# SCRATCH

Kevin O'Brien

January 13, 2017

# Contents

(a) What is Agreement?

(b) Repeatability

(c)

(d)

(e)

2. Bland Altman Single Observations

(a)

(b)

(c)

(d)

(e)

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 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
  -
  -
- Introduction to LME Models
  -
  -
  -
- Roy's Hypothesis Tests
  -

  - Bendix Carstensens's Approach
  -

- Simulation Studies
  - Reconstruction of Blood Data
  -
  -

- Profile Likelihood
  - Douglas Bates Comments on Interval Estimation
  -
  -

## 0.2 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.

### 0.2.1 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.

# Influence Analysis

The basic rationale behind measuring influential cases is that when iteratively single units are omitted from the data, models based on these data should not produce substantially different estimates.

## Well Known Influence Measures

*"Regression Diagnostics: Identifying Influential Data and Source of Collinearity (1980)"* by Belsley,Kuh,& Welsch is a landmark text in the field of residual diagnostics, and provides a foundation for much of the subsequent work.

**Cook's Distance** Cooks Distance is a measure indicating to what extent model parameters are influenced by (a set of) influential data on which the model is based.

**DFBETAS** DFBETAS (standardized difference of the beta) is a measure that standardizes the absolute difference in parameter estimates between a (mixed effects) regression model based on a full set of data, and a model from which a (potentially influential) subset of data is removed. A value for DFBETAS is calculated for each parameter in the model separately.

# Influence Measures with *influence.ME*

influence.ME calculates measures of influence for mixed effects models estimated with the lme4 R package. The basic rationale behind measuring influential cases is that when iteratively single units are omitted

from the data, models based on these data should not produce substantially different estimates.

Calculating measures of influential data for an LME model requires the re-estimation of this model for each set of potentially influential data separately. The `estex(` function does this, and returns the altered estimates resulting from each re-estimation.

The main function in the influence.ME package is the `influence()`.

> Based on a priorly estimated mixed effects regression model (estimated using lme4), the `influence()` function iteratively modifies the mixed effects model to neutralize the effect a grouped set of data has on the parameters, and which returns returns the fixed parameters of these iteratively modified models. These are used to compute measures of influential data. ()

## Using the influence.ME package

Influence Analysis can onnly be carried out with LME models fitted using the functions in the **lme4** package. Such models are known as `mer` objects. Hence the `estex()` function only works on LME models of class `mer`. The package developers advise that it is required that the `mer` model was estimated using a factor variable to indicate group levels. When using something similar to `+ (1 | as.factor(variable))`, the function is not able of identifying the correct grouping factors, and returns an error.

Executing this procedure can be computationally highly demanding, because `estex()` entails the re-estimation of the provided mixed effects model for each level of the spec-

ified grouping factor (after alteration of the data).

## Functionality of the influence.ME pacakge

To standardize the assessment of how influential an observation (or group of observations)is, several measures of influence are used by influence.ME.

- DFBETAS is a standardized measure of the absolute difference between the estimate with a particular case included and the estimate without that particular case.
- Cooks distance provides an overall measurement of the change in all parameter estimates, or a selection thereof.

The `estex()` command computes revised estimates can subsequently be entered to the `cooks.distance` and `dfbetas` commands, to calculate Cooks Distance and the DFBETAS (standardized difference of the beta) measures.

### The `pchange` command

The `pchange` command computes the percentile change, as a measure of influential data. This unstandardized measure can serve to help interpret the magnitude of the influence single or combined grouping levels exert on mixed effects models.

The percentage change in parameter estimates between an LME model based on a full set of data, and a model from which a (potentially influential) subset of data is removed. A value of percentage change is calculated for each parameter in the model separately, based on the information returned by the `estex()` function.

### sigtest

The `sigTest` function can test for changes in the level of statistical significance resulting from the deletion of potentially influential observations

## 0.2.2 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.

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.

## 0.3 Measures 2

### 0.3.1 Cook's Distance

- For variance components $\gamma$

Diagnostic tool for variance components

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

### 0.3.2 Variance Ratio

- For fixed effect parameters $\beta$.

### 0.3.3 Variance Ratio

- For fixed effect parameters $\beta$.

### 0.3.4 Cook-Weisberg statistic

- For fixed effect parameters $\beta$.

### 0.3.5 Andrews-Pregibon statistic

- For fixed effect parameters $\beta$.

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

## 0.4 Measures 2

### 0.4.1 Cook's Distance

- For variance components $\gamma$

Diagnostic tool for variance components

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

## 0.5 Zewotir Measures of Influence in LME Models

Zewotir 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

**Random Effects**

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

**linear functions**

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

### 0.5.1 Information Ratio

## 0.6 Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components $\sigma_j^2$, 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 $u\boldsymbol{I}$

## 0.7 Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

Influence arises at two stages of the LME model. Firstly when $V$ is estimated by $\hat{V}$, and subsequent estimations of the fixed and random regression coefficients $\beta$ and $u$, given $\hat{V}$.

