Contents

Chapter 1

Model Diagnostics

1.1 Model Validation

Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. Often the validation of a model seems to consist of nothing more than quoting the R^2 statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model). Unfortunately, a high R^2 value does not guarantee that the model fits the data well. Use of a model that does not fit the data well will not provide any meaningful insights to the underlying research questions.

In classical linear models, 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. Schabenberger (2004) discusses the state of LME diagnostics tools, providing a useful summary of established measures.

Schabenberger (2004) describes the examination of model-data agreement as com-

prising several elements; residual analysis, goodness of fit, collinearity diagnostics and influence analysis.

1.2   Residual diagnostics

Residual analysis is a widely used model validation technique. For classical linear models, residual diagnostics are typically implemented as a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations. 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 noticeable patterns or trends. If some sort of nonrandom trend is evident in the model, then the model can be considered to be poorly fitted. For classical linear models, residual diagnostics are typically implemented as a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations. Statistical software environments, such as the `R` programming language, provides a suite of tests and graphical procedures for appraising a fitted linear model, with several of these procedures analysing the model residuals. However, for LME models the matter of residual is more complex, both from a theoretical point of view and from the practical matter of implementing a comprehensive analysis using statistical software. As the LME model can be tailored to the needs of the particular research question, the rationale behind the model appraisal must follow accordingly.

For an LME model, the *raw* residuals at level i are obtained by subtracting the fitted levels at that level from the response vector If we require a *Pearson* type residual, we further dviide this by the estimated within-group standard error. The Pearson residual is the raw residual divided by the square root of the the within-group standard error for both methods. The fitted values at level i are obtained by adding together

6

the population fitted values (based only on the fixed effects estimates) and the estimated contributions of the random effects to the fitted values at grouping levels less or equal to i. In addition to Raw and Pearson residuals, there are normalized residuals : standardized residuals pre-multiplied by the inverse square-root factor of the estimated error correlation matrix. For the forthcoming examples, we will use the raw residuals.

Preliminary Residual Analysis

The underlying assumptions for LME models are similar to those of classical linear mdoels. There are two key techniques: a residual plot and the normal probability plot. Using the nlme package it is possible to create plots specific to each method. This is useful in determine which methods 'disagree' with the rest. Analysis of the residuals would 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. Erroneous measurements are incorrect measurements that indicate disagreement between methods that would otherwise be in agreement. Once the residuals are computed, they can be used to make an assessment about the model fit. For the LME model described in Chapter 2, we can plot the residuals against the fitted values, to assess the assumption of constant variance. f the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

LME models assume that the residuals of the model are normally distributed. The residuals can be divided according to groups according to the method of measurement. In the following examples, we seperately assess normality the J method residuals (the first 255 residuals) and S method residuals (the remaining 255). Importantly the residuals from the J method are normally distributed, but there is non-normality of the residuals according to the S method.

