Contents

Chapter 1

Method Comparison Studies

1.1 Introduction

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

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm gun and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels 'Fotobalk', 'Counter' and 'Terma'.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 1.1.1: Velocity measurement from the three chronographs (Grubbs 1973).

An important aspect of the these data is that all three methods of measurement are assumed to have an attended measurement error, and the velocities reported in Table 1.1 can not be assumed to be ‘true values’ in any absolute sense.

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown ‘true value’. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy. The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The smaller the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimate of the inter-method bias is given by the differences between pairs of measurements, for example, Table 1.1.2 is a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. A cursory inspection of the table will indicate a systematic tendency for the Counter method to result in higher measurements than the Fotobalk method.

The absence of inter-method bias is, by itself, not sufficient to establish that two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than those of the other, they can not be considered to be in agreement. Hence, method comparison studies are required to take account of both inter-method bias and difference in precision of measurements.

Round	Fotobalk (F)	Counter (C)	Difference (F-C)
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.1.2: Difference between Fotobalk and Counter measurements.

Chapter 2

Introduction to Method Comparison Studies

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 2.0.1: Difference between Fotobalk and Counter measurements.

Chapter 3

Extending Current Methodologies

3.1 Extension of Roy's methodology

Roy's methodology is constructed to compare two methods in the presence of replicate measurements. Necessarily it is worth examining whether this methodology can be adapted for different circumstances.

An implementation of Roy's methodology, whereby three or more methods are used, is not feasible due to computational restrictions. Specifically there is a failure to reach convergence before the iteration limit is reached. This may be due to the presence of additional variables, causing the problem of non-identifiability. In the case of two variables, it is required to estimate two variance terms and four correlation terms, six in all. For the case of three variabilities, three variance terms must be estimated as well as nine correlation terms, twelve in all. In general for n methods has $2 \times T_n$ variance terms, where T_n is the triangular number for n, i.e. the addition analogue of the factorial. Hence the computational complexity quite increases substantially for every increase in n.

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

The tests would not be informative as to whether any two of those three were interchangeable, or equivalently if one method in particular disagreed with the other two. Indeed it is easier to perform three pair-wise comparisons separately and then to combine the results.

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector y_i, as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

3.2 Conclusion

Carstensen et al. (2008) and Roy (2009a) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009a) presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

3.3 BXC - Model Terms

- Let y_{mir} be the response of method m on the ith subject at the $r-$th replicate.
- Let $\boldsymbol{y}_{ir}$ be the 2×1 vector of measurements corresponding to the $i-$th subject at the $r-$th replicate.
- Let $\boldsymbol{y}_i$ be the $R_i \times 1$ vector of measurements corresponding to the $i-$th subject, where R_i is number of replicate measurements taken on item i.
- Let $\alpha_m i$ be the fixed effect parameter for method for subject i.

- Formally Roy uses a separate fixed effect parameter to describe the true value μ_i, but later combines it with the other fixed effects when implementing the model.

- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i.

- $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.

- $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to Roy's first test.

Hamlett and Lam

Lam et al. (1999) used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

The true correlation ρ_{xy} is repeated measurements can be considered as having two components: between subject and within-subject correlation. The usefulness of estimating repeated measure correlation coefficients is the calculation of between-method and within-method variabilities are produced as by-products.

3.3.1 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.

3.4 LME

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

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{3.1}$$

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;

$$v_{ij} \sim N(0, \Sigma), \tag{3.2}$$

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;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{3.3}$$

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.

3.5 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.

Chapter 4

LME Likelihood

4.1 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.

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

4.2 One Way ANOVA

4.2.1 Page 448

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

$$\text{var}(\hat{\beta}) = (X'V^{-1}X)^{-}1 \tag{4.1}$$

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

$$\begin{aligned} V^{-1}X &= \Sigma^{1}X - Z()Z'\Sigma^{-1}X && (4.2) \\ &= \Sigma^{-1}(X - Zb_x) && (4.3) \end{aligned}$$

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.

4.2.2 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^2_b. The average of the the variances of the groups is the initial estimate of the σ^2_e.

