SCRATCH

Kevin O'Brien

February 3, 2017

Contents

1. Agreement and Method Comparison Studies

   (a) What is Agreement?

(b) Repeatability

(c)

(d)

(e)

2. Bland Altman Single Observations

(a)

(b)

3. Alternative Methods

(a) Deming Regression

(b) Mountain Plot

(c) Bartko's Ellipse

(d) Formal Tests and Procedures

4. Replicate Observations

5. LME models

6. Estimation and Algorithms

(a) ML and REML estimation

(b) MINQUE

(c)

7. Residual Diagnostics

(a) Marginal and Conditional Diagnostics

(b) Scaled Residuals

8. Influence Diagnostics

(a) Underlying Concepts

(b) Managing the Covariance Parameters

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

(d) Leverage

(e) Internally and Externally Studentized Residuals

(f) DFFITs and MDFFITs

(g) Covariance Ratio and Trace

(h) Likelihood Distance

(i) Non-iterative Update Procedures

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

Notwithstanding previous remarks about 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),

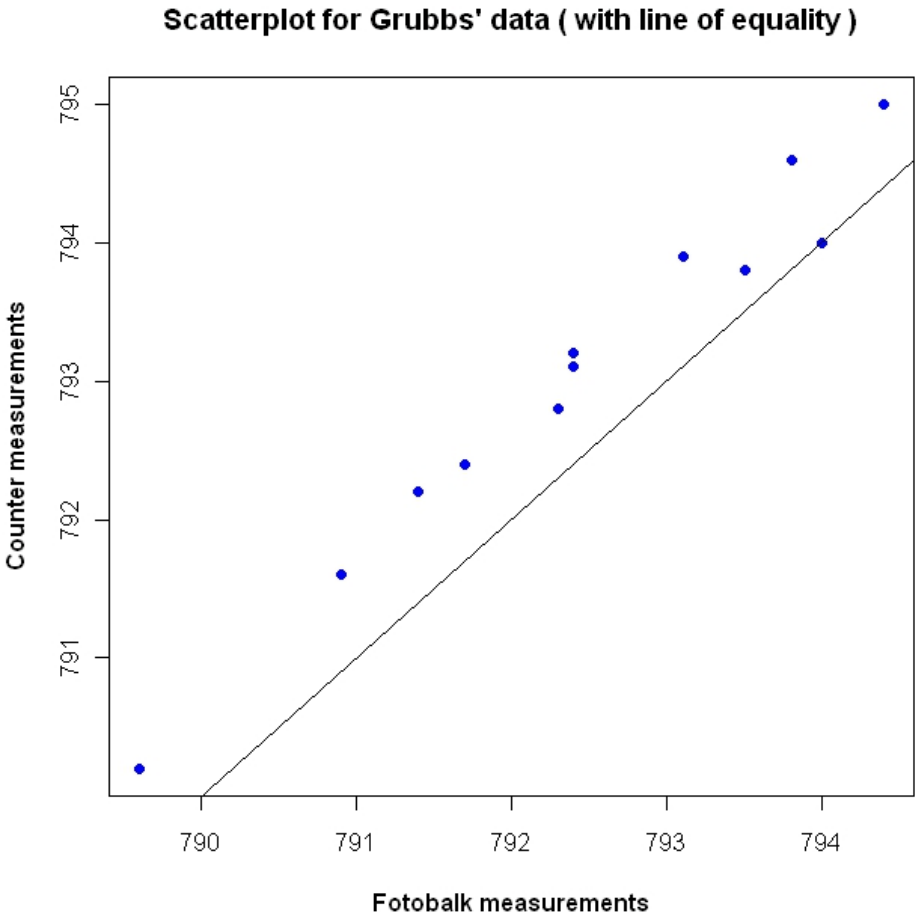

Figure 1: Scatter plot For Fotobalk and Counter Methods.

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 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: 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 2: Fotobalk and Terma methods: differences and averages.

0.1.1 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.

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. This is known as proportional bias. In both Figures 1.4 and 1.5, the assumptions

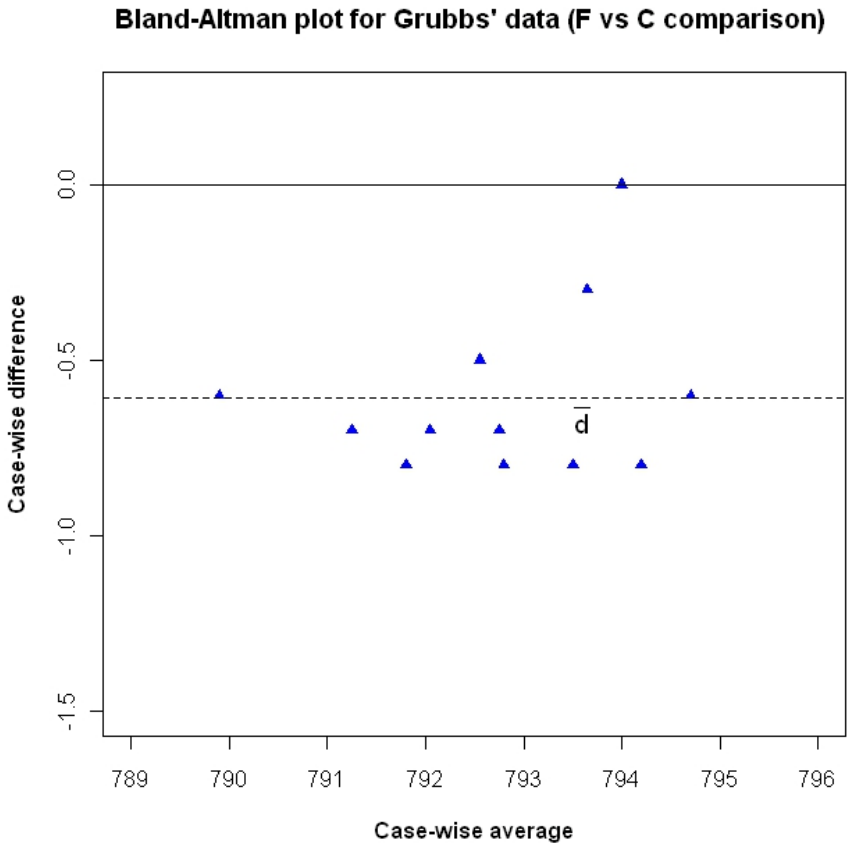

Figure 2: Bland-Altman plot For Fotobalk and Counter methods.

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

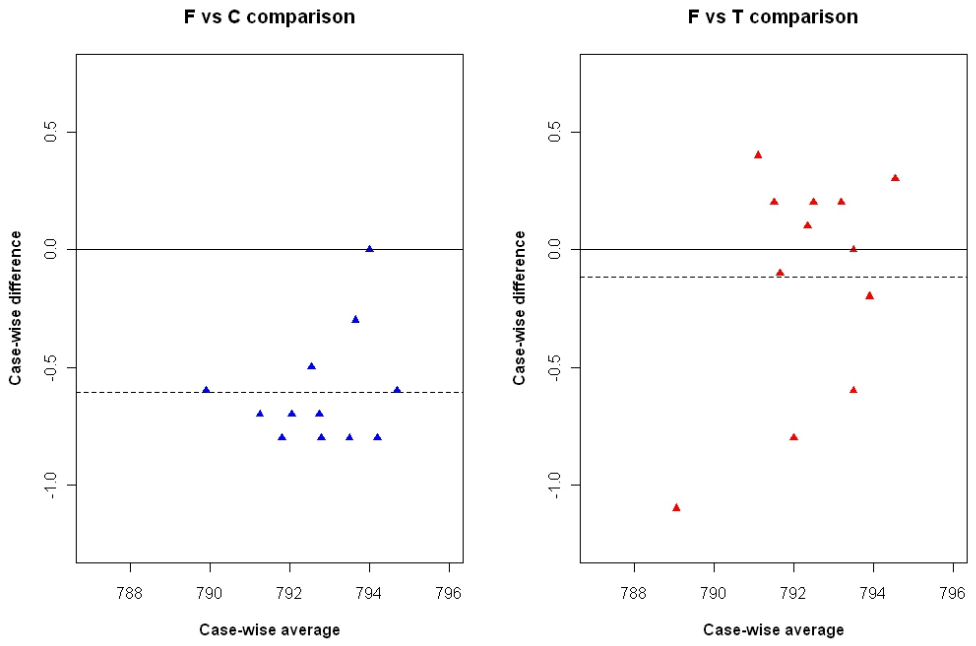

Figure 3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

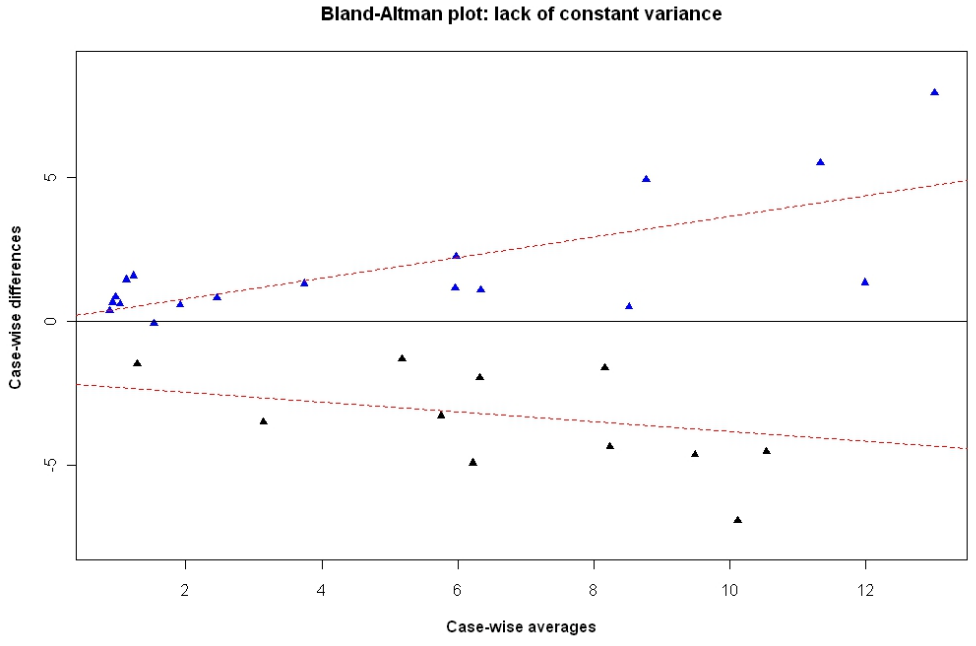

Figure 4: Bland-Altman plot demonstrating the increase of variance over the range.

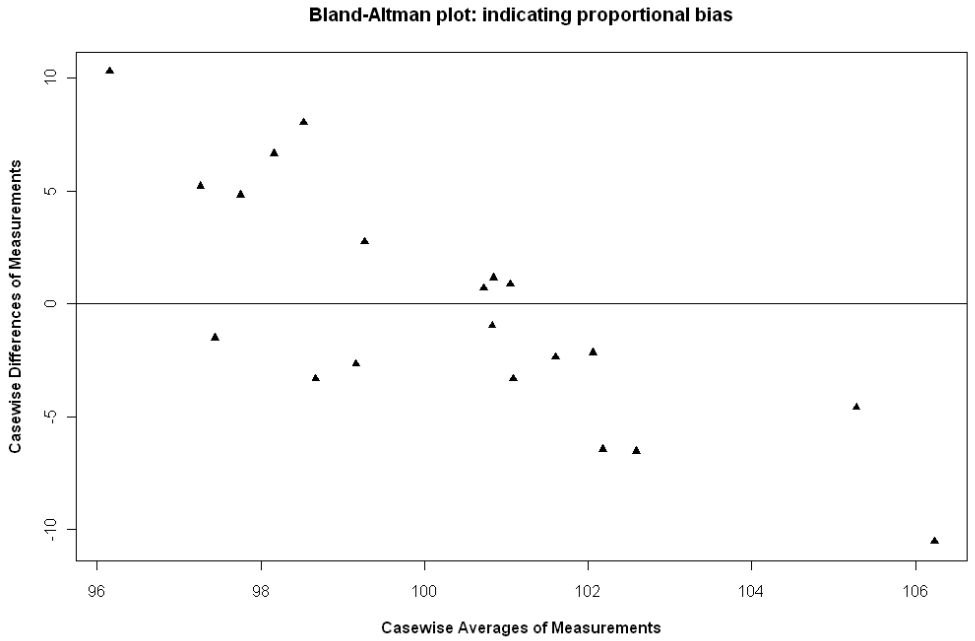

Figure 5: Bland-Altman plot indicating the presence of proportional bias.

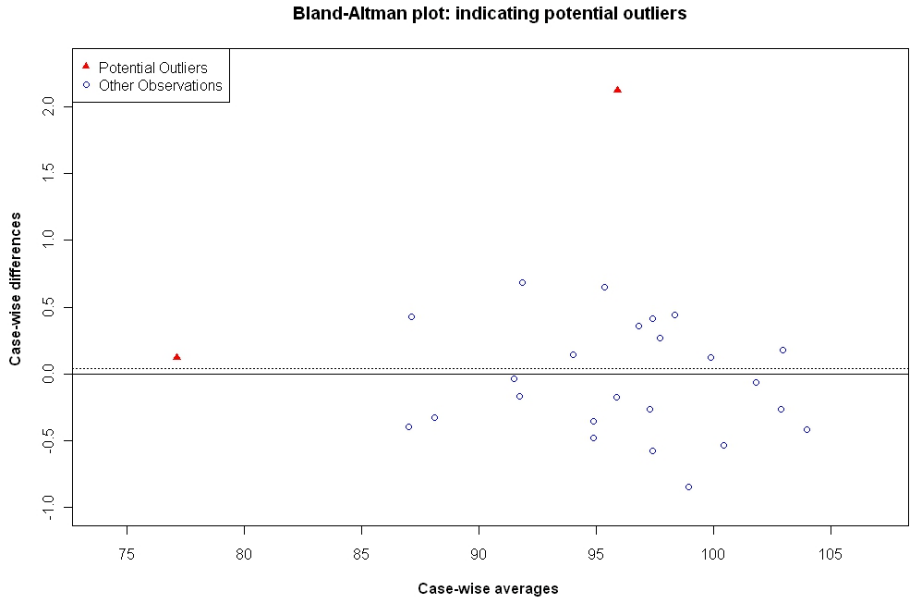

Figure 6: Bland-Altman plot indicating the presence of potential outliers.