7
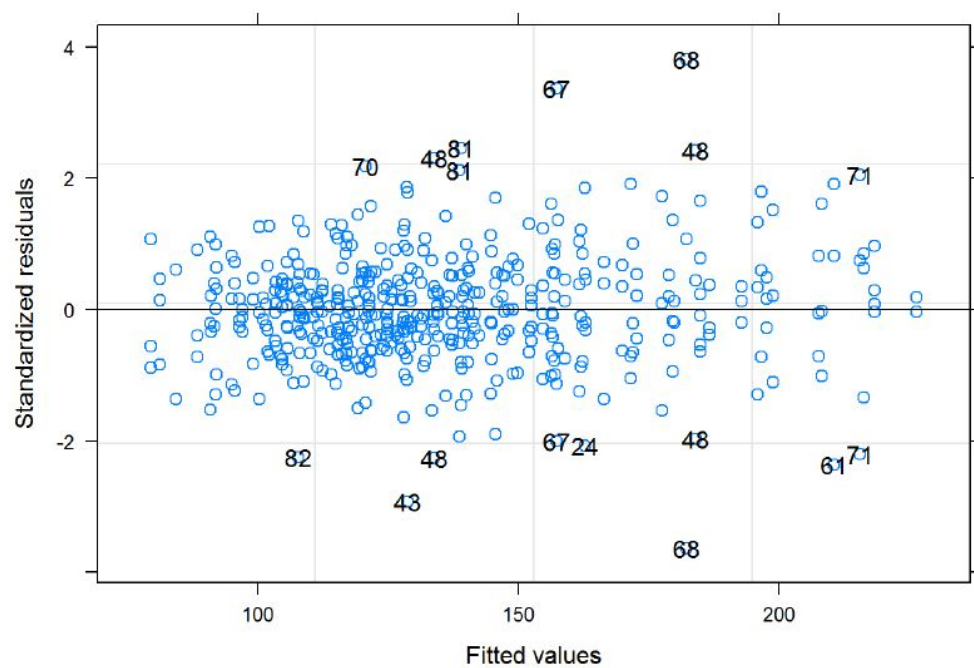
Figure 1.2.1:

```
> shapiro.test(resid(JS.roy1)[1:255])


Shapiro-Wilk normality test


data:  resid(JS.roy1)[1:255]
W = 0.9931, p-value = 0.2852
```

```
> shapiro.test(resid(JS.roy1)[256:510])


Shapiro-Wilk normality test
```

```
data:  resid(JS.roy1)[256:510]
W = 0.9395, p-value = 9.503e-09
```


Method J   **Method S**
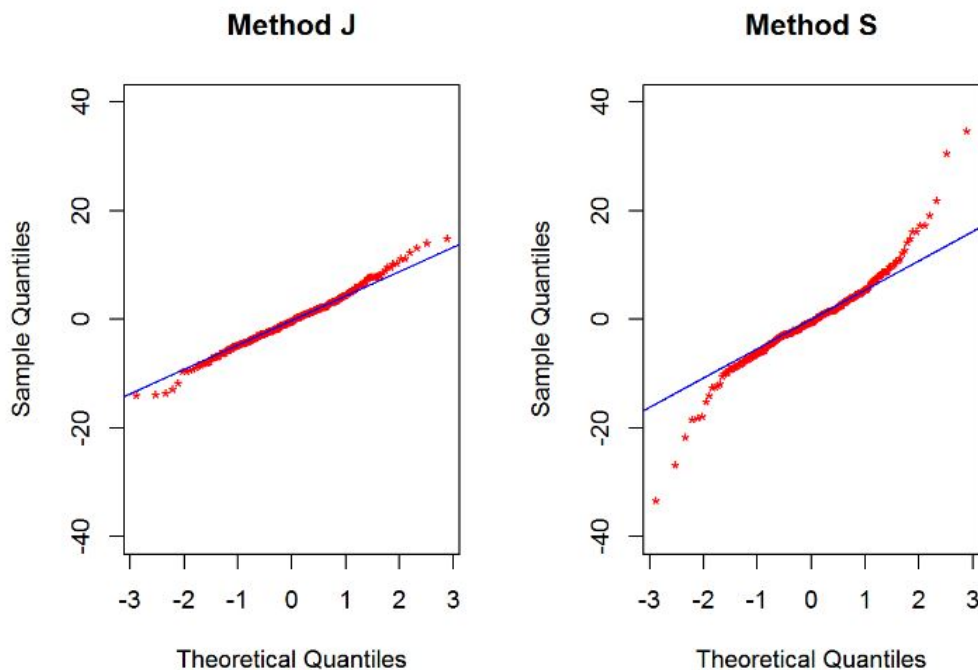
Normal probability plots can be rendered for each level of the random effects. LME models assume that not only the within-cluster residuals are normally distributed, but that each level of the random effects are as well.