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,

$$\lambda = \frac{\sigma^2_b}{\sigma^2_e}$$

$$\mu = \frac{1}{N}\sum_{ij}(y_{ij} - b_i)$$
$$b_i = \frac{n(\bar{y}_i - \mu)}{n + \lambda}$$

The second step is to updat $sigma^2_e$

$$\sigma^2_e = \frac{e'e}{N - df} \tag{4.4}$$

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

$$\sigma^2_b = \frac{1}{q}\sum_{i=1}^{q} b^2_1 + (\frac{n}{\sigma^2_e} + \frac{1}{\sigma^2_b})^{-1} \tag{4.5}$$

Worked Example

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

n=16, q=5

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

At convergene the following estimates are obtained,

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

4.3 Sampling

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this)

4.4 Conclusion

Carstensen et al. (2008) and Roy (2009b) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009b) presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into ARoy2009's methodology.

Permutation Test, Power Tests and Missing Data

This section explores topics such as dependent variable simulation and power analysis, introduced by Galecki & Burzykowski (2013), and implementable with their ***nlmeU*** `R` package.

Using the ***predictmeans*** `R` package, it is possible to perform permutation t-tests for coefficients of (fixed) effects and permutation F-tests.

The matter of missing data has not been commonly encountered in either Method Comparison Studies or Linear Mixed Effects Modelling. However ARoy2009 (2009) deals with the relevant assumptions regrading missing data.

Galecki & Burzykowski (2013) approaches the subject of missing data in LME Modelling. 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*".

4.4.1 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.

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 Nobre and Singer (2007). In Table 2 we summarize diagnostic techniques involving residuals discussed in Nobre and Singer (2007).

Chapter 5

General Appendices

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

5.1 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.

5.1.1 Estimation

$$\hat{\beta} = X^T \tag{5.1}$$

$$\hat{\gamma} = G(\hat{\theta})Z^T \tag{5.2}$$

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.

5.1.2 Zewotir-Cook's Distance

Diagnostic tool for variance components

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

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

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

For fixed effects parameter estimates in LME models, the Cook's distance can be extended to measure influence on these fixed effects.

$$\mathrm{CD}_i(\beta) = \frac{(c_{ii} - r_{ii}) \times t_i^2}{r_{ii} \times p}$$

For random effect estimates, the Cook's distance is

$$\mathrm{CD}_i(b) = g\prime_{(i)}(I_r + \mathrm{var}(\hat{b})D)^{-2}\mathrm{var}(\hat{b})g_{(i)}.$$

Large values for Cook's distance indicate observations for special attention.

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

Mean Square Prediction Error

$$MSPR = \frac{\sum(y_i - \hat{y}_i)^2}{n^*} \tag{5.3}$$

5.1.3 Leverage

Leverage can be defined through the projection matrix that results from a transformation of the model with the inverse of the Cholesky decomposition of $\boldsymbol{V}$, or an oblique projector: $\boldsymbol{Y} = \boldsymbol{H}\hat{\boldsymbol{Y}}$.

While H is idempotent, it is generally not symmetric and thus not a projection matrix in the narrow sense.

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

The trace of $\boldsymbol{H}$ equals the rank of $\boldsymbol{X}$. If V_{ij} denotes the element in row i, column j of $\boldsymbol{V}^{-1}$, then for a model containing only an intercept the diagonal elements of $\boldsymbol{H}$.

$$h_{ii} = \frac{\sum v_{ij}}{\sum\sum v_{ij}}$$

PRESS

Schabenberger (2004) descibes the use of the $PRESS$ and $DFFITS$ in determining influence.

The $PRESS$ residual is the difference between the observed value and the predicted (marginal)value.

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{5.4}$$

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.

$$PRESS = \sum(y - y^{(k)})^2 \tag{5.5}$$

$$e_{-Q} = y_Q - x_Q\hat{\beta}^{-Q}$$
$$PRESS = \sum(y - y^{-Q})^2$$
$$PRESS_{(U)} = y_i - x\hat{\beta}_{(U)}$$

PRESS Residuals and PRESS Statistic

The predicted residual sum of squares (PRESS) statistic is a form of cross-validation used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is calculated as the sums of squares of the prediction residuals for those observations.

A fitted model having been produced, each observation in turn is removed and the model is refitted using the remaining observations. The out-of-sample predicted value is calculated for the omitted observation in each case, and the PRESS statistic is calculated as the sum of the squares of all the resulting prediction errors:[4]

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

Given this procedure, the PRESS statistic can be calculated for a number of candidate model structures for the same dataset, with the lowest values of PRESS indicating the best structures. Models that are over-parameterised (over-fitted) would tend to give small residuals for observations included in the model-fitting but large residuals for observations that are excluded.

5.1.4 Local Influence

? developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted regression problem (conditional on the estimated covariance matrix) for fixed effects.

Bibliography

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

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.

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.

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

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

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

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.

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

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

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

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.

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

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

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

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

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

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

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

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI*, Volume 29, pp. 189–29.

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

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

In the graph above, you can predict non-zero values for the residuals based on the fitted value. For example, a fitted value of 8 has an expected residual that is negative. Conversely, a fitted value of 5 or 11 has an expected residual that is positive.

The non-random pattern in the residuals indicates that the deterministic portion (predictor variables) of the model is not capturing some explanatory information that is leaking into the residuals. The graph could represent several ways in which the model is not explaining all that is possible.

Possibilities include:

- A missing variable
- A missing higher-order term of a variable in the model to explain the curvature
- A missing interction between terms already in the model

Identifying and fixing the problem so that the predictors now explain the information that they missed before should produce a good-looking set of residuals.

In addition to the above, here are two more specific ways that predictive information can sneak into the residuals:

The residuals should not be correlated with another variable. If you can predict the residuals with another variable, that variable should be included in the model. In Minitabs regression, you can plot the residuals by other variables to look for this problem.

Autocorrelation

Adjacent residuals should not be correlated with each other (**autocorrelation**). If you can use one residual to predict the next residual, there is some predictive information present that is not captured by the predictors. Typically, this situation involves time-ordered observations. For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated. You can include a variable that captures the relevant time-related information, or use a time series analysis.

In Minitabs regression, you can perform the ***Durbin-Watson*** test to test for autocorrelation.

5.2 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 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.

Chapter 6

Residual Diagnostics

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.

6.0.1 Case-Deletion Diagnostics

Influence diagnostics are formal techniques that assess the influence of observations on parameter estimates for β and θ. A common technique is to refit the model with an observation or group of observations omitted.

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

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

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

6.1 Demidenk Case Deletion Diagnostics

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

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

6.2 LME diagnostic measures

6.2.1 Andrews-Pregibon statistic

- For fixed effect parameters β.

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.

6.2.2 Cook's Distance

- For variance components γ

Diagnostic tool for variance components

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

6.2.3 Variance Ratio

- For fixed effect parameters β.

6.2.4 Cook-Weisberg statistic

- For fixed effect parameters β.

6.2.5 Andrews-Pregibon statistic

- For fixed effect parameters β.

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.

Chapter 7

BA99

Chapter 8

Random Effects and MCS

8.1 Random Effects and MCS

The methodology comprises two calculations. The second calculation is for the standard deviation of means Before the modified Bland and Altman method can be applied for repeated measurement data, a check of the assumption that the variance of the repeated measurements for each subject by each method is independent of the mean of the repeated measures. This can be done by plotting the within-subject standard deviation against the mean of each subject by each method. Mean Square deviation measures the total deviation of a

8.1.1 Random coefficient growth curve model

(Chincilli 1996) Random coefficient growth curve model, a special type of mixed model have been proposed a single measure of agreement for repeated measurements.

$$\mathbf{d} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{8.1}$$

The distributional asummptions also require **d** to **N**

8.2 Random effects Model

Myles (2007) proposes the use of Random effects models to address the issue of repeated measurement.

Myles proposes a formulation of the BlandAltman plot, using the within-subject variance estimated by the random effects model, with the time of the measurement taken as a random effect. He states that *random effects models account for the dependent nature of the data, and additional explanatory variables, to provide reliable estimates of agreement in this setting.*

Agreement between methods is reflected by the between-subject variation.The Random Effects Model takes this into account before calculating the within-subject standard deviation.

8.2.1 Myers Random Effects Model

The presentation of the 95% limits of agreement is for visual judgement of how well two methods of measurement agree. The smaller the range between the two, the better the agreement is The question of small is small is a question of clinical judgement

Repeated measurements for each subjects are often used in clinical research.

8.2.2 Random Effects Modelling

Random effects models are used to examine the within-subject variation after adjusting for known and unknown variables, in which each subject has a different intercept and slope over a time period period.

Myles (2007) remarks that the random effects model is an extension of the analysis of variance method, accounting for more covariates.

