# Transfer Report

Kevin O'Brien

February 5, 2015

# Contents

# Chapter 1

# Introduction

## 1.1 Introduction

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

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured simultaneously (and independently) by three chronographs devices, 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: 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 method of measurement should ideally be both accurate and precise. 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'. 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.

The FDA define precision as the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under prescribed conditions. Barnhart et al. (2007) describes precision as being further subdivided as either within-run, intra-batch precision or repeatability (which assesses precision during a single analytical run), or between-run, inter-batch precision or repeatability (which measures precision over time)

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than that of the other, they can not be considered to be in agreement. With this in mind a methodology is required that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

| Round | Fotobalk (F) | Counter (C) | 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 1.2: Difference between Fotobalk and Counter measurements.

## 1.2 Bland-Altman Plots

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of paired sample t-test, correlation coefficients or simple linear regression. Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983). Furthermore they proposed their simple methodology specifically constructed for method compar-

ison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be 'explained to non-statisticians'.

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

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, ..n$). These differences and averages are then plotted. This methodology, now commonly known as the 'Bland-Altman Plot', has proved very successful. Bland and Altman (1986), which further develops the methodology, was found to be the sixth most cited paper of all time by the Ryan and Woodall (2005). Dewitte et al. (2002) also commented on the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. The Bland-Altman Plot has since become expected, and often 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.

The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$. The variances around this bias is estimated by the standard deviation of the differences $S(d)$. This inter-method bias is represented with a line on

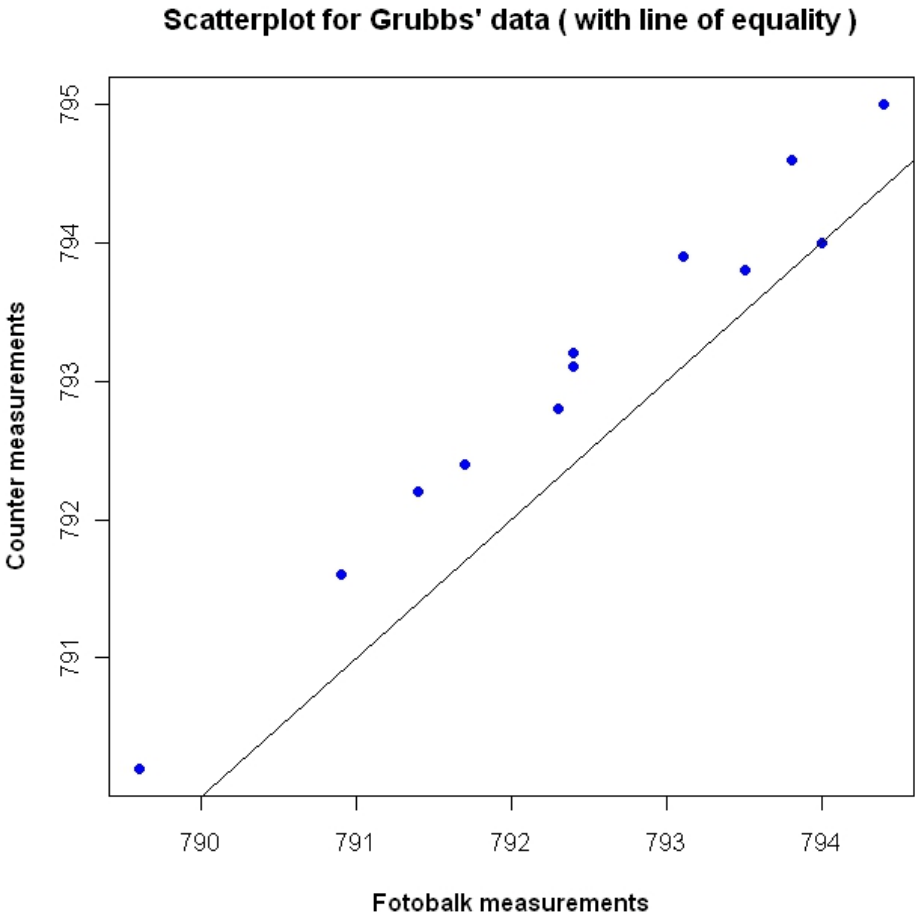


Figure 1.1: Scatter plot For Fotobalk and Counter Methods.