1.3   Standardization and Studentization

To alleviate the problem caused by inconstant variance, the residuals can be scaled (i.e. divided) by their standard deviations. This results in a 'standardized residual'. A random variable is said to be standardized if the difference from its mean is scaled by its standard deviation. The residuals have mean zero but their variance is unknown, it depends on the true values of θ. Standardization is not possible in practice. Because
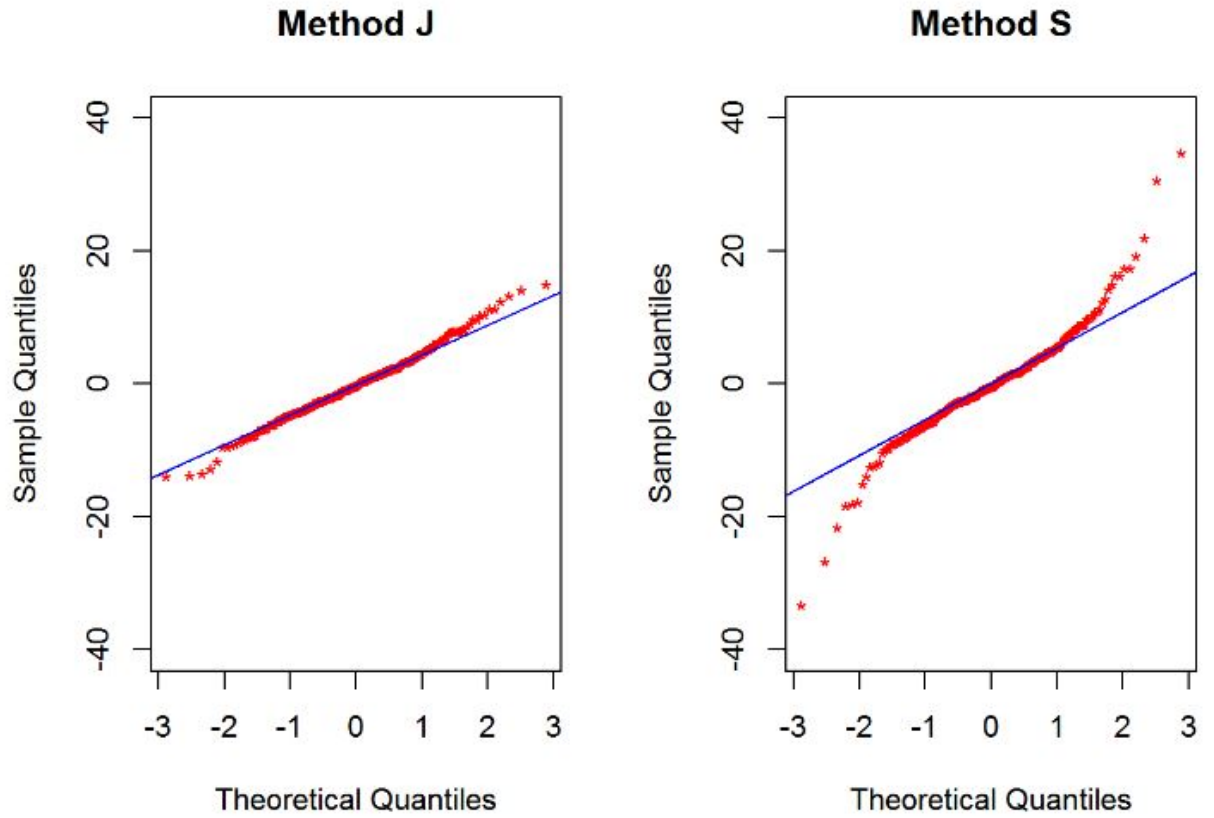
Figure 1.2.2:

true standard deviations are frequently unknown, one can instead divide a residual by the estimated standard deviation to obtain the 'studentized residual. If that estimate is independent of the $i-$th observation, the process is termed 'external studentization'. This is usually accomplished by excluding the $i-$th observation when computing the estimate of its standard error. If the observation contributes to the standard error computation, the residual is said to be internally studentized. Externally studentized residual require iterative influence analysis or a profiled residuals variance (**CITE**).