### 0.7.1 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. It is closely related to the studentized residual.

$$DFFITS = \frac{\widehat{y_i} - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}}$$

### 0.7.2 Influence Statistics for LME models

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

## 0.8 Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

Influence arises at two stages of the LME model. Firstly when $V$ is estimated by $\hat{V}$, and subsequent estimations of the fixed and random regression coefficients $\beta$ and $u$, given $\hat{V}$.

### 0.8.1 DFBETA

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \quad (1)$$

$$= B(Y - Y_{\bar{a}} \quad (2)$$

### 0.8.2 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. It is closely related to the studentized residual.

$$DFFITS = \frac{\widehat{y_i} - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}}$$

## 0.9 Likelihood Distance

The likelihood distance is a global summary measure that expresses the joint influence of the subsets of observations, $U$, on all parameters in $\phi$ that were subject to updating. **?** points out the likelihood distance gives the amount by which the log-likelihood of the model fitted from the full data changes if one were to estimate the model from a reduced-data estimates. Importantly $LD(\psi_{(U)})$ 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.

$$LD((\boldsymbol{U})) = 2[l(\hat{\boldsymbol{\phi}}) - l\hat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}]$$

$$RLD((\boldsymbol{U})) = 2[l_R(\hat{\boldsymbol{\phi}}) - l_R(\hat{\boldsymbol{\phi}})_{\boldsymbol{\omega}}]$$

## 0.10 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 $\psi$ that were subject to updating.

### 0.10.1 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 $\phi$ that were subject to updating.

Krouwer and Monti (29) presented a graphical method for evaluation of laboratory assays (a mountain plot). They computed the percentile for each ranked difference between the two methods, and by turning at the 50th percentile produced a histogram-like function (the mountain). This method is relevant for detecting large infrequent errors (differences) but lacks the aspect of concentration relationship. These investigators, therefore, recommend use of their plot together with difference plots. Introduction of analytical quality specifications in the mountain plots may be useful in method evaluations.

Davis (1989) proposes the TAM model, which suggests an hypothesis as to why users may adopt particular technologies, and not others. According to this theory, when users are presented with a new technology, two important factors will influence their decision about how and when they will adopt it.

**Perceived usefulness (PU)** - This was defined by Fred Davis as "the degree to which a person believes that using a particular system would enhance his or her job performance".

**Perceived ease-of-use (PEOU)** - Davis defined this as "the degree to which a person believes that using a particular system would be free from effort"

Davis's explanations of these term can be rephrased for application to statistical analysis. Perceived Use could refer to the degree to which an user would deem a particular statistical method would properly establish the results of an analaysis. In the case of method comparison studies, proper indication of agreement, or lack thereof.

Perceived ease-of-use requires only applying the context of a

A very modest statistical skill set is the only prerequistive for constructing a Bland-Altman plot, and computing limits of agreement. The main building blocks are simple descriptive, statistics and a knowledge of the normal distribution. These are topics that feature in almost every undergraduate statistics courses.

In short, the user perceives the Bland-Altman methodology to be an easy-to-implement technique, that will properly address the question of agreement.

Conversely the Survival plot is a derivative of the Kaplan-Meier Curve, a non-parametric graphical technique that features in Survival Analysis. This subject area is a well known domain of statistics, but would be encountered on curriculums of specialist courses. The Mountain Plot is formally called the empirical folder cumulative distribution plot. Currently there is only one software implementation , medcalc.be toolkot (FIX)

The ROC curve is a plot that is commonly used in the appraisal of a statistical analytics systems. Interpretation of the plot, the nearer the curve is to the top left corner of the plot, the better the statistical method is at making predicting outcomes.

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.

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

### 0.10.2 Limits of agreement for Carstensen's data

Carstensen et al. (2008) describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

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

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

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 $(-12.18, 17.12)$.

## 0.11 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`

## 0.12 Method Comparison Stduies with `R`

### 0.12.1 Accuracy and Precision

An important consideration in discussing methods of measurement are the issues of accuracy and precision.

