Contents

Chapter 1

Method Comparison Studies

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.

Abstract

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

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

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

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.

1.1 Outline of Thesis

In the first chapter the study of method comparison is introduced, while the second chapter provides a review of current methodologies. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models. Chapter three shall describes linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the `R` programming language.

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

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

1.2 Test for inter-method bias

Bias is determinable by examination of the 't-table'. Estimate for both methods are given, and the bias is simply the difference between the two. Because the R implementation does not account for an intercept term, a $p-$value is not given. Should a $p-$value be required specifically for the bias, and simple restructuring of the model is required wherein an intercept term is included. Output from a second implementation will yield a $p-$value.

1.3 LME

Consistent with the conventions of mixed models, **?** formulates the measurement $y_{ij}$from method i on individual j as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.1}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{1.2}$$

These vectors are assumed to be independent for different is, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.3}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate x is the key step in how this methodology assesses agreement.

1.4 Remarks

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.

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.

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.

What is required is the computation of the variance ratios of within-item and between-item standard deviations.

A naive approach would be to compute the variance ratios by relevant F distribution quantiles. However, the question arises as to the appropriate degrees of freedom.

Limits of agreement are easily computable using the LME framework. While we will not be considering this analysis, a demonstration will be provided in the example.

1.5 PRESS

An (unconditional) predicted value is $\hat{y}_i = x_i'\hat{\boldsymbol{\beta}}$, where the vector x_i is the ith row of $\boldsymbol{X}$.

An (unconditional) predicted value is $\hat{y}_i = x_i'\hat{\boldsymbol{\beta}}$, where the vector x_i is the ith row of $\boldsymbol{X}$. The (raw) residual is given as $\varepsilon_i = y_i - \hat{y}_i$. The PRESS residual is similarly constructed, using the predicted value for observation i with a model fitted from reduced data.

$$\varepsilon_{i(U)} = y_i - x_i'\hat{\boldsymbol{\beta}}_{(U)}$$

1.6 One Way ANOVA

1.6.1 Page 448

Computing the variance of $\hat{\beta}$

$$\text{var}(\hat{\beta}) = (X'V^{-1}X)^{-}1 \tag{1.4}$$

It is not necessary to compute V^{-1} explicitly.

$$V^{-1}X = \Sigma^1 X - Z()Z'\Sigma^{-1}X \tag{1.5}$$

$$= \Sigma^{-1}(X - Zb_x) \tag{1.6}$$

The estimate b_x is the same term obtained from the random effects model; $X = Zb_x + e$, using X as an outcome variable. This formula is convenient in applications where b_x can be easily computed. Since X is a matrix of p columns, b_x can simple be computed column by column. according to the columns of X.

1.6.2 Page 448- simple example

Consider a simple model of the form;

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}.$$

The iterative procedure is as follows Evaluate the individual group mean $\bar{y}_i$ and variance $\hat{Sigma^2}_i$. Then use the variance of the group means as an estimate of the σ_b^2. The average of the the variances of the groups is the initial estimate of the σ_e^2.

Iterative procedure

The iterative procedure comprises two steps, with 0 as the first approximation of b_i.

The first step is to compute λ, the ratio of variabilities,

$$\lambda = \frac{\sigma_b^2}{\sigma_e^2}$$

$$\mu = \frac{1}{N}\sum_{ij}(y_{ij} - b_i)$$

$$b_i = \frac{n(\bar{y}_i - \mu)}{n + \lambda}$$

The second step is to updat $sigma_e^2$

$$\sigma_e^2 = \frac{e'e}{N - df} \tag{1.7}$$

where e is the vector of $e_{ij} = y_{ij} - \mu - b_i$ and $df = qn/n + \lambda$ and

$$\sigma_b^2 = \frac{1}{q}\sum_{i=1}^{q} b_1^2 + (\frac{n}{\sigma_e^2} + \frac{1}{\sigma_b^2})^{-1} \tag{1.8}$$

Worked Example

Further to [pawitan 17.1] the initial estimates for variability are $\sigma_b^2 = 1.7698$ and $\sigma_e^2 = 0.3254$. At convergence the following results are obtained.

n=16, q=5

$$\hat{\mu} = \bar{y} = 14.175$$
$$\hat{\sigma}^2 = 0.325$$
$$\hat{\sigma}_b^2 = 1.395$$
$$\sigma = 0.986$$

At convergene the following estimates are obtained,

$$\hat{\mu} = 14.1751$$
$$\hat{b} = (-0.6211, 0.2683, 1.4389, -1.914, 0.8279)$$
$$\hat{\sigma}_b^2 = 1.3955$$
$$\hat{\sigma}_e^2 = 0.3254$$

1.7 BA - BXC 2009 Discussion

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.

1.7.1 EBLUPS-Diagnostics for Random Effects

West et al. (2007) recommends the empirical Bayes predictor, also known as EBLUPS as a diagnostic tool for Random effects. Checking EBLUPS for normality is of limited value.

