Contents

1 Techniques for Method Comparison **3**

1.1 The Bland-Altman Approach to Method Comparison 3

1.1.1 Inspecting Method Comparison Data 5

1.2 Limits of Agreement 10

1.2.1 Interpretation of Limits Of Agreement 11

1.2.2 Precision of Limits of Agreement 13

1.3 Detection of Outliers in the Bland-Altman Framework 13

1.3.1 Bartko's Ellipse 15

1.3.2 Grubbs' Test for Outliers 16

1.4 Prevalence of the Bland-Altman Plot 17

1.5 Criticism of Limits of Agreement 18

1.6 Limits of Agreement for Replicate Measurements 19

1.7 Formal Models and Tests 20

1.7.1 Kinsella's Model 20

1.7.2 Pitman-Morgan Testing 23

1.7.3 Regression-Based Testing Techniques 24

1.8 Regression-Based Methods 25

1.8.1 Deming Regression 26

1.8.2 Kummel's Estimates 27

1.8.3 Inferences for Deming Regression 27

1.8.4 Worked Example of Deming Regression 28

1.9 Structural Equation Modelling 29
1.10 Model for Replicate Measurements 32
1.10.1 Carstensen's Model for Replicate Measurements 32

Chapter 1

Techniques for Method Comparison

1.1 The Bland-Altman Approach to Method Comparison

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 recommending the use of graphical techniques to assess agreement.

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 technique 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.

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

1.1.1 Inspecting Method Comparison Data

The first step recommended, which the authors argue should be mandatory, is construction of an identity plot, which is a simple scatter-plot approach of measurements for both methods on either axis. The line of equality (the $X = Y$ line, i.e. the 45 degree line through the origin) should also be shown, as it is necessary to give the correct interpretation of how both methods compare.

This plot can gives the analyst a cursory examination of how well the measurement methods agree. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown on the left in Figure 1.1.1. 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.

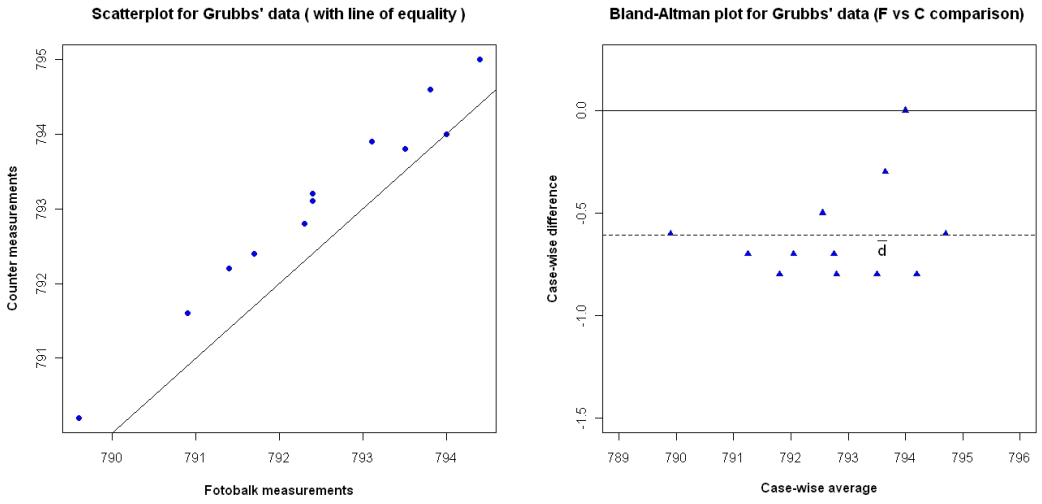

Figure 1.1.1: Identity Plot and Bland-Altman Plot For Fotobalk and Counter methods.

However scatter-plots, such as these, are not sufficient 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.

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 on the right in

Figure 1.1.1, using data from Table 1.1.1. The dashed line in the Bland-Altman plot alludes to the inter-method bias between the two methods, estimated by calculating the average of the differences. In the case of Grubbs data the inter-method bias is -0.6083 metres per second. By inspection of the plot, one would notice that the differences tend to increase as the averages increase.

Round	Fotobalk [F]	Counter [C]	Terma [T]	[F-C]	[(F+C)/2]	[F-T]	[(F+T)/2]
1	793.8	794.6	793.2	-0.8	794.2	0.6	793.5
2	793.1	793.9	793.3	-0.8	793.5	-0.2	793.2
3	792.4	793.2	792.6	-0.8	792.8	-0.2	792.5
4	794.0	794.0	793.8	0.0	794.0	0.2	793.9
5	791.4	792.2	791.6	-0.8	791.8	-0.2	791.5
6	792.4	793.1	791.6	-0.7	792.8	0.8	792.0
7	791.7	792.4	791.6	-0.7	792.0	0.1	791.6
8	792.3	792.8	792.4	-0.5	792.5	-0.1	792.3
9	789.6	790.2	788.5	-0.6	789.9	1.1	789.0
10	794.4	795.0	794.7	-0.6	794.7	-0.3	794.5
11	790.9	791.6	791.3	-0.7	791.2	-0.4	791.1
12	793.5	793.8	793.5	-0.3	793.6	0.0	793.5

Table 1.1.1: Fotobalk : Differences and Averages with Counter and Terma.