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 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,

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

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.

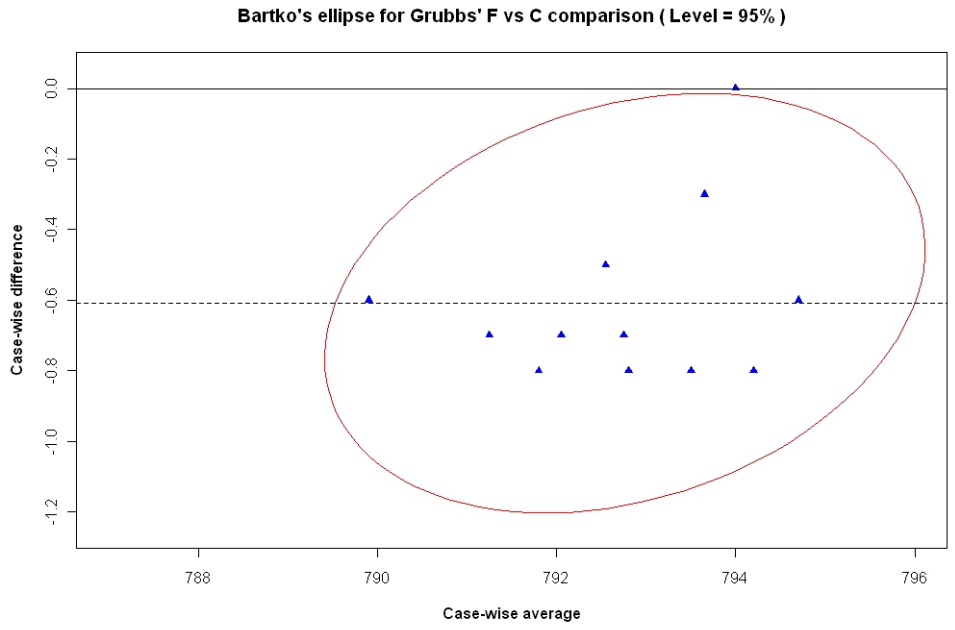

Figure 7: Bartko's Ellipse For Grubbs' Data.

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.

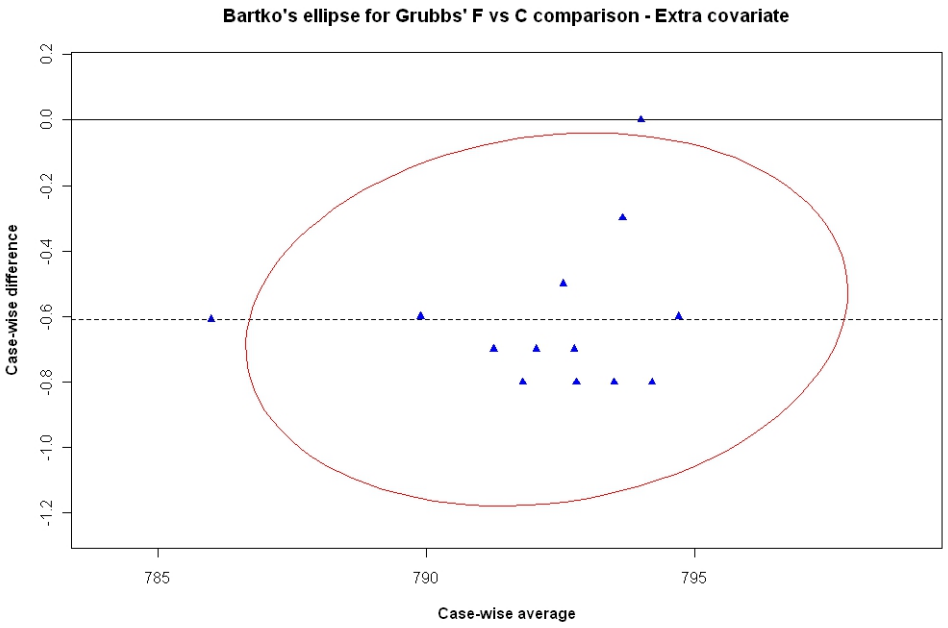

Figure 8: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

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.

0.1.2 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'.

0.1.3 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

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

0.1.4 Replicate Measurements

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

0.2 MCS Data Sets

1. Blood Data
2. Cardiac Data
3. Nadler Hurley

- Introduction to Method Comparison Studies
  - Accuracy and Precision
  - Repeatability (Bland Altman 1999)
  - Barnharts Paper
  -
- Bland and Altman Plot
  - Bland and Altman 1983 and 86
  - Limits of Agreement
  -
  -

0.3 Introduction

Outliers and detection of influent observations is an important step in the analysis of a data set. There are several ways of evaluating the influence of perturbations in the data set and in the model given the parameter estimates.

0.3.1 Overview of R implementations

Further to previous material, an appraisal of the current state of development (or lack thereof) for current implemenations for LME models, particularly for `nlme` and `lme4` fitted models.

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

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

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

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

Important Consideration for MCS

The key issue is that `nlme` allows for the particular specification of Roy's Model, specifically direct specification of the VC matrices for within subject and between subject residuals. The `lme4` package does not allow for Roy's Model, for reasons that will identified shortly. To advance the ideas that eminate from Roys' paper, one is required to use the `nlme` context. However, to take advantage of the infrastructure already

provided for `lme4` models, one may change the research question away from that of Roy's paper. To this end, an exploration of what textbfinfluence.ME can accomplished is merited.

0.4 Computation and Notation

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

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

0.5 Liao Shaio

Lai et Shiao is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. The Data Set used in their examples is unavailable for independent use. Therefore, for the sake of consistency, a data set will be simulated based on the Blood Data that will allow for extra variables.

A Study of the Bland-Altman Plot and its Associated Methodology

Joseph G. Voelkel Bruce E. Siskowski

0.6 Limits of agreement for Carstensen's data

? describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

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

For the 'Fat' data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the 'Oximetry' data set. When the interaction term is omitted,

the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

0.7 Hamlett and Lam

The methodology proposed by **?** is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999).

The desired outcome of this research is to

- Formulate a methodology that represents Best practice in Method Comparison Studies. Indeed the methodology is envsiaged to advance what is considered best practice, inter alia, by making diagnostics procedures a standard part of MCS.
- Provide for ease of use such that non-statisticians can master and implement the method, with a level of training that one would expect as part of a Professional CPD programe.

Apropos of the matter of ease-of-use, certain assumptions must be made.

The user has a reasonable amount of computer literacy. The user would have a reasonable understanding of statistics, consistent with an undergraduate statistics module. That is to say, that the user is acquainted with the idea of $p-$values.

Easy to follow set of instructions to properly implement the method.

Linear Mixed Effects Models can be implemented by using one of the following R packages. lme4 nlme

The first package to be introducted was nlme, developled by Jose Pinheiro and Douglas Bates (Authors of the the companion textbook, NAME)

As this package has been under ongoing development for quite a long time, it is now allows for a lot of complex LME implementations. Furthermore, nlme is one of the base R packages. That is to say, when one downloads and installs R, nlme is automatically installed also, and can be called immediately.

Having said that, the authors have pointed to several limitations of the overall methodology thrugh R. The original developers have both left the project, but other

statisticians have taken over the development, and indeed a new version of nlme was released.

LME4 is a more recent package. at a glance, the syntax is easier, but the development is less advanced. There are several functionalities that can not be implemented with lme4 yet. As an example - CHAP5 in PB - has no equivalent in LME4. Indeed no textbook exists to co-incide with LME4.

The main author, Douglas Bates, has turned his attention to development of LME models in the Julia programming language.

The nlmeU package is described by its authors as an extesntion of the nlme package, and indeed provides for additionally functionality. The package is also useful as it serves as a companion piece to the book by Galecki and Burzwhatski.

The nlme package also allows for the specification of GLS models.

Objects and Classes

The main nlme object is an `nlme` model.

The main lme4 object is called an `lmer` model

The lattice package is used for graphical methods.

Model Diagnostics with `nlme`

0.7.1 Inappropriate Techniques for MCS

0.7.2 Links and Papers

`Westgard Statistics - http://www.westgard.com/lesson23.htm`

Measurement Systems Analysis

The topic of measurement sensitivity anaylysis (MSA, also known as Gauge R&R) is prevalent in industrial statistics (i.e Six Sigma).

There is extensive literature that covers the area. For the sake of brevity, we will use Cano et al.

For sake of clarity, Cano's definitions of repeatability and reproducibility are listed, with added emphasis.

Reproducibility is rarely, if ever, discussed in the domain of Method Comparison Studies. This may be due to the fact that prevalent methodologies can be used for the problem.However the methodologies proposed by this research can easily be extended.

Chapter 1

Introduction

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

Notwithstanding previous remarks about 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.

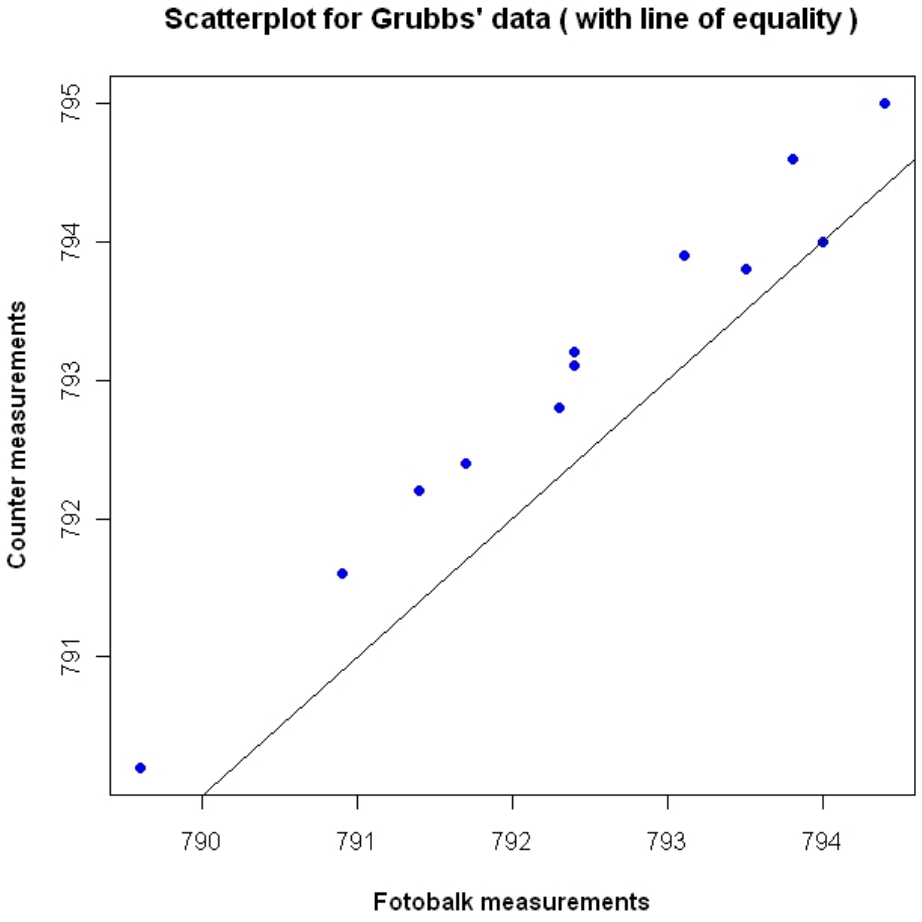

Figure 1.1: Scatter plot For Fotobalk and Counter Methods.

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 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 1.1: 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 1.2: Fotobalk and Terma methods: differences and averages.

1.1.1 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.

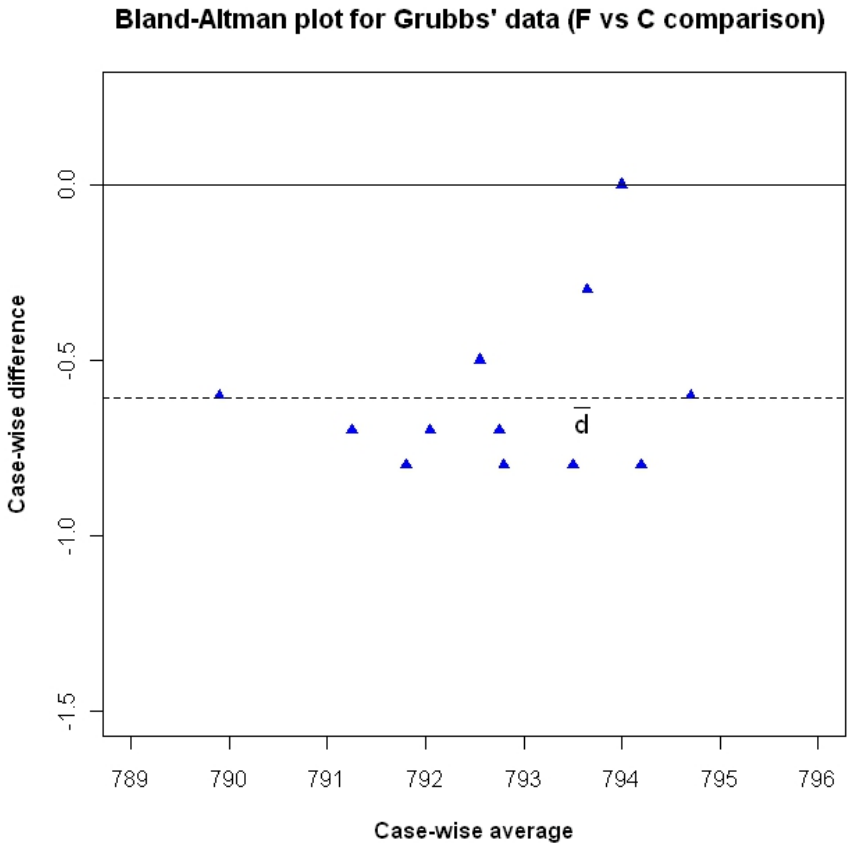

Figure 1.2: Bland-Altman plot For Fotobalk and Counter methods.