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

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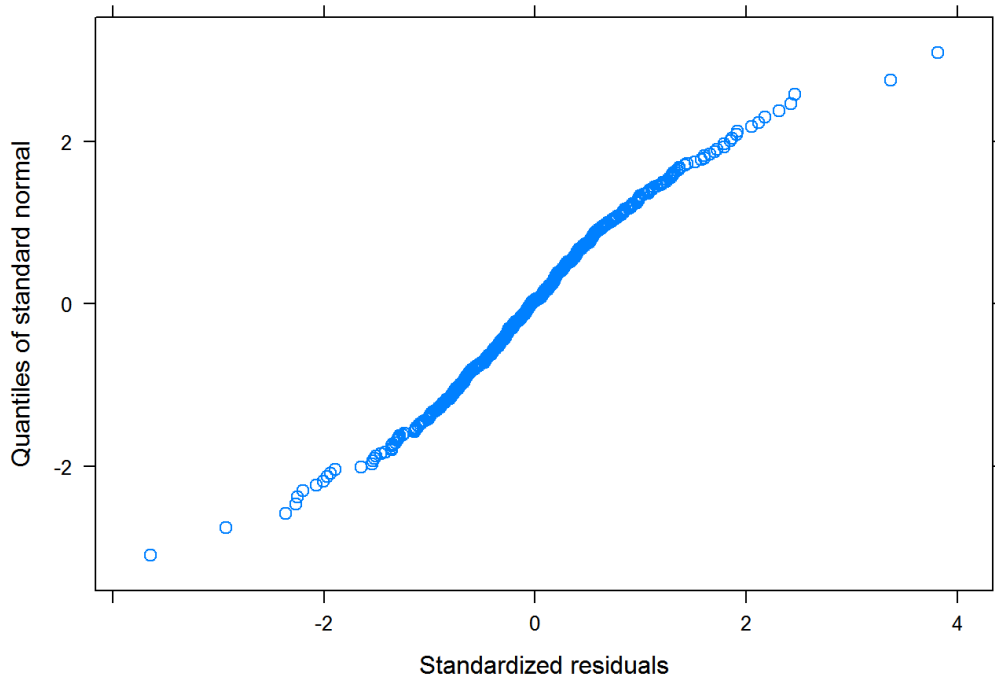
1.3.1 Pearson Residual

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

In this chapter, we will look at residual analysis and diagnostic toold for LME models, and discuss how they can be applied to the Method Comparison Problem. In classical linear models, a residual is the difference between an observed value and its estimated or predicted value. For LME models, the topic is more complicated, and still a matter of active research.
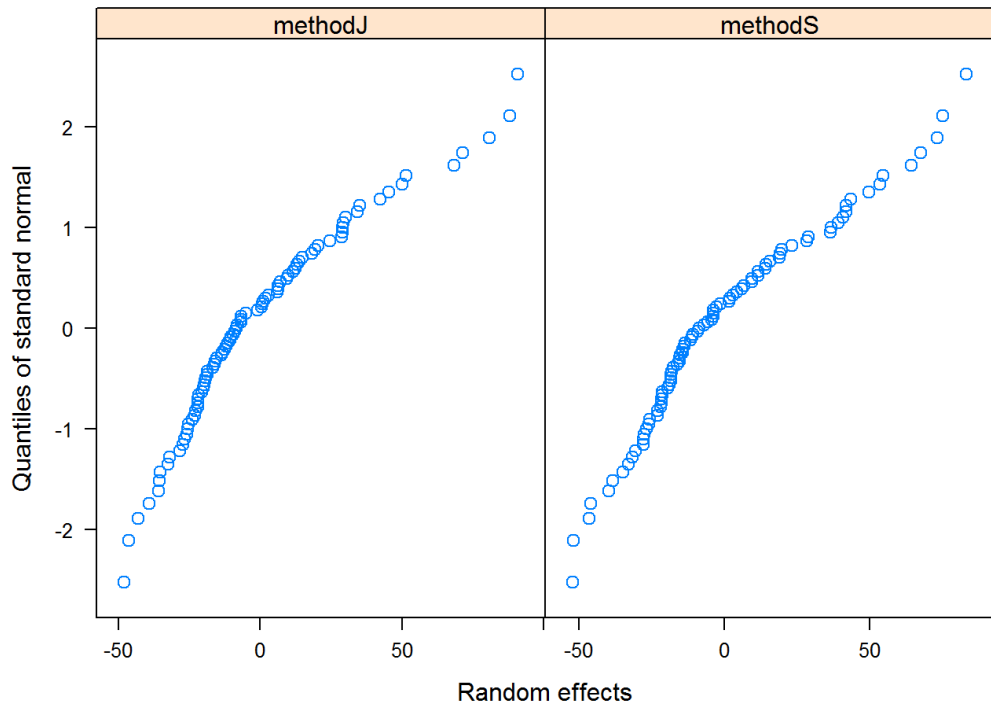
Figure 1.2.3:

1.4 Distinction From Linear Models (Schabenberger page 3)

The differences between perturbation and residual analysis in the linear model and the linear mixed model are connected to the important facts that b

and b depend on the estimates of the covariance parameters, that b

has the form of an (estimated) generalized least squares (GLS) estimator, and that is a random vector. In a mixed model, you can consider the data in a conditional and an unconditional sense. If you imagine a particular realization of the random effects, then you are considering the conditional distribution Y— If you are interested in quantities averaged over all possible values of the random effects, then you are interested in Y; this is called the marginal formulation. In a clinical trial, for example, you may be interested in drug efficacy for a particular patient. If random effects vary by patient,

that is a conditional problem. If you are interested in the drug efficacy in the population of all patients, you are using a marginal formulation. Correspondingly, there will be conditional and marginal residuals, for example. The estimates of the fixed effects

depend on the estimates of the covariance parameters. If you are interested in determining the influence of an observation on the analysis, you must determine whether this is influence on the fixed effects for a given value of the covariance parameters, influence on the covariance parameters, or influence on both. Mixed models are often used to analyze repeated measures and longitudinal data. The natural experimental or sampling unit in those studies is the entity that is repeatedly observed, rather than each individual repeated observation. For example, you may be analyzing monthly purchase records by customer. An influential data point is then not necessarily a single purchase. You are probably more interested in determining the influential customer. This requires that you can measure the influence of sets of observations on the analysis, not just influence of individual observations. The application of well-known concepts in model-data diagnostics to the mixed model can produce results that are at first counter-intuitive, since our understanding is steeped in the ordinary least squares (OLS) framework. As a consequence, we need to revisit these important concepts, ask whether they are portable to the mixed model, and gain new appreciation for their changed properties. An important example is the ostensibly simple concept of leverage.

1.5   Taxonomy of LME Residuals

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 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.

Pinheiro and Bates (1994) describes three types of residual that describe the variabilities present in LME models

1. marginal residuals, $\hat{\xi}$, which predict marginal errors,

2. conditional residuals, $\hat{\epsilon}$, which predict conditional errors,

3. the BLUP,$\boldsymbol{Z}\hat{\boldsymbol{b}}$, that predicts random effects.

Each type of residual is useful to evaluates some assumption of the model.

According to Hilton & Minton (1995), a residual is considered pure for a specfic type fo error if it depends only on the fixed components and on the error that it is supposed to predict. Residuals that depend on other types of error are known as 'confounded errors'.

1.6 Residuals in LME Models : Marginal, Conditional and Scaled

1.6.1 Internally and Externally Studentized Residuals

The computation of internally studentized residuals relies on the diagonal values of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) - \boldsymbol{Q}(\hat{\boldsymbol{\theta}})$ Externally studentized residuals require iterative influece analysis or a profiled residual variance.

Cook's Distance

$$\boldsymbol{\delta}_{(U)} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(U)}$$

A DFFIT measures the change in predicted values due to the removal of data points. (Belsey, Kuh and Welsch (1980))

14

$\boldsymbol{D}(\boldsymbol{\beta}) = \boldsymbol{\delta}'_{(U)}\boldsymbol{\delta}_{(U)}/rank(\boldsymbol{X})$ Cook's D can be calibrated according to a chi-square distribution with degress of freedom equal to the rank of $\boldsymbol{X}$?.