In Figure 1.1.2 Bland-Altman plots for the 'F vs C' and 'F vs T' comparisons are shown, where 'F vs T' refers to the comparison of the 'Fotobalk' and 'Terma' methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the 'F vs C' comparison than in the 'F vs T' comparison. Conversely there appears to be less precision in 'F vs T' comparison, as indicated by the greater dispersion of covariates.

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 Bland-Altman plot.

Figure 1.1.3 and Figure 1.1.4 show two Bland-Altman plots derived from simulated

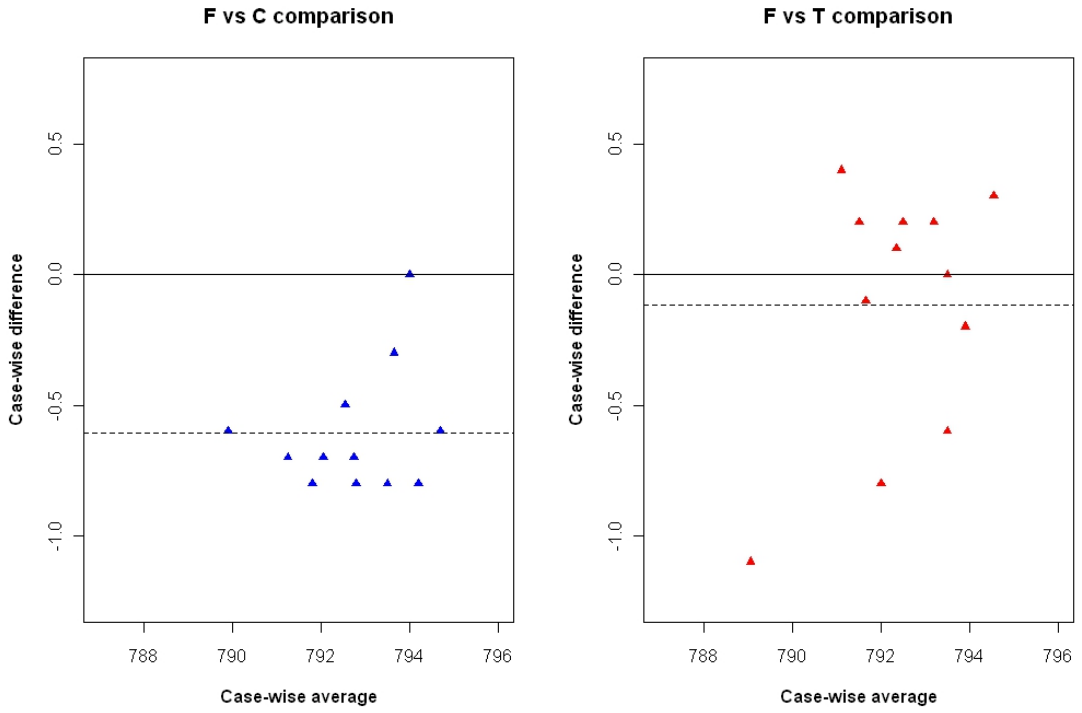

Figure 1.1.2: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

data, each for the purpose of demonstrating how the plot would inform an analyst of trends that would adversely affect use of the recommended approach. Additionally the procedure is not properly constructed to deal with outliers, which shall be reverted to later.

Figure 1.1.3 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 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 are advisable, for example, the Benjamini-

Hochberg Test (Benjamini and Hochberg, 1995).

Figure 1.1.4 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.

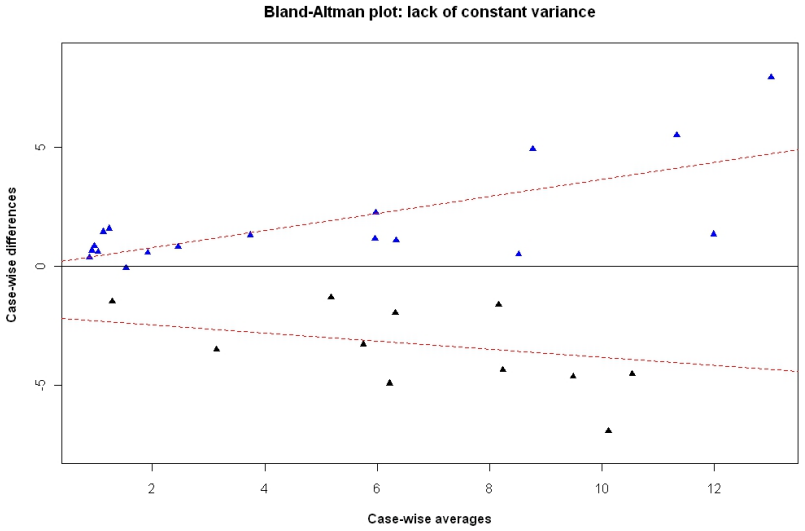

Figure 1.1.3: Bland-Altman Plot demonstrating the increase of variance over the range