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.

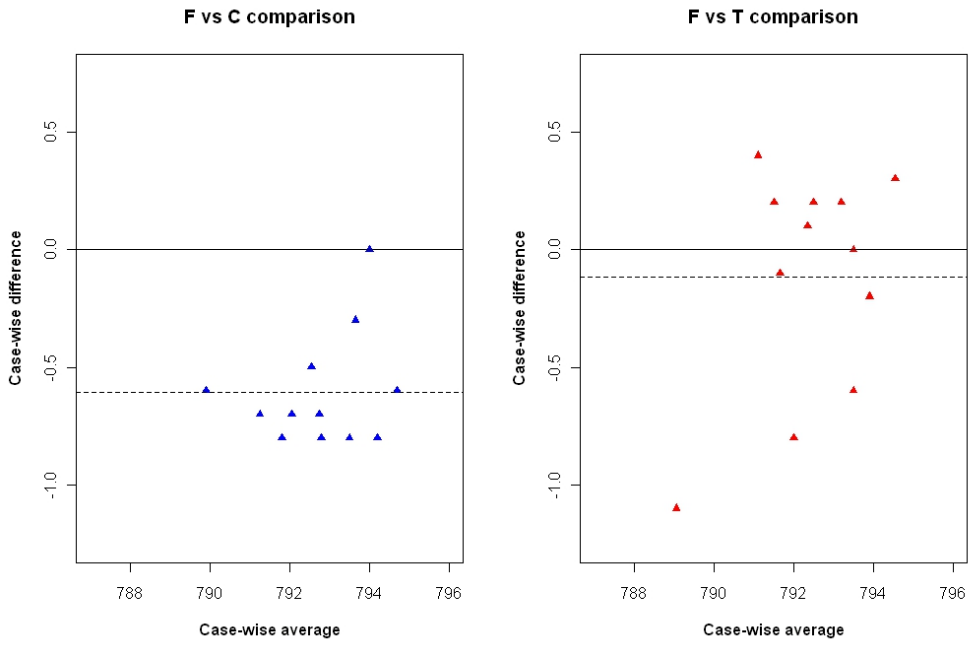

Figure 1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

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

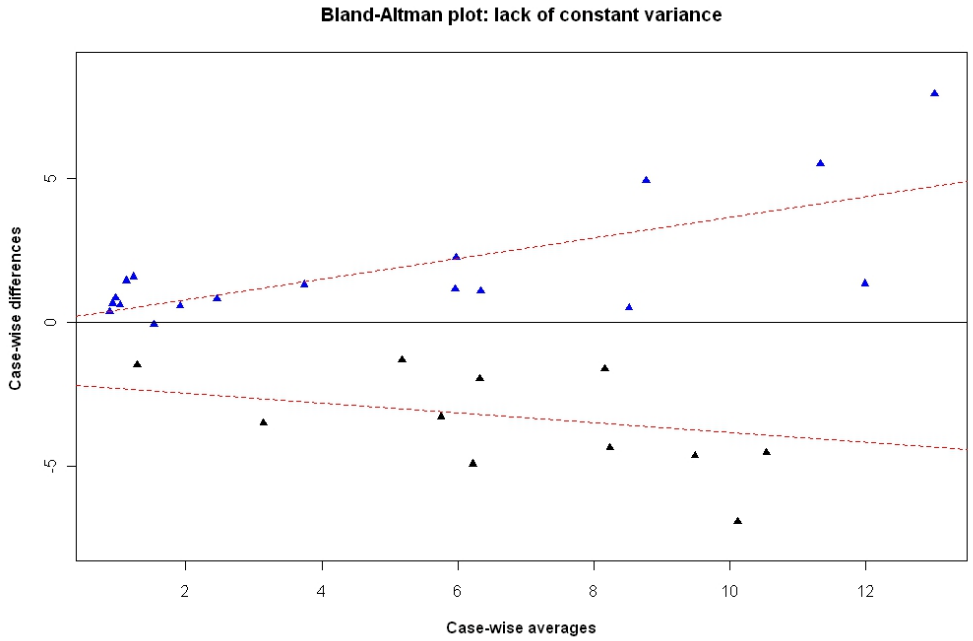

Figure 1.4: Bland-Altman plot demonstrating the increase of variance over the range.

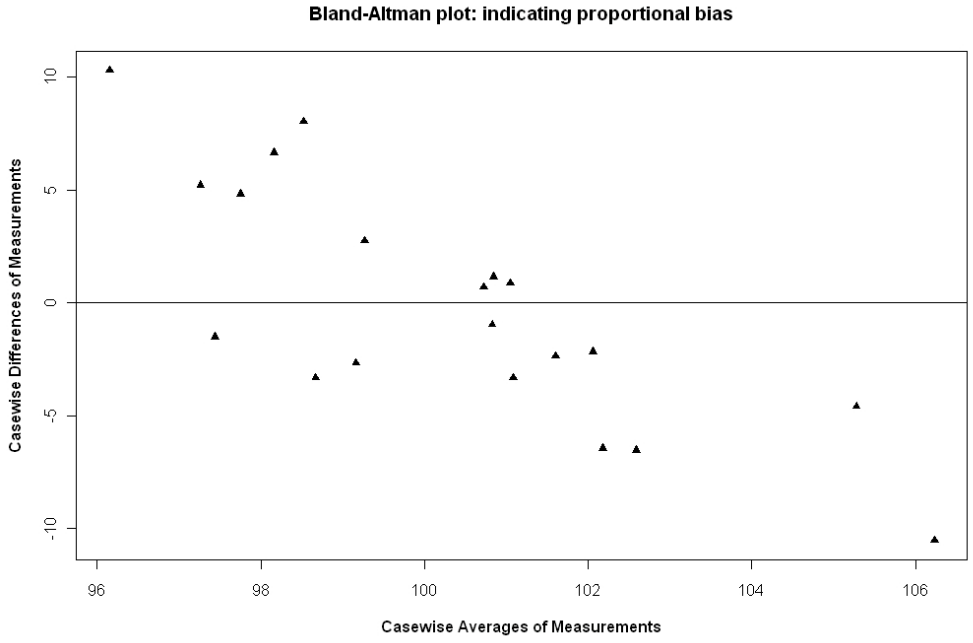

Figure 1.5: Bland-Altman plot indicating the presence of proportional bias.

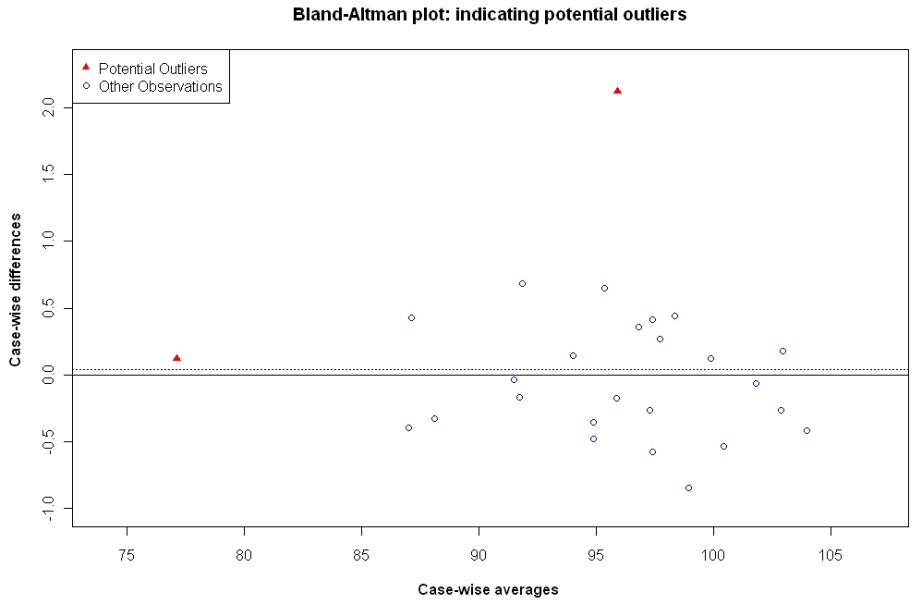

Figure 1.6: Bland-Altman plot indicating the presence of potential outliers.

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.

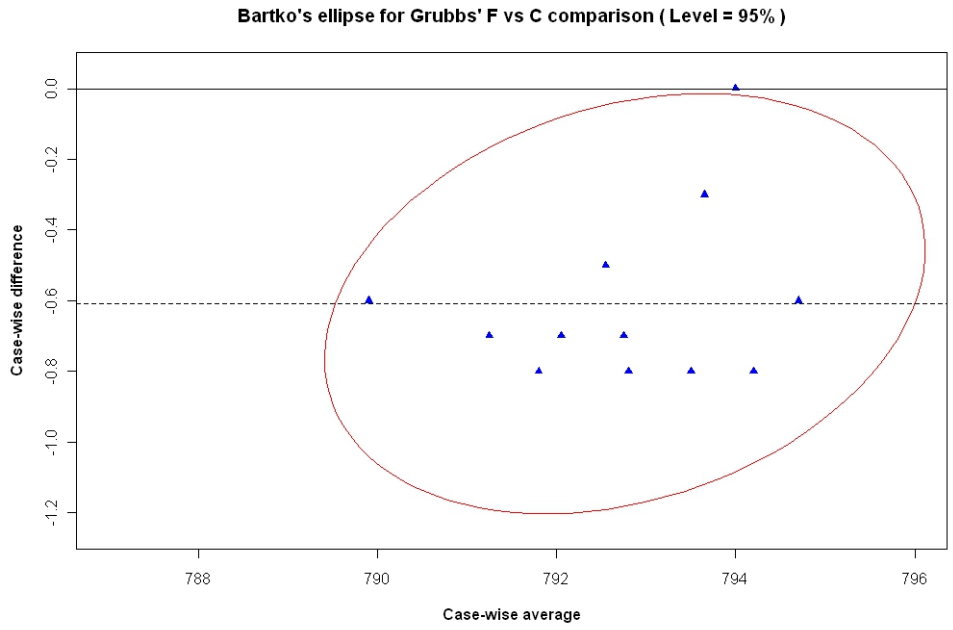

Figure 1.7: Bartko's Ellipse For Grubbs' Data.

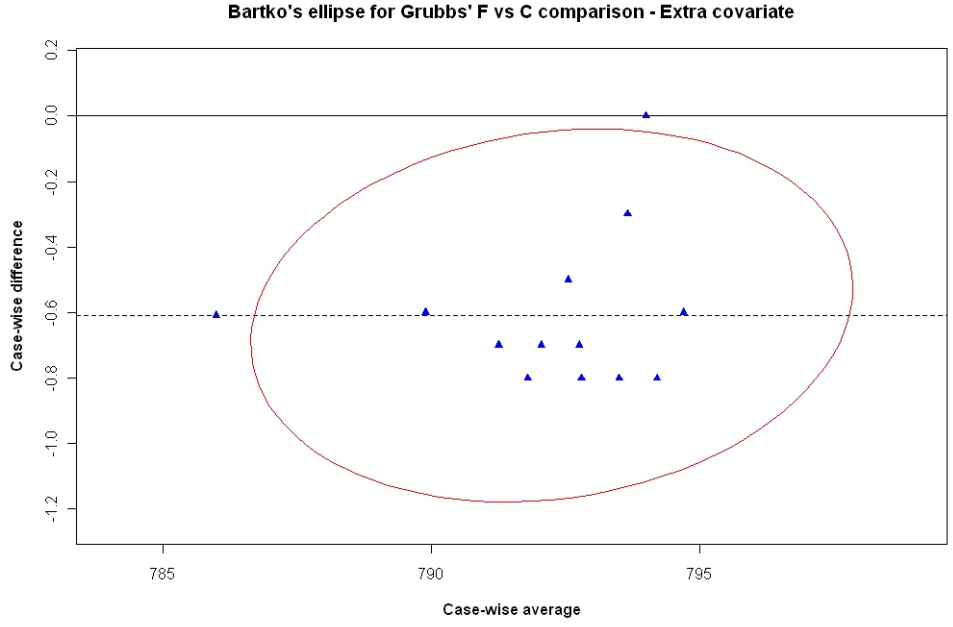

Figure 1.8: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

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.

1.1.2 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'.

1.1.3 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

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

1.1.4 Replicate Measurements

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

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.

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

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

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).

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

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

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

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

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

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.

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

Haslett, J. and D. Dillane (2004). Application of 'delete = replace' to deletion diagnostics for variance component estimation. *Journal of the Royal Statistical Society (Series B) 66*, 131–143.

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

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.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

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

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

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.

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

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

Roy, A. (2009). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

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

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

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

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

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.

1.2 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.