the Bland-Altman plot. These estimates are only meaningful if there is uniform inter-bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot. In the case of Grubbs data the inter-method bias is $-0.61$ metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

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

Table 1.3: Fotobalk and Counter methods: differences and averages.

| Round | Fotobalk [F] | Terma [T] | Differences [F-T] | Averages [(F+T)/2] |
|---|---|---|---|---|
| 1 | 793.80 | 793.20 | 0.60 | 793.50 |
| 2 | 793.10 | 793.30 | -0.20 | 793.20 |
| 3 | 792.40 | 792.60 | -0.20 | 792.50 |
| 4 | 794.00 | 793.80 | 0.20 | 793.90 |
| 5 | 791.40 | 791.60 | -0.20 | 791.50 |
| 6 | 792.40 | 791.60 | 0.80 | 792.00 |
| 7 | 791.70 | 791.60 | 0.10 | 791.65 |
| 8 | 792.30 | 792.40 | -0.10 | 792.35 |
| 9 | 789.60 | 788.50 | 1.10 | 789.05 |
| 10 | 794.40 | 794.70 | -0.30 | 794.55 |
| 11 | 790.90 | 791.30 | -0.40 | 791.10 |
| 12 | 793.50 | 793.50 | 0.00 | 793.50 |

Table 1.4: Fotobalk and Terma methods: differences and averages.

### 1.2.1 Using Bland-Altman Plots

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

> "From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase 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 Bland-Altman plot is simply a scatterplot of the case-wise averages and differences of two methods of measurement. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are particularly. Later it will be shown that case-wise differences are the sole component of the next part of the methodology, the limits of agreement.

For creating plots, the case wise-averages fulfil several functions, 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.

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

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.

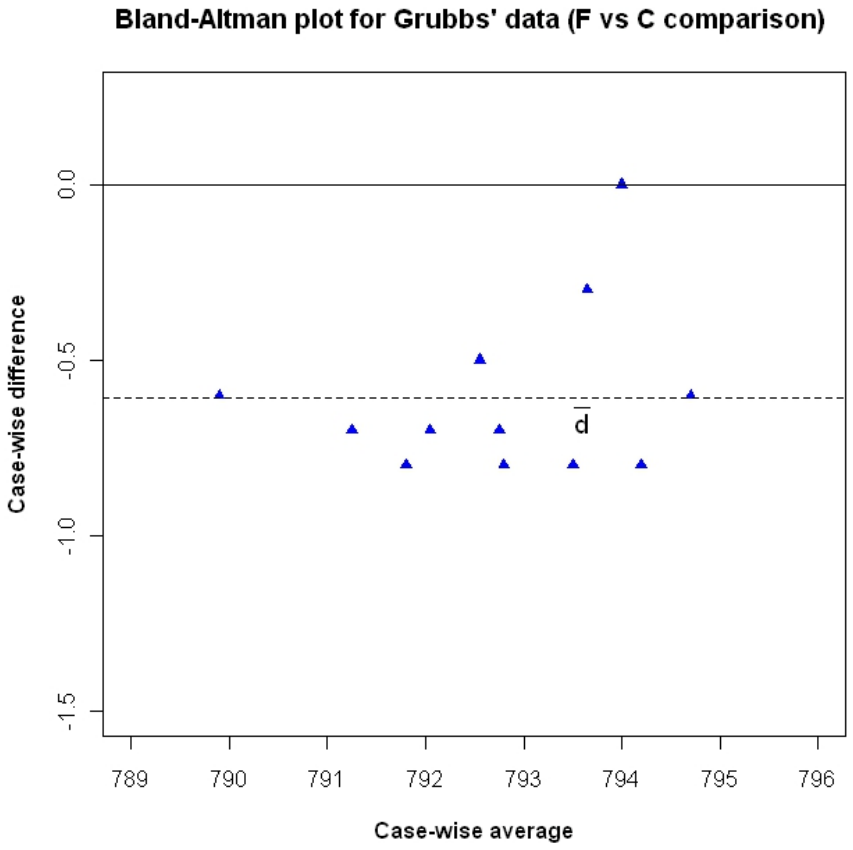


Figure 1.2: Bland-Altman plot For Fotobalk and Counter methods.

By inspection, there exists a larger inter-method bias in the 'F vs C' comparison than in the 'F vs T' comparison. Conversely there appears to be less precision in'F vs T' comparison, as indicated by the greater dispersion of co-variates.