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

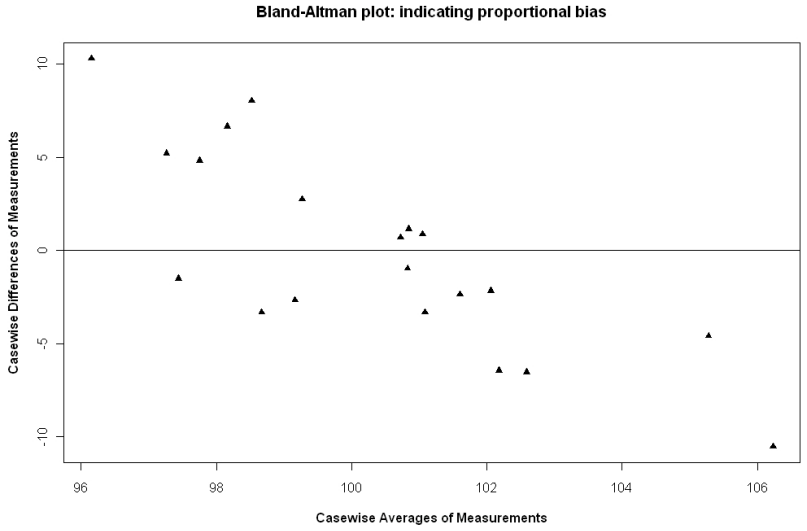

Figure 1.1.4: Bland-Altman Plot indicating the presence of proportional bias

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*'.

1.2 Limits of Agreement

A third element of the Bland-Altman approach, an interval known as limits of agreement is introduced in Bland and Altman (1986) (sometimes referred to in literature as 95% limits of agreement). These limits centre on the average difference line, and are computed as $LOA = \bar{d} \pm 1.96 s_d$ with $\bar{d}$ as the estimate of the inter method bias and s_d the standard deviation of the differences. 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. The limits of agreement requires an assumption of 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.*

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.

Calculation of the limits of agreement relies on the assumption that the case-wise differences are normally distributed, although the measurements themselves are not assumed to follow any distribution. This assumption is justified because variation

between subjects has been removed, leaving measurement error, which is likely to be normally distributed (Bland and Altman, 1986). Bland and Altman (1999) remark that this assumption is easy to check using commonly used methods, i.e. a normal probability plot.

For the Grubbs 'F vs C' comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure 1.2.5 shows the resultant Bland-Altman plot, with the limits of agreement shown in dotted lines.

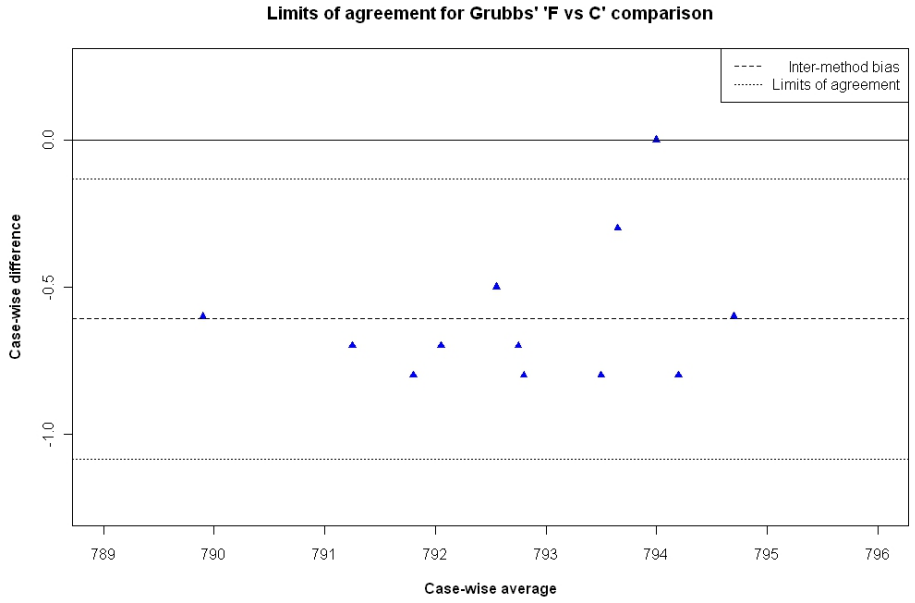

Figure 1.2.5: Bland-Altman plot with limits of agreement

1.2.1 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 technique 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

$$\bar{d} \pm t_{(0.025,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{1.1}$$

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).

1.2.2 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

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

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

$$\mathrm{Var}(LOA) \approx 1.71^2 \frac{s_d^2}{n},$$

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

1.3 Detection of Outliers in the Bland-Altman Framework

The Bland-Altman plot can be used to identify outliers. Here we use a simple definition of an outlier as 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 mechanism that produces the data.. Figure 1.3.6 is a Bland-Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively. In the Bland-Altman plot depicted in Figure 1.3.6, consider the

