagnostics as '*cluster-deletion*' diagnostics. Suppose we have two methods of measurement X and Y, each with three measurements for a specific case: $(x_1, x_2, x_3, y_1, y_2, y_3)$

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

The natural sampling unit is the item or subject, similar to the example provided by Schabenberger (2004). Hence, the third option, henceforth, referred to as "Leave subject Out" will be the option used.

5.2.6 Local Influence

Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models, introducing methods for local influence assessment for classical linear models. These methods provide a powerful tool for examining perturbations in the assumption of a model, particularly the effects of local perturbations of parameters of observations. The local-influence approach to influence assessment is quite different from the case deletion approach, comparisons are of interest.

Beckman et al. (1987) applied the local influence method of Cook (1986) to the analysis of the LME model. Other authors such as Lesaffre and Verbeke (1998) have also extended these idea to LME models. While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known. As such the local influence approach are not particularly useful in the context of Method Comparison, and so will not be considered further.

5.2.7 Comparing Influence and Residual Analysis

Nieuwenhuis et al. (2012) compares residual analysis and influence analysis. Cases with high residuals (defined as the difference between the observed and the predicted scores

on the dependent variable) or with high standardized residuals (defined as the residual divided by the standard deviation of the residuals) are indicated as outliers.

However, an influential case is not necessarily an outlying residual. On the contrary: a strongly influential case dominates the regression model in such a way, that the estimated regression line lies closely to this case. The analysis of residuals cannot be used for the detection of influential cases (Crawley, 2012).

5.2.8 Iterative and Non-Iterative Influence Analysis

For linear models, the implementation of influence analysis is straightforward, but for LME models the process is more complex. Schabenberger (2004) examines the use and implementation of influence measures in LME models. Schabenberger (2004) highlights some of the issue regarding implementing LME model diagnostics, describing the choice between iterative influence analysis and non-iterative influence analysis. Schabenberger (2004) considers several important aspects of the use and implementation of influence measures in LME models, noting that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates. Closed-form expressions for computing the change in important model quantities might not be available.

On a related matter, Schabenberger (2004) describes the scenario wherein a data point is removed and the new estimate of the D matrix is not positive definite. This may occur if a variance component estimate now falls on the boundary of the parameter space (Schabenberger, 2004).

For classical linear models, it is not necessary to refit the model after removing a data point in order to measure the impact of an observation on the model. The change in fixed effect estimates, residuals, residual sums of squares, and the variance-covariance matrix of the fixed effects can be computed based on the fit to the full data alone, using update formulas (Sherman and Morrison, 1950; Hager, 1989).

However, in LME models several important complications arise. Data points can

affect not only the fixed effects but also the covariance parameter estimates on which the fixed-effects estimates depend.

When applied to LME models, such update formulas are available only if one assumes that the covariance parameters are not affected by the removal of the observation in question. However, this is rarely a reasonable assumption. For LME models, non-iterative methods are computationally efficient, but require the rather strong assumption that all covariance parameters are known, and thus are not updated, with the exception of the profiled residual variance. Update formulas for "leave-U-out" estimates typically fail to account for changes in covariance parameters. As the influence that each item would have on the variance estimate of a method comparison model is crucial, this substanitally negates their usefulness for Roy's Model.

Iterative influence diagnostics requiring fitting the model without the observations in question. Computation execution time is substantially longer, although this is balanced by algorithmic simplicity, with no assumptions beyond those used for the original model. A measure of total influence requires updates of all model parameters. This can only be achieved in general is by omitting observations or cases, then refitting the model.

An iterative analysis may seem computationally expensive. Computing iterative influence diagnostics for n observations requires $n + 1$ mixed models to be fitted iteratively. The execution times for iterative procedures are longer relative to non-iterative procedures, but are not so long that they would dissuade an analyst from using them. Despite the addition execution time of iteratives approaches, they are preferable for Method Comparison problems, as they can facilitate several complementary analyses concurrently.