1.2.1 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 λ, Kummel (1879) derived the following estimate for the Demimg regression slope parameter. (α 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 **?**, 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 1.3: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

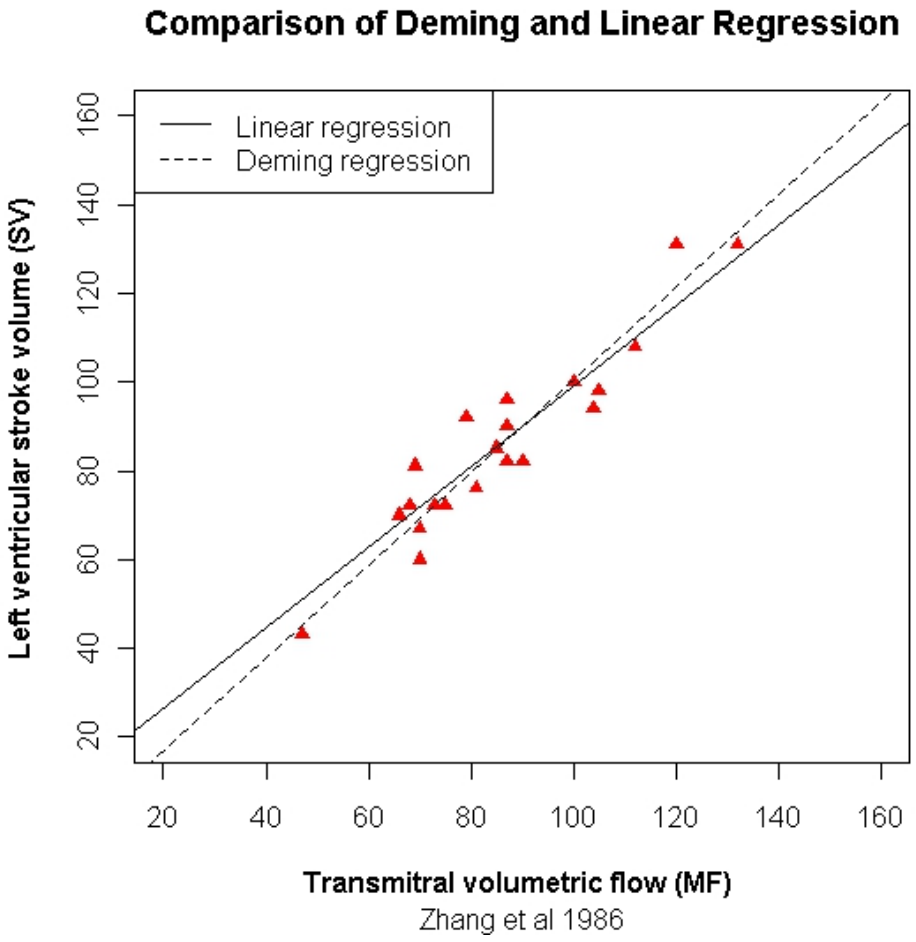

Figure 1.9: Deming Regression For Zhang's Data

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 (λ, in their paper as η) is correctly specified ,but in practice this is often not the case, with the λ being underestimated.

Chapter 2

Model Diagnostics

Contents

Abstract

This chapter is broken into two parts. The first part is a review of diagnostics methods for linear models, intended to acquaint the reader with the subject, and also to provide a basis for material covered in the second part. Particular attention is drawn to graphical methods.

The second part of the chapter looks at diagnostics techniques for LME models, firsly covering the theory, then proceeding to a discussion on implementing these using `R` code.

While a substantial body of work has been developed in this area, there is still areas worth exploring. In particular the development of graphical techniques pertinent to LME models should be looked at.

2.1 Model Validation Framework

In statistical modelling, the process of model validation is a critical step of model fitting process, but also a step that is too often overlooked. A very simple procedure is to examine commonly-used metrics, such as the R^2 value. However, using a small handful of simple measures and methods is insufficient to properly assess the quality of a fitted model. To do so properly, a full and comprehensive analysis that tests of all of the assumptions, as far as possible, must be carried out.

? describes the model validatin framework as comprised of the following tasks

- overall measures of goodness-of-fit
- the informal, graphical examination of estimates of model errors to assess the quality of distributional assumptions: residual analysis
- the quantitative assessment of the inter-relationship of model components; for example, collinearity diagnostics
- the qualitative and quantitative assessment of influence of cases on the analysis, i.e. influence analysis.

Residual analysis is a widely used model validation technique. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect. that is to say, if the residuals behave randomly, with no discernible trend, the model has fitted the data well. If some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted.

Statistical software environments, such as the `R` Programming language, provides a suite of tests and graphical procedure sfor appraising a fitted linear model, with several

of these procedures analysing the model residuals.

For LME models the matter of residual is more complex. **?** describes two types of residuals, marginal residuals and conditional residuals. A marginal residual is the difference between the observed data and the estimated marginal mean. A conditional residual is the difference between the observed data and the predicted value of the observation. In a model without random effects, both sets of residuals coincide. We shall revert to this matter in due course.

Further to the analysis of residuals, **?** recommends the examination of the following questions.

- Does the model-data agreement support the model assumptions?
- Should model components be refined, and if so, which components? For example, should regressors be added or removed, and is the covariation of the observations modeled properly?
- Are the results sensitive to model and/or data? Are individual data points or groups of cases particularly influential on the analysis?

2.1.1 Outliers and Leverage

The question of whether or not a point should be considered an outlier must also be addressed. An outlier is an observation whose true value is unusual given its value on the predictor variables. The leverage of an observation is a further consideration. Leverage describes an observation with an extreme value on a predictor variable is a point with high leverage. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence can be thought of as the product of leverage and outlierness. An observation is said to be influential if removing the observation substantially changes the

estimate of the regression coefficients. The R programming language has a variety of methods used to study each of the aspects for a linear model. While linear models and GLMS can be studied with a wide range of well-established diagnostic technqiues, the choice of methodology is much more restricted for the case of LMEs.

For classical linear models, residual diagnostics are typically conducted using a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations.

2.2 Case Deletion Diagnostics

Since the pioneering work of Cook in 1977, deletion measures have been applied to many statistical models for identifying influential observations. Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of β and σ^2, which exclude the $i-$th observation, can be computed without re-fitting the model.

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called '*observation-diagnostics*'. For multiple observations, Preisser describes the diagnostics as '*cluster-deletion*' diagnostics. When applied to LME models, such update formulas are available only if one assumes that the covariance parameters are not affected by the removal of the observation in question. However, this is rarely a reasonable assumption.

2.2.1 Extension of Diagnostic Methods to LME models

? noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

? develops case deletion diagnostics, in particular the equivalent of Cook's distance, a well-known metric, for diagnosing influential observations when estimating the fixed effect parameters and variance components. Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models. We shall provide a fuller discussion of Cook's distance in due course.

Demidenko (2004) extends several regression diagnostic techniques commonly used in linear regression, such as leverage, infinitesimal influence, case deletion diagnostics, Cook's distance, and local influence to the linear mixed-effects model. In each case, the proposed new measure has a direct interpretation in terms of the effects on a parameter of interest, and reduces to the familiar linear regression measure when there are no random effects.

The new measures that are proposed by Demidenko (2004) are explicitly defined functions and do not require re-estimation of the model, especially for cluster deletion diagnostics. The basis for both the cluster deletion diagnostics and Cook's distance is a generalization of Miller's simple update formula for case deletion for linear models. Furthermore Demidenko (2004) shows how Pregibon's infinitesimal case deletion diagnostics is adapted to the linear mixed-effects model.

Demidenko (2004) proposes two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

Cook's distance

In the study of Linear model diagnostics, Cook proposed a measure that combines the information of leverage and residual of the observation, now known simply as the Cook's Distance. **?** would later adapt the Cook's distance measure for the analysis of LME models.

2.3 Analysis of Influence

2.3.1 Influence Analysis for LME Models

The linear mixed effects model is a useful methodology for fitting a wide range of models. However, linear mixed effects models are known to be sensitive to outliers. **?** advises that identification of outliers is necessary before conclusions may be drawn from the fitted model.

Standard statistical packages concentrate on calculating and testing parameter estimates without considering the diagnostics of the model.The assessment of the effects of perturbations in data, on the outcome of the analysis, is known as statistical influence analysis. Influence analysis examines the robustness of the model. Influence analysis methodologies have been used extensively in classical linear models, and provided the basis for methodologies for use with LME models. Computationally inexpensive diagnostics tools have been developed to examine the issue of influence (**?**).

2.4 Zewotir Measures of Influence in LME Models

? describes a number of approaches to model diagnostics, investigating each of the following;

- Variance components
- Fixed effects parameters
- Prediction of the response variable and of random effects
- likelihood function

? lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

2.5 Matrix Notation for Case Deletion

2.5.1 Case deletion notation

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i, j)-$th element of $\boldsymbol{A}$.

2.5.2 Further Assumptions of Linear Models

As with fitted models, the assumption of normality of residuals and homogeneity of variance is applicable to LMEs also.

Homoscedascity is the technical term to describe the variance of the residuals being constant across the range of predicted values. Heteroscedascity is the converse scenario : the variance differs along the range of values.

On occasion, quantification is not possible. Assume, for example, that a data point is removed and the new estimate of the G matrix is not positive definite. This may occur if a variance component estimate now falls on the boundary of the parameter space. Thus, it may not be possible to compute certain influence statistics comparing the full-data and reduced-data parameter estimates. However, knowing that a new singularity was encountered is important qualitative information about the data points influence on the analysis.

The basic procedure for quantifying influence is simple:

1. Fit the model to the data and obtain estimates of all parameters.

2. Remove one or more data points from the analysis and compute updated estimates of model parameters.

3. Based on full- and reduced-data estimates, contrast quantities of interest to determine how the absence of the observations changes the analysis.

We use the subscript (U) to denote quantities obtained without the observations in the set U. For example, (U) denotes the fixed-effects ***leave-U-out*** estimates. Note that the set U can contain multiple observations.

If the global measure suggests that the points in U are influential, you should next determine the nature of that influence. In particular, the points can affect

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

It is important to further decompose the initial finding to determine whether data points are actually troublesome. Simply because they are influential somehow, should not trigger their removal from the analysis or a change in the model. For example, if points primarily affect the precision of the covariance parameters without exerting much influence on the fixed effects, then their presence in the data may not distort hypothesis tests or confidence intervals about β.

2.5.3 Summary of Paper

Standard residual and influence diagnostics for linear models can be extended to LME models. The dependence of the fixed effects solutions on the covariance parameters has important ramifications on the perturbation analysis. Calculating the studentized residuals-And influence statistics whereas each software procedure can calculate both conditional and marginal raw residuals, only SAs Proc Mixed is currently the only program that provide studentized residuals Which ave preferred for model diagnostics.

The conditional Raw residuals ave not well suited to detecting outliers as are the studentized conditional residuals. (schabenbege r)

LME are flexible tools for the analysis of clustered and repeated measurement data. LME extend the capabilities of standard linear models by allowing unbalanced and missing data, as long as the missing data are MAR. Structured covariance matrices for both the random effects G and the residuals R. missing at Random.

A conditional residual is the difference between the observed valve and the predicted valve of a dependent variable- Influence diagnostics are formal techniques that allow the identification observation that heavily influence estimates of parameters. To alleviate the problems with the interpretation of conditional residuals that may have unequal variances, we consider sealing. Residuals obtained in this manner ave called studentized residuals.

2.6 Schabenberger: Summary and Conclusions

- Standard residual and inuence diagnostics for linear models can be extended to linear mixed models. The dependence of xed-effects solutions on the covariance parameter estimates has important ramications in perturbation analysis.

- To gauge the full impact of a set of observations on the analysis, covariance parameters need to be updated, which requires retting of the model.

- The experimental INFLUENCE option of the MODEL statement in the MIXED procedure (SAS 9.1) enables you to perform iterative and noniterative inuence analysis for individual observations and sets of observations.

- The conditional (subject-specic) and marginal (population-averaged) formulations in the linear mixed model enable you to consider conditional residuals that

use the estimated BLUPs of the random effects, and marginal residuals which are deviations from the overall mean.

- Residuals using the BLUPs are useful to diagnose whether the random effects components in the model are specied correctly, marginal residuals are useful to diagnose the xed-effects components.

- Both types of residuals are available in SAS 9.1 as an experimental option of the MODEL statement in the MIXED procedure.

- It is important to note that influence analyses are performed under the assumption that the chosen model is correct. Changing the model structure can alter the conclusions. Many other variance models have been t to the data presented in the repeated measures example. You need to see the conclusions about which model component is affected in light of the model being fit.

Leave-One-Out Diagnostics with `lmeU`

Galecki et al provide a brief the matter of LME influence diagnostics in their book.

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

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

The addition of an extra factor

Interaction terms are featured in ANOVA designs.

My search just now found no mention of Cook's distance or influence measures.

The closest I found was an unanswered question on this from April 2003 (http://finzi.psych.upenn.e

Beyond that, there is an excellent discussion of "Examining a Fitted Model" in Sec. 4.3 (pp. 174-197) of Pinheiro and Bates (2000) Mixed-Effects Models in S and S-Plus (Springer).

Pinheiro and Bates decided NOT to include plots of Cook's distance among the many diagnostics they did provide. However, 'plot(fit.lme)' plots 'standardized residuals' vs. predicted or 'fitted values'. Wouldn't points with large influence stand apart from the crowd in terms of 'fitted value'?

Of course, there are many things other one could do to get at related information, including reading the code for 'influence' and 'lme', and figure out from that how to write an 'influence' method for an 'lme' object.

2.7 Paired T tests

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

Components in assessing agreement

1. The degree of linear relationship between the two sets

2. The amount of bias as represented by the difference in the means

3. The Differences in the two variances.

2.8 Methods of assessing agreement

1. Pearson's Correlation Coefficient
2. Intraclass correlation coefficient
3. Bland Altman Plot
4. Bartko's Ellipse (1994)
5. Blackwood Bradley Test
6. Lin's Reproducibility Index
7. Luiz Step function

Bland and Altman attend to the issue of repeated measures in 1996.

Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual.

Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

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

2.8.1 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits

of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

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

Chapter 3

Appendix

Bayesian BA - Philip J Schluter

Bayesian Bland Altman Approaches A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies

http://www.biomedcentral.com/1471-2288/9/6

Background

Assessing agreement in method comparison studies depends on two fundamentally important components; validity (the between method agreement) and reproducibility (the within method agreement).

The Bland-Altman limits of agreement technique is one of the favoured approaches in medical literature for assessing between method validity. However, few researchers have adopted this approach for the assessment of both validity and reproducibility.