Figures 1.4, 1.5 and 1.6 are three prototype Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range.

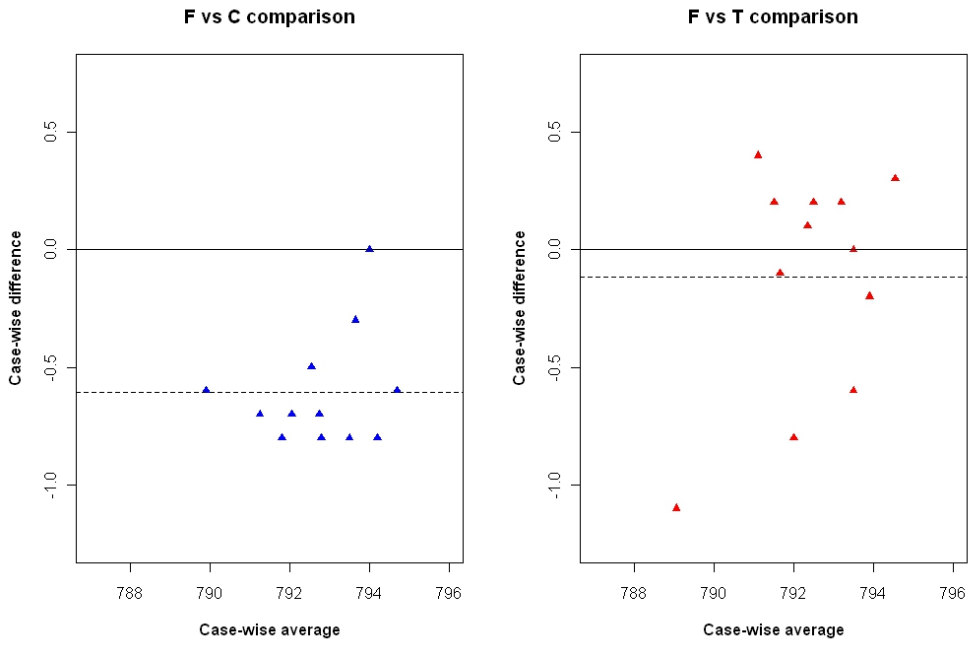


Figure 1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

This is known as proportional bias. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

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

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. Classification of outliers can be determined with numerous

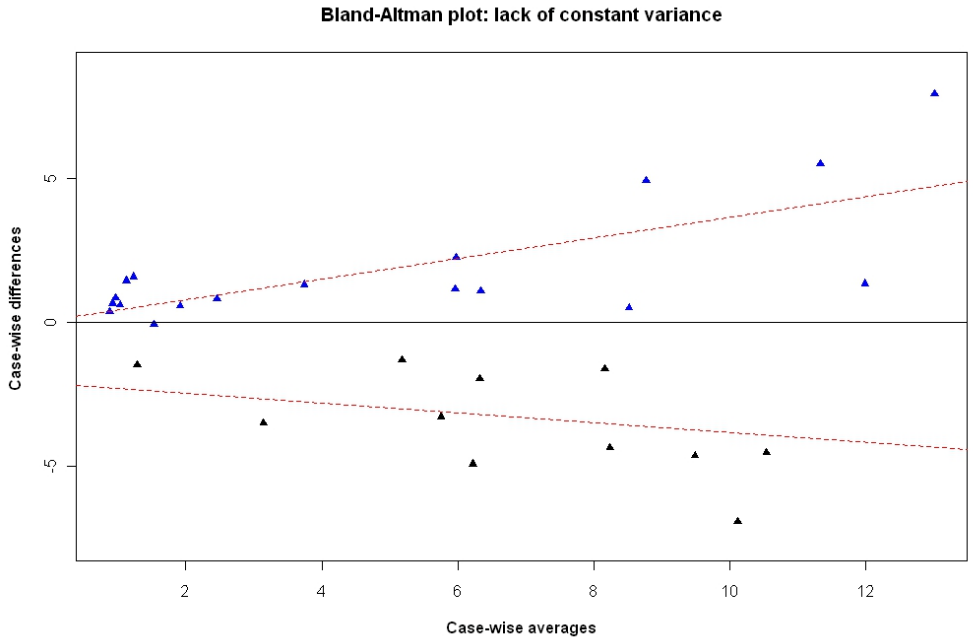


Figure 1.4: Bland-Altman plot demonstrating the increase of variance over the range.