A random effect (in Myles's case, time of measurement) is chosen to reflect the different intercept and slope for each subject with respect to their change of measurements over the time period.

In Myles's methodology, the standard deviation of difference between the means of the repeated measurements can be calculated based on the within-subject standard deviation estimates.

A random effects model (also variance components model)is a type of hierarchical linear model. Hierarchical linear modelling (HLM) is a more advanced form of simple linear regression and multiple linear regression. HLM is appropriate for use with nested data.

Faraway comments that the random effects approach is *more ambitious than the LME model in that it attempts to say something about the wider population beyond the particular sample.*

8.3 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.

8.4 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. Agreement and Method Comparison Studies
   (a) What is Agreement?
   (b) Repeatability
   (c)
   (d)
   (e)
2. Bland Altman Single Observations
   (a)
   (b)
3. Alternative Methods
   (a) Deming Regression
   (b) Mountain Plot
   (c) Bartko's Ellipse

    (d) Formal Tests and Procedures

4. Replicate Observations

5. LME models

6. Estimation and Algorithms

    (a) ML and REML estimation

    (b) MINQUE

    (c)

7. Residual Diagnostics

    (a) Marginal and Conditional Diagnostics

    (b) Scaled Residuals

8. Influence Diagnostics

    (a) Underlying Concepts

    (b) Managing the Covariance Parameters

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

    (d) Leverage

    (e) Internally and Externally Studentized Residuals

    (f) DFFITs and MDFFITs

    (g) Covariance Ratio and Trace

    (h) Likelihood Distance

    (i) Non-iterative Update Procedures

8.5 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
  -
  -

8.6 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.

8.6.1 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, speciifically 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.

8.7 Computation and Notation

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

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

8.8 Lai Shiao

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

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

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

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

Single fixed effect

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

$$\begin{aligned}\text{fixed effects}: \ & 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \quad (8.2)\\ (\text{p-values}: \ & = 0.0054, < 0.0001, < 0.0001)\end{aligned}$$

$$\begin{aligned}\text{random effects}: \ & u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma^2_e = 0.1525, \rho = 0.6978)\\ (\text{p-values}: \ & = 0.8113, < 0.0001, < 0.0001)\end{aligned}$$

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

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

Two fixed effects

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

$$\begin{aligned}\text{fixed effects}: \ & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl}\\ (\text{p-values}: \ & = 0.8113, < 0.0001, < 0.0001)\end{aligned}$$

$$\begin{aligned}\text{random effects}: \ & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma^2_e = 0.0920, \rho = 0.5577)\\ (\text{p-values}: \ & = 0.0446, < 0.0001, < 0.0001) \quad (8.3)\end{aligned}$$

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

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

8.9 Liao Shaio

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

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

Joseph G. Voelkel Bruce E. Siskowski

8.10 Hamlett and Lam

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

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`

8.10.1 Inappropriate Techniques for MCS

8.10.2 Links and Papers

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

Measurement Systems Analysis

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

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

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

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

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

8.11 Escaramis

8.11.1 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.

8.11.2 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.

8.11.3 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.

8.11.4 Conclusions

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

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

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

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

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

Tolerance Interval around the TDI estimate

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

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

8.12 Schabenberger

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

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

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

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

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

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

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

schab remarks that the concept of critiquing the model-data agreement applies in mixed models in the same way as in linear fixed-effects models. In fact, because of the more complex model structure, you can argue that model and data diagnostics are even more important. For example, you are not only concerned with capturing the important variables in the model. You are also concerned with “distributing them correctly between the fixed and random components of the model. The mixed model structure presents unique and interesting challenges that prompt us to reexamine the traditional ideas of influence and residual analysis.

8.13 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”.

8.14 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.

8.14.1 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.

8.14.2 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.

8.15 Work List

1. ML v REML

2. Nested Models and LRTs

3. Generalized Lease Squares

4. Diagnostics

5. Simplifying GLS

6. Paper progression

8.16 Diagnostics

8.16.1 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.

8.16.2 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)

8.17 Covariance Parameters

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

8.18 Carstensen Model (mir model)

A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m) \tag{8.4}$$

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

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

Using Carstensen's notation, a measurement y_{mi} by method m on individual i the measurement y_{mir} is the rth replicate measurement on the ith item by the mth method, where $m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m), \quad c_{mi} \sim \mathcal{N}(0, \tau^2_m). \tag{8.5}$$

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{8.6}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = d^2_m$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = d_{12}$.