Iterative methods retain the potential for useful analyses, if applied at different stage of the modelling process. Diagnostic measures, specifically the DFBETA, have characteristics that would make them very useful at the exploratory stage of the method comparison process. Implicitly various assumptions about variance are used, but simultaneously an approach based on DFBETA can be used to assess if these assumptions

are valid.

5.2.9 Likelihood Distance

An overall influence statistic measures the change in the objective function being minimized. For example, in classical linear, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance (Cook and Weisberg, 1983).

The likelihood distance is a global summary measure that expresses the joint influence of the subsets of observations, U, on all parameters that were subject to updating. Schabenberger (2004) points out that the likelihood distance $LD(\psi_{(U)})$ is not the log-likelihood obtained by fitting the model to the reduced data set. Instead it is obtained by evaluating the likelihood function based on the full data set (containing all n observations) at the reduced-data estimates.

The procedure requires the calculation of the full data estimates $\hat{\psi}$ and estimates based on the reduced data set $\hat{\psi}_{(U)}$. The likelihood distance is given by determining

$$LD_{(U)} = 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\}$$

$$RLD_{(U)} = 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\}$$

Large values indicate that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ differ considerably.

West et al. (2007) examines a group of methods that examine various aspects of influence diagnostics for LME models. For overall influence, the most common approaches are the *likelihood distance* and the *restricted likelihood distance*.

5.3 Model Diagnostics for Roy's Models

Further to previous work, this section revisits case-deletion and residual diagnostics, and explores how approaches devised by Gałecki and Burzykowski (2013) can be used

to appraise Roy's model. These authors specifically look at Cook's Distances and Likelihood Distances.

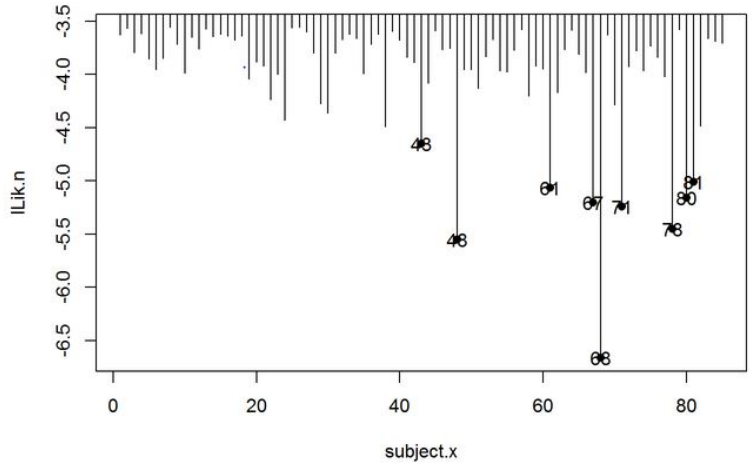


Figure 5.3.3:

Case Deletion Diagnostics for Variance Ratios

Taking the core principals of his methods, and applying them to the Method Comparison problem, case deletion diagnostics are used on the variance components of the Roy's model, specifically the ratio of between subject variances and the within subject covariances respecitvely.

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

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

The Grubbs' Test for Outliers is a commonly used technique for assessing outlier in a univariate data set, of which there are several variants. As there may be several outliers present, the Grubbs test is not practical. However an indication that a point being beyond the fences according to Tukey's specification for boxplots will suffice.

The WSVR values are plotted against the corresponding BSVR values, with commonly used bivariate methods may be applied jointly to the both sets of data sets, e.g Mahalanobis distances. Confidence ellipses can be superimposed over the plot with minimal effort. Two ellipses are generated by this technique, a 50 % and 97.5% confidence ellipse respectively. Outlying cases are idenified by the plot. Subject 68 is the most prominent case.

The subjects were ranked by Mahalanobis distance, with the top 10 being presented in the following table. Both sets of ratio are addtionally expressed as a ratio of the full model variance ratios.