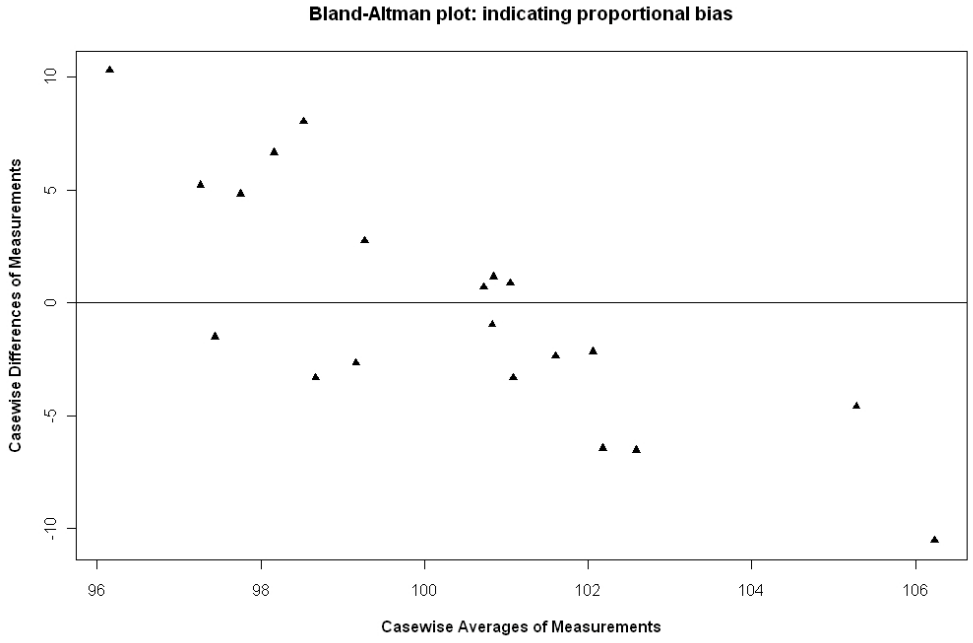


Figure 1.5: Bland-Altman plot indicating the presence of proportional bias.

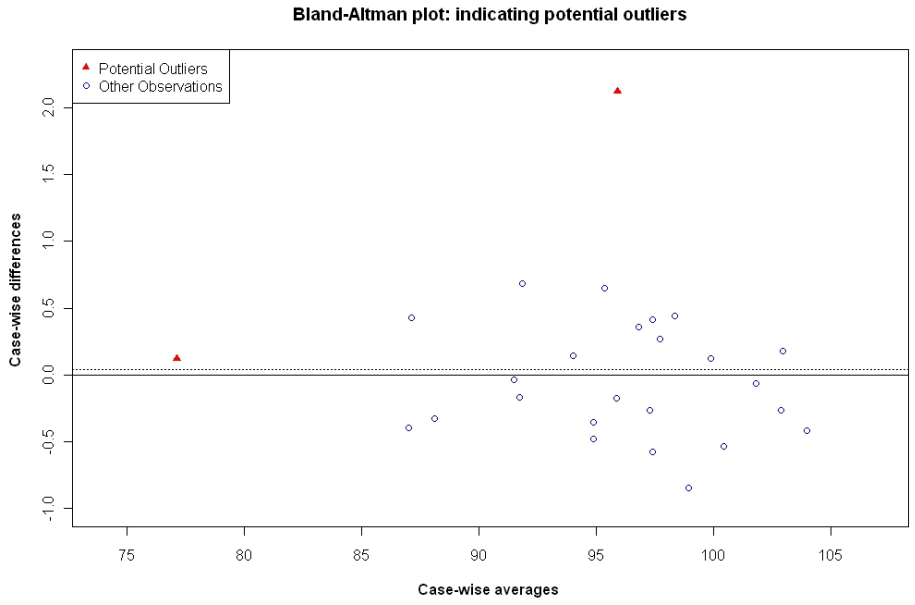


Figure 1.6: Bland-Altman plot indicating the presence of potential outliers.

established approaches, such as the Grubb's test, but always classification must be informed by the logic of the data's formulation. Figure 1.6 is a Bland-Altman plot with two potential outliers.