1.8 EBLUP

The EBLUP is useful to identify outlier subjects given that it represents the distance between the population mean value and the value predicted for the ith subject. A way of using the EBLUP to search for outliers subjects is to use the Mahalanobis distance (see Waternaux et al., 1989), FORMULA

It is also possible to use the EBLUP to verify the random effects normality assumption. For more information; see **?**. In Table 2 we summarize diagnostic techniques involving residuals discussed in

Nobre and Singer (2007).

$\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L}$

1.9 Unknown Material

To standardize the assessment of how influential data is, several measures of influence are commonly used, such as DFBETAS and Cooks Distance.

Although influential cases thus have extreme values on one or more of the variables, they can be onliers rather than outliers.

To account for this, the (standardized) deleted residual is defined as the difference between the observed score of a case on the dependent variable, and the predicted score from the regression model fitted from data when that case is omitted.

Just as influential cases are not necessarily outliers, outliers are not necessarily influential cases.

This also holds for deleted residuals. The reason for this is that the amount of influence a case exerts on the regression slope is not only determined by how well its (observed) score is fitted by the specified regression model, but also by its score(s) on the independent variable(s). The degree to which the scores of a case on the independent variable(s) are extreme is indicated by the leverage of this case.

1.9.1 Estimation

$$\hat{\beta} = X^T \tag{1.9}$$

$$\hat{\gamma} = G(\hat{\theta}) Z^T \tag{1.10}$$

The difference between perturbation and residual analysis between the linear and LME models. The estimates of the fixed effects β depend on the estimates of the covariance parameters.

1.10 RSquared for LME models

As a complement to this, one can also consider how to properly employ the R^2 measure, in the context of Methoc Comparison Studies, further to the work by **?**, namely "An R^2 statistic for fixed effects in

the linear mixed model".

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

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

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

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

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

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

1.11 Roy's Candidate models

The original Bland Altman Method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for repeated measures data. However, as a

nave analysis, it may be used to explore the data because of the simplicity of the method. Myles states that such misuse of the standards Bland Altman method is widespread in Anaesthetic and critical care literature.

Bland and Altman have provided a modification for analysing repeated measures under stable or chaninging conditions, where repeated data is collected over a period of time. Myers proposes an alternative Random effects model for this purpose.

with repeated measures data, we can calculate the mean of the repeated measurements by each method on each individuals. *The pairs of means can then be used to compare the two methods based on the 95% limits of agreement for the difference of means. The bias between the two methods will not be affected by averaging the repeated measurements.*.However the variation of the differences will be underestimated by this practice because the measurement error is, to some extent, removed. Some advanced statistical calculations are needed to take into account these measurement errors. *Random effects models can be used to estimate the within-subject variation after accounting for other observed and unobserved variations, in which each subject has a different intercept and slope over the observation period .On the basis of the within-subject variance estimated by the random effects model, we can then create an appropriate Bland Altman Plot.*The sequence or the time of the measurement over the observation period can be taken as a random effect.

1.12 Other Approaches : Marginal Modelling

(Diggle 2002) proposes the use of marginal models as an alternative to mixed models.m Marginal models are appropriate when interences about the mean response are of specific interest.

1.13 Other Approaches

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and

that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

1.14 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.

1.15 Computation and Notation

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

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

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

Joseph G. Voelkel Bruce E. Siskowski

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.

Bayesian BA - Philip J Schluter

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

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

Background

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

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

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

Methods

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

Results

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

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

Bayesian Approach

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

- Assessing agreement in method comparison studies depends on two fundamentally important components; validity (the between method agreement) and reproducibility (the within method agreement).
- The Bland-Altman limits of agreement technique is one of the f

1.16 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.

1.17 Hutson et al

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".

1.18 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.

1.19 Diagnostics

1.19.1 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.

1.19.2 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 et al., 2007).

1.20 Covariance Parameters

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

Cook's distance

In the study of Linear model diagnostics, Cook proposed a measure that combines the information of leverage and residual of the observation, now known simply as the Cook's Distance. Christensen et al. (1992) would later adapt the Cook's distance measure for the analysis of LME models.

1.21 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.

1.21.1 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.

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.

1.21.2 Further Assumptions of Linear Models

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

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

Leave-One-Out Diagnostics with `lmeU`

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

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

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

The addition of an extra factor

Interaction terms are featured in ANOVA designs.

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

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

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.

1.22 Covariance Parameters

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

1.23 Roy2013

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

Testing the Equality of Mean Vectors for Paired Doubly Multivariate Observations

Example 2. (Mineral Data): This data set is taken from Johnson and Wichern (2007, p. 43). An investigator measured the mineral content of bones (radius, humerus and ulna) by photon absorptiometry to examine whether dietary supplements would slow bone loss in 25 older women. Measurements were

recorded for three bones on the dominant and nondominant sides. Thus, the data is doubly multivariate and clearly u = 2 and q = 3. The bone mineral contents for the rst 24 women one year after their participation in an experimental program is given in Johnson and Wichern (2007, p. 353).