Residuals

The raw residuals r_{mi} and r_{ci} are usually not well suited for these purposes.

- Conditional Residuals r_{ci}

- Marginal Residuals r_{mi}

Chapter 2

Residuals diagnostics in LME Models

Residuals are used to examine model assumptions and to detect outliers and potentially influential data point. In LME models, there are two types of residuals, marginal residuals and conditional residuals. A marginal residual is the difference between the observed data and the estimated marginal mean. A conditional residual is the difference between the observed data and the predicted value of the observation. In a model without random effects, both sets of residuals coincide. Schabenberger (2004) provides a useful summary. In a model without random effects, both sets of residuals coincide.

The definitions of both marginal residuals (r_m) and conditional residuals (r_c) follow from the definitions of marginal and conditional means in the LME model $E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ and $E[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{X\beta} + \boldsymbol{Zu}$, respectively.

A marginal residual is the difference between the observed data and the estimated (marginal) mean,

$$r_{mi} = y_i - x_0'\hat{b} = x_i^T \hat{\beta}$$

$$y - X\beta = Z\eta + \epsilon$$

A conditional residual is the difference between an observed value y_i and the con-

ditional predicted value $\hat{y}_i$,

$$r_{ci} = y_i - x'_i\hat{b} - z'_i\hat{\gamma}$$

$$y - X\beta - Z\eta = \epsilon$$

Conditional residuals include contributions from both fixed and random effects, whereas marginal residuals include contribution from only fixed effects. Marginal residuals should have mean of zero, but may show grouping structure. Also they may not be homoscedastic. Marginal residuals are good for checking fixed effects. Plots of the elements of the marginal residual vector versus the explanatory variables in X can be used to check the linearity of $\boldsymbol{y}$ in a similar manner to the residual plots used in linear models. Conditional residuals should have mean of zero with no grouping structure They should be homoscedastic. Conditional residuals are useful for checking normality of outliers

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

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

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

2.0.2 Marginal Residuals

The marginal residuals are defined according to

$$\hat{\xi} = y - X\hat{\beta} = M^{-1}Qy.$$

2.0.3 Conditional Residuals

A conditional probability is the difference between the observed value and the predicted value of the dependent variable.

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

In general conditional residuals are not well suited for verifying model assumptions and detecting outliers. Even if the true model residuals are uncorrelated and have equal variance, conditional variances will tend to be correlated and their variances may be different for different subgroups of individuals (West et al., 2007).

2.1 Residual diagnostics

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

2.2 Conditional and Marginal Residuals

Conditional residuals include contributions from both fixed and random effects, whereas marginal residuals include contribution from only fixed effects.

Suppose the linear mixed-effects model lme has an $n \times p$ fixed-effects design matrix $\boldsymbol{X}$ and an $n \times q$ random-effects design matrix $\boldsymbol{Z}$.

Also, suppose the p-by-1 estimated fixed-effects vector is $\hat{\beta}$, and the q-by-1 estimated best linear unbiased predictor (BLUP) vector of random effects is $\hat{b}$. The fitted conditional response is

$$\hat{y}_{Cond} = X\hat{\beta} + Z\hat{b}$$

and the fitted marginal response is

$$\hat{y}_{Mar} = X\hat{\beta}$$

residuals can return three types of residuals:

- raw,

- Pearson, and

- standardized.

For any type, you can compute the conditional or the marginal residuals. For example, the conditional raw residual is

$$r_{Cond} = y - X\hat{\beta} - Z\hat{b}$$

and the marginal raw residual is

$$r_{Mar} = y - X\hat{\beta}$$

2.3   Pearson and Deviance Residuals

The **deviance residual** is the measure of deviance contributed from each observation and is given by

$$r_{Di} = \text{sign}(r_i)\sqrt{d_i}$$

where d_i is the individual deviance contribution. The deviance residuals can be used to check the model fit at each observation for generalized linear models.