This may be partly due to a lack of a flexible, easily implemented and readily available statistical machinery to analyse repeated measurement method comparison data.

Methods

Adopting the Bland-Altman framework, but using Bayesian methods, we present this statistical machinery. Two multivariate hierarchical Bayesian models are advocated, one which assumes that the underlying values for subjects remain static (exchangeable replicates) and one which assumes that the underlying values can change between repeated measurements (non-exchangeable replicates).

Results

We illustrate the salient advantages of these models using two separate datasets that have been previously analysed and presented; (i) assuming static underlying values analysed using both multivariate hierarchical Bayesian models, (ii) assuming each subject's underlying value is continually changing quantity and analysed using the non-exchangeable replicate multivariate hierarchical Bayesian model.

Conclusion These easily implemented models allow for full parameter uncertainty, simultaneous method comparison, handle unbalanced or missing data, and provide estimates and credible regions for all the parameters of interest. Computer code for the analyses in also presented, provided in the freely available and currently cost free software package WinBUGS. ¡hr¿

Bayesian Approach

A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies PJ Schluter - BMC medical research methodology, 2009 - biomedcentral.com

- Assessing agreement in method comparison studies depends on two fundamentally important components; validity (the between method agreement) and reproducibility (the within method agreement).

- The Bland-Altman limits of agreement technique is one of the f

3.1 Escaramis

3.1.1 Background

In an agreement assay, it is of interest to evaluate the degree of agreement between the different methods (devices, instruments or observers) used to measure the same characteristic. We propose in this study a technical simplification for inference about the total deviation index (TDI) estimate to assess agreement between two devices of normally-distributed measurements and describe its utility to evaluate inter- and intra-rater agreement if more than one reading per subject is available for each device.

3.1.2 Methods

We propose to estimate the TDI by constructing a probability interval of the difference in paired measurements between devices, and thereafter, we derive a tolerance interval (TI) procedure as a natural way to make inferences about probability limit estimates. We also describe how the proposed method can be used to compute bounds of the coverage probability.

3.1.3 Results

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

ods, showing that the TI approximation produces accurate empirical confidence levels which are reasonably close to the nominal confidence level.

3.1.4 Conclusions

The method proposed is straightforward since the TDI estimate is derived directly from a probability interval of a normally-distributed variable in its original scale, without further transformations. Thereafter, a natural way of making inferences about this estimate is to derive the appropriate TI. Constructions of TI based on normal populations are implemented in most standard statistical packages, thus making it simpler for any practitioner to implement our proposal to assess agreement.

Lin defined the TDI as the boundary, κ_P which capyures a large proportion p of paired based differences from two devices or observers within the boundary.

The value of κ_P that yeilds $P(|D| < \kappa_p) = p$ where D is the paired-difference variate.

$$\kappa_P = F^{-1}(p) = \sigma_D \sqrt{\chi^2(p, 1, \mu_D^2/\sigma_d^2)}$$

$$\kappa_P = Z_{\frac{1+p}{2}} \|\varepsilon\|$$

Tolerance Interval around the TDI estimate

$$\hat{\kappa_p} = \hat{\mu}_D = Z_{p_i} \sigma_d$$

Coverage Probability is another user friendly measure of agrre,ment which is related to the computation of the TDI.

3.2 Schabenberger

schab examines the use and implementation of influence measures in LME models.

Influence is understood to be the ability of a single or multiple data points, through their presences or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model (*schabenberger*).

Outliers are the most noteworthy data points in an analysis, and an objective of influence analysis is how influential they are, and the manner in which they are influential.

schab describes a simple procedure for quantifying influence. Firstly a model should be fitted to the data, and estimates of the parameters should be obtained. The second step is that either single of multiple data points, specifically outliers, should be omitted from the analysis, with the original parameter estimates being updated.

This is known as '*leave one out leave k out*' analysis. The final step of the procedure is comparing the sets of estimates computed from the entire and reduced data sets to determine whether the absence of observations changed the analysis.

schabenberger notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates.

In recent years, mixed models have become invaluable tools in the analysis of experimental and observational data. In these models, more than one term can be subject to random variation. Mixed model technology enables you to analyze complex experimental data with hierarchical random processes, temporal, longitudinal, and spatial data, to name just a few important applications.

schab remarks that the concept of critiquing the model-data agreement applies in mixed models in the same way as in linear fixed-effects models. In fact, because of

the more complex model structure, you can argue that model and data diagnostics are even more important. For example, you are not only concerned with capturing the important variables in the model. You are also concerned with "distributing them correctly between the fixed and random components of the model. The mixed model structure presents unique and interesting challenges that prompt us to reexamine the traditional ideas of influence and residual analysis.

3.3 Hawkins : Diagnostics for conformity of paired quantitative measurements

- Matched pairs data arise in many contexts in case-control clinical trials, for example, and from cross-over designs. They also arise in experiments to verify the equivalence of quantitative assays. This latter use (which is the main focus of this paper) raises difficulties not always seen in other matched pairs applications.

- Since the designs deliberately vary the analyte levels over a wide range, issues of variance dependent on mean, calibrations of differing slopes, and curvature all need to be added to the usual model assumptions such as normality.

- Violations in any of these assumptions invalidate the conventional matched pairs analysis.

- A graphical method, due to Bland and Altman, of looking at the relationship between the average and the difference of the members of the pairs is shown to correspond to a formal testable regression model.

- Using standard regression diagnostics, one may detect and diagnose departures from the model assumptions and remedy them for example using variable trans-

formations. Examples of different common scenarios and possible approaches to handling them are shown.

A multi-Rate nonparametric test of agreement and corresponding agreement plot - Published in: Computational Statistics and Data Analysis 54(2010)109-119 - Author: Alan D. Hutson, University of Buffalo

This approach takes advantage of readily avilable tests of uniformity found in most statistical software packages. Such tests include the KS d statistic, the Anderson Darling Statistic and the Cramer-Von Mises statistical test for univariate data.

An important aspect of this approach is the "Agreement Region".

Roy Test

Roys Tests (Roy 2009) Roy 2009 devised an LME based Testing approach to the MCS problem, based on earlier work by Hamlett et al. Roy 2009 presents a series of three formal hypothesis tests for assessing agreement between two methods of measurement. Roy also alludes to some of the current shortcomings of the approach.

Comparing different model specifications with LRT tests

- Roy 2007 - Roy 2009 - Hamlett et al. - Roy Leiva 2011

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat,   #Symm , Symm#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corSymm(form=~1 | item/repl),
+     method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model.

```
Nested Model (Between-Item Variability)
> NMB.fit  = lme(y ~ meth-1, data = dat,   #CS , Symm#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corSymm(form=~1 | item/repl),
+     method="ML")

Nested Model (Within item Variability)
> NMW.fit = lme(y ~ meth-1, data = dat,   #Symm , CS#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="ML")
```

Nested Model (Overall Variability) Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.

```
> NMO.fit = lme(y ~ meth-1, data = dat,   #CS , CS#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="ML")
```

ANOVAs for Original Fits The likelihood Ratio test is very simple to implement in R. All that is required it to specify the reference model and the relevant nested mode as arguments to the command anova(). The figure below displays the three tests described by Roy (2009).

```
> testB    = anova(Ref.Fit,NMB.fit)                    # Between-Subject Vari
> testW   = anova(Ref.Fit,NMW.fit)                   # Within-Subject Variabil
> testO     = anova(Ref.Fit,NMO.fit)                   # Overall Variabilities
```

3.4 Profile Function with "lmer"

The profile() function for lmer models is now available in the latest version of lme4, to be installed by typing:

install.packages("lme4",repos="http://r-forge.r-project.org")

also

The mle function from the stats4 package is a wrapper of optim, which makes it quite easy to produce profile likelihood computations.

See help("profile,mle-method", package = "stats4") for more information.

http://people.upei.ca/hstryhn/stryhn208.pdf

The profile likelihood (or likelihood or likelihood ratio) methid is applicable to all likelihood based statstical analysis and is generally less sensitive to the difficulties encountered by walkd-Tyoe CIs.

3.5 Turkan's LMEs

The linear mixed model is considerably sensitive to outliers and influential observations. It is known that outliers and influential observations affect substantially the results of analysis. So it is very important to be aware of these observations.

Some diagnostics which are analogue of diagnostics in multiple linear regression were developed to detect outliers and influential observations in the linear mixed model. *In this paper, the new diagnostic measure which is analogue of the Pena's influence statistic is developed for the linear mixed model.*

Estimation and Building blacks in LME models

$$\hat{u} = DZ^T H^{-1}(y - X\hat{\beta})$$

$$\hat{y} = (I_n - H^{-1})y + H^{-1}X\hat{\beta}$$

The proposed diagnostic Measure.

3.5.1 Ordinary Least Product Regression

Ludbrook (1997) states that the grouping structure can be straightforward, but there are more complex data sets that have a hierarchical(nested) model.

Observations between groups are independent, but observations within each groups are dependent because they belong to the same subpopulation. Therefore there are two sources of variation: between-group and within-group variance. Mean correction is a method of reducing bias.

3.5.2 A regression based approach based on Bland Altman Analysis

Lu et al used such a technique in their comparison of DXA scanners. They also used the Blackwood Bradley test. However it was shown that, for particular comparisons, agreement between methods was indicated according to one test, but lack of agreement was indicated by the other.

3.6 Measurement Error Models

Dunn (2002) proposes a measurement error model for use in method comparison studies. Consider n pairs of measurements X_i and Y_i for $i = 1, 2, ...n$.

$$X_i = \tau_i + \delta_i \tag{3.1}$$

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

In the above formulation is in the form of a linear structural relationship, with τ_i and $\beta\tau_i$ as the true values , and δ_i and ϵ_i as the corresponding measurement errors. In the case where the units of measurement are the same, then $\beta = 1$.

$$E(X_i) = \tau_i \tag{3.2}$$

$$E(Y_i) = \alpha + \beta\tau_i$$

$$E(\delta_i) = E(\epsilon_i) = 0$$

The value α is the inter-method bias between the two methods.

$$z_0 = d = 0 \tag{3.3}$$

$$z_{n+1} = z_n^2 + c \tag{3.4}$$

3.7 Work List

1. ML v REML
2. Nested Models and LRTs
3. Generalized Lease Squares

4. Diagnostics

5. Simplifying GLS

6. Paper progression

3.8 Diagnostics

3.8.1 Identifying outliers with a LME model object

The process is slightly different than with standard LME model objects, since the ***influence*** function does not work on lme model objects. Given ***mod.lme***, we can use the plot function to identify outliers.

3.8.2 Diagnostics for Random Effects

Empirical best linear unbiased predictors EBLUPS provide the a useful way of diagnosing random effects.

EBLUPs are also known as "shrinkage estimators" because they tend to be smaller than the estimated effects would be if they were computed by treating a random factor as if it was fixed (West etal)

3.9 Iterative and non-iterative influence analysis

? highlights some of the issue regarding implementing mixed model diagnostics.

3.9.1 Iterative Influence Analysis

For linear models, the implementation of influence analysis is straightforward. However, for LME models, the process is more complex. Update formulas for the fixed effects are available only when the covariance parameters are assumed to be known. A measure of total influence requires updates of all model parameters. This can only be achieved in general is by omitting observations, then refitting the model.

? describes the choice between iterative influence analysis and non-iterative influence analysis.

3.9.2 Iterative vs Non-Iterative Influence Analysis

While the basic idea of influence analysis is straightforward, the implementation in mixed models can be tricky. For example, update formulas for the fixed effects are available only when the covariance parameters are assumed to be known. At most the profiled residual variance can be updated without refitting the model.

A measure of total influence requires updates of all model parameters, and the only way that this can be achieved in general is by removing the observations in question and refitting the model.

Because this **bruteforce** method involves iterative reestimation of the covariance parameters, it is termed ***iterative influence analysis***. Reliance on closed-form update formulas for the fixed effects without updating the (un-profiled) covariance parameters is termed a noniterative influence analysis.

An iterative analysis seems like a costly, computationally intensive enterprise. If

you compute iterative influence diagnostics for all n observations, then a total of $n+1$ mixed models are fit iteratively. This does not imply, of course, that the procedures execution time increases n-fold. Keep in mind that

- iterative reestimation always starts at the converged full-data estimates. If a data point is not influential, then its removal will have little effect on the objective function and parameter estimates. Within one or two iterations, the process should arrive at the reduced-data estimates.

- if complete reestimation does require many iterations, then this is important information in itself. The likelihood surface has probably changed drastically, and the reduced-data estimates are moving away

from the full-data estimates.

3.10 Two-tailed testing

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

3.11 One Tailed Testing

The approach proposed by Roy deals with the question of agreement, and indeed interchangeability, as developed by Bland and Altman's corpus of work. In the view of Dunn, a question relevant to many practitioners is which of the two methods is more precise.

The relationship between precision and the within-item and between-item variability must be established. Roy establishes the equivalence of repeatability and within-item variability, and hence precision. The method with the smaller within-item variability can be deemed to be the more precise.

3.12 Enabling One Tailed Testing

A useful approach is to compute the confidence intervals for the ratio of within-item standard deviations (equivalent to the ratio of repeatability coefficients), which can be interpreted in the usual manner (or alternatively, the ratio of the variances). In fact, the ratio of within-item standard deviations, with the attendant confidence interval, can be determined using a single `R` command: `intervals()`.

Pinheiro and Bates (pg 93-95) give a description of how confidence intervals for the variance components are computed. Furthermore a complete set of confidence

intervals can be computed to complement the variance component estimates. However , to facilitate one tailed testing, What is required is the computation of the variance ratios of within-item and between-item standard deviations.