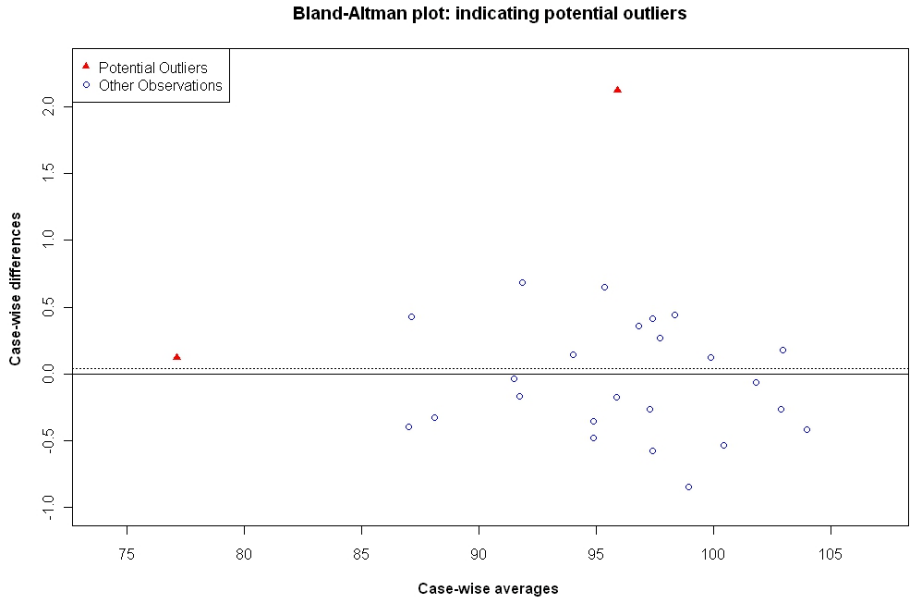

Figure 1.3.6: Bland-Altman plot indicating the presence of outliers

covariate located on the extreme left of the plot. Ordinarily we would conclude that this point due to it's horizontal displacement from the main cluster of points. However

this horizontal displacement is supported by two independent measurements and is very close to the inter-method bias, i.e. very close to its expected value. Therefore that observation, should not be considered an outlier at all.

Conversely the observation located at the top of the plot, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations. There are no mitigating factors.

1.3.1 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, with Altman (1978) providing the relevant calculations. 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.

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 between the variance of the means $\mathrm{Var}(a)$ and the variance of the differences $\mathrm{Var}(d)$. If $\mathrm{Var}(a)$ is greater than $\mathrm{Var}(d)$, the orientation of the ellipse is horizontal. Conversely if $\mathrm{Var}(a)$ is less than $\mathrm{Var}(d)$, 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-

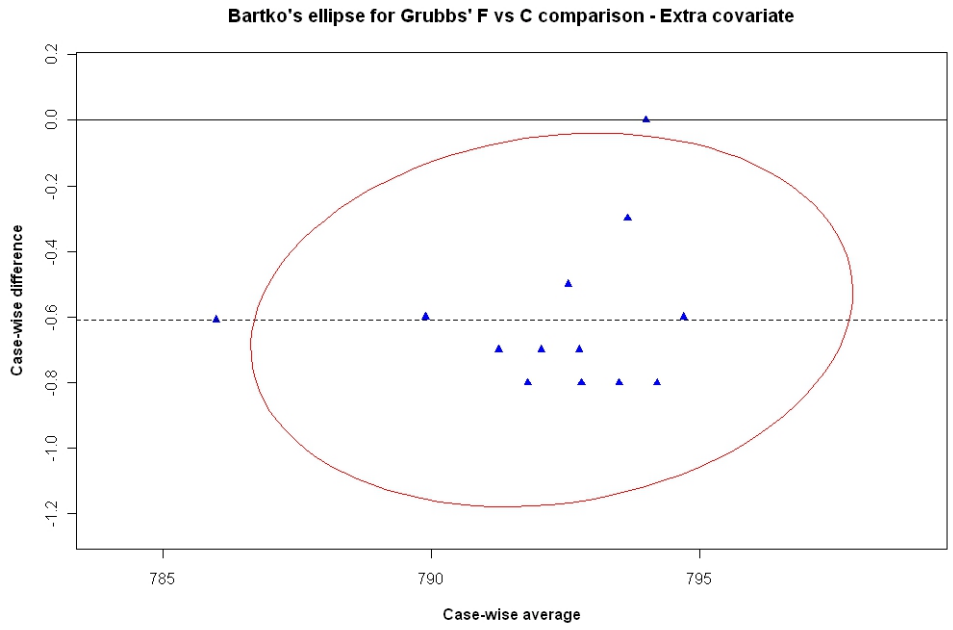

Figure 1.3.7: Bartko's ellipse for Grubbs data

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 1.3.7. However, both observation that were previously considered as potential outliers (i.e. the extreme left and the uppermost) are shown to be outside the bounds of the ellipse, indicating both to be outliers.

1.3.2 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,

$$G = \max_{i=1,\ldots,n} \frac{\left|d_i - \bar{d}\right|}{S_d}. \tag{1.2}$$

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 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.

1.4 Prevalence of the Bland-Altman Plot

Bland and Altman (1986), which further develops the Bland-Altman approach, 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.

1.5 Criticism of Limits of Agreement

The Bland-Altman approach 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' or the presence of an outlier.

However the approach comes in for criticism in a number of respects. In the first instance, some caution must be given to the inter-method bias estimate. If one method is sometimes higher, or sometimes lower, the average of the differences will be close to zero. If the inter-method bias is close to zero, there be an indication that the two measurement methods are in agreement, when in fact they 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. 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. According to the authors, a knowledgeable practitioner in the field should 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. Furthermore Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

While in essence they are similar to confidence intervals, limits of agreement are not constructed as such; they are designed for future values. Lack of clarity in this regards can give rise to confusion, and incorrect interpretations.