### 0.12.2 What is Agreement

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. (Bland and Altman 1999)

### 0.12.3 Inappropriate Techniques for MCS

### 0.12.4 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.

### 0.12.5 PRESS

The prediction residual sum of squares (PRESS) is an value associated with this calculation. When fitting linear models, PRESS can be used as a criterion for model

selection, with smaller values indicating better model fits.

$$PRESS = \sum (y - y^{(k)})^2 \tag{3}$$

- $e_{-Q} = y_Q - x_Q \hat{\beta}^{-Q}$
- $PRESS_{(U)} = y_i - x\hat{\beta}_{(U)}$

### 0.12.6 DFBETA

$$\begin{aligned} DFBETA_a &= \hat{\beta} - \hat{\beta}_{(a)} && (4) \\ &= B(Y - Y_{\bar{a}} && (5) \end{aligned}$$

### 0.12.7 Influential Observations : DFBeta and DFBetas

Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. dfbeta refers to how much a parameter estimate changes if the observation in question is dropped from the data set. Note that with k covariates, there will be k+1 dfbetas (the intercept,$\beta_0$, and 1 $\beta$ for each covariate). Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

## 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.

## 0.13 Bland Altman Methodology

### 0.13.1 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* [3] and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

## 0.14 The Bland Altman Plot

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

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

$$D_i = X_i - Y_i \tag{6}$$

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

### 0.14.1 Criticism of Bland Altman Plot

Unfortunately the Bland-Altman plot has a fatal flaw: it indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one

has been calibrated against the other. (Hopkins)

### 0.14.2 Treatment of Outliers

Bland and Altman attend to the issue of outliers in their 1986 paper, wherein they present a data set with an extreme outlier

## 0.15 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.

## 0.16 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.

### 0.16.1 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.

## 0.17 Bland Altman Plots In Literature

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

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

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

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

### 0.17.1 Gold Standard

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

## 0.18 Discussion on Method Comparison Studies

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

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

Indications on how to deal with outliers in Bland Altman plots

We wish to determine how outliers should be treated in a Bland Altman Plot

In their 1983 paper they merely state that the plot can be used to 'spot outliers'.

In their 1986 paper, Bland and Altman give an example of an outlier. They state that it could be omitted in practice, but make no further comments on the matter.

In Bland and Altmans 1999 paper, we get the clearest indication of what Bland and Altman suggest on how to react to the presence of outliers. Their recommendation is to recalculate the limits without them, in order to test the difference with the calculation where outliers are retained.

The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large reduction.

However, they do not recommend removing outliers. Furthermore, they say:

We usually find that this method of analysis is not too sensitive to one or two large outlying differences.

We ask if this would be so in all cases. Given that the limits of agreement may or may not be disregarded, depending on their perceived suitability, we examine whether

it would possible that the deletion of an outlier may lead to a calculation of limits of agreement that are usable in all cases?

Should an Outlying Observation be omitted from a data set? In general, this is not considered prudent.

Also, it may be required that the outliers are worthy of particular attention themselves.

Classifying outliers and recalculating We opted to examine this matter in more detail.

The following points have to be considered

how to suitably identify an outlier (in a generalized sense)

Would a recalculation of the limits of agreement generally results in a compacted range between the upper and lower limits of agreement?

### 0.18.1 Agreement

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

### 0.18.2 Lack Of Agreement

1. Constant Bias

2. Proportional Bias

#### Constant Bias

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

**Proportional Bias**

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

## 0.19 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.
Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

### 0.19.1 Bland Altman plots using 'Gold Standard' raters

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

### 0.19.2 Bias Detection

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

# Chapter 1

# The Bland Altman Plot

## 1.1 Bland Altman Plots

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

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

The first step recommended which the authors argue should be mandatory is construction of a simple scatter plot of the data. The line of equality ($X = Y$) should also

be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1). The averages of the two measurements is considered by Bland and Altman to the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2). 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."

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

Table 1.1: Fotobalk and Counter Methods: Differences and Averages

### 1.1.1 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this.

Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

### 1.1.2 Criticism of Bland Altman Plot

Hopkins[8] argues that the plot indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one has been calibrated against the other.

An Evaluation of the correlation between the difference and means complement the analysis.

Bland and Altman caution that the calculations are based on the assumption that the data is normally distributed. This can be verified by using a histogram. If Data is not normally distributed, it can be transformed.

### 1.1.3 Introduction

- Comparing two methods of measurement is normally done by computing limits of agreement (LoA), i.e. prediction limits for a future difference between measurements with the two methods. When the difference is not constant it is not clear what this means, since the difference between the methods depends on the average; hence, unlike the case where the difference is constant, LoA cannot directly be translated into a prediction interval for a measurement by one method given that of another.

- The main point in the paper by Bland and Altman [1] is however different from the outlook in this paper; Bland and Altman mainly discuss whether two methods of measurement can be used interchangeably and how to assess this with the help of proper statistical methods to derive LoA, i.e. prediction limits for differences between two methods. This paper takes as starting point that the classical LoA can be converted to a prediction interval for one method given a measurement by the other (details in the next section). This sort of relationship can be shown

in a plot as a line with slope 1 and prediction limits as lines also with slope 1; applicable for the prediction both from method 1 to method 2 and vice versa. In the case of non-constant difference it would be desirable to be able to produce a similar plot, usable both ways. Thus, the aim of this paper is to produce a conversion from one method to another that also applies in the case where the difference between methods is not constant.