Bland and Altman (1999) do not recommend excluding outliers from analyses, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that 'we usually find that this method of analysis is not too sensitive to one or two large outlying differences'.

In classifying whether a observation from a univariate data set is an outlier, Grubbs' outlier test is widely used. In assessing whether a co-variate in a Bland-Altman plot is an outlier, this test is useful when applied to the difference values treated as a univariate data set. For Grubbs' data, this outlier test is carried out on the differences, yielding the following results.

The null and alternative hypotheses is the absence and presence of at least one outlier respectively. Grubbs' outlier test statistic $G$ is the largest absolute deviation from the sample mean divided by the standard deviation of the differences. For the 'F vs C' comparison, $G = 3.6403$. The critical value is calculated using Student's $t$ distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}. \tag{1.1}$$

For this test $U = 0.7501$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with $p - value = 0.002799$.

As a complement to the Bland-Altman plot, Bartko (1994) proposes the use of a bivariate confidence ellipse, constructed for a predetermined level.

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.Altman (1978) provides the relevant calculations for the ellipse. Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers (furthermore Bartko (1994) proposes formal testing procedures, that shall be discussed in due course). Inspection of Figure 1.7 shows that the fourth observation is outside the bounds of the ellipse, concurring with the conclusion that it is an outlier.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko's ellipse. A co-variate 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. Table 1.8 depicts a 95% confidence ellipse for this enhanced data set. By inspection of the confidence interval, a conclusion would be reached that this extra co-variate is an outlier, in spite of the fact that this observation is consistent with the intended conclusion of the Bland-Altman plot.

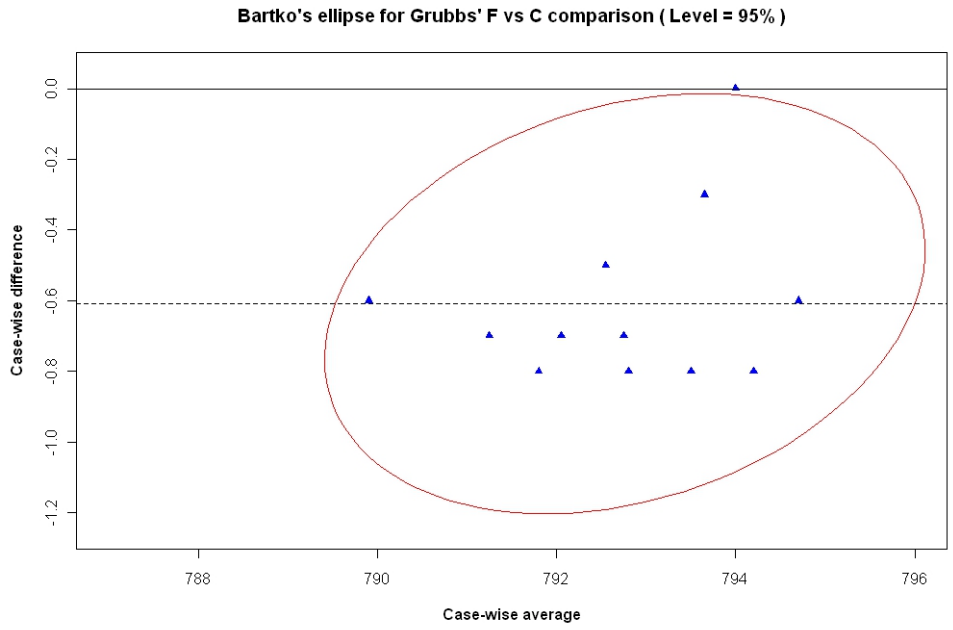


Figure 1.7: Bartko's Ellipse For Grubbs' Data.