The standardized and studentized deviance residuals are

$$r_{Dsi} = \frac{r_{Di}}{\sqrt{\hat{\phi}(1 - h_i)}}$$

$$r_{Dti} = \frac{r_{Di}}{\sqrt{\hat{\phi}_{(i)}(1 - h_i)}}$$

Chapter 3

Influence Diagnostics

3.1   What is Influence

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

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

3.2   Introduction to Influence analysis

Model diagnostic techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations. In classical linear models model diagnostics have been become a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers that model

diagnostics do not often accompany LME model analyses.

For linear models for uncorrelated data, 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. By contrast, in mixed 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.

3.2.1   Importance of Influence

The influence of an observation can be thought of in terms of how much the predicted values for other observations would differ if the observation in question were not included in the model fit.

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

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.

3.3   Influence Diagnostics: Basic Idea and Statistics

Broadly defined, "*influence* is understood as the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the

analysis, yield qualitatively different inferences, or violate assumptions of the statistical model.

The goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis. Outliers, for example, may be the most noteworthy data points in an analysis. They can point to a model breakdown and lead to development of a better model.

The general idea of quantifying the influence of one or more observations relies on computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation.

3.3.1 Diagnostic Methods for OLS models

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.

The idea of influence diagnostics for a given observation is to quantify the effect of omission of this observation from the data on the results of the model fit. To this aim, the concept of likelihood displacement is used.

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's Distance , denoted as $D_{(i)}$, is a well known diagnostic technique used in classical linear models, used as an overall measure of the combined impact of the ith case of all estimated regression coefficients. Cook's key observation was the effects of deleting each observation in turn could be calculated with little additional computation. That is to say, $D_{(i)}$ 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.

The focus of this analysis is related to the estimation of point estimates (i.e. regression coefficients). 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.

As well as individual observations, Cook's distance can be used to analyse the influence of observations in subset U on a vector of parameter estimates (Cook, 1977).

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{3.1}$$

$$\delta_{(U)} = \hat{\beta} - \hat{\beta}_{(U)} \tag{3.2}$$

3.3.2   Cook's 1986 paper on Local Influence

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

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be calculated with little additional computation. That is to say, $D_{(i)}$ 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. Cook proposed a measure that combines the information of leverage and residual of the observation, now known simply as the Cook's Distance. Cook's Distance , denoted as $D_{(i)}$, is a well known diagnostic technique used in classical linear models, used as an overall measure of the combined impact of the $i-$th case of all estimated regression coefficients.

The local-influence approach to influence assessment is quitedifferent from the case deletion approach, comparisons are of interest. Cook (1986) introduces powerful tools for local-influence assessment and examining perturbations in the assumptions of a model. In particular the effect of local perturbations of parameters or observations are examined

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

3.4   Influence analysis for 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.

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

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

3.5   Influence Diagnostics: Basic Idea and Statistics

The general idea of quantifying the influence of one or more observations relies on computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation.

Influence statistics can be coarsely grouped by the aspect of estimation that is their primary target:

- overall measures compare changes in objective functions: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2)

- influence on parameter estimates: Cooks (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

Furthermore, closed-form expressions for computing the change in important model quantities might not be available. This section provides background material for the various influence diagnostics available with the MIXED procedure. See the section Mixed Models Theory for relevant expressions and definitions. The parameter vector denotes all unknown parameters in the and matrix. The observations whose influence is being ascertained are represented by the set and referred to simply as "the observations in ." The estimate of a parameter vector, such as , obtained from all observations except those in the set is denoted . In case of a matrix , the notation represents the matrix with the rows in removed; these rows are collected in . If is symmetric, then notation implies removal of rows and columns. The vector comprises the responses of the data points being removed, and is the variance-covariance matrix of the remaining observations. When , lowercase notation emphasizes that single points are removed, such as .

3.6   A Procedure for Quantifying Influence

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.