The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma^2_m$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, d^2_1 andd^2_2, and σ^2_1 andσ^2_2.

Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2\colon \sigma^2_1 = \sigma^2_2$ and $\mathrm{H}_3\colon d^2_1 = d^2_2$ hold simultaneously. Roy (2009b) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Additionally, Roy combines H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4\colon \omega^2_1 = \omega^2_2$, where $\omega^2_m = \sigma^2_m + d^2_m$ represent the overall variability of method m.

Here the terms α_m and μ_i represent the fixed effect for method m and a true value for item i respectively. The random effect terms comprise an interaction term c_{mi} and the residuals ε_{mir}. The c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau^2_m$.

Carstensen specifies the variance of the interaction terms as being univariate normally distributed.

As such, $\text{Cov}(c_{mi}, c_{m'i}) = 0$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

The presence of the true value term μ_i gives rise to an important difference between Carstensen's and Roy's models. Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. In other words, Roy considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

8.19 Carstensen's Mixed Models

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

$$y_{mir} = \alpha_m + \beta_m\mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma^2_m), c_{mi} \sim N(0, \tau^2_m)) \tag{8.7}$$

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

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

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

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{8.8}$$

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

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

8.19.1 Using LME models to create Prediction Intervals

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma^2_m)) \tag{8.9}$$

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma^2_m), c_{mi} \sim N(0, \tau^2_m)) \tag{8.10}$$

Chapter 9

Model Diagnostics

Contents

Abstract

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

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

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

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.

9.1 Matrix Notation for Case Deletion

9.1.1 Case deletion notation

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

9.1.2 Further Assumptions of Linear Models

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

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

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

The basic procedure for quantifying influence is simple:

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

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

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

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

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

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

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

9.1.3 Summary of Paper

Standard residual and influence diagnostics for linear models can be extended to LME models. The dependence of the fixed effects solutions on the covariance parameters has important ramifications on the perturbation analysis. Calculating the studentized residuals-And influence statistics whereas each software procedure can calculate both conditional and marginal raw residuals, only SAs Proc Mixed is currently the only program that provide studentized residuals Which ave preferred for model diagnostics. The conditional Raw residuals ave not well suited to detecting outliers as are the studentized conditional residuals. (schabenbege r)

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

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

9.2 Schabenberger: Summary and Conclusions

- Standard residual and influence diagnostics for linear models can be extended to linear mixed models. The dependence of fixed-effects solutions on the covariance parameter estimates has important ramifications in perturbation analysis.
- To gauge the full impact of a set of observations on the analysis, covariance parameters need to be updated, which requires refitting of the model.
- The experimental INFLUENCE option of the MODEL statement in the MIXED procedure (SAS 9.1) enables you to perform iterative and noniterative influence analysis for individual observations and sets of observations.
- The conditional (subject-specific) and marginal (population-averaged) formulations in the linear mixed model enable you to consider conditional residuals that use the estimated BLUPs of the random effects, and marginal residuals which are deviations from the overall mean.
- Residuals using the BLUPs are useful to diagnose whether the random effects components in the model are specified correctly, marginal residuals are useful to diagnose the fixed-effects components.
- Both types of residuals are available in SAS 9.1 as an experimental option of the MODEL statement in the MIXED procedure.

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

Leave-One-Out Diagnostics with `lmeU`

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

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

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

The addition of an extra factor

Interaction terms are featured in ANOVA designs.

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

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

9.3 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.

9.4 Covariance Parameters

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

Chapter 10

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 first 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 first 24 women in the first 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 first three variables, that is, the variables in the first location ($u = 1$) and then the mineral contents for the non-dominant side of the same bones ($u = 2$)

10.1 Outlier Testing

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

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

10.2 Lorelia

Method comparison studies are performed in order to prove equivalence between two measurement methods or instruments. The identification of outliers is an important part of data analysis as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in method comparison studies usually do not meet these assumptions. To overcome this problem, different data transformation methods are proposed in the literature. However, they will only be applicable if the random errors can be described by simple additive or multiplicative models. In this work, a new outlier test based on robust linear regression is proposed which provides a general solution to the above problem. The LORELIA (LOcal RELIAbility) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. Outlier limits are estimated from the actual data situation without making assumptions on the underlying error variance model. The performance of the new test is demonstrated in examples and simulations.