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

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

3.13 Profile Likelihood

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

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

3.14 Implementation of PL Confidence Intervals

The suitable calculation of confidence limits for this variance ratio are to be computed using the profile likelihood approach. The `R` package `profilelikelihood` will

be assessed for feasibility, particularly the command `profilelikelihood.lme()`

3.15 Extension of techniques to LME Models

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

Beckman, Nachtsheim and Cook (1987) Beckman et al. (1987) applied the local influence method of Cook (1986) to the analysis of the linear mixed model.

While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known.

If the global measure suggests that the points in U are influential, the nature of that influence should be determined. In particular, the points in U can affect the following

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

3.15.1 Residuals diagnostics in mixed models

The marginal and conditional means in the linear mixed model are $E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ and $E[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{X\beta} + \boldsymbol{Zu}$, respectively.

A residual is the difference between an observed quantity and its estimated or predicted value. In the mixed model you can distinguish marginal residuals r_m and conditional residuals r_c.

3.15.2 Marginal and Conditional Residuals

A marginal residual is the difference between the observed data and the estimated (marginal) mean, $r_{mi} = y_i - x_0'\hat{b}$ A conditional residual is the difference between the observed data and the predicted value of the observation, $r_{ci} = y_i - x_i'\hat{b} - z_i'\hat{\gamma}$

In linear mixed effects models, diagnostic techniques may consider 'conditional' residuals. A conditional residual is the difference between an observed value y_i and the conditional predicted value $\hat{y}_i$.

$$\hat{epsilon}_i = y_i - \hat{y}_i = y_i - (X_i\hat{beta} + Z_i\hat{b}_i)$$

However, using conditional residuals for diagnostics presents difficulties, as they tend to be correlated and their variances may be different for different subgroups, which can lead to erroneous conclusions.

$$r_{mi} = x_i^T\hat{\beta} \tag{3.5}$$

3.15.3 Marginal Residuals

$$\begin{aligned}\hat{\beta} &= (X^TR^{-1}X)^{-1}X^TR^{-1}Y \\ &= BY\end{aligned}$$

3.16 Standardized and studentized residuals

To alleviate the problem caused by inconstant variance, the residuals are scaled (i.e. divided) by their standard deviations. This results in a 'standardized residual'. Because true standard deviations are frequently unknown, one can instead divide a residual by the estimated standard deviation to obtain the 'studentized residual.

3.16.1 Standardization

A random variable is said to be standardized if the difference from its mean is scaled by its standard deviation. The residuals above have mean zero but their variance is unknown, it depends on the true values of θ. Standardization is thus not possible in practice.

3.16.2 Studentization

Instead, you can compute studentized residuals by dividing a residual by an estimate of its standard deviation.

3.16.3 Internal and External Studentization

If that estimate is independent of the $i-$th observation, the process is termed 'external studentization'. This is usually accomplished by excluding the $i-$th observation when computing the estimate of its standard error. If the observation contributes to the standard error computation, the residual is said to be internally studentized.

Externally studentized residual require iterative influence analysis or a profiled residuals variance.

3.16.4 Computation

The computation of internally studentized residuals relies on the diagonal entries of $\boldsymbol{V}(\hat{\theta})$ - $\boldsymbol{Q}(\hat{\theta})$, where $\boldsymbol{Q}(\hat{\theta})$ is computed as

$$\boldsymbol{Q}(\hat{\theta}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{Q}(\hat{\theta})^{-1}\boldsymbol{X})\boldsymbol{X}^{-1}$$

3.16.5 Pearson Residual

Another possible scaled residual is the 'Pearson residual', whereby a residual is divided by the standard deviation of the dependent variable. The Pearson residual can be used when the variability of $\hat{\beta}$ is disregarded in the underlying assumptions.

3.17 Covariance Parameters

The unknown variance elements are referred to as the covariance parameters and collected in the vector θ.

3.17.1 Methods and Measures

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

? lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

3.18 Computation and Notation

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

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

3.19 Measures 2

3.19.1 Cook's Distance

- For variance components γ

Diagnostic tool for variance components

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

Chapter 4

Linear Mixed Effects Models

4.1 LME models in Method comparison

4.2 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';

$$\text{fixed effects}: 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \tag{4.1}$$

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

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

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

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;

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

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

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

$$(\text{p-values}: = 0.0446, < 0.0001, < 0.0001) \tag{4.2}$$

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.

4.3 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 rth replicate measurement on subject i with method m.

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

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

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

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m} \mu_i + c_{mi} \tag{4.4}$$

. Under the assumption that the μs are the true item values, this would be sufficient

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

Carstensen 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.

4.3.1 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{4.5}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{4.6}$$

Chapter 5

Appendices

5.1 The Hat Matrix

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{5.1}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{5.2}$$

The variances of Y and R can be expressed as:

$$\text{var}(Y) = H\sigma^2$$

$$\text{var}(R) = (I - H)\sigma^2 \tag{5.3}$$

Updating techniques allow an economic approach to recalculating the projection

matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

5.2 Sherman Morrison Woodbury Formula

The 'Sherman Morrison Woodbury' Formula is a well known result in linear algebra;

$$(A + a^T B)^{-1} \quad = \quad A^{-1} - A^{-1}a^T(I - bA^{-1}a^T)^{-1}bA^{-1} \tag{5.4}$$

This result is highly useful for analyzing regression diagnostics, and for matrices inverses in general. Consider a $p \times p$ matrix X, from which a row x_i^T is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{5.5}$$

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{5.6}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{5.7}$$

The variances of Y and R can be expressed as:

$$\text{var}(Y) = H\sigma^2$$

$$\text{var}(R) = (I - H)\sigma^2 \tag{5.8}$$

Updating techniques allow an economic approach to recalculating the projection matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

5.2.1 Hat Values for MCS regression

With A as the averages and D as the casewise differences.

```
fit = lm(D~A)
```

$$H = A\left(A^\top A\right)^{-1} A^\top,$$

Chapter 6

Model Diagnostics

6.1 Introduction

In classical linear models model diagnostics have been become a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers that model diagnostics do not often accompany LME model analyses.

6.1.1 Checking model assumptions

In classical linear regression, it is important to carry out model diagnostic techniques to determine whether or not the distributional assumptions are satisfied. Model diagnostics are also used to determine the influence of unusual observations.

Schabenberger (2004) describes the examination of model-data agreement as comprising several elements; residual analysis, goodness of fit, collinearity diagnostics and influence analysis.

6.1.2 Influence Diagnostics: Basic Idea and Statistics

The general idea of quantifying the influence of one or more observations relies on computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation. Influence statistics can be coarsely grouped by the aspect of estimation that is their primary target:

- overall measures compare changes in objective functions: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2)
- influence on parameter estimates: Cook's (Cook 1977, 1979), MDFFITS (Belsley, Kuh, and Welsch 1980, p. 32)
- influence on precision of estimates: CovRatio and CovTrace
- influence on fitted and predicted values: PRESS residual, PRESS statistic (Allen 1974), DFFITS (Belsley, Kuh, and Welsch 1980, p. 15)
- outlier properties: internally and externally studentized residuals, leverage

6.1.3 Introduction

The linear mixed effects model is a useful methodology for fitting a wide range of models. However, linear mixed effects models are known to be sensitive to outliers. Christensen et al. (1992) advises that identification of outliers is necessary before conclusions may be drawn from the fitted model.

Standard statistical packages concentrate on calculating and testing parameter estimates without considering the diagnostics of the model.

The assessment of the effects of perturbations in data, on the outcome of the analysis, is known as statistical influence analysis. Influence analysis examines the robustness

of the model.

Influence analysis methodologies have been used extensively in classical linear models, and provided the basis for methodologies for use with LME models.

Computationally inexpensive diagnostics tools have been developed to examine the issue of influence (Tewomir and Galpin, 2005). Studentized residuals, error contrast matrices and the inverse of the response variance covariance matrix are regular components of these tools.

6.2 Outline of Thesis

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

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

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

6.3 What is Influence

Broadly defined, "influence"is understood as the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model. The goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis (Schabenberger, 2004).

Influence is defined as the 'ability of a single or multiple data points, through their presence or absence

6.3.1 Quantifying Influence

The basic procedure for quantifying influence is simple as follows:

- Fit the model to the data and obtain estimates of all parameters.
- Remove one or more data points from the analysis and compute updated estimates of model parameters.
- Based on full- and reduced-data estimates, contrast quantities of interest to determine how the absence of the observations changes the analysis.

6.4 Residual diagnostics

For classical linear models, residual diagnostics are typically implemented as a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations.

In linear mixed effects models, diagnostic techniques may consider 'conditional' residuals. A conditional residual is the difference between an observed value y_i and the conditional predicted value $\hat{y}_i$.

$$\hat{epsilon}_i = y_i - \hat{y}_i = y_i - (X_i \hat{beta} + Z_i \hat{b}_i)$$

However, using conditional residuals for diagnostics presents difficulties, as they tend to be correlatedand their variances may be different for different subgroups, which can lead to erroneous conclusions.

6.4.1 Residuals

The computation of internally studentized residuals relies on the diagonal entries of $\boldsymbol{V}(\hat{\theta})$ - $\boldsymbol{Q}(\hat{\theta})$, where $\boldsymbol{Q}(\hat{\theta})$ is computed as

$$\boldsymbol{Q}(\hat{\theta}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{Q}(\hat{\theta})^{-1}\boldsymbol{X})\boldsymbol{X}^{-1}$$

Externally studentized residual require iterative influence analysis or a profiled residuals variance.

6.4.2 Residuals diagnostics in mixed models

A residual is the difference between an observed quantity and its estimated or predicted value. In the mixed model you can distinguish marginal residuals rm and conditional

residuals rc. A marginal residual is the difference between the observed data and the estimated (marginal) mean.

Cook (1986) introduces powerful tools for local-influence assessment and examining perturbations in the assumptions of a model. In particular the effect of local perturbations of parameters or observations are examined.

6.5 Extension of technique to LME Models

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

Beckman, Nachtsheim and Cook (1987) applied the local influence method of Cook (1986) to the analysis of the linear mixed model.

While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known.

If the global measure suggests that the points in U are influential, the nature of that influence should be determined. In particular, the points in U can affect

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

6.6 Marginal and Conditional Residuals

The marginal and conditional means in the linear mixed model are $E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ and $E[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{X\beta} + \boldsymbol{Zu}$, respectively.

A residual is the difference between an observed quantity and its estimated or predicted value. In the mixed model you can distinguish marginal residuals r_m and conditional residuals r_c. A marginal residual is the difference between the observed data and the estimated (marginal) mean, $r_{mi} = y_i - x_0'\hat{b}$ A conditional residual is the difference between the observed data and the predicted value of the observation, $r_{ci} = y_i - x_i'\hat{b} - z_i'\hat{\gamma}$

6.6.1 Marginal and Conditional Residuals

$$r_{mi} = x_i^T\hat{\beta} \tag{6.1}$$

6.6.2 Marginal Residuals

$$\begin{aligned} \hat{\beta} &= (X^T R^{-1} X)^{-1} X^T R^{-1} Y \\ &= BY \end{aligned}$$

6.7 Standardized and studentized residuals

To alleviate the problem caused by inconstant variance, the residuals are scaled (i.e. divided) by their standard deviations. This results in a 'standardized residual'. Because true standard deviations are frequently unknown, one can instead divide a residual by the estimated standard deviation to obtain the 'studentized residual.

Another possible scaled residual is the 'Pearson residual' whereby a residual is divided by the standard deviation of the dependent variable. The Pearson residual can be used when the variability of $\hat{\beta}$ is disregarded in the underlying assumptions.

6.7.1 Studentization

A random variable is said to be standardized if the difference from its mean is scaled by its standard deviation. The residuals above have mean zero but their variance is unknown, it depends on the true values of θ. Standardization is thus not possible in practice. Instead, you can compute studentized residuals by dividing a residual by an estimate of its standard deviation. If that estimate is independent of the ith observation, the process is termed external studentization. This is usually accomplished by excluding the $i-$th observation when computing the estimate of its standard error. If the observation contributes to the standard error computation, the residual is said to be ***internally studentized***.

6.8 Case Deletion Diagnostics

Christensen, Pearson and Johnson (1992) studied case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

6.8.1 Case Deletion Diagnostics

Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

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

6.9 Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{6.2}$$

6.9.1 Deletion Diagnostics

Since the pioneering work of Cook in 1977, deletion measures have been applied to many statistical models for identifying influential observations.

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models.

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may

identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

6.9.2 Case Deletion Diagnostics for Mixed Models

? notes the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect.

? develops these techniques in the context of REML

6.9.3 Methods and Measures

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

? lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

6.10 Influence analysis

Likelihood based estimation methods, such as ML and REML, are sensitive to unusual observations. Influence diagnostics are formal techniques that assess the influence of observations on parameter estimates for β and θ. A common technique is to refit the model with an observation or group of observations omitted.

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

6.10.1 Cook's 1986 paper on Local Influence

Cook 1986 introduced methods for local influence assessment. These methods provide a powerful tool for examining perturbations in the assumption of a model, particularly the effects of local perturbations of parameters of observations.

The local-influence approach to influence assessment is quitedifferent from the case deletion approach, comparisons are of interest.

6.10.2 Overall Influence

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

6.11 Likelihood Distance

The likelihood distance gives the amount by which the log-likelihood of the full data changes if one were to evaluate it at the reduced-data estimates. The important point is that $l(\psi_{(U)})$ is not the log-likelihood obtained by fitting the model to the reduced data set.

It is obtained by evaluating the likelihood function based on the full data set (containing all n observations) at the reduced-data estimates.

The likelihood distance is a global, summary measure, expressing the joint influence of the observations in the set U on all parameters in ψ that were subject to updating.

6.11.1 Likelihood Distance

The likelihood distance is a global, summary measure, expressing the joint influence of the observations in the set U on all parameters in ϕ that were subject to updating.

6.12 Haslett's Analysis

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

6.13 Iterative and non-iterative influence analysis

Schabenberger (2004) highlights some of the issue regarding implementing mixed model diagnostics.

A measure of total influence requires updates of all model parameters.

however, this doesnt increase the procedures execution time by the same degree.