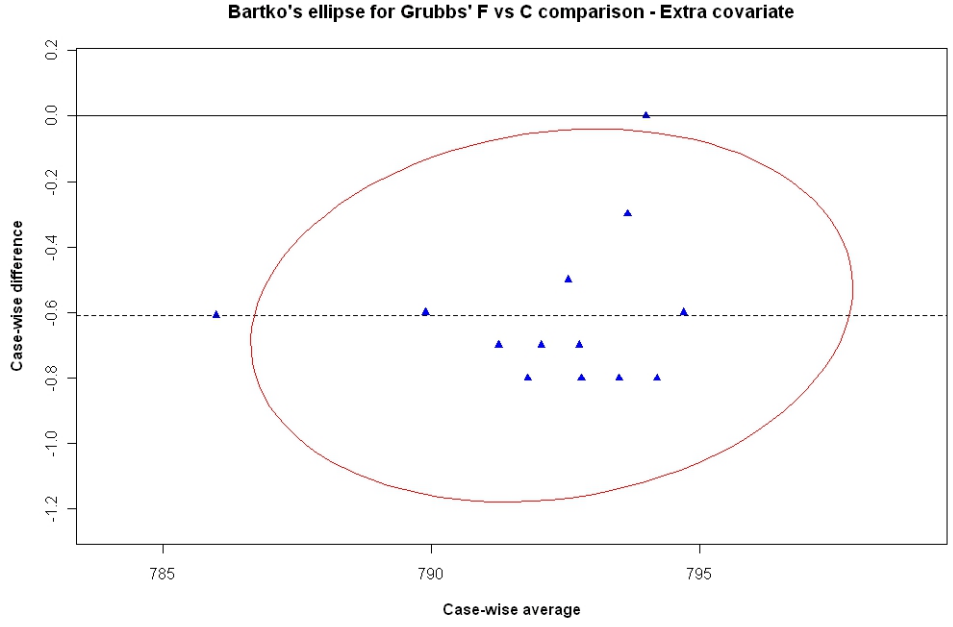


Figure 1.8: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Any observation should not be considered

an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra co-variate. 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.

Bartko's ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a_i)$ is greater than $\text{var}(d_i)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a_i)$ is less than $\text{var}(d_i)$, the orientation of the ellipse is vertical.

### 1.2.2 Variations of the Bland-Altman Plot

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 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 inappropriate for. The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages. This will remove the need for log 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'.

### 1.2.3 Regression-based Limits of Agreement

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

the value $\bar{d}$.

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

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

$$\hat{d} \pm 1.96\sqrt{\frac{\pi}{2}}\hat{R} = \hat{d} \pm 2.46\hat{R} \tag{1.2}$$

### 1.2.4 Replicate Measurements

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

Bland and Altman (1986) address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

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

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

all measurements. Incorrect conclusions would be caused by such a misinterpretation. Carstensen et al. (2008) demonstrates how the limits of agreement calculated using the mean of replicates are 'much too narrow as prediction limits for differences between future single measurements'. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the 'mean of replicates' approach.

The approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means ($\rho_{ad}$). According to the authors, this test is equivalent to the 'Pitman Morgan Test'. For the Grubbs data, the correlation coefficient estimate ($r_{ad}$) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers 'r to z' transformation (Cohen, Cohen, West, and Aiken, Cohen et al.). The null hypothesis ($\rho_{ad}$ =0)would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has no been no further mention of this particular test in Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman's rank correlation coefficient. Bland and Altman (1999) comments 'we do not see a place for methods of analysis based on hypothesis testing'. Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

Dunn (2002) highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example $\alpha$ may take the value of the inter-method bias estimate from Bland-Altman methodology. Another assumption is that the precision ratio $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\delta^2}$ may be known.

Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement charac-

teristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

Dunn (2002) recommends the following approach for analyzing method comparison data. Firstly he recommends conventional Bland-Altman methodology; plotting the scatterplot and the Bland-Altman plot, complemented by estimate for the limits of agreement and the correlation coefficient between the difference and the mean. Additionally boxplots may be useful in considering the marginal distributions of the observations. The second step is the calculations of summary statistics; the means and variances of each set of measurements, and the covariances.

When both methods measure in the same scale (i.e. $\beta = 1$), Dunn (2002) recommends the use of Grubbs estimators to estimate error variances, and to test for their equality. A test of whether the intercept $\alpha$ may be also be appropriate.

## 1.3 Regression Methods

Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as 'Model I regression'(Cornbleet and Cochrane, 1979; Ludbrook, 1997). A key feature of Model I models is that the independent variable is assumed to be measured without error. As often pointed out in several papers (Altman and Bland, 1983; Ludbrook, 1997), this assumption invalidates simple linear regression for use in method comparison studies, as both methods must be assumed to be measured with error.

The use of regression models that assumes the presence of error in both variables $X$ and $Y$ have been proposed for use instead. (Cornbleet and Cochrane, 1979; Ludbrook, 1997), These methodologies are collectively known as 'Model II regression'. They differ in the method used to estimate the parameters of the regression.