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

1.24 Outlier Testing

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

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

1.25 Lorelia

Method comparison studies are performed in order to prove equivalence between two measurement methods or instruments. The identification of outliers is an important part of data analysis as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in method comparison studies usually do not meet these assumptions. To overcome this problem, different data transformation methods are proposed in the literature. However, they will only

be applicable if the random errors can be described by simple additive or multiplicative models. In this work, a new outlier test based on robust linear regression is proposed which provides a general solution to the above problem. The LORELIA (LOcal RELIAbility) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. Outlier limits are estimated from the actual data situation without making assumptions on the underlying error variance model. The performance of the new test is demonstrated in examples and simulations.

1.26 Note on Roy's paper

1. Basic model:

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

   Assumptions are made about homoskedasticity.

2. General model:

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

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

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

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

$$\boldsymbol{y}_i = (y_i^{j1}, y_i^{Jj2}, y_i^{j3}, y_i^{s1}, y_i^{s2}, y_i^{s3})\boldsymbol{\prime} \tag{1.11}$$

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

1.27 Regression Of Differences On Averages

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

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

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

1.28 Residual diagnostics

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

Appendix to Section 4

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

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

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

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

Important Consideration for MCS

The key issue is that `nlme` allows for the particular specification of Roy's Model, speciifically direct spefiication of the VC matrices for within subject and between subject residuals. The `lme4` package does not allow for this. To advance the ideas that eminate from Roys' paper, one is required to use the `nlme` context. However, to take advantage of the infrastructure already provided for `lme4` models, one may change the research question away from that of Roy's paper. To this end, an exploration of what

textitinfluence.ME can accomplished is merited. As a complement to this, one can also consider how to properly employ the R^2 measure, in the context of Methoc Comparison Studies, further to the work by Edwards et al, namely "An R^2 statistic for fixed effects in the linear mixed model".

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

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

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

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

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

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

The `nlme` package

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

The `lme4` package

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

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

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

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

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

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

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

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

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

$$r_{mi} = x_i^T \hat{\beta} \tag{1.12}$$

1.28.1 Note 1: Coefficient of Repeatability

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

computation of the coefficients of repeatability for both methods is straightforward.

1.28.2 Note 2: Carstensen model in the single measurement case

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.13}$$

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

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

1.28.3 Note 3: Model terms

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

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

- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

1.29 Modelling Agreement with LME Models

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

Four candidates models are fitted to the data.

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

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

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

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

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

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

Carstensen's model is a mixed effects ANOVA.

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

This model includes a method by item iteration term.

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

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

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

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

1.30 Covariance Parameters

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

1.31 Missing Data in Method Comparison Studies

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

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

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

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

- This chapter is broken into two parts. The first part is a review of diagnostics methods for linear models, intended to acquaint the reader with the subject, and also to provide a basis for material covered in the second part. Particular attention is drawn to graphical methods.
- The second part of the chapter looks at diagnostics techniques for LME models, firsly covering the theory, then proceeding to a discussion on implementing these using `R` code.
- While a substantial body of work has been developed in this area, ther are still area worth exploring. In particular the development of graphical techniques pertinent to LME models should be looked at.

1.32 LME - Pankaj Choudhury

Consistent with the conventions of mixed models, (**?**) formulates the measurement $y_{ij}$from method i on individual j as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.14}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{1.15}$$

These vectors are assumed to be independent for different is, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.16}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate x is the key step in how this approach assesses agreement.

1.33 Modelling Agreement with LME Models

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

Four candidates models are fitted to the data.

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

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

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

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

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

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

Carstensen's model is a mixed effects ANOVA.

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

This model includes a method by item iteration term.

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

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

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

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

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

- *The previous Section (Section 4) is a literary review of residual diagnostics and influence procedures for Linear Mixed Effects Models, drawing heavily on Schabenberger and Zewotir.*
- *Section 4 begins with an introduction to key topics in residual diagnostics, such as influence, leverage, outliers and Cook's distance. Other concepts such as DFFITS and DFBETAs will be introduced briefly, mostly to explain why the are not particularly useful for the Method Comparison context, and therefore are not elaborated upon.*
- *In brief, Variable Selection is not applicable to Method Comparison Studies, in the commonly used used context. Testing a rather simplisticy specificied model against one with more random effects terms is tractable, but this research question is of secondary importance.*

1.34 Remarks on the Multivariate Normal Distribution

Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution. ARoy2009's model is specified using the bivariate normal distribution. This gives rises to a key difference between the two model, in that a bivariate model accounts for covariance between the variables of interest. The multivariate normal distribution of a k-dimensional random vector $X = [X_1, X_2, \ldots, X_k]$ can be written in the following notation:

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

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

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

with k-dimensional mean vector

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

and $k \times k$ covariance matrix

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

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

   (a)

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

   (b)

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

Bibliography

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

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

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.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

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

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