6.13.1 Iterative Influence Analysis

For linear models, the implementation of influence analysis is straightforward. However, for LME models, the process is more complex. Update formulas for the fixed effects are available only when the covariance parameters are assumed to be known. A measure of total influence requires updates of all model parameters. This can only be achieved in general is by omitting observations, then refitting the model.

Schabenberger (2004) describes the choice between iterative influence analysis and non-iterative influence analysis.

6.14 Matrix Notation for Case Delection

6.14.1 Case deletion notation

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i, j)-$th element of $\boldsymbol{A}$.

6.14.2 Partitioning Matrices

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

6.15 CPJ's Three Propositions

Proposition 1

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \nu^{ii} & \lambda_i' \\ \lambda_i & \Lambda_{[i]} \end{bmatrix}$$

$$\boldsymbol{V}_{[i]}^{-1} = \boldsymbol{\Lambda}_{[i]} - \frac{\lambda_i \lambda_i'}{\lambda_i}$$

6.15.1 Proposition 2

(i) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{X}_{[i]} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$

(ii) $= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y})^{-1}$

(iii) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{Y}_{[i]} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$

6.15.2 Proposition 3

This proposition is similar to the formula for the one-step Newtown Raphson estimate of the logistic regression coefficients given by pregibon (1981) and discussed in Cook Weisberg.

6.16 Augmented GLMs

Generalized linear models are a generalization of classical linear models.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{6.3}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{6.4}$$

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{6.5}$$

$$y_a = T\delta + e^* \tag{6.6}$$

Weighted least squares equation

6.17 Covariance Parameters

The unknown variance elements are referred to as the covariance parameters and collected in the vector θ.

6.18 Terminology for Case Deletion diagnostics

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called 'observation-diagnostics'. For multiple observations, Preisser describes the diagnostics as 'cluster-deletion' diagnostics.

6.19 The CPJ Paper

6.19.1 Case-Deletion results for Variance components

?examines case deletion results for estimates of the variance components, proposing the use of one-step estimates of variance components for examining case influence. The method describes focuses on REML estimation, but can easily be adapted to ML or other methods.

Christensen developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted resgression problem (conditional on the estimated covariance matrix) for fixed effects. Lesaffre's approach accords with that proposed by Christensen et al when applied in a repeated measurement context, with a large sample size.

6.19.2 CPJ Notation

$$\boldsymbol{C} = \boldsymbol{H}^{-1} = \begin{bmatrix} c_{ii} & \boldsymbol{c}_i' \\ \boldsymbol{c}_i & \boldsymbol{C}_{[i]} \end{bmatrix}$$

? noted the following identity:

$$\boldsymbol{H}_{[i]}^{-1} = \boldsymbol{C}_{[i]} - \frac{1}{c_{ii}} \boldsymbol{c}_{[i]} \boldsymbol{c}_{[i]}'$$

? use the following as building blocks for case deletion statistics.

- $\breve{x}_i$
- $\breve{z}_i$
- $\breve{z}_i j$

- $\breve{y}_i$
- $p_i i$
- m_i

All of these terms are a function of a row (or column) of $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}_{[i]}$

Chapter 7

Roy2013

http://business.utsa.edu/wps/MSS/0017MSS-253-2013.pdf

Testing the Equality of Mean Vectors for Paired Doubly Multivariate Observations

Example 2. (Mineral Data): This data set is taken from Johnson and Wichern (2007, p. 43). An investigator measured the mineral content of bones (radius, humerus and ulna) by photon absorptiometry to examine whether dietary supplements would slow bone loss in 25 older women. Measurements were recorded for three bones on the dominant and nondominant sides. Thus, the data is doubly multivariate and clearly u = 2 and q = 3. The bone mineral contents for the rst 24 women one year after their participation in an experimental program is given in Johnson and Wichern (2007, p. 353).

Thus, for our analysis we take only rst 24 women in the rst data set. We test whether there has been a bone loss considering the data as doubly multivariate and has BCS structure. We rearrange the variables in the data set by grouping together the mineral content of the dominant sides of radius, humerus and ulna as the rst three variables, that is, the variables in the rst location (u = 1) and then the mineral contents for the non-dominant side of the same bones (u = 2)

7.1 Outlier Testing

A new outlier identification test for method comparison studies based on robust regression.

The identification of outliers in method comparison studies (MCS) is an important part of data analysis, as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature usually require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in MCS usually do not meet these assumptions. In this work, a new outlier test based on robust linear regression is proposed to overcome these special problems. The LORELIA (local reliability) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. The new test is compared to a standard test proposed in the literature by a Monte Carlo simulation. Its performance is illustrated in examples.

7.2 Lorelia

Method comparison studies are performed in order to prove equivalence between two measurement methods or instruments. The identification of outliers is an important part of data analysis as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in method comparison studies usually do not meet these assumptions. To overcome this problem, different data transformation methods are proposed in the literature. However, they will only be applicable if the random errors can be described by simple additive or multiplicative models. In this work, a new outlier test based on robust linear regression is proposed which provides a general solution to the above problem. The LORELIA

(LOcal RELIAbility) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. Outlier limits are estimated from the actual data situation without making assumptions on the underlying error variance model. The performance of the new test is demonstrated in examples and simulations.

7.3 Note on Roy's paper

1. Basic model:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, n$$
$$\boldsymbol{Z_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma^2 I})$$

   Assumptions are made about homoskedasticity.

2. General model:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, n$$
$$\boldsymbol{Z_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma^2 \Lambda})$$

   Assumptions about homoskedasticity are relaxed (Pinheiro and Bates, 1994, pg.202).

3. $\sigma^2\boldsymbol{\Lambda}$ is the general form for the VC structure for residuals.

4. The response vector $\boldsymbol{y}_i$ comprises the observations of the subject, as measured by two methods, taking three measurements each. Hence a 6×1 random vector corresponding to the ith subject.

$$\boldsymbol{y}_i = (y_i^{j1}, y_i^{Jj2}, y_i^{j3}, y_i^{s1}, y_i^{s2}, y_i^{s3})\prime \tag{7.1}$$

5. The number of replicates is p. A subject will have up to $2p$ measurements, for the two instrument case, i.e. $Max(n_i) = 2p$. (Let k denote number of instruments, which is assumed to be 2 unless stated otherwise.) For the blood pressure data $p = 3$.

Chapter 8

Residual Analysis

8.1 Introduction to Residual Analysis

Residual analysis is a widely used model validation technique. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect. that is to say, if the residuals behave randomly, with no discernible trend, the model has fitted the data well. If some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted.

Statistical software environments, such as the `R` Programming language, provides a suite of tests and graphical procedure sfor appraising a fitted linear model, with several of these procedures analysing the model residuals.

8.2 Framework for Model Validation using Residual Diagnostics

In statistical modelling, the process of model validation is a critical step, but also a step that is too often overlooked. A very simple procedure is to examine commonly encountered metrics, such as the R^2 value. However, using a small handful of simple measures and methods is insufficient to properly assess the quality of a fitted model. To do so properly, a full and comprehensive analysis that tests of all of the assumptions, as far as possible, must be carried out. A statistical model, whether of the fixed-effects or mixed-effects variety, represents how you think your data were generated. Following model specification and estimation, it is of interest to explore the model-data agreement by raising questions such as

- Does the model-data agreement support the model assumptions?
- Should model components be refined, and if so, which components? For example, should regressors be added or removed, and is the covariation of the observations modeled properly?
- Are the results sensitive to model and/or data? Are individual data points or groups of cases particularly influential on the analysis?

8.2.1 Residual Analysis

A residual is the difference between an observed quantity and its estimated or predicted value. Residual analysis is a widely used model validation technique. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect.

that is to say, if the residuals behave randomly, with no discernible trend, the model has fitted the data well. If some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted. Statistical software environments, such as the R Programming language, provides a suite of tests and graphical procedure sfor appraising a fitted linear model, with several of these procedures analysing the model residuals.

In classical linear models, an examination of model-data agreement has traditionally revolved around

The second part of the chapter looks at diagnostics techniques for LME models, firsly covering the theory, then proceeding to a discussion on implementing these using R code.

While a substantial body of work has been developed in this area, there is still areas worth exploring. In particular the development of graphical techniques pertinent to LME models should be looked at.

8.2.2 Outliers and Leverage

The question of whether or not a point should be considered an outlier must also be addressed. An outlier is an observation whose true value is unusual given its value on the predictor variables. The leverage of an observation is a further consideration. Leverage describes an observation with an extreme value on a predictor variable is a point with high leverage. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence can be thought of as the product of leverage and outlierness. An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. The R programming language has a variety of methods used to study each of the aspects for a linear model. While linear models and

GLMS can be studied with a wide range of well-established diagnostic technqiues, the choice of methodology is much more restricted for the case of LMEs.

For classical linear models, residual diagnostics are typically conducted using a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations.

8.2.3 Matrix Notation for Case Deletion

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i,j)-$th element of $\boldsymbol{A}$.

8.2.4 Extension of Diagnostic Methods to LME models

When similar notions of statistical influence are applied to mixed models, things are more complicated. Removing data points affects fixed effects and covariance parameter estimates. Update formulas for *leave-one-out* estimates typically fail to account for changes in covariance parameters.

? noted the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

? noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML. ¿¿¿¿¿¿¿ origin/master

Demidenko (2004) extends several regression diagnostic techniques commonly used in linear regression, such as leverage, infinitesimal influence, case deletion diagnostics, Cook's distance, and local influence to the linear mixed-effects model. In each case,

the proposed new measure has a direct interpretation in terms of the effects on a parameter of interest, and reduces to the familiar linear regression measure when there are no random effects.

The new measures that are proposed by Demidenko (2004) are explicitly defined functions and do not require re-estimation of the model, especially for cluster deletion diagnostics. The basis for both the cluster deletion diagnostics and Cook's distance is a generalization of Miller's simple update formula for case deletion for linear models. Furthermore Demidenko (2004) shows how Pregibon's infinitesimal case deletion diagnostics is adapted to the linear mixed-effects model.

Demidenko (2004) proposes two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

8.3 Model Validation Framework

In statistical modelling, the process of model validation is a critical step of model fitting process, but also a step that is too often overlooked. A very simple procedure is to examine commonly-used metrics, such as the R^2 value. However, using a small handful of simple measures and methods is insufficient to properly assess the quality of a fitted model. To do so properly, a full and comprehensive analysis that tests of all of the assumptions, as far as possible, must be carried out.

? describes the model validatin framework as comprised of the following tasks

- overall measures of goodness-of-fit
- the informal, graphical examination of estimates of model errors to assess the quality of distributional assumptions: residual analysis
- the quantitative assessment of the inter-relationship of model components; for example, collinearity diagnostics
- the qualitative and quantitative assessment of influence of cases on the analysis, i.e. influence analysis.

The sensitivity of a model is studied through measures that express its stability under perturbations. You are not interested in a model that is either overly stable or overly sensitive. Changes in the data or model components should produce commensurate changes in the model output. The difficulty is to determine when the changes are substantive enough to warrant further investigation, possibly leading to a reformulation of the model or changes in the data (such as dropping outliers). This paper is primarily concerned with stability of linear mixed models to perturbations of the data; that is, with influence analysis.

8.3.1 Residual Analysis

Residual analysis is a widely used model validation technique. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect. that is to say, if the residuals behave randomly, with no discernible trend, the model has fitted the data well. If some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted.

Statistical software environments, such as the `R` Programming language, provides a suite of tests and graphical procedure sfor appraising a fitted linear model, with several of these procedures analysing the model residuals.

8.3.2 Outliers and Leverage

The question of whether or not a point should be considered an outlier must also be addressed. An outlier is an observation whose true value is unusual given its value on the predictor variables. The leverage of an observation is a further consideration. Leverage describes an observation with an extreme value on a predictor variable is a point with high leverage. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence can be thought of as the product of leverage and outlierness. An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. The `R` programming language has a variety of methods used to study each of the aspects for a linear model. While linear models and GLMS can be studied with a wide range of well-established diagnostic technqiues, the choice of methodology is much more restricted for the case of LMEs.

8.4 Regression Of Differences On Averages

Further to Carstensen, we can formulate the two measurements y_1 and y_2 as follows:

$y_1 = \alpha + \beta\mu + \epsilon_1$

$y_2 = \alpha + \beta\mu + \epsilon_2$

8.4.1 Note 1: Coefficient of Repeatability

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

8.4.2 Note 2: Carstensen model in the single measurement case

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{8.1}$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$.

For the replicate case, an interaction term c is added to the model, with an associated variance component.

8.4.3 Note 3: Model terms

It is important to note the following characteristics of this model.

- Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.
- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.
- $\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.
- $\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.
- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.
- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.
- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.
- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

Roys uses and LME model approach to provide a set of formal tests for method comparison studies.

Four candidates models are fitted to the data.

These models are similar to one another, but for the imposition of equality constraints.

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

Roy's model uses fixed effects $\beta_0 + \beta_1$ and $\beta_0 + \beta_1$ to specify the mean of all observationsby
methods 1 and 2 respectuively.

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

Allocation of fixed effects and random effects are very different in each model

Carstensen's interest lies in the difference between the population from which they were drawn.

Carstensen's model is a mixed effects ANOVA.

$$Y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad c_{mi} \sim \tau^{\in}_{\Updownarrow}, \qquad e_{mir} \sim \sigma^{\in}_{\Updownarrow},$$

This model includes a method by item iteration term.

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen. Carstensen makes some interesting remarks in this regard.

> The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

8.4.4 Note 3: Model terms

It is important to note the following characteristics of this model.

- Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.
- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.
- $\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.
- $\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.
- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.
- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.
- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.
- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

8.5 Regression Of Differences On Averages

Further to Carstensen, we can formulate the two measurements y_1 and y_2 as follows:

$y_1 = \alpha + \beta\mu + \epsilon_1$

$y_2 = \alpha + \beta\mu + \epsilon_2$

For classical linear models, residual diagnostics are typically conducted using a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations.