Regression estimates depend on formulation of the model. A formulation with one method considered as the $X$ variable will yield different estimates for a formulation where it is the $Y$ variable. With Model I regression,the models fitted in both cases will entirely different and inconsistent. However with Model II regression, they will be consistent and complementary.

### 1.3.1 Deming's Regression

The most commonly known Model II methodology is known as Deming's Regression, (also known an Ordinary Least Product regression). Deming regression is recommended by Cornbleet and Cochrane (1979) as the preferred Model II regression for use in method comparison studies. As previously noted, the Bland Altman Plot is uninformative about the comparative influence of proportional bias and fixed bias. Deming's regression provides independent tests for both types of bias.

For a given $\lambda$, Kummel (1879) derived the following estimate for the Demimg regression slope parameter. ($\alpha$ is simply estimated by using the identity $\bar{Y} - \hat{\beta}\bar{X}$.)

$$\hat{\beta} = \frac{S_{YY} - \lambda S_{XX} + [(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2]^{1/2}}{2S_{XY}} \tag{1.3}$$

As with conventional regression methodologies, Deming's regression calculates an estimate for both the slope and intercept for the fitted line, and standard errors thereof. Therefore there is sufficient information to carry out hypothesis tests on both estimates, that are informative about presence of fixed and proportional bias.

A 95% confidence interval for the intercept estimate can be used to test the intercept, and hence fixed bias, is equal to zero. This hypothesis is accepted if the confidence interval for the estimate contains the value 0 in its range. Should this be, it can be concluded that fixed bias is not present. Conversely, if the hypothesis is rejected, then it is concluded that the intercept is non zero, and that fixed bias is present.

Testing for proportional bias is a very similar procedure. The 95% confidence interval for the slope estimate can be used to test the hypothesis that the slope is equal to 1. This hypothesis is accepted if the confidence interval for the estimate contains the value 1 in its range. If the hypothesis is rejected, then it is concluded that the slope is significant different from 1 and that a proportional bias exists.

For convenience, a new data set shall be introduced to demonstrate Demings regression. Measurements of transmitral volumetric flow (MF) by doppler echocardiography, and left ventricular stroke volume (SV) by cross sectional echocardiography in 21 patients withour aortic valve disease are tabulated in Zhang et al. (1986). This data set features in the discussion of method comparison studies in Altman (1991, p.398) .

| Patient | MF ($cm^3$) | SV ($cm^3$) | Patient | MF ($cm^3$) | SV ($cm^3$) | Patient | MF ($cm^3$) | SV ($cm^3$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 47 | 43 | 8 | 75 | 72 | 15 | 90 | 82 |
| 2 | 66 | 70 | 9 | 79 | 92 | 16 | 100 | 100 |
| 3 | 68 | 72 | 10 | 81 | 76 | 17 | 104 | 94 |
| 4 | 69 | 81 | 11 | 85 | 85 | 18 | 105 | 98 |
| 5 | 70 | 60 | 12 | 87 | 82 | 19 | 112 | 108 |
| 6 | 70 | 67 | 13 | 87 | 90 | 20 | 120 | 131 |
| 7 | 73 | 72 | 14 | 87 | 96 | 21 | 132 | 131 |