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. 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. Finally the adaptation of the approach to deal with replicate measurements, as specified by Bland and Altman (1999), is flawed.

1.6 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. Contrary 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.

1.7 Formal Models and Tests

While the Bland-Altman plot is a simple technique for comparing measurements, Kinsella (1986) noted the lack of formal testing offered by that approach, with it relying on the practitioner's opinion to judge the outcome. Altman and Bland (1983) proposed a formal test on the Pearson correlation coefficient of case-wise differences and means which, according to the authors, is equivalent to the 'Pitman-Morgan Test', a key contribution to method comparison studies that shall discussed shortly (Morgan, 1939; Pitman, 1939). There has 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) remarked that'*we do not see a place for methods of analysis based on hypothesis testing*', while also stating that they consider structural equation models to be inappropriate.

1.7.1 Kinsella's Model

Kinsella (1986) presented a simple model to describe a measurement by each method, describing the relationship with its real value. Only the non-replicate case is considered, as this is the context of the Bland-Altman plots. Other authors, such as Carstensen (2004); Carstensen et al. (2008), present similar formulations of the same model, as well as modified models to account for multiple measurements by each methods on each item, known as replicate measurements.

Kinsella (1986) formulates a model for single measurement observations for a method comparison study as a linear mixed effects model, i.e. model that additively combine fixed effects and random effects.

$$Y_{ij} = \quad \mu + \beta_j + u_i + \epsilon_{ij} \qquad i = 1, \ldots, n \qquad j = 1, 2$$

The true value of the measurement is represented by μ while the fixed effect due to method j is β_j. For simplicity these terms can be combined into single terms; $\mu_1 = \mu + \beta_1$ and $\mu_2 = \mu + \beta_2$. The inter-method bias is the difference of the two fixed effect terms, $\mu_d = \beta_1 - \beta_2$. Each of the i items are assumed to give rise to random error, represented by u_i. This random effects terms is assumed to have mean zero and be normally distributed with variance σ^2. There is assumed to be an attendant error for each measurement on each item, denoted ϵ_{ij}, which is also assumed to have mean zero. The variance of measurement error for both methods are not assumed to be identical for both methods variance, hence it is denoted σ_j^2. The set of observations (x_i, y_i) by methods X and Y are assumed to follow the bivariate normal distribution with expected values $E(x_i) = \mu_1$ and $E(y_i) = \mu_2$ respectively. The variance covariance of the observations Σ is given by

$$\Sigma_{(X,Y)} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma_2^2 \end{bmatrix}.$$

The case-wise differences and means are calculated as $d_i = x_i - y_i$ and $a_i = (x_i + y_i)/2$ respectively. Both d_i and a_i are assumed to follow a bivariate normal distribution with $E(d_i) = \mu_d = \mu_1 - \mu_2$ and $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$. Constructively, the paired measurements can be expressed as

$$d_i = x_i - y_i \sim \mathcal{N}(\mu_d, \sigma_1^2 + \sigma_2^2).$$

The variance matrix $\Sigma_{(A,D)}$ is

$$\Sigma_{(A,D)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \tag{1.3}$$

In some types of analysis, such as the conversion problems described by Lewis et al. (1991), measurements made by methods X and Y may be denominated in different units, and an estimate for the proportionality, i.e. a scaling factor, must be determined. Using amended notation, for comparing two methods X and Y, for the measurement of item i is formulated as

$$X_i = \tau_i + \epsilon_{i1}, \qquad \epsilon_{i1} \sim \mathcal{N}(0, \sigma_1^2), \tag{1.4}$$

$$Y_i = \alpha + \lambda\tau_i + \epsilon_{i2}, \qquad \epsilon_{i2} \sim \mathcal{N}(0, \sigma_2^2). \tag{1.5}$$

Here the unknown 'true value' is τ_i, α represents the inter-method bias, and the scaling factor is denoted here as λ. For the time being, we will restrict ourselves to problems where λ is assumed to be 1, but will revert back to this conversion problem later.

Kinsella (1986) demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for both. The measurement model introduced by Grubbs (1948, 1973) provides a formal procedure for estimate the variances σ^2, σ_1^2 and σ_2^2 devices. Grubbs (1948) offers maximum likelihood estimates, commonly known as Grubbs estimators, for the various variance components,

$$\hat{\sigma^2} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = Sxy,$$
$$\hat{\sigma_1^2} = \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2x - Sxy,$$
$$\hat{\sigma_2^2} = \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2y - Sxy.$$

Thompson (1963) presents confidence intervals for the relative precisions of the measurement methods, $\Delta_j = \sigma_S^2/\sigma_j^2$ (where $j = 1, 2$), as well as the variances σ_S^2, σ_1^2 and σ_2^2,

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}}. \tag{1.6}$$

Thompson (1963) defines Δ_j to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma^2/\sigma_j^2$. Thompson also demonstrates how to make statistical

inferences about Δ_j. Based on the following identities,

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ |A| &= C_x \times C_y - (C_{xy})^2, \end{aligned}$$

the confidence interval limits of Δ_1 are

$$\Delta_1 < \frac{C_{xy} - t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}} \tag{1.7}$$

$$\Delta_1 > \frac{C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}} \tag{1.8}$$

The value t is the $100(1-\alpha/2)\%$ upper quantile of Student's t distribution with $n-2$ degrees of freedom (Kinsella, 1986). The confidence limits for Δ_2 are found by substituting C_y for C_x in 1.7 and 1.8. Negative lower limits are replaced by the value 0.