- In this paper, I set up a proper model for data for method comparison studies which in the case of constant difference between methods leads to the classical LoA, and in the case of linear bias gives a simple formula for the prediction. The paper only addresses the situation where only one measurement by each method is available, although replicate measurements by each method are desirable whenever possible [2]. Moreover, the situation with non-constant variance over the range of measurements is not covered either.

### 1.1.4 Discussion

I have here proposed a simple twist to the results from regression of the differences on the sums in the case of a linear relationship between two methods of measurement. It is consistent with the obvious underlying model, and exploits the fact that although the parameters of the model cannot be estimated, those functions of the parameters that are needed for creating predictions can be estimated. The prediction limits provided have the attractive property that if the prediction line with limits is drawn in a coordinate system, the chart will apply in both ways; hence, both the line and the limits are symmetric. Precisely as the prediction intervals derived from the classical LoA are in the case where the difference between methods is constant. The drawback is that the regression of the differences on the means ignores that the averages are correlated with the residuals (i.e. the error terms), and therefore gives biased estimates if the slope linking the two methods is far from 1 or the residual variances are very different. However, both of these are rather uncommon in method comparison studies, so the method proposed here is widely applicable. When considering LoA, the only feasible transformation is the log-transform, which gives LoA for the ratio of measurements, which is immediately understandable. If, for example, the measurements are fractions where some are close to either 0 or 1 a logit transform may be adequate.

LoA would then be for (log) odds-ratios, not very easily understood. For other more arbitrarily chosen transformation the situation may be even worse. But if a plot with conversion lines and limits are constructed, then the plot is readily back-transformed to the original scale for practical use.

# Chapter 2

# REGRESSION

## 2.1 Model II Regression

### 2.1.1 Simple Linear Regression

Simple Linear Regression is well known statistical technique , wherein estimates for slope and intercept of the line of best fit are derived according to the Ordinary Least Square (OLS) principle.This method is known to Cornbleet and Cochrane as Model I regression.

In Model I regression, the independent variable is assumed to be measured without error. For method comparison studies, both sets of measurement must be assumed to be measured with imprecision and neither case can be taken to be a reference method. Arbitrarily selecting either method as the reference will yield two conflicting outcomes. A fitting based on '$X$ on $Y$' will give inconsistent results with a fitting based on '$Y$ on $X$'. Consequently model I regression is inappropriate for such cases.

Conversely, Cornbleet Cochrane state that when the independent variable $X$ is a pre-

cisely measured reference method, Model I regression may be considered suitable. They qualify this statement by referring the $X$ as *the 'correct' value*, tacitly implying that there must still be some measurement error present. The validity of this approach has been disputed elsewhere.

### 2.1.2 Model II regression

Cochrane and Cornbleet argue for the use of methods that based on the assumption that both methods are imprecisely measured ,and that yield a fitting that is consistent with both '$X$ on $Y$' and '$Y$ on $X$' formulations. These methods uses alternatives to the OLS approach to determine the slope and intercept.

They describe three such alternative methods of regression; Deming , Mandel, and Bartlett regression. Collectively the authors refer to these approaches as Model II regression techniques.

### 2.1.3 Distribution of Maxima

It is possible to use Order Statistics theory to assess conditional probabilities. With two random variables $T_0$ and $T_1$, we define two variables $Z$ and $W$ such that they take the maximum and minimum values of the pair of $T$ values.

### 2.1.4 Plot of the Maxima against the Minima

In Figure 1, The Maximas are plotted against their corresponding minima. The Critical values of the Maxima and Minima are displayed in the dotted lines.The Line of Equality depicts the obvious logical constraint of the each Maximum value being greater than its corresponding minimum value.

The scientific question at hand is the correct approach to assessing whether two

methods can be used interchangeably. Bland and Altman (1999) expresses this as follows:

> We want to know by how much (one) method is likely to differ from the (other), so that if it not enough to cause problems in the mathematical interpretation we can ... use the two interchangeably.

Consequently, of the categories of method comparison study, comparison studies, the second category, is of particular importance, and the following discussion shall concentrate upon it. Less emphasis shall be place on the other three categories.

Further to Bland and Altman (1986), 'equivalence' of two methods expresses that both can be used interchangeably. Dunn (2002, p.49) remarks that this is a very restrictive interpretation of equivalence, and that while agreement indicated equivalence, equivalence does not necessarily reflect agreement.

The main difference between Myers proposed method and the Bland Altman is that the random effects model is used to estimate the within-subject variance after adjusting for known and unknown variables. The Bland Altman approach uses one way analysis of variance to estimate the within subject variance. In general, the random effects model is an extension of the analysis of the ANOVA method and it can adjust for many more covariates than the ANOVA method

### 2.1.5 Criticism of Bland Altman Plots

An Evaluation of the correlation between the difference and means complement the analysis.