Subject (u)	MD	$WSVR_{(u)}$	WSVR (%)	$BSVR_{(u)}$	BSVR (%)
68	44.7284	1.3615	0.9132	1.0353	0.9849
30	16.7228	1.5045	1.0092	1.1024	1.0487
71	11.5887	1.5210	1.0202	1.0932	1.0400
80	11.0326	1.4796	0.9925	1.0114	0.9621
38	10.3671	1.5011	1.0069	1.0917	1.0385
67	10.1940	1.4308	0.9598	1.0514	1.0002
43	7.6932	1.4385	0.9649	1.0511	0.9999
72	4.7350	1.4900	0.9995	1.0262	0.9762
48	4.4321	1.4950	1.0028	1.0280	0.9779
29	4.3005	1.4910	1.0001	1.0769	1.0244

From this table one may conclude that subjects 72, 48 and 29 are not particularly influential. Interestingly Subject 78, which was noticeable in the case deletion diagnostics for fixed effects, does not feature in this table.

Variance Ratios

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

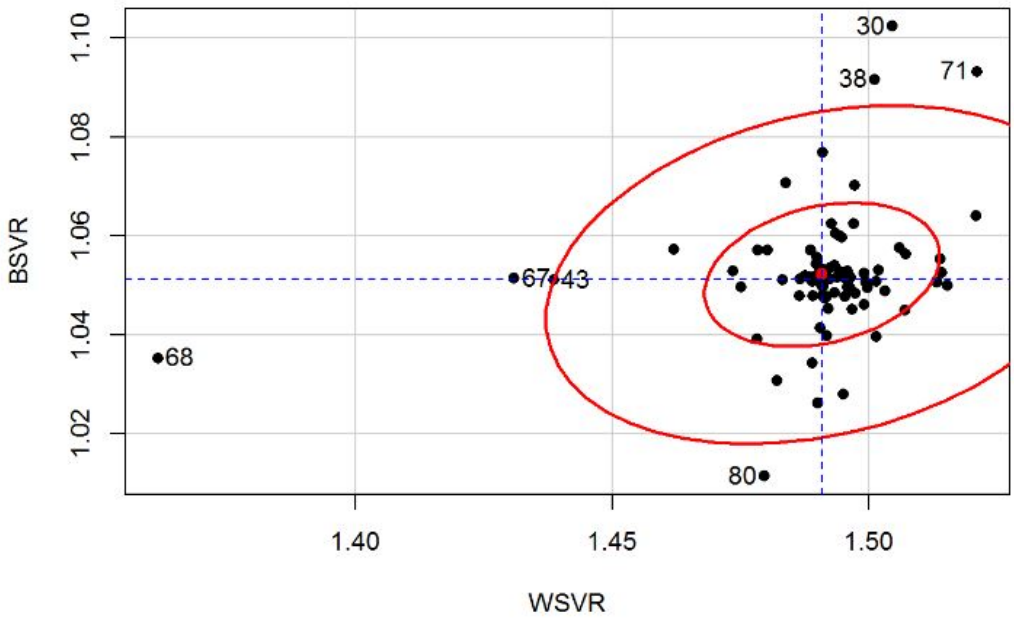


Figure 5.3.4:

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.

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. Bootstrap methods for computing confidence intervals may be considered.

5.4 Using DFBETAs from LME Models to Assess Agreement

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. DFBETA and DFFITS are well known measures of influence. Emphasis shall be placed on DFBETA, but a brief discussion of DFFITS is merited as it potentially provides for useful techniques in method comparison. Schabenberger (2004) provides a mathematical desciption of both.

DFBETAS is a standardized measure of the absolute difference between the estimate with a particular case included and the estimate without that particular case,, thus measuring the impact each observation has on a particular predictor (Belsley et al., 2005). For LME models, the DFBETA is a measure that standardizes the absolute difference in parameter estimates between an LME model based on a full set of data, and a model from reduced data.

In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley et al. (2005) recommend 2 as a general cutoff value to indicate influential observations and as a size-adjusted cutoff. There is no agreement as to the critical threshold for DFBETAs. The cut-off value for DFBETAs is $\frac{2}{\sqrt{n}}$, where n is the number of observations. However, another cut-off is to look for observations with a value greater than 1.00. Here cutoff means, "this observation could be overly influential on the estimated coefficient".