1.7.2 Pitman-Morgan Testing

An early contribution to formal testing in method comparison was devised concurrently by Pitman (1939) and Morgan (1939) in separate contributions.

The classical Pitman-Morgan test can be adapted as a hypothesis test of equal variance for both methods, based on the correlation value between differences and means $\rho_{a,d}$. This is a test statistic for the null hypothesis of equal variances given bivariate normality ;

$$\rho(a,d) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}}. \tag{1.9}$$

These authors noted that the correlation coefficient depends upon the difference $\sigma_1^2 - \sigma_2^2$, being zero if and only if $\sigma_1^2 = \sigma_2^2$. The hypothesis test $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a

test of the hypothesis $H : \rho(a, d) = 0$. This corresponds to the well-known $t-$test for a correlation coefficient with $n - 2$ degrees of freedom.

Bartko (1994) describes the Pitman-Morgan test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, a result that can be derived using straightforward algebra. The Pitman-Morgan test is equivalent to the marginal test of the slope estimate in Bradley and Blackwood (1989).

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 a t random variable with $n-1$ degrees of freedom. Only if the two methods show comparable precision then the paired sample $t-$test is appropriate for testing the inter-method bias. Therefore, it should only be used in succession to the Pitman-Morgan test. Furthermore, these tests are only valid in the case of non-replicate measurements.

1.7.3 Regression-Based Testing Techniques

Bradley and Blackwood (1989) have developed a regression based procedure for assessing the agreement. This approach performs a simultaneous test for the equivalence of means and variances of two paired data sets.

Bradley and Blackwood (1989) construct a linear model which fits D on S, which are the case-wise differences and sums of a pair of measurements respectively, creating estimates for intercept and slope, β_0 and β_1:

$$D = \beta_0 + \beta_1 S.$$

The null hypothesis of this test is that the mean (μ) and variance (σ^2) of both data sets are equal if the slope and intercept estimates are equal to zero (i.e $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$ if and only if $\beta_0 = \beta_1 = 0$). The test is conducted using an $F-$test, calculated from the results of the regression of D on S. Bartko (1994) amends this approach for use in method comparison studies, using the averages of the pairs, as opposed to the sums. This approach can facilitate simultaneous usage of test with the Bland-Altman technique.

Bartko's test statistic is then calculated from the regression analysis of variance values (Bradley and Blackwood, 1989) and is distributed as 'F' random variable:

$$F^* = \frac{(\Sigma d^2) - SSReg}{2MSReg}.$$

The degrees of freedom are $\nu_1 = 2$ and $\nu_1 = n - 2$ (where n is the number of pairs). The critical value is chosen for $\alpha\%$ significance with those same degrees of freedom.

For the Grubbs data, $\Sigma d^2 = 5.09$, $SSReg = 0.60$ and $MSreg = 0.06$ Therefore the test statistic is 37.42, with a critical value of 4.10. Hence the means and variance of the Fotobalk and Counter chronometers are assumed to be simultaneously equal.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Averages	1	0.04	0.04	0.74	0.4097
Residuals	10	0.60	0.06		

Table 1.7.2: Regression ANOVA of case-wise differences and averages for Grubbs Data

Importantly, this approach determines whether there is both inter-method bias and precision present, or alternatively if there is neither present. It has previously been demonstrated that there is a inter-method bias present, but as this procedure does not allow for separate testing, no conclusion can be drawn on the comparative precision of both methods.

1.8 Regression-Based 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 these 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. Additionally one method must be arbitrarily identified as the independent variable.

Cornbleet and Cochrane (1979) argue for the use of alternatives to the OLS approach, that based on the assumption that both methods are imprecisely measured, and that yield a fitting that is consistent with both 'X on Y' and 'Y on X' formulations.

Errors-in-variables models assume the presence of error in both variables X and Y have been proposed for use instead (Cornbleet and Cochrane, 1979; Ludbrook, 1997). These models are collectively known as 'Model II regression'. These approaches suitable for method comparison studies, but are more difficult to implement.

1.8.1 Deming Regression

The most commonly known Model II methodology is known as Deming's Regression, and is recommended by Cornbleet and Cochrane (1979) as the preferred Model II regression for use in method comparison studies. The Bland-Altman plot is uninformative about the comparative influence of proportional bias and fixed bias. However Deming regression can provide independent tests for both types of bias.

The measurement error is specified with measurement error variance related as $\lambda = \sigma_y^2/\sigma_x^2$, where σ_x^2 and σ_y^2 is the measurement error variance of the X and Y variables respectively.

The Deming regression method calculates a line of best fit for two sets of data. This derivation results in the best fit to simultaneously minimize the sum of the squares of the perpendicular distances from the data points at an angle specified by the ratio λ. For OLS Models, the distances are minimized in the vertical direction (Linnet, 1999). When λ is one, the angle is 45 degrees. Normally distributed error of both variables is assumed, as well as a constant level of imprecision throughout the range of measurements.