Bland and Altman caution that the calculations are based on the assumption that the data is normally distributed. This can be verified by using a histogram. If Data is not normally distributed, it can be transformed.

Luiz *et al* remarks that that Bland Altman Plot should be used only for small data sets, as the use of an index will be of little value to the analysis.

## 2.2 Conclusions about Existing Methodologies

Scatterplots are recommended by Altman and Bland (1983) for an initial examination of the data, facilitating an initial judgement and helping to identify potential outliers. They are not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation.

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

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

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

# Chapter 3

# Appendix

### 3.0.1 Contention

Several papers have commented that this approach is undermined when the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined. Outliers are a source of error in regression estimates.In method comparison studies, the X variable is a precisely measured reference method. Cornbleet Gochman (1979) argued that criterion may be regarded as the correct value. Other papers dispute this.

### 3.0.2 Least Product Regression

Least Product Regression , also known as 'Model II regression' caters for cases in which random error is attached to both dependent and independent variables. Ludbrook cites this methodology as being pertinent to Method comparison studies.

The sum of the products of the vertical and horizontal deviations of the x,y values from the line is minimized.

Least products regression analysis is considered suitable for calibrating one method against another.Ludbrook comments that it is also a sensitive technique for detecting and distinguishing fixed and proportional bias between methods.

Proposed as an alternative to Bland & Altman methodology ,this method is also known as 'Geometric Mean Regression' and 'Reduced Major Axis Regression'.

**Difference with Least Squares Regression**

Least-products regression can lead to inflated SEEs and estimates that do not tend to their true values an N approaches infinity (Draper and Smith, 1998).

# 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 re-

producibility (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 methods, 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, $\kappa_P$ which capyures a large proportion $p$ of paired based differences from two devices or observers within the boundary.

The value of $\kappa_P$ 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 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 $\beta$ and $\theta$. A common technique is to refit the model with an observation or group of observations omitted.

*west* 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'.

### 3.2.1 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.

### 3.2.2 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 ].

### 3.2.3 Influence

*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.

### 3.2.4 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. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis. Outliers, for example,

may be the most noteworthy data points in an analysis. They can point to a model breakdown and lead to development of a better model.

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 transformations. 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 Quiroz Burdick

Assessmet of individual agreements with repeated measurements based on Generalized Confidence intervals.

Bootstrap confidence intervals. Coverage probability (CP) Equivalence Studies Individual agreements Generalized Confidence intervals (GCI) Total deviation index (TDI) Variance components

Proposing an equivalence test for assessing individual agreement based on TDI and CP. The bounds used in the tests are constucted using a bootstrap approach and generalized confidence intervals (GCI).

Equivalence testing is an approach commonly used to determine the acceptability of a new method against a reference method.

Both the TDI and CP are attractive criteria as they are easy to interpret.

Bootstrap approach was later applied to mixed models with repeated measurements by Choudhary (2007)

$T$ for test measurement, $R$ for reference measurement

$\otimes$ is the Kroneckor Product operator.

$$\Sigma_{MS} = \begin{bmatrix} \sigma^2_{TS} & 0 \\ 0 & \sigma^2_{RS} \end{bmatrix}$$

## 3.6 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.

### 3.6.1 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.

### 3.6.2 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.

## 3.7 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 $\tau_i$ and $\beta\tau_i$ as the true values , and $\delta_i$ and $\epsilon_i$ 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 $\alpha$ is the inter-method bias between the two methods.

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

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

## 3.8 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.

Christensen et al. (1992) describes three propositions that are required for efficient case-deletion in LME models. The first proposition decribes how to efficiently update

$V$ when the $i$th element is deleted.

$$V_{[i]}^{-1} = \Lambda_{[i]} - \frac{\lambda\lambda\prime}{\nu ii} \tag{3.5}$$

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

$$X'_{[i]}V_{[i]}^{-1}X_{[i]} = X'V^{-1}X - \frac{\hat{x}_i\hat{x}'_i}{s_i} \tag{3.6}$$

$$(X'_{[i]}V_{[i]}^{-1}X_{[i]})^{-1} = (X'V^{-1}X)^{-1} + \frac{(X'V^{-1}X)^{-1}\hat{x}_i\hat{x}'_i(X'V^{-1}X)^{-1}}{s_i - \bar{h}_i} \tag{3.7}$$

$$X'_{[i]}V_{[i]}^{-1}Y_{[i]} = X\prime V^{-1}Y - \frac{\hat{x}_i\hat{y}'_i}{s_i} \tag{3.8}$$

Schabenberger (2004) 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, 2004).

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.

Schabenberger (2004) 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.