Appendix to Section 4

As an appendix to section 4, an appraisal of the current state of development (or lack thereof) for current implemenations for LME models, particularly for `nlme` and `lme4` fitted models.

Crucially, a review of internet resources indicates that almost all of the progress in this regard has been done for `lme4` fitted models, specifically the *Influence.ME* `R` package. (Nieuwenhuis et 2012)

Conversely there is very little for `nlme` models. To delve into this mor, one would immediately investigate the current development workflow for both packages.

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

Important Consideration for MCS

The key issue is that `nlme` allows for the particular specification of Roy's Model, specifically direct spefiication of the VC matrices for within subject and between subject

residuals. The `lme4` package does not allow for this. To advance the ideas that eminate from Roys' paper, one is required to use the `nlme` context. However, to take advantage of the infrastructure already provided for `lme4` models, one may change the research question away from that of Roy's paper. To this end, an exploration of what textitinfluence.ME can accomplished is merited. As a complement to this, one can also consider how to properly employ the R^2 measure, in the context of Methoc Comparison Studies, further to the work by Edwards et al, namely "An R^2 statistic for fixed effects in the linear mixed model".

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

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

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

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

Furthermore, the R2 statistic leads immediately to a natural definition of a partial R2 statistic. A mixed model in which ethnicity gives a

> very small p-value as a longitudinal predictor of blood pressure (BP) compellingly illustrates the value of the statistic.
>
> In sharp contrast to the extreme p-value, a very small R^2 , a measure of statistical and scientific importance, indicates that ethnicity has an almost negligible association with the repeated BP outcomes for the study.

The `nlme` package

With regards to `nlme`, the torch has been passed to Galecki Galecki & Burzykowski (UMich. and Hasselt respecitely). Galecki & Burzykowski published *Linear Mixed Effects Models using* `R`. Also, the accompanying `R` package, nlmeU package is under current development, with a version being released XXXX.

The `lme4` package

The `lme4` package is also under active development, under the leadership of Ben Bolker (McMaster University). According to CRAN, the LME4 package, fits linear and generalized linear mixed-effects models

> The models and their components are represented using S4 classes and methods. The core computational algorithms are implemented using the Eigen C++ library for numerical linear algebra and RcppEigen "glue". (CRAN)

The key issue is that `nlme` allows for the particular specification of Roy's Model, speciifically direct spefiication of the VC matrices for within subject and between subject residuals. The `lme4` package does not allow for this. To advance the ideas that

eminate from Roys' paper, one is required to use the `nlme` context. However, to take advantage of the infrastructure already provided for `lme4` models, one may change the research question away from that of Roy's paper. To this end, an exploration of what textitinfluence.ME can accomplished is merited. As a complement to this, one can also consider how to properly employ the R^2 measure, in the context of Methoc Comparison Studies, further to the work by Edwards et al, namely "An R^2 statistic for fixed effects in the linear mixed model".

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

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

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

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

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

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

$$r_{mi} = x_i^T \hat{\beta} \tag{8.2}$$

8.5.1 Marginal Residuals

$$\begin{aligned} \hat{\beta} &= (X^T R^{-1} X)^{-1} X^T R^{-1} Y \\ &= BY \end{aligned}$$

8.6 Covariance Parameters

The unknown variance elements are referred to as the covariance parameters and collected in the vector θ.

8.6.1 Methods and Measures

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

? lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

8.7 Missing Data in Method Comparison Studies

The matter of missing data has not been commonly encountered in either Method Comparison Studies or Linear Mixed Effects Modelling. However Roy (2009) deals with the relevant assumptions regrading missing data.

Galecki & Burzykowski (2013) tackles the subject of missing data in LME Modelling.

Furthermore the nlmeU package includes the `patMiss` function, which "allows to compactly present pattern of missing data in a given vector/matrix/data frame or combination of thereof".

8.8 Leave-One-Out Diagnostics with `lmeU`

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

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

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

- `R` command and `R` object - Typewriter Font
- `R` Package name - Italics
- Selected Acronyms and Proper Nouns - Italics

- This chapter is broken into two parts. The first part is a review of diagnostics

methods for linear models, intended to acquaint the reader with the subject, and also to provide a basis for material covered in the second part. Particular attention is drawn to graphical methods.

- The second part of the chapter looks at diagnostics techniques for LME models, firsly covering the theory, then proceeding to a discussion on implementing these using `R` code.

- While a substantial body of work has been developed in this area, ther are still area worth exploring. In particular the development of graphical techniques pertinent to LME models should be looked at.

8.9 Introduction

Chapter 9

Model Diagnostics

Abstract

This chapter is broken into two parts. The first part is a review of diagnostics methods for linear models, intended to acquaint the reader with the subject, and also to provide a basis for material covered in the second part. Particular attention is drawn to graphical methods.

9.1 Case Deletion Diagnostics for LME models

Christensen (19XX) describes three propositions that are required for efficient case-deletion in LME models. The first proposition decribes how to efficiently update V when the ith element is deleted.

$$V_{[i]}^{-1} = \Lambda_{[i]} - \frac{\lambda\lambda\prime}{\nu ii} \tag{9.1}$$

The second of Christensen's propostions is the following set of equations, which are

variants of the Sherman Wood bury updating formula.

$$X'_{[i]}V^{-1}_{[i]}X_{[i]} = X'V^{-1}X - \frac{\hat{x}_i\hat{x}'_i}{s_i} \tag{9.2}$$

$$(X'_{[i]}V^{-1}_{[i]}X_{[i]})^{-1} = (X'V^{-1}X)^{-1} + \frac{(X'V^{-1}X)^{-1}\hat{x}_i\hat{x}'_i(X'V^{-1}X)^{-1}}{s_i - \bar{h}_i} \tag{9.3}$$

$$X'_{[i]}V^{-1}_{[i]}Y_{[i]} = X\prime V^{-1}Y - \frac{\hat{x}_i\hat{y}'_i}{s_i} \tag{9.4}$$

In LME models, fitted by either ML or REML, an important overall influence measure is the likelihood distance (**?**). The procedure requires the calculation of the full data estimates $\hat{\psi}$ and estimates based on the reduced data set $\hat{\psi}_{(U)}$. The likelihood distance is given by determining

$$LD_{(U)} = 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\} \tag{9.5}$$

$$RLD_{(U)} = 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\} \tag{9.6}$$

9.1.1 Residual Diagnostics in LME models

- A **residual** is the difference between the observed quantity and the predicted value. In LME models a distinction is made between marginal residuals and conditional residuals.

- A **Marginal residual** is the difference between the observed data and the estimated marginal mean (Schabenberger pg3) The computation of case deletion diagnostics in the classical model is made simple by the fact that important estimates can be computed without refitting the model.

- Such update formulae are available in the mixed model only if you assume that the covariance parameters are not affect by the removal of the observation in question. Schabenberger remarks that this is not a reasonable assumption.

Basic procedure for quantifying influence is simple

1. Fit the model to the data

2. Remove one or more data points from the analysis and compute updated estimates of model parameters

3. Based on the full and reduced data estimates, contrast quantities of interest to determine how the absence of the observations changed the analysis.

The likelihood distance is a global summary measure expressing the joint influence of the observations in the set U on all parameters in Ψ that were subject to updating.

9.2 Case Deletion Diagnostics for LME models

Haslett and Dillane (2004) remark that linear mixed effects models didn't experience a corresponding growth in the use of deletion diagnostics, adding that McCullough and Searle (2001) makes no mention of diagnostics whatsoever.

? describes three propositions that are required for efficient case-deletion in LME models. The first proposition decribes how to efficiently update V when the ith element is deleted.

$$V_{[i]}^{-1} = \Lambda_{[i]} - \frac{\lambda\lambda\prime}{\nu ii} \tag{9.7}$$

The second of christensen's propostions is the following set of equations, which are variants of the Sherman Wood bury updating formula.

$$X'_{[i]}V_{[i]}^{-1}X_{[i]} = X'V^{-1}X - \frac{\hat{x}_i\hat{x}'_i}{s_i} \tag{9.8}$$

$$(X'_{[i]}V_{[i]}^{-1}X_{[i]})^{-1} = (X'V^{-1}X)^{-1} + \frac{(X'V^{-1}X)^{-1}\hat{x}_i\hat{x}'_i(X'V^{-1}X)^{-1}}{s_i - \bar{h}_i} \tag{9.9}$$

$$X'_{[i]}V_{[i]}^{-1}Y_{[i]} = X\prime V^{-1}Y - \frac{\hat{x}_i\hat{y}'_i}{s_i} \tag{9.10}$$

In LME models, fitted by either ML or REML, an important overall influence measure is the likelihood distance (**?**). The procedure requires the calculation of the full data estimates $\hat{\psi}$ and estimates based on the reduced data set $\hat{\psi}_{(U)}$. The likelihood distance is given by determining

$$LD_{(U)} = 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\} \tag{9.11}$$

$$RLD_{(U)} = 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\} \tag{9.12}$$

9.3 Carstensen

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{9.13}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The β terms can be gathered together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \beta_1 = \beta_2$ and $\mathrm{H}_2\colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3\colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009) proposes a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Let $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall

variability of method m. Roy also integrates H_2 and H_3 into a single testable hypothesis $H_4 : \omega_1^2 = \omega_2^2$. CONCERNS?

? demonstrates how to implement a method comparison study further to model (1) using the SAS proc mixed package. Carstensen et al. (2008) demonstrates how to construct limits of agreement using SAS, STATA and R. In the case of SAS, the PROC MIXED procedure is used. Implementation in R is performed using the nlme package (**?**).

Carstensen et al. (2008) remarks that the implementation using R is quite "arcane".

As R is freely available, this paper demonstrates an implementation of Roy's model using R.

The R statistical software package is freely available.

The LME model is very easy to implement using PROC MIXED of SAS and the results are also easy to interpret. The SAS proc mixed procedure has very simple syntax.

As the required code to fit the models is complex, R code necessary to fit the models is provided.

A demonstration is provided on how to use the output to perform the tests, and to compute limits of agreement.

We assume the data are formatted as a dataset with four columns named:

meth, method of measurement, the number of methods being M, item, items (persons, samples) measured by each method, of which there are I, repl, replicate indicating repeated measurement of the same item by the same method, and y, the measurement.

9.3.1 Remarks on the Multivariate Normal Distribution

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

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

or to make it explicitly known that X is k-dimensional,

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

with k-dimensional mean vector

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

and $k \times k$ covariance matrix

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

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

   (a)

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

(b)

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

9.4 Modelling Agreement with LME Models

Roys uses and LME model approach to provide a set of formal tests for method comparison studies.

Four candidates models are fitted to the data.

These models are similar to one another, but for the imposition of equality constraints.

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

Roy's model uses fixed effects $\beta_0 + \beta_1$ and $\beta_0 + \beta_1$ to specify the mean of all observationsby
methods 1 and 2 respectuively.

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

Allocation of fixed effects and random effects are very different in each model

Carstensen's interest lies in the difference between the population from which they were drawn.

Carstensen's model is a mixed effects ANOVA.

$$Y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad c_{mi} \sim \tau^{\in}_{\Updownarrow}, \qquad e_{mir} \sim \sigma^{\in}_{\Updownarrow},$$

This model includes a method by item iteration term.

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen. Carstensen makes some interesting remarks in this regard.

> The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

It is also desirable to measure the influence of the case deletions on the covariance matrix of $\hat{\beta}$.

- *The previous Section (Section 4) is a literary review of residual diagnostics and influence procedures for Linear Mixed Effects Models, drawing heavily on Schabenberger and Zewotir.*
- *Section 4 begins with an introduction to key topics in residual diagnostics, such as influence, leverage, outliers and Cook's distance. Other concepts such as DFFITS and DFBETAs will be introduced briefly, mostly to explain why the are*

not particularly useful for the Method Comparison context, and therefore are not elaborated upon.

- *In brief, Variable Selection is not applicable to Method Comparison Studies, in the commonly used used context. Testing a rather simplisticy specificied model against one with more random effects terms is tractable, but this research question is of secondary importance.*

9.4.1 Matrix Notation for Case Deletion

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i,j)-$th element of $\boldsymbol{A}$.

9.5 Extension of Diagnostic Methods to LME models

When similar notions of statistical influence are applied to mixed models, things are more complicated. Removing data points affects fixed effects and covariance parameter estimates. Update formulas for *leave-one-out* estimates typically fail to account for changes in covariance parameters.

? noted the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

? noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

Demidenko (2004) extends several regression diagnostic techniques commonly used in linear regression, such as leverage, infinitesimal influence, case deletion diagnostics, Cook's distance, and local influence to the linear mixed-effects model. In each case, the proposed new measure has a direct interpretation in terms of the effects on a parameter of interest, and reduces to the familiar linear regression measure when there are no random effects.

The new measures that are proposed by Demidenko (2004) are explicitly defined functions and do not require re-estimation of the model, especially for cluster deletion diagnostics. The basis for both the cluster deletion diagnostics and Cook's distance is a generalization of Miller's simple update formula for case deletion for linear models. Furthermore Demidenko (2004) shows how Pregibon's infinitesimal case deletion diagnostics is adapted to the linear mixed-effects model.

Demidenko (2004) proposes two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

9.5.1 Remarks on the Multivariate Normal Distribution

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

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

or to make it explicitly known that X is k-dimensional,

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

with k-dimensional mean vector

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

and $k \times k$ covariance matrix

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

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

   (a)

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

   (b)

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

9.6 Repeated measurements in LME models

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

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

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

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

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

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

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

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.

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

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

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).

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

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

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

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

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

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.

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

Haslett, J. and D. Dillane (2004). Application of 'delete = replace' to deletion diagnostics for variance component estimation. *Journal of the Royal Statistical Society (Series B) 66*, 131–143.

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

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.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

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

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

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.

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

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

Roy, A. (2009). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

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

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

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

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

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.