DFFITS is a diagnostic meant to show how influential a point is in a statistical regression. It is defined as the change, in the predicted value for a point, obtained when that point is left out of the regression, divided by the estimated standard deviation of the fit at that point:

DFBETAs for Method Comparison

For LME models, a value for DFBETAS is calculated for each of the k fixed effects, and for each of the n item. Correctly there will be $p+1$ DFBETAs (the intercept, β_0, and one β for each covariate). When the LME model is specified without an intercept term, as in Roy's Model, there is a set of DFBETAs corresponding to each measurement method, hence an $n \times p$ matrix.

In the case of method comparison studies, a series of scatterplots can be constructed to compare each pair of measurement methods. Furthermore 95% confidence ellipse can be constructed around these scatterplots.

The LME approach proposed by Roy (2009) is constrained by computational tractability. Consequently a simpler LME formulation is used, one similar to that of Carstensen et al. (2008). However one constraint that can be dispensed with is the restriction to two methods of measurement: we can now use any number of methods. The benefit of using this model is that metrics such as Cook's Distance and DFBETAs can be computed also.

Furthermore, these measures form the basis of the analysis, rather than the estimates derived from the model. In the context of method comparison, these variables are the methods of measurement. Agreement will considered in the context of inter-method bias and the within-item variance ratio. Between-item variance ratio is not considered for this analysis.

For a Method Comparison study, DFBETAs can be used as a proxy measurement, allowing simple techniques to be used for assessing agreement. Suppose an LME model was formulated to model agreement for two or more methods of measurement, specifically with replicate measurements. If the methods are to be agreement, the DFBETAs for each case would be the same for both methods. As such, agreement between any two methods can be determined by a simple scatterplot of the DFBetas.

If the lack of agreement is caused, in part or in full, by differing within-item variances, there would be differing DFBETAs for each pair of methods. If the points align

along the line of equality, then both methods can be said to be in agreement for within item variance. However DFBETAs are not useful for determining inter-method bias. If there is good agreement between methods, or if lack of agreement is caused by inter-method bias only, the DFBETA values will be almost identical for each subject in the data set.

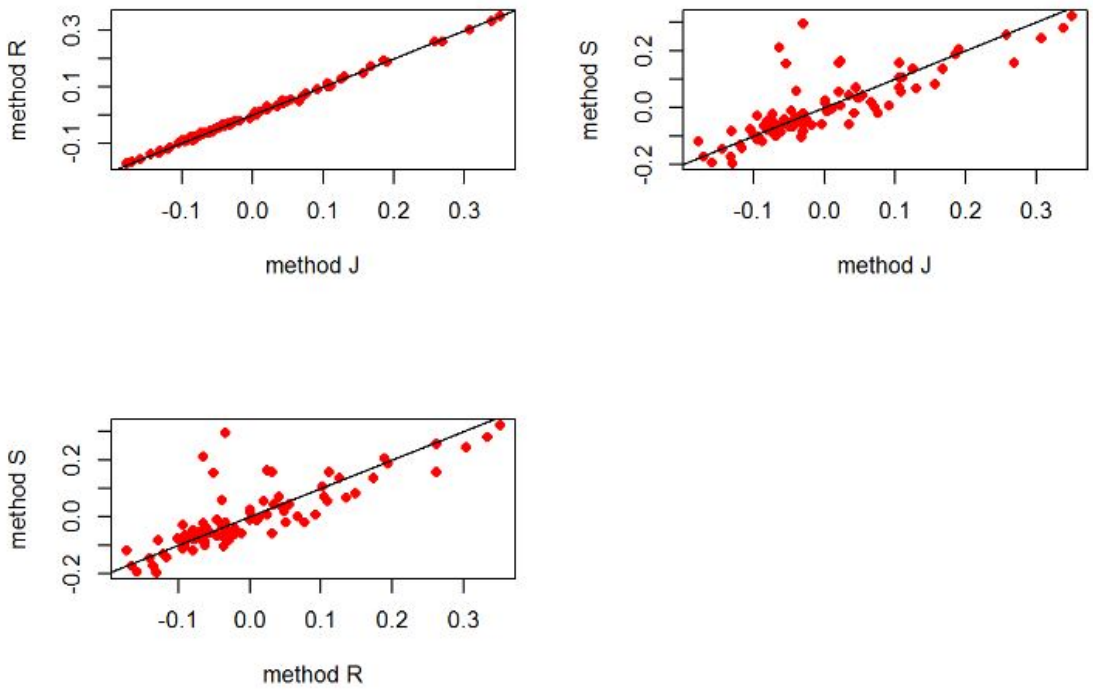
Following the idea proposed by Bland and Altman (1986), an identity plot to visually inspect this relationship between sets of DFBETAs. Modern statistical software usually allows for the creation of co-plots, so a grid of identity plots may be easily rendered for comparing each pair of methods. Used in conjection with a Bland-Altman plot, this co-plot can quickly determine agreement and indicate the source of lack of agreement.

For an LME model fitted to the Blood data, the results tabulated below can be produced. Cases can be ranked by the Cook's Distance, such that the most divergent DFBETA are highlighted, with the top 6 being presented below). The remaining columns are the DFBeta for each of the fixed effects, for each of the 85 subject.

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

For DFBETA identity plots are presented below. This set of plots indicate agreement between methods J and R in terms of within-item variance, while severe lack of agreement exists between these methods and the third method S, as is the conclusion of Roy (2009).

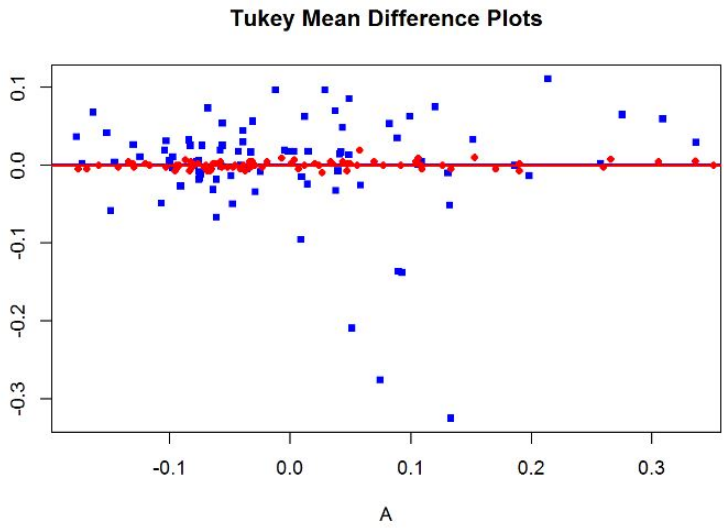
If lack of agreement is indicated, a subsequent analysis using a technique proposed by Roy (2009) can be used to identify the specific cause for this lack of agreement.



Other analyses may be used to complement these plots. The Pearson Correlation coefficient of the DFBETAs can be used in conjection with this analysis. A high correlation confirms good agreement, thouhg no threshold value for agreement is suggested.

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

As an alternative to scatterplots, a mean difference plot could be used to assess agreement of with-item variance. This mean-difference plot differs from the Bland-Altman plot in that the plot is denominated in terms of DFBETA values, and not in measurement units. Here two of the three pairs of methods are compared on the same plot, red points indicate the J-R comparison while blue points are for the J-S comparison.


Tukey Mean Difference Plots
0.1
0.0
-0.1
-0.2
-0.3
-0.1
0.0
0.1
0.2
0.3
A

Bibliography

(1975). Precision of test methods 1: Guide for the determination and reproducibility for a standard test method. Technical Report BS 597, Part 1, British Standards Institute, London.

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. (1978). Plotting probability ellipses. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 27*(3), 347–349.

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.