A residual is the difference between an observed quantity and its estimated or

predicted value. In LME models, there are 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.

Schabenberger (2004) notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates. Haslett and Dillane (2004) offers an procedure to assess the influences for the variance components within the linear model, complementing the existing methods for the fixed components. The essential problem is that there is no useful updating procedures for $\hat{V}$, or for $\hat{V}^{-1}$. Haslett and Dillane (2004) propose an alternative , and computationally inexpensive approach, making use of the 'delete=replace' identity.

Haslett (1999) considers the effect of 'leave k out' calculations on the parameters $\beta$ and $\sigma^2$, using several key results from Haslett and Hayes (1998) on partioned matrices.

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

$$LD_{(U)} = 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\} \tag{3.9}$$

$$RLD_{(U)} = 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\} \tag{3.10}$$

### 3.8.1 Case Deletion Diagnostics

**?** develops case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

### 3.8.2 Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{3.11}$$

### 3.8.3 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

A general method for comparing nested models fit by maximum liklihood is the liklihood ratio test. This test can be used for models fit by REML (restricted maximum liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method=ML must be employed (ML = maximum liklihood).

Example of a liklihood ratio test used to compare two models:

!"

The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.

Generally, liklihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.

A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple anova function.

Example: " !"

will give the most reliable test of the fixed effects included in model1.

### 3.8.4 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.

### 3.8.5 Matrix Notation for Case Deletion

### 3.8.6 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 $\boldsymbol{A}$.

### 3.8.7 Partitioning Matrices

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

### 3.8.8 Case Deletion Diagnostics

**?** develops case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

### 3.8.9 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

### 3.8.10 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.

### 3.8.11 Case Deletion Diagnostics

**?** develops case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

### 3.8.12 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.

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 $\beta$ and $\sigma^2$, 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.

### 3.8.13 Terminology for Case Deletion diagnostics

Preisser(19XX) 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.

## 3.9 Work List

1. ML v REML
2. Nested Models and LRTs
3. Generalized Lease Squares
4. Diagnostics
5. Simplifying GLS
6. Paper progression

## 3.10 Linear mixed effects models

These models are used when there are both fixed and random effects that need to be incorporated into a model.

Fixed effects usually correspond to experimental treatments for which one has data for the entire population of samples corresponding to that treatment.

Random effects,on the other hand, are assigned in the case where we have measurements on a group of samples, and those samples are taken from some larger sample pool, and are presumed to be representative.

As such, linear mixed effects models treat the error for fixed effects differently than the error for random effects.

## 3.11 Diagnostics

### 3.11.1 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.

### 3.11.2 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 )

## 3.12 Iterative and non-iterative influence analysis

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

### 3.12.1 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.

### 3.12.2 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.

# Deming Regression

Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies

Application of Deming regression analysis to interpret method comparison data presupposes specification of the squared analytical error ratio ($\lambda$, but in cases involving only single measurements by each method, this ratio may be unknown and is often assigned a default value of one.

On the basis of simulations, this practice was evaluated in situations with real error ratios deviating from one. Comparisons of two electrolyte methods and two glucose methods were simulated.

In the first case, misspecification of $\lambda$ produced a bias that amounted to two-thirds of the maximum bias of the ordinary least-squares regression method. Standard errors and the results of hypothesis-testing also became misleading. In the second situation, a misspecified error ratio resulted only in a negligible bias.

Thus, given a short range of values in relation to the measurement errors, it is important that $\lambda$ is correctly estimated either from duplicate sets of measurements or, in the case of single measurement sets, specified from quality-control data. However, even with a misspecified error ratio, Deming regression analysis is likely to perform better than least-squares regression analysis.

## 3.13 Generalized Least Squares

### 3.13.1 Introduction to Generalized Least Squares

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{3.12}$$

Estimation under this model has been studied extensively in the linear regression model.

### 3.13.2 Introduction

#### Robinson's (1991) review

*Robinson's (1991) review of best linear unbiased prediction (BLUP), together with the subsequent discussion, has emphasized the very considerable range of models that may be addressed via the general least squares (GLS) solution to the general linear model* $Y = X\beta + \varepsilon$, *where* $E(\varepsilon) = 0$ *and* $var(\varepsilon) = V$. *These include linear mixed models, geostatistics, time series and multivariate regression.*

*The texts by Christensen (1996, 1991) and the connections to modern topics of image analysis, quality analysis, Bayesian methods, and splines (all in Robinson and discussion) make it an eminently suitable topic for teaching in any course concerning statistical linear models.*

*Nevertheless some of the matrix algebra that results from solving the normal equations for individual specifications of the general linear model will be daunting, and far from intuitive for many students, even those who are at home in linear space. The conventional approach to prediction and estimation from data* $Y$ *associated with covariates* $X$ *via the general linear model* $Y = X\beta + \varepsilon$ *is essentially a two-stage process.*

The first stage is to determine the best,in the GLS sense, estimator $\hat{\beta}$ of $\beta$ and subsequently to determine everything else from this.