3.7   Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. DFBETA and DFFITS are well known measures of influence. The measure DFBETA is the studentized value of this difference. DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. DFFITS is closely related to the studentized residual.

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \tag{3.3}$$

$$= B(Y - Y_{\bar{a}} \tag{3.4}$$

$$DFFITS = \frac{\widehat{y_i} - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}} \tag{3.5}$$

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.

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

3.8   Influence Statistics for LME models

Influence statistics can be coarsely grouped by the aspect of estimation that is their primary target:

- overall measures compare changes in objective functions: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2)

- influence on parameter estimates: Cook's (Cook 1977, 1979), MDFFITS (Belsley, Kuh, and Welsch 1980, p. 32)

- influence on precision of estimates: CovRatio and CovTrace

- influence on fitted and predicted values: PRESS residual, PRESS statistic (Allen 1974), DFFITS (Belsley, Kuh, and Welsch 1980, p. 15)

- outlier properties: internally and externally studentized residuals, leverage

3.8.1 Cook's Distance

- For variance components γ

Diagnostic tool for variance components

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

3.8.2 Variance Ratio

- For fixed effect parameters β.

3.8.3 Cook-Weisberg statistic

- For fixed effect parameters β.

3.8.4 Zewotir Measures of Influence in LME Models

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

- Variance components

- Fixed effects parameters

- Prediction of the response variable and of random effects

- likelihood function

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

- Cook's distance for LME models,

- likelihood distance,

- the variance (information) ration,

- the Cook-Weisberg statistic,

- the Andrews-Prebigon statistic.

3.8.5 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.

3.8.6 Cook's Distance

- For variance components γ: $CD(\gamma)_i$,

- For fixed effect parameters β: $CD(\beta)_i$,

- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,

- For linear functions of $\hat{beta}$: $CD(\psi)_i$

Diagnostic tool for variance components

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

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

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

3.9 Influence Diagnostics

a Overall Measures that compare changes in objectives functions; (restricted) maximum likelihood (Cook Weisberg, 1982)

b Influence on parameter estimates: Cook's Distance, MDFFITs

c Influence on precision of estimates: CovTrace and CovRatio

d Influence on fitted and predicted values: PRESS residuals, PRESS statistics, DFFITs

e Outlier properties : internally and externally studentized residuals, leverage

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

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

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

3.10 Overall Influence

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

3.10.1 Iterative Influence Analysis

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

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

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

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

3.11 Local Influence

Beckman et al. (1987) applied the local influence method of Cook (1986) to the analysis of the LME model. While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known.

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

- the estimates of fixed effects,

- the estimates of the precision of the fixed effects,

- the estimates of the covariance parameters,

- the estimates of the precision of the covariance parameters,

- fitted and predicted values.

Cook 1986 introduced methods for local influence assessment. These methods provide a powerful tool for examining perturbations in the assumption of a model, particularly the effects of local perturbations of parameters of observations. The local-influence approach to influence assessment is quite different from the case deletion approach, comparisons are of interest.

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.

3.12   Iterative and non-iterative influence analysis

Schabenberger (2004) highlights some of the issue regarding implementing mixed model diagnostics. A measure of total influence requires updates of all model parameters. However, this doesnt increase the procedures execution time by the same degree.

estimation

$$\hat{\beta} \quad = \quad X^T \tag{3.6}$$

$$\hat{\gamma} \quad = \quad G(\hat{\theta})Z^T \tag{3.7}$$

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.

3.12.1   Influence Analysis for LME Models

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

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

Studentized residuals, error contrast matrices and the inverse of the response variance covariance matrix are regular components of these tools.

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

3.12.2   Computation Matters

Key to the implementations of influence diagnostics for LME Models is the attempt to quantify influence, where possible, by drawing on the basic definitions of the various statistics in the classical linear model. 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.

3.12.3   Extension of techniques to LME Models

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

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

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

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

- the estimates of fixed effects,

- the estimates of the precision of the fixed effects,

- the estimates of the covariance parameