Chapter 1

Residual Analysis and Influence Diagnostics for Method Comparison

Model validation and model appraisal are vital parts of the modelling process, yet are too often overlooked. Using a small set of simple measures and methods, such as the AIC and R^2 measures, is insufficient to properly assess the usefulness of a fitted model. In classical linear models model diagnostics are now considered a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. A full and comprehensive analysis that comprises residual analysis and influence analysis for testing model assumptions, should be carried out. However it has been noted by several papers (Christensen et al., 1992; Schabenberger, 2004) that model diagnostics do not often accompany LME model analyses. Furthermore, a suite of dagnostic procedures designed for method comparison should be adopted.

1.1 Residual Analysis

Model diagnostics techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations, and have been become a required part of any statistical analysis. Well established methods are commonly

available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers that model diagnostics do not often accompany LME model analyses.

A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. As with classical models, there are two key techniques for LME models: a residual plot and the normal probability plot. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect. If the residuals behave randomly, with no discernible trend, the model has fitted the data well. Conversely, if some sort of non-random trend is evident in the model, then the model can be considered to be poorly fitted.

The underlying assumptions for LME models are similar to those of classical linear models. However, for LME models the matter of residuals are more complex, both from a theoretical point of view and from the practicalities of implementing a comprehensive analysis using statistical software. Schabenberger (2004) discusses residuals for LME model, providing a useful summary of various techniques. Prominent in literature is the taxonomy of residuals for LME Models, distinguishing between condition residuals, marginal residuals and EBLUPS, including ?Schabenberger (2004); West et al. (2007); Nobre and Singer (2007).

Statistical software environments, such as the `R` programming language, provides a suite of tests and graphical procedures for appraising a fitted LME model, with several of these procedures analysing the model residuals. Texts such as Pinheiro and Bates (1994); West et al. (2007); Gałecki and Burzykowski (2013) describe what can be implemented for LME residual analyses with statistical software, such as `R` and `SAS`.

In the context of method comparison, a residual analysis would be carried out just as any other LME model would, testing normality. There is little scope for adding additional insights, other than to say that it is possible to create plots specific to each method. The figures on the next page depict the residual analysis for the *Blood* data, which can be used to indicate which methods disagree with the rest, but these would

be a confirmation of something detected previously.

Analysis of the residuals could determine if the methods of measurement disagree systematically, or whether or not erroneous measurements associated with a subset of the cases are the cause of disagreement. The figure depicts residual plot for the Systolic Blood Pressure example used in Bland and Altman (1999). Points are labelled by subjects, with cases 67, 68 and 71 being among the prominent cases. Prominent cases warrant further investigation, but an analyst should procede to influence diagnostics beforehand.

The next figure depicts residual plot for the Systolic Blood Pressure example, panelled by the various measurement methods. It serves to confirm agreement between methods J and R, with lack of agreement between those two methods and method S. However, little insight can be gained as to what actually causes lack of agreement here.

1.2 Influence Diagnostics

Model diagnostic techniques can determine whether or not the distributional assumptions are satisfied, but also to assess the influence of unusual observations. Following model specification and estimation, it is of interest to explore the model-data agreement by raising pertinent questions. **?** provide some insight into how to compute and interpret model diagnostic plots for LME models. Unfortunately this aspect of LME theory is not as expansive as the corresponding body of work for Linear Models. Their particular observations will be reverted to shortly. Further to the analysis of residuals, Schabenberger (2004) recommends the examination of the following questions:

- Does the model-data agreement support the model assumptions?
- Should model components be refined, and if so, which components? For example, should certain explanatory variables be added or removed, and is the covariance of the observations properly specified?
- Are the results sensitive to model and/or data? Are individual data points or

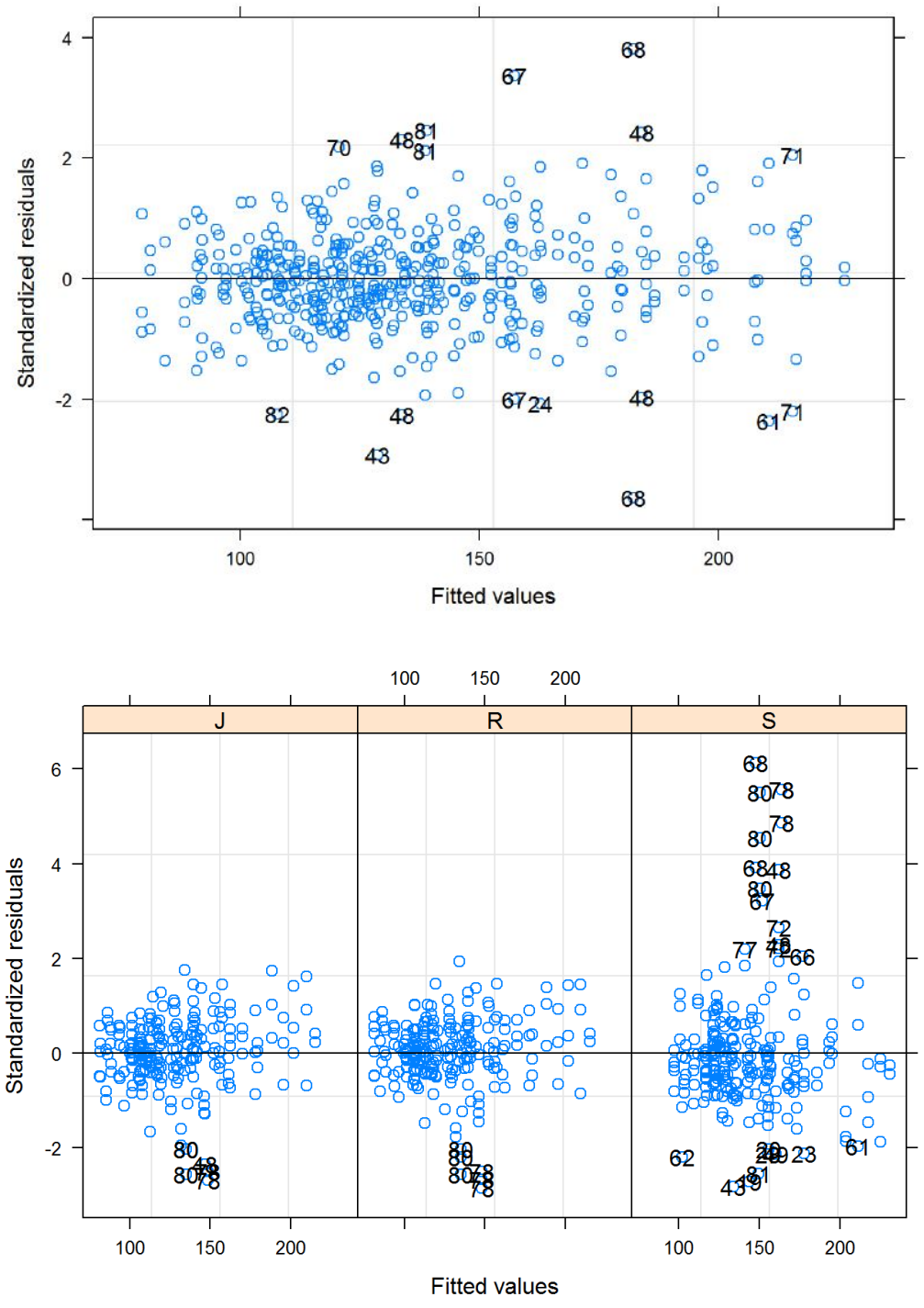


Figure 1.1.1: LME Residuals by Method (Blood Pressure Data)

groups of cases particularly influential on the analysis?

The last of these three questions, regarding influential points, is of particular interest in the context of Method Comparison. After fitting an LME model, it is important to carry put model diagnostics to check whether distributional assumptions for the residuals as satisfied and whether the fit the model is sensitive to unusual assumptions. The process of carrying out model diagnostic involves several informal and formal techniques, which will mentioned throughout the chapter.

Influential points have a large influence on the fit of the model. Influential points are a set of one or more observations whose removal would cause a different conclusion in the analysis, e.g. substantially changes the estimate of the regression coefficients. West et al. (2007) remarks that influence diagnostics play an important role in the interpretation of results, because influential data can negatively influence the statistical model and generalizability of the model. Schabenberger (2004) 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 (West et al., 2007).

1.2.1 A Procedure for Quantifying Influence

Influence can be thought of as consequence of leverage and outlierness. 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. They can point to a model breakdown and lead to development of a better model. Schabenberger (2004) describes a simple procedure for quantifying influence for LME Models. Firstly a model should be fitted to the data, and estimates of the parameters should be obtained. The second step is that either single or 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"* or *"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.

The LME model is a useful methodology for fitting a wide range of models. However, linear mixed effects models are known to be sensitive to outliers. Specifically likelihood based estimation techniques, such as ML and REML, are sensitive to outliers. Christensen et al. (1992) advises that identification of outliers is necessary before conclusions may be drawn from the fitted model. The leverage of an observation is a further consideration.

An observation with an extreme, but not unusual, 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. In general. a high leverage point means a extreme value for the one or more of the independent variables, and a greater potential of overly influencing the final fitted model. However, if a case has extreme values for the independent variables but is fitted very well by a regression model, this case is not necessarily overly influential.

In classical models, leverages are the diagonal elements h_{ii} of the Projection matrix, also known as the Hat Matrix H. Schabenberger (2004) describes two analogues of H for LME models. However the practical use for either approach is not made clear.

1.2.2 Analyzing Influence in LME Models

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

Influence diagnostics are formal techniques allowing for the identification of observations that exert substantial influence on the estimates of fixed effects and variance covariance parameters. 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. However influence diagnostics for LME Models is an

area of active research. Research on diagnostic analyses for LME models are presented in **?**, Christensen et al. (1992), **?**, Lesaffre and Verbeke (1998), **?**, **?**, Demidenko (2004), Zewotir and Galpin (2005), Zewotir (2008) and Nobre and Singer (2007, 2011).

Schabenberger (2004) states that goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis.

1.2.3 Measuring of Influence for LME Models

Influence analysis methodologies have been used extensively in classical linear models, and provided the basis for methodologies for use with LME models. Computationally inexpensive diagnostics tools have been developed to examine the issue of influence (Zewotir and Galpin, 2005).

Zewotir and Galpin (2005) remarks the development of efficient computational formulas is crucial making deletion diagnostics useable, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model. A number of approaches to model diagnostics are described, including variance components, dixed effects parameters, prediction of the response variable and of random effects, and the likelihood function. Influence statistics can be grouped by the aspect of estimation that is their primary target:

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

- **outlier properties**: internally and externally studentized residuals, leverage

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, and the Andrews-Prebigon statistic.

The subscript (U) is used to denote quantities computed from data with subset of cases U omitted. 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

For example, if observations 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. Schabenberger (2004) notes that removing observations or sets of observations affects fixed effects and covariance parameter estimates.

1.2.4 Deletion Diagnostics

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on parameters inferences for a fitted model. For classical linear models, Cook (1977) greatly expands the study of residuals and influence 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. Cook's key observation was the effects of deleting each observation in turn could be calculated with little additional computation. Cook proposed a measure that combines the information of leverage and residual of the observation, now known simply as the Cook's Distance, $D_{(i)}$, which can be calculated without fitting a new regression coefficient each time an observation is deleted. Consequently deletion diagnostics have become an integral part of assessing linear models.

It must be pointed out that the effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's distance, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large.

Christensen et al. (1992) notes the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. Christensen et al. (1992) developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted resgression problem (conditional on the estimated covariance matrix) for fixed effects.

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

1.2.5 Cook's Distance

As previously described, Cooks Distance (D_i) is a diagnostic technique used in classical linear models, that functions as an overall measure of the influence of an observation that is a measure of aggregate impact of each observation on the group of regression

coefficients, as well as the group of fitted values. Cook's Distance as a measure of the influence of observations in subset U on a vector of parameter estimates is given below (Cook, 1977)

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)}.$$

Observations, or sets of observations, that have high Cook's distance usually have high residuals, although this is not necessarily the case.

If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Large values for Cook's distance indicate observations for special attention. Cook's distance can be used in several ways: to indicate data points that are particularly worth checking for validity; to indicate regions of the design space where it would be good to be able to obtain more data points.

Use of threshold values for Cook's Distance is discouraged (**?**). However, informal heuristics do exist for OLS models; Obervations for which Cook's distance is higher than 1 are usually considered as influential. Another informal threshold of $4/n$ or $4/(n-k-1)$, where n is the number of observations and k the number of explanatory variables.

? advises the use of diagnostic plotting and to examine in closer details the points with *"values of D that are substantially larger than the rest"*, and that thresholds should feature only to enhance graphical displays.

The effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's distance, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large.

Christensen et al. (1992) develops case deletion diagnostics, in particular the equivalent of Cook's distance for diagnosing influential observations when estimating the fixed effect parameters and variance components, adapting the Cook's Distance measure for the analysis of LME models. For LME models, two formulations exist; a Cook's

distance that examines the change in fixed fixed parameter estimates, and another that examines the change in random effects parameter estimates. The outcome of either Cook's distance is a scaled change in either β or θ. Zewotir and Galpin (2005) gives a detailed discussion of the various formulation for Cook's distances for LME Models.

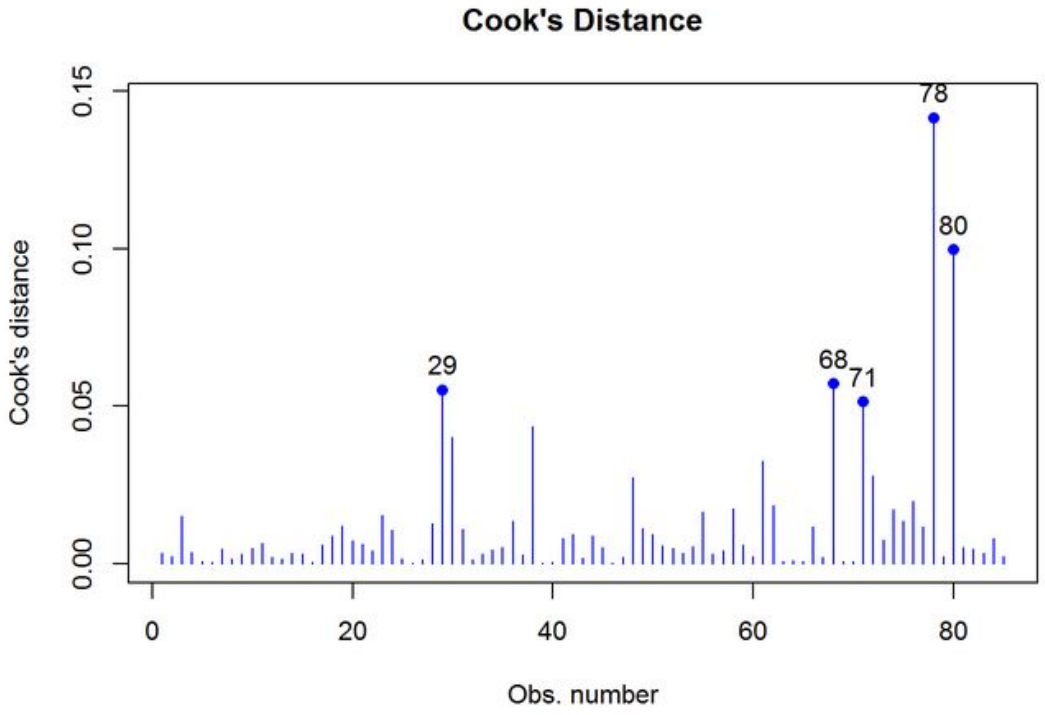


Figure 1.2.2: Cook's Distance Plot for the JS Comparison

Consideration of how leave-U-out diagnostics would work in the context of Method Comparison problems is required. There are several scenarios. Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called '*observation-diagnostics*'. For multiple observations, Preisser describes the diagnostics as '*cluster-deletion*' diagnostics. Suppose we have two methods of measurement X and Y, each with three measurements for a specific case: $(x_1, x_2, x_3, y_1, y_2, y_3)$

- Leave One Out - one observation is omitted (e.g. x_1)
- Leave Pair Out - one pair of observation is omitted (e.g. x_1 and y_1)

- Leave Case (or Item or Subject) Out - All observations associated with a particular case or subject are omitted. (e.g. $\{x_1, x_2, x_3, y_1, y_2, y_3\}$)

The natural sampling unit is the item or subject, similar to the example provided by Schabenberger (2004). Hence, the third option, henceforth, referred to as "Leave subject Out" will be the option used.

1.2.6 Local Influence

Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models, introducing methods for local influence assessment for classical linear models. These methods provide a powerful tool for examining perturbations in the assumption of a model, particularly the effects of local perturbations of parameters of observations. The local-influence approach to influence assessment is quite different from the case deletion approach, comparisons are of interest.

? applied the local influence method of Cook (1986) to the analysis of the LME model. Other authors such as Lesaffre and Verbeke (1998) have also extended these idea to LME models. While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known. As such the local influence approach are not particularly useful in the context of Method Comparison, and so will not be considered further.

1.2.7 Comparing Influence and Residual Analysis

? compares residual analysis and influence analysis. Cases with high residuals (defined as the difference between the observed and the predicted scores on the dependent variable) or with high standardized residuals (defined as the residual divided by the standard deviation of the residuals) are indicated as outliers.

However, an influential case is not necessarily an outlying residual. On the contrary: a strongly influential case dominates the regression model in such a way, that the estimated regression line lies closely to this case. The analysis of residuals cannot be used for the detection of influential cases (**?**).

1.2.8 Iterative and Non-Iterative Influence Analysis

For linear models, the implementation of influence analysis is straightforward, but for LME models the process is more complex. Schabenberger (2004) examines the use and implementation of influence measures in LME models. Schabenberger (2004) highlights some of the issue regarding implementing LME model diagnostics, describing the choice between iterative influence analysis and non-iterative influence analysis. Schabenberger (2004) considers several important aspects of the use and implementation of influence measures in LME models, noting that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates. Closed-form expressions for computing the change in important model quantities might not be available.

On a related matter, Schabenberger (2004) describes the scenario wherein a data point is removed and the new estimate of the D matrix is not positive definite. This may occur if a variance component estimate now falls on the boundary of the parameter space (Schabenberger, 2004).

For classical linear models, it is not necessary to refit the model after removing a data point in order to measure the impact of an observation on the model. The change in fixed effect estimates, residuals, residual sums of squares, and the variance-covariance matrix of the fixed effects can be computed based on the fit to the full data alone, using update formulas (**?**Hager, 1989).

However, in LME models several important complications arise. Data points can affect not only the fixed effects but also the covariance parameter estimates on which the fixed-effects estimates depend.

When applied to LME models, such update formulas are available only if one assumes that the covariance parameters are not affected by the removal of the observation in question. However, this is rarely a reasonable assumption. For LME models, non-iterative methods are computationally efficient, but require the rather strong assumption that all covariance parameters are known, and thus are not updated, with the exception of the profiled residual variance. Update formulas for "leave-U-out" estimates typically fail to account for changes in covariance parameters. As the influence that each item would have on the variance estimate of a method comparison model is crucial, this substanitally negates their usefulness for Roy's Model.

Iterative influence diagnostics requiring fitting the model without the observations in question. Computation execution time is substantially longer, although this is balanced by algorithmic simplicity, with no assumptions beyond those used for the original model. A measure of total influence requires updates of all model parameters. This can only be achieved in general is by omitting observations or cases, then refitting the model.

An iterative analysis may seem computationally expensive. Computing iterative influence diagnostics for n observations requires $n + 1$ mixed models to be fitted iteratively. The execution times for iterative procedures are longer relative to non-iterative procedures, but are not so long that they would dissuade an analyst from using them. Despite the addition execution time of iteratives approaches, they are preferable for Method Comparison problems, as they can facilitate several complementary analyses concurrently.

Iterative methods retain the potential for useful analyses, if applied at different stage of the modelling process. Diagnostic measures, specifically the DFBETA, have characteristics that would make them very useful at the exploratory stage of the method comparison process. Implicitly various assumptions about variance are used, but simultaneously an approach based on DFBETA can be used to assess if these assumptions are valid.

1.2.9 Likelihood Distance

An overall influence statistic measures the change in the objective function being minimized. For example, in classical linear, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance (Cook and Weisberg, 1983).

The likelihood distance is a global summary measure that expresses the joint influence of the subsets of observations, U, on all parameters that were subject to updating. Schabenberger (2004) points out that the likelihood distance $LD(\psi_{(U)})$ is not the log-likelihood obtained by fitting the model to the reduced data set. Instead it is obtained by evaluating the likelihood function based on the full data set (containing all n observations) at the reduced-data estimates.

The procedure requires the calculation of the full data estimates $\hat{\psi}$ and estimates based on the reduced data set $\hat{\psi}_{(U)}$. The likelihood distance is given by determining

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

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

Large values indicate that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ differ considerably.

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

1.3 Model Diagnostics for Roy's Models

Further to previous work, this section revisits case-deletion and residual diagnostics, and explores how approaches devised by Gałecki and Burzykowski (2013) can be used to appraise Roy's model. These authors specifically look at Cook's Distances and Likelihood Distances.

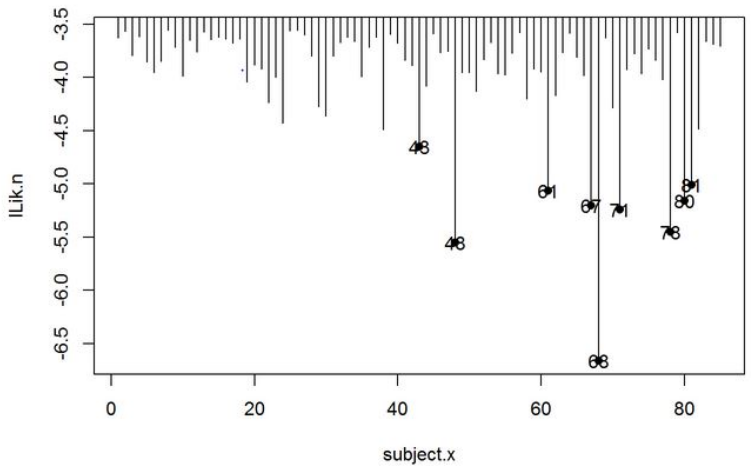


Figure 1.3.3:

Case Deletion Diagnostics for Variance Ratios

Taking the core principals of his methods, and applying them to the Method Comparison problem, case deletion diagnostics are used on the variance components of the Roy's model, specifically the ratio of between subject variances and the within subject covariances respecitvely.

$$\text{BSVR} = \frac{\sigma_2^2}{\sigma_2^2} \qquad\qquad \text{WSVR} = \frac{d_2^2}{d_2^2}$$

These variance ratios are re-computed for each case removed, and may be analysed seperately or jointly for outliers.

The Grubbs' Test for Outliers is a commonly used technique for assessing outlier in a univariate data set, of which there are several variants. As there may be several outliers present, the Grubbs test is not practical. However an indication that a point being beyond the fences according to Tukey's specification for boxplots will suffice.

The WSVR values are plotted against the corresponding BSVR values, with commonly used bivariate methods may be applied jointly to the both sets of data sets, e.g Mahalanobis distances. Confidence ellipses can be superimposed over the plot with minimal effort. Two ellipses are generated by this technique, a 50 % and 97.5% confi-

dence ellipse respectively. Outlying cases are idenified by the plot. Subject 68 is the most prominent case.

The subjects were ranked by Mahalanobis distance, with the top 10 being presented in the following table. Both sets of ratio are addtionally expressed as a ratio of the full model variance ratios.

Subject (u)	MD	$WSVR_{(u)}$	WSVR (%)	$BSVR_{(u)}$	BSVR (%)
68	44.7284	1.3615	0.9132	1.0353	0.9849
30	16.7228	1.5045	1.0092	1.1024	1.0487
71	11.5887	1.5210	1.0202	1.0932	1.0400
80	11.0326	1.4796	0.9925	1.0114	0.9621
38	10.3671	1.5011	1.0069	1.0917	1.0385
67	10.1940	1.4308	0.9598	1.0514	1.0002
43	7.6932	1.4385	0.9649	1.0511	0.9999
72	4.7350	1.4900	0.9995	1.0262	0.9762
48	4.4321	1.4950	1.0028	1.0280	0.9779
29	4.3005	1.4910	1.0001	1.0769	1.0244

From this table one may conclude that subjects 72, 48 and 29 are not particularly influential. Interestingly Subject 78, which was noticeable in the case deletion diagnostics for fixed effects, does not feature in this table.

Variance Ratios

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.

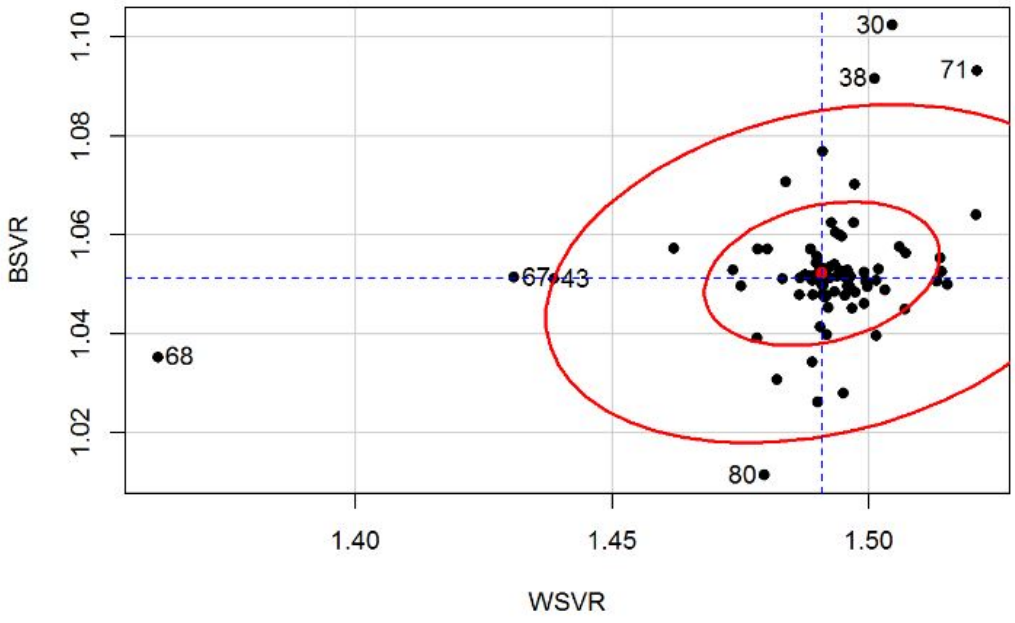


Figure 1.3.4:

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. Bootstrap methods for computing confidence intervals may be considered.

1.4 Using DFBETAs from LME Models to Assess Agreement

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the

estimated coefficient when the particular observation is deleted. DFBETA and DFFITS are well known measures of influence. Emphasis shall be placed on DFBETA, but a brief discussion of DFFITS is merited as it potentially provides for useful techniques in method comparison. Schabenberger (2004) provides a mathematical desciption of both.

DFBETAS is a standardized measure of the absolute difference between the estimate with a particular case included and the estimate without that particular case,, thus measuring the impact each observation has on a particular predictor (**?**). For LME models, the DFBETA is a measure that standardizes the absolute difference in parameter estimates between an LME model based on a full set of data, and a model from reduced data.

In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. **?** recommend 2 as a general cutoff value to indicate influential observations and as a size-adjusted cutoff. There is no agreement as to the critical threshold for DFBETAs. The cut-off value for DFBETAs is $\frac{2}{\sqrt{n}}$, where n is the number of observations. However, another cut-off is to look for observations with a value greater than 1.00. Here cutoff means, "this observation could be overly influential on the estimated coefficient".

DFFITS is a diagnostic meant to show how influential a point is in a statistical regression. It is defined as the change, in the predicted value for a point, obtained when that point is left out of the regression, divided by the estimated standard deviation of the fit at that point:

DFBETAs for Method Comparison

For LME models, a value for DFBETAS is calculated for each of the k fixed effects, and for each of the n item. Correctly there will be $p+1$ DFBETAs (the intercept, β_0, and one β for each covariate). When the LME model is specified without an intercept term, as in Roy's Model, there is a set of DFBETAs corresponding to each measurement method, hence an $n \times p$ matrix.

In the case of method comparison studies, a series of scatterplots can be constructed to compare each pair of measurement methods. Furthermore 95% confidence ellipse can be constructed around these scatterplots.

The LME approach proposed by Roy (2009) is constrained by computational tractability. Consequently a simpler LME formulation is used, one similar to that of Carstensen et al. (2008). However one constraint that can be dispensed with is the restriction to two methods of measurement: we can now use any number of methods. The benefit of using this model is that metrics such as Cook's Distance and DFBETAs can be computed also.

Furthermore, these measures form the basis of the analysis, rather than the estimates derived from the model. In the context of method comparison, these variables are the methods of measurement. Agreement will considered in the context of inter-method bias and the within-item variance ratio. Between-item variance ratio is not considered for this analysis.

For a Method Comparison study, DFBETAs can be used as a proxy measurement, allowing simple techniques to be used for assessing agreement. Suppose an LME model was formulated to model agreement for two or more methods of measurement, specifically with replicate measurements. If the methods are to be agreement, the DFBETAs for each case would be the same for both methods. As such, agreement between any two methods can be determined by a simple scatterplot of the DFBetas.

If the lack of agreement is caused, in part or in full, by differing within-item variances, there would be differing DFBETAs for each pair of methods. If the points align along the line of equality, then both methods can be said to be in agreement for within item variance. However DFBETAs are not useful for determining inter-method bias. If there is good agreement between methods, or if lack of agreement is caused by inter-method bias only, the DFBETA values will be almost identical for each subject in the data set.

Following the idea proposed by Bland and Altman (1986), an identity plot to visually inspect this relationship between sets of DFBETAs. Modern statistical software

usually allows for the creation of co-plots, so a grid of identity plots may be easily rendered for comparing each pair of methods. Used in conjection with a Bland-Altman plot, this co-plot can quickly determine agreement and indicate the source of lack of agreement.

For an LME model fitted to the Blood data, the results tabulated below can be produced. Cases can be ranked by the Cook's Distance, such that the most divergent DFBETA are highlighted, with the top 6 being presented below). The remaining columns are the DFBeta for each of the fixed effects, for each of the 85 subject.

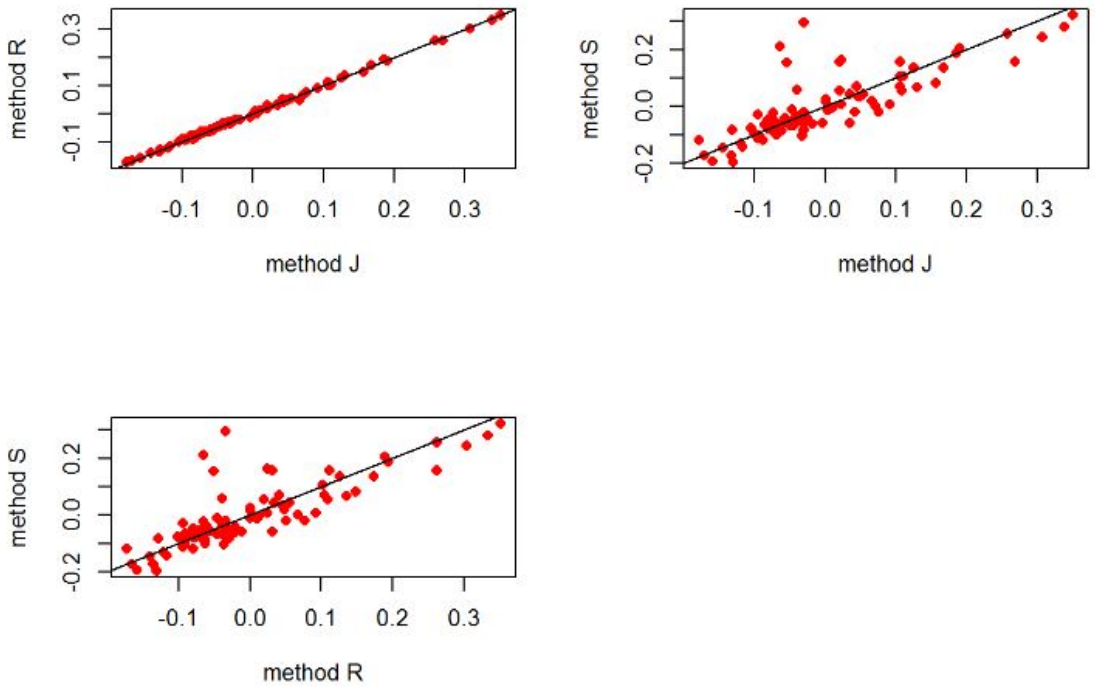
Subject	Cook's D	Method J	Method R	Method S
78	0.61557407	-0.02934556	-0.03387780	0.2954937
80	0.41590973	-0.06305026	-0.06515241	0.2123881
68	0.22536651	-0.05334867	-0.05062375	0.1555187
72	0.09348500	0.02388626	0.02419887	0.1617474
48	0.08706988	0.02147541	0.03145273	0.1581591
30	0.07118415	0.26925807	0.26215970	0.1581569