10.3 Note on Roy's paper

1. Basic model:

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

   Assumptions are made about homoskedasticity.

2. General model:

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

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

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

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

$$\boldsymbol{y}_i = (y_i^{j1}, y_i^{Jj2}, y_i^{j3}, y_i^{s1}, y_i^{s2}, y_i^{s3})\prime \tag{10.1}$$

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

10.3.1 Outliers and Leverage

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

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

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

10.3.2 Matrix Notation for Case Deletion

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

10.3.3 Extension of Diagnostic Methods to LME models

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

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

Christensen et al. (1992) noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. Christensen et al. (1992) develops these techniques in the context of REML. ¿¿¿¿¿¿¿ origin/master

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

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

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

10.4 Regression Of Differences On Averages

Further to Carstensen, we can formulate the two measurements y_1 and y_2 as follows:
$y_1 = \alpha + \beta\mu + \epsilon_1$
$y_2 = \alpha + \beta\mu + \epsilon_2$

10.4.1 Note 1: Coefficient of Repeatability

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

10.4.2 Note 2: Carstensen model in the single measurement case

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m) \tag{10.2}$$

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.

10.4.3 Note 3: Model terms

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

- Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.

- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.

- $\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.

- $\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.

- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.

- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.

- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.

- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)

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

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

Four candidates models are fitted to the data.

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

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

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

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

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

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

Carstensen's model is a mixed effects ANOVA.

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

This model includes a method by item iteration term.

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

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

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

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

10.4.4 Note 3: Model terms

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

- Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.

- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.

- $\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.

- $\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.

- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.

- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.

- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.

- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)

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

10.5 Regression Of Differences On Averages

Further to Carstensen, we can formulate the two measurements y_1 and y_2 as follows:
$y_1 = \alpha + \beta\mu + \epsilon_1$
$y_2 = \alpha + \beta\mu + \epsilon_2$

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

Appendix to Section 4

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

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

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

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

Important Consideration for MCS

The key issue is that `nlme` allows for the particular specification of Roy's Model, 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.

$$r_{mi} = x_i^T \hat{\beta} \tag{10.3}$$

10.5.1 Marginal Residuals

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

10.6 Covariance Parameters

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

10.6.1 Methods and Measures

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

Zewotir and Galpin (2005) 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.

10.7 Missing Data in Method Comparison Studies

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

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

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

10.8 Leave-One-Out Diagnostics with `lmeU`

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

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

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

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

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

Chapter 11

Model Diagnostics

11.1 Carstensen

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{11.1}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The β terms can be gathered together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g^2_m$ and $\mathrm{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma^2_m$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \beta_1 = \beta_2$ and $\mathrm{H}_2\colon \sigma^2_1 = \sigma^2_2$ and $\mathrm{H}_3\colon g^2_1 = g^2_2$ hold simultaneously. Roy (2009a) proposes a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Let $\omega^2_m = \sigma^2_m + g^2_m$ represent the overall variability of method m. Roy also integrates H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4\colon \omega^2_1 = \omega^2_2$. CONCERNS?

? demonstrates how to implement a method comparison study further to model (1) using the SAS proc mixed package. Carstensen et al. (2008) demonstrates how to construct limits of agreement using

SAS, STATA and R. In the case of SAS, the PROC MIXED procedure is used. Implementation in R is performed using the nlme package (**?**).

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

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

The R statistical software package is freely available.

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

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

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

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

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

11.1.1 Remarks on the Multivariate Normal Distribution

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

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

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

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

with k-dimensional mean vector

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

and $k \times k$ covariance matrix

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

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

   (a)

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

   (b)

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

11.2 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.*

11.2.1 Matrix Notation for Case Deletion

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

11.3 Extension of Diagnostic Methods to LME models

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

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

Christensen et al. (1992) noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. Christensen et al. (1992) develops these techniques in the context of REML.

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

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

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

11.4 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)

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

11.4.1 Remarks on the Multivariate Normal Distribution

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

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

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

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

with k-dimensional mean vector

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

and $k \times k$ covariance matrix

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

1. Univariate Normal Distribution

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

2. Bivariate Normal Distribution

(a)

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

(b)

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

Bibliography

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

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.

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.

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

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

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

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.

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

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

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

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.

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

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

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

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

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

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

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

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI*, Volume 29, pp. 189–29.

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

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