The estimator is said to be best if it minimizes the generalization of the sum of squares $\hat{e}^t V^{-1} \hat{e}$, where $\hat{e} = Y - X\hat{\beta}$

It is straightforward to show that $\hat{\beta} = (X^t V^{-l} X)^{-l} X^t V^{-l} Y = BY$ and at the minimum the sum of squares is $Y^t(V^{-l} - V^{-l}(X^t V^{-l} X)^{-l} X^t V^{-l})Y = Y^t QY$.

*The purpose of this note is to give emphasis to one derivation, based on Lagrange multipliers, which leads to a system of equations that is very intuitive and lends itself readily to specialization. This approach is in fact standard in the geostatistical treatment of* ***kriging*** *(see Matheron 1962; Journel and Huijbregts 1981; Ripley 1981; Cressie 1993). In the genetics literature it is associated with the name of Henderson (1983); or in the classical statistical literature Hocking (1996, p. 73) is a suitable reference.*

*The approach based on Lagrange multipliers deemphasizes the explicit determination of $\hat{\beta}$ and leads to a clearer understanding of the complementary (but for some confusing) tasks known as best linear unbiased estimation (BLUE) and best linear unbiased prediction (BLUP). Regrettably, Robinson-despite offering four derivations, and having as his main concern the interplay of BLUP and BLUE-gives it little prominence.*

It has recently been discussed by Searle (1997, p; 278) who said that it makes another approach (Searle, Casella, and McCulloch 1992, p. 271) seem "obtuse and unnecessarily complicated." By contrast, our treatment emphasizes the fact that it leads to a single set of equations whose solution sheds simplifying light on very many issues in general least squares.

The American Statistician's Teacher's Corner (e.g., McLean, Sanders, and Stroup 1991; Puntanen and Styan 1989) has already played host to previous attempts to simplify the explanation of such topics. Various authors (CPJ, Haslett Hayes ,Martin ) have visited the more specialized area of diagnostics and have developed ***down-dating*** (leave-$k$-out) formulas.

The conventional approach here is via tricky identities based on the inverses of

partitioned matrices. Here again the Lagrange system of equations leads to a much simplified and-we claim-much more intuitive derivation of these more technical results.

*The essence of the approach is to seek that linear combination of the available data Y which is best for the estimation of Z among those linear estimators which are constrained to be unbiased. We adopt therefore a constrained minimization approach, using Lagrange multipliers. By best we mean that combination* $\hat{Z}(Y) = \lambda_z^t Y$ *which has least mean square error* $E(Z-\lambda_z^t Y)^2$*, and by unbiased we mean* $E(Z-\lambda_z^t Y)) = 0$. Here $Z$ denotes that scalar which is to be the objective of the estimation. This estimator is written as $\hat{Z}(Y)$ to make its dependence on $Y$ explicit. Note that the term "best" is applied in the context of minimizing the prediction variance $var(Z - Z(Y))$. We shall see that Z may be used to denote either a random variable or an unknown parameter, and that it will be sufficient to specify Z via $E[Z]$ and $cov(Z, Y)$. If $Z$ is not a random variable then of course the latter is zero and $E[Z] = Z$. We establish-very simply, as below-a general solution in terms of A and cov(Z, Y) and achieve particular tasks by identification of these. Our presentation is for a scalar Z, but the notation facilitates generalization to vector Z.

### 3.13.3 Predictors and Estimators

*We note that Robinson (1991) stated "A convention has somehow developed that estimators of random effects are called predictors while estimators of fixed effects are called estimators." We agree that this distinction is confusing and indeed unnecessary.*

We seek $\hat{Z}(Y) = \lambda_z^t Y$, where $\lambda_z^t$, is an $n \times 1$ vector of estimation coefficients. It is convenient to specify $E[Z] = A\beta$ for known $A$. In this context $A$ denotes a row vector, but we generalize this in the following. The constraint requiring $\hat{Z}(Y)$

to be unbiased now reduces to $(A - \lambda_z^t X) = 0$. A solution is found by minimizing $var(Z - \lambda_z^t Y) + \gamma_z^t(X^t\lambda_z - A^t)$, where $\gamma_z$ is a $p \times 1$ vector of Lagrange multipliers, where $p$ is the length of the parameter vector $\beta$. Setting to zero the derivatives with respect to $\lambda_z$ and $\gamma_z$ yields the system.

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{3.13}$$

If the inverse exists we have that

$$\begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix}^{-1} \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{3.14}$$

so that

$$\hat{Z}(Y) = \begin{pmatrix} \lambda_z^t & \gamma_z^t \end{pmatrix} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

In terms of the estimation problem being considered the square matrix on the left-hand side of (1) concerns "what we have," namely, the data plus constraints.

The matrix does not depend on Z and consequently need only be constructed once before application to a range of problems. The right- hand side contains the term $cov(Z, Y)$ and can be specified for whatever Z is being considered.

It is this feature of system (1) that makes a generic approach to estimation possible.

### 3.13.4 Two Options

- Wald Type CIs
- PL Type CIs

### 3.13.5 Profile Likelihood Confidence Intervals

The Profile-likelihood based confidence intervals methods is described in Venzon and Moolgavkar, Journal of the Royal Statistical Society, Series C vol 37, no.1, 1988, pp. 87-94.

Profile likelihood confidence intervals can be computed for real parameter estimates.

The default confidence intervals for real parameter estimates in the 0-1 interval are based on the standard error and the logit transformation. That is, a 95% confidence interval is computed on the logit estimate, and then these intervals are transformed to the real scale.

## 3.14 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.

## 3.15 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.

## 3.16 Enabling One Tailed Testing

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

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

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

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

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

## 3.17 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.

## 3.18 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()`

## 3.19 Quiroz-Burdick PEFR Example

The data consist of two paired measurements on the same subject made with the large Wright peak flow meter and a mini Wright meter.

Paired differences of less than 101/min are considered of no practical clinical significance. That is to say, it would have no bearing on any decision related to a clinical matter.

A serious error would be declare that the mini-meter is as effective as the large meter when in fact it is not.

$$H_0 : \kappa_{0.90} \geq 10$$

$$H_A : \kappa_{0.90} < 10$$

# Chapter 4

# Application to Method Comparison Studies

## 4.1 Application to MCS

Let $\hat{\beta}$ denote the least square estimate of $\beta$ based upon the full set of observations, and let $\hat{\beta}^{(k)}$ denoted the estimate with the $k^{th}$ case excluded.

## 4.2 Grubbs' Data

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $k = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

$$Y^{-Q} = \hat{\beta}^{-Q} X^{-Q} \tag{4.1}$$

When considering the regression of case-wise differences and averages, we write $D^{-Q} = \hat{\beta}^{-Q} A^{-Q}$

| | F | C | D | A |
|---|---|---|---|---|
| 1 | 793.80 | 794.60 | -0.80 | 794.20 |
| 2 | 793.10 | 793.90 | -0.80 | 793.50 |
| 3 | 792.40 | 793.20 | -0.80 | 792.80 |
| 4 | 794.00 | 794.00 | 0.00 | 794.00 |
| 5 | 791.40 | 792.20 | -0.80 | 791.80 |
| 6 | 792.40 | 793.10 | -0.70 | 792.75 |
| 7 | 791.70 | 792.40 | -0.70 | 792.05 |
| 8 | 792.30 | 792.80 | -0.50 | 792.55 |
| 9 | 789.60 | 790.20 | -0.60 | 789.90 |
| 10 | 794.40 | 795.00 | -0.60 | 794.70 |
| 11 | 790.90 | 791.60 | -0.70 | 791.25 |
| 12 | 793.50 | 793.80 | -0.30 | 793.65 |

$$Y^{(k)} = \hat{\beta}^{(k)} X^{(k)} \tag{4.2}$$

Consider two sets of measurements , in this case F and C , with the vectors of case-wise averages $A$ and case-wise differences $D$ respectively. A regression model of differences on averages can be fitted with the view to exploring some characteristics of the data.

When considering the regression of case-wise differences and averages, we write

$$D^{-Q} = \hat{\beta}^{-Q} A^{-Q} \tag{4.3}$$

Let $\hat{\beta}$ denote the least square estimate of $\beta$ based upon the full set of observations, and let $\hat{\beta}^{(k)}$ denoted the estimate with the $k^{th}$ case excluded.

For the Grubbs data the $\hat{\beta}$ estimated are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. Leaving the fourth case out, i.e. $k = 4$ the corresponding estimates are $\hat{\beta}_0^{-4}$ and $\hat{\beta}_1^{-4}$

$$Y^{(k)} = \hat{\beta}^{(k)} X^{(k)} \tag{4.4}$$

Consider two sets of measurements , in this case F and C , with the vectors of case-wise averages $A$ and case-wise differences $D$ respectively. A regression model of differences on averages can be fitted with the view to exploring some characteristics of the data.

```
Call: lm(formula = D ~ A)


Coefficients: (Intercept)            A
-37.51896      0.04656
```

### 4.2.1 Influence measures using R

R provides the following influence measures of each observation.

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

# Bibliography

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.

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.

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.

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.

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. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society (Series B) 61*, 603–609.

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.

Haslett, J. and K. Hayes (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society (Series B) 60*, 201–215.

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.

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

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

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

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

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.

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

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