For DFBETA identity plots are presented below. This set of plots indicate agreement between methods J and R in terms of within-item variance, while severe lack of agreement exists between these methods and the third method S, as is the conclusion of Roy (2009).

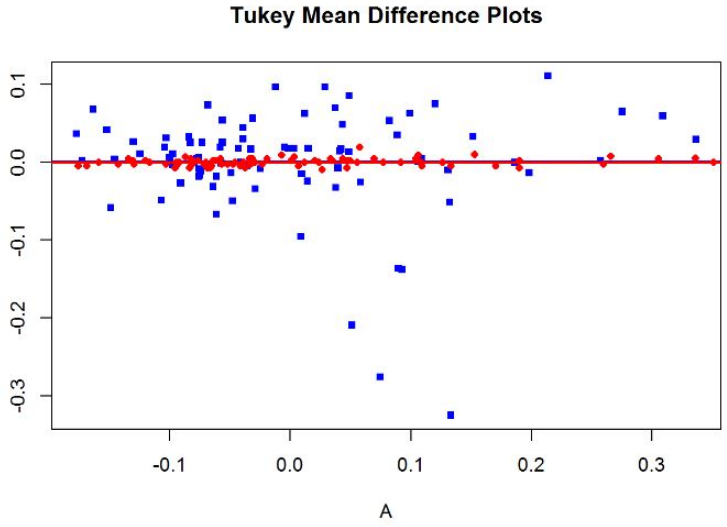
If lack of agreement is indicated, a subsequent analysis using a technique proposed by Roy (2009) can be used to identify the specific cause for this lack of agreement.

Other analyses may be used to complement these plots. The Pearson Correlation coefficient of the DFBETAs can be used in conjection with this analysis. A high correlation confirms good agreement, thouhg no threshold value for agreement is suggested.

The Bonferroni Outlier Test and Cook's Distance values can be used to identify unusual cases, when the relationship between sets of DFBETA is modelled as a (classical) linear model. In this model, the covariates should be homoskedastic. A test for non-constant variance may be used to verify this.



As an alternative to scatterplots, a mean difference plot could be used to assess agreement of with-item variance. This mean-difference plot differs from the Bland-Altman plot in that the plot is denominated in terms of DFBETA values, and not in measurement units. Here two of the three pairs of methods are compared on the same plot, red points indicate the J-R comparison while blue points are for the J-S comparison.


Tukey Mean Difference Plots
0.1
0.0
-0.1
-0.2
-0.3
-0.1
0.0
0.1
0.2
0.3
A

Bibliography

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

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

Carstensen, B., 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.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics 19*, 15–18.

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

Cook, R. D. and S. Weisberg (1983). Diagnostics for heteroscedasticity in regression. *Biometrika 70*(1), 1–10.

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

Gałecki, A. and T. Burzykowski (2013). *Linear mixed-effects models using R: A step-by-step approach*. Springer Science & Business Media.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review 31*(2), 221–239.

Lesaffre, E. and G. Verbeke (1998). Local influence in linear mixed models. *Biometrics*, 570–582.

Nobre, J. S. and J. M. Singer (2007). Residual analysis for linear mixed models. *Biometrical Journal 49*(6), 863–875.

Nobre, J. S. and J. M. Singer (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics 38*(5), 1063–1072.

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

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

Roy, A. (2009). An application of 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.

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

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in StatisticsTheory and Methods 37*(7), 1071–1084.

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*(2), 153–177.