Table 1.5: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

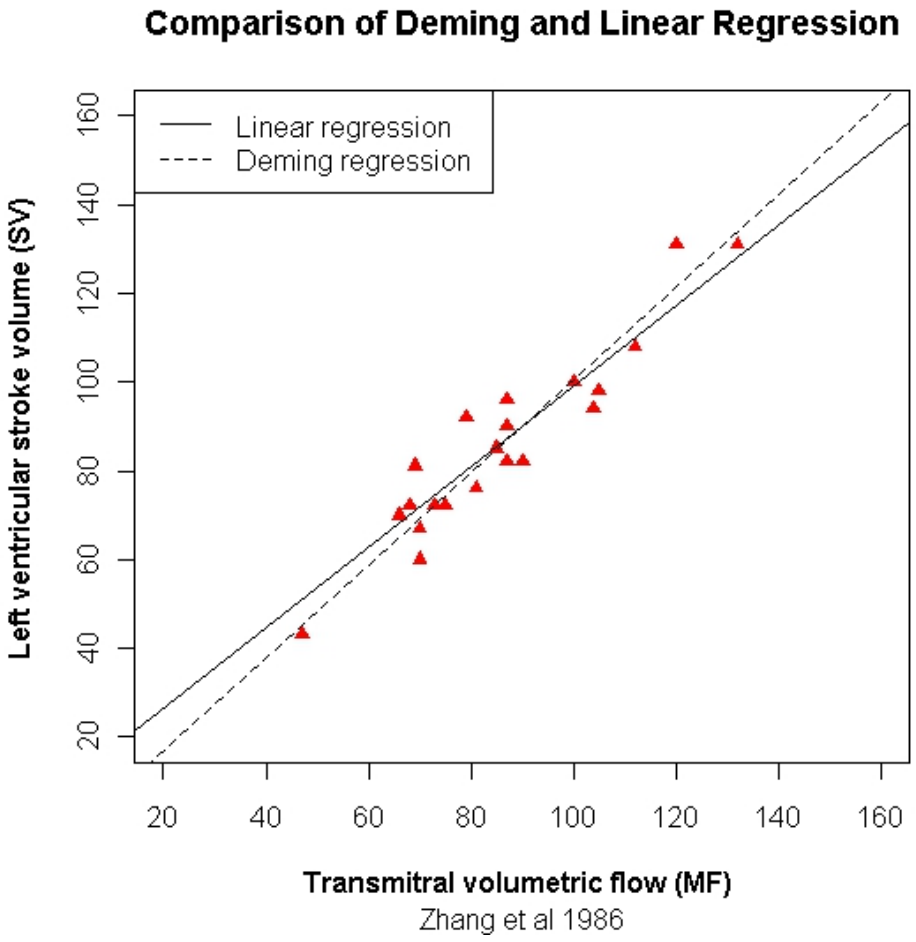


Figure 1.9: Deming Regression For Zhang's Data

Deming's Regression suffers from some crucial drawback. Firstly it is computationally complex, and it requires specific software packages to perform calculations.Secondly it is uninformative about the comparative precision of two methods of measurement. Most importantly Carroll and Ruppert (1996) states that Deming's regression is acceptable only when the precision ratio ($\lambda$, in their paper as $\eta$) is correctly specified ,but in practice this is often not the case, with the $\lambda$ being underestimated.

# Chapter 2

# Linear Mixed Effects Models

## 2.1 LME models in Method comparison

## 2.2 Lai Shiao

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

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

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

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

**Single fixed effect**

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

$$\text{fixed effects}: 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \tag{2.1}$$

$$(\text{p-values}: = 0.0054, < 0.0001, < 0.0001)$$

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

$$(\text{p-values}: = 0.8113, < 0.0001, < 0.0001)$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of $SO_2$. Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

**Two fixed effects**

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\text{fixed effects}: -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl}$$

$$(\text{p-values}: = 0.8113, < 0.0001, < 0.0001)$$

$$\text{random effects}: u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577)$$

$$(\text{p-values}: = 0.0446, < 0.0001, < 0.0001) \tag{2.2}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidisation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

## 2.3 Carstensen's Mixed Models

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model ( in the authors own notation) is formulated as follows, where $y_{mir}$ is the $r$th replicate measurement on subject $i$ with method $m$.

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

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

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

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

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m} \mu_i + c_{mi} \tag{2.4}$$

. Under the assumption that the $\mu$s are the true item values, this would be sufficient

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

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

### 2.3.1 Using LME models to create Prediction Intervals

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement $y_{mi}$ by method $m$ on individual $i$ is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{2.5}$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$ For the replicate case, an interaction term $c$ is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{2.6}$$

# Bibliography

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.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall.

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

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

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

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

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

Carroll, R. and D. Ruppert (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician 50*(1), 1–6.

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

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).

Cohen, J., P. Cohen, S. West, and L. Aiken. *Applied multiple regression / correlation analysis for the behavioral sciences* (Third ed.). Laurence Erlbaum Associates.

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.

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.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

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

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

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

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

Kummel, C. (1879). Reduction of observation equations which contain more than one observed quantity. *The Analyst 6*, 97–105.

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.

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

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

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.

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

Zhang, Y., S. Nitter-Hauge, H. Ihlen, K. Rootwelt, and E. Myhre (1986). Measurement of aortic regurgitation by doppler echocardiography. *British Heart Journal 55*, 32–38.