Banerjee, M. and E. W. Frees (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association 92*(439), 999–1005.

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.

Beckman, R., C. Nachtsheim, and R. Cook (1987). Diagnostics for mixed-model analysis of variance. *Technometrics 29*(4), 413–426.

Belsley, D. A., E. Kuh, and R. E. Welsch (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, Volume 571. John Wiley & Sons.

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.

Bland, J. M. and D. G. Altman (2003). Applying the right statistics: analyses of measurement studies. *Ultrasound in obstetrics & gynecology 22*(1), 85–93.

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).

Christensen, R., L. M. Pearson, and W. Johnson (1992). Case-deletion diagnostics for mixed models. *Technometrics 34*(1), 38–45.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics 19*, 15–18.

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

Cook, R. D. and S. Weisberg (1983). Diagnostics for heteroscedasticity in regression. *Biometrika 70*(1), 1–10.

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.

Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

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

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

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.

Eksborg, S. (1981). Evaluation of method-comparison data [letter]. *Clinical Chemistry 27*, 1311–1312.

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

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc.

Fung, W.-K., Z.-Y. Zhu, B.-C. Wei, and X. He (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 565–579.

Gałecki, A. and T. Burzykowski (2013). *Linear mixed-effects models using R: A step-by-step approach*. Springer Science & Business Media.

Giavarina, D. (2015). Understanding bland altman analysis. *Biochemia medica 25*(2), 141–151.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall Ltd.

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

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

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review 31*(2), 221–239.

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

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

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

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

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

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

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

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

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

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models*. Ph. D. thesis, University of California Los Angeles.

Hollis, S. (1996). Analysis of method comparison studies. *Ann Clin Biochem 33*, 1–4.

IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. `http://goldbook.iupac.org/R05293.html`.

Kelly, G. E. (1985). Use of the structural equations model in assessing the reliability of a new measurement technique. *Applied Statistics*, 258–263.

Kozak, M. and A. Wnuk (2014). Including the tukey mean-difference (bland–altman) plot in a statistics course. *Teaching Statistics 36*(3), 83–87.

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.

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

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models (Disc: P656-678). *Journal of the Royal Statistical Society, Series B: Methodological 58*, 619–656.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lesaffre, E. and G. Verbeke (1998). Local influence in linear mixed models. *Biometrics*, 570–582.

Lewis, P., P. Jones, J. Polak, and H. Tillotson (1991). The problem of conversion in method comparison studies. *Applied Statistics*, 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.

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.

Nieuwenhuis, R., H. te Grotenhuis, and B. Pelzer (2012). Influence. me: tools for detecting influential data in mixed effects models.

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

Nobre, J. S. and J. M. Singer (2007). Residual analysis for linear mixed models. *Biometrical Journal 49*(6), 863–875.

Nobre, J. S. and J. M. Singer (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics 38*(5), 1063–1072.

O'Brien, E., J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. Altman, M. Bland, A. Coats, and N. Atkins (1990). The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension 8*, 607–619.

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

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

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.

Preisser, J. S. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika 83*(3), 551–5562.

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

Roy, A. (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.

Roy, A., C. D. Fuller, D. I. Rosenthal, and C. R. Thomas Jr (2015). Comparison of measurement methods with a mixed effects procedure accounting for replicated evaluations (com 3 pare): method comparison algorithm implementation for head and neck igrt positional verification. *BMC medical imaging 15*(1), 35.

Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics 32*(5), 461 – 474.

Sanchez, M. M. and B. S. Binkowitz (1999). Guidelines for measurement validation in clinical trial design. *Journal of biopharmaceutical statistics 9*(3), 417–438.

Schabenberger, O. (2004). Mixed model influence diagnostics. 18929.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics 21*(1), 124–127.

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.

Voelkel, J. G. and B. E. Siskowski (2005). Center for quality and applied statistics kate gleason college of engineering rochester institute of technology technical report 2005–3.

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

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in StatisticsTheory and Methods 37*(7), 1071–1084.

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