In cases involving only single measurements by each method, λ may be unknown and is therefore assumes a value of one. While this will produce biased estimates, they

are less biased than ordinary linear regression.

1.8.2 Kummel's Estimates

The appropriate estimates were derived by Kummel (1879), but were popularized in the context of medical statistics and clinical chemistry by Deming (1943). For a given λ, Kummel (1879) derived the following estimate that would later be used for the Deming regression slope parameter.

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2]^{1/2}}{2S_{xy}}, \tag{1.10}$$

with λ as the variance ratio. The intercept estimate α is simply estimated in the same way as in conventional linear regression, by using the identity $\bar{Y} - \hat{\beta}\bar{X}$.

This approach would be appropriate when errors in y and x are both caused by measurements, and the accuracy of measurement systems are known. In cases involving only single measurements by each method, λ may be unknown and is therefore assumes a value of one. While this will bias the estimates, it is less biased than ordinary linear regression. Deming regression assumes that the variance ratio λ is known. When λ is defined as one, (i.e. equal error variances), the approach is known as orthogonal regression. Several candidate models, with varying variance ratios may be fitted, and estimates of the slope and intercept are produced. However no model selection information is available to determine the best fitting model.

1.8.3 Inferences for Deming Regression

As with classical regression models, Deming regression calculates an estimate for both the slope and intercept for the fitted line, and standard errors thereof Cornbleet and Cochrane (1979). Standard errors and confidence intervals can be estimated using the Bootstrap techniques. Authors such as Carpenter and Bithell (2000) and Johnson (2001) provide relevant insights.

Therefore there is sufficient information to carry out hypothesis tests on both estimates, that are informative about presence of constant and proportional bias. The

test for the intercept estimate acts as a test for the presence of constant bias between both measurement methods. Similarly the test for the slope estimate can be used to formally test proportional bias between the two methods.

One of the assumptions that underline Deming regression is constancy of the measurement errors throughout the range of values. However the author point out that *clinical laboratory measurements usually increase in absolute imprecision when larger values are measured.*

Model selection and diagnostic technique are well developed for classical linear regression methods. Typically an implementation of a linear model fit will be accompanied by additional information, such as the coefficient of determination and likelihood and information criterions, and a regression ANOVA table. Such additional information has not, as yet, been implemented for Deming regression.

1.8.4 Worked Example of Deming Regression

For convenience, a new data set shall be introduced to demonstrate Deming regression. Measurements of transmitral volumetric flow (MF) by doppler echocardiography, and left ventricular stroke volume (SV) by cross sectional echocardiography in 21 patients with 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).

This ratio can be estimated if multiple measurements were taken with each method, but if only one measurement was taken with each method, it can be assumed to be equal to one.

Deming regression is undermined by several factors. Firstly it is computationally complex, and it requires specific software packages to perform calculations. Secondly, in common with all regression methods, Deming regression is vulnerable to outliers. Lastly, Deming regression 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 (λ, in their paper as

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

η) is correctly specified, but in practice this is often not the case, with the λ being underestimated. This underestimation leads to an overcorrection for attenuation.

Several candidate models, with varying variance ratios may be fitted, and estimates of the slope and intercept are produced. However no model selection information is available to determine the best fitting model.

As noted before, Deming regression is an important and informative methodology in method comparison studies. For single measurement method comparisons, Deming regression offers a useful complement to LME models.

1.9 Structural Equation Modelling

Structural equation modelling is a statistical technique used for testing and estimating causal relationships using a combination of statistical data and qualitative causal assumptions. Carrasco (2004) describes the structural equation model is a regression approach that allows to estimate a linear regression when independent variables are measured with error. The structural equations approach avoids the biased estimation

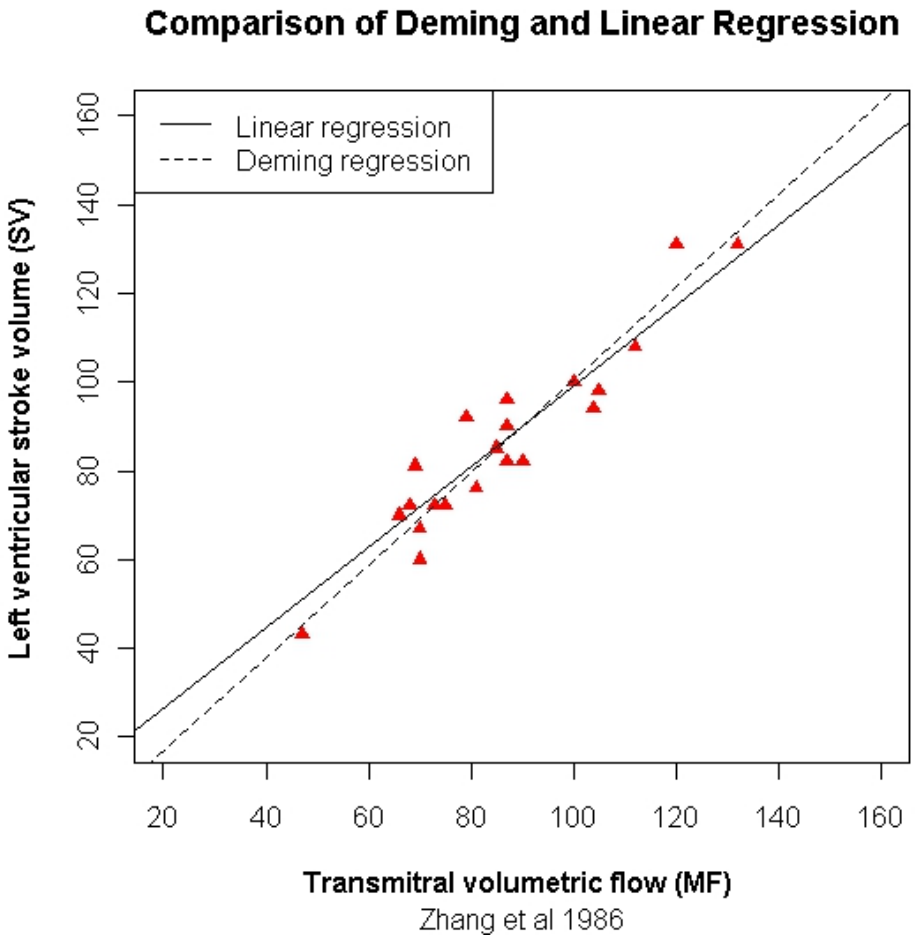

Figure 1.8.8: Deming Regression For Zhang's Data

of the slope and intercept that occurs in ordinary least square regression.

Several authors, such as Lewis et al. (1991), Kelly (1985), Voelkel and Siskowski (2005) and Hopkins (2004) advocate the use of SEM methods for method comparison. In Hopkins (2004), a critique of the Bland-Altman plot he makes the following remark:

> *What's needed for a comparison of two or more measures is a generic approach more powerful even than regression to model the relationship and error structure of each measure with a latent variable representing the true value.*

Hopkins also adds that he himself is collaborating in research utilising SEM and

mixed effects modelling. Kelly (1985) advised that *the Structural equations model is used to estimate the linear relationship between new and standards method. The Delta method is used to find the variance of the estimated parameters* (Kelly, 1985).

Conversely Bland and Altman (1999) also states that consider structural equation models to be inappropriate. However Altman et al. (1987) contends that it is unnecessary to perform elaborate statistical analysis, while also criticizing the SEM approach on the basis that it offers insights on inter-method bias only, and not the variability about the line of equality.

> *However, it is quite wrong to argue solely from a lack of bias that two methods can be regarded as comparable... Knowing the data are consistent with a structural equation with a slope of 1 says something about the absence of bias but nothing about the variability about $Y = X$ (the difference between the measurements), which, as has already been stated, is all that really matters.*

Dunn (2002) highlights an important issue regarding using models such as structural equatio modelling; 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, the ratio of the precision of both methods $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$ must often be assumed to be equal to 1 (Linnet, 1998).

Dunn (2002) considers techniques based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods, simply because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires an invasive medical procedure.

1.10 Model for Replicate Measurements

The single measurement model can be generalized to the replicate measurement case, by additionally specifying replicate values. Let y_{mir} be the $r-$th replicate measurement for item i made by method m. Further to Barnhart et al. (2007) fixed effect can be expressed with a single term α_{mi}, which incorporate the true value μ_i.

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

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

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

The following assumptions are required e_{mir} is independent of the fixed effects with mean $\mathrm{E}(e_{mir}) = 0$. Further to Barnhart et al. (2007) between-item and within-item variances $\mathrm{Var}(\alpha_{mi}) = \sigma^2_{Bm}$ and $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$

1.10.1 Carstensen's Model for Replicate Measurements

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. For the replicate case, an interaction term c is added to the model, with an associated variance component. Their model describing y_{mir}, again the rth replicate measurement on the ith item by the mth method ($m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{1.11}$$

The fixed effects α_m and μ_i represent the intercept for method m and the 'true value' for item i respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau^2_m)$, and model error terms $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi^2_m)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item i, a_{ir} can be removed.

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ_m^2 can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{1.12}$$

Based on this model, Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}. \tag{1.13}$$

This provides the basis of a modified approach to computing LOAs that will be reverted to later.

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.

Altman, D. G., J. M. Bland, and G. E. Kelly (1987). Letters to the editors.

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

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

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

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

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

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

Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician 43*(4), 234–235.

Carpenter, J. and J. Bithell (2000). Bootstrap con" dence intervals: when, which, what? a practical guide for medical statisticians.

Carrasco, J. L. (2004). Structural equation model. In *Encyclopedia of Biopharmaceutical Statistics, Second Edition*, pp. 1–7. Taylor & Francis.

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).

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. (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.

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

Hopkins, W. G. (2004). Bias in bland-altman but not regression validity analyses. *Sportscience 8*(4).

Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics 23*(2), 49–54.

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

Kinsella, A. (1986). Estimating method precision. *The Statistician 35*, 421–427.

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

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

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

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

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

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

Linnet, K. (1999). Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry 45*(6), 882–894.

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.

Morgan, W. A. (1939). A test for the signicance of the difference between two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

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.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

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

Thompson, W. (1963). Precision of simultaneous measurement procedures. *Journal of American Statistical Association 58*, 474–479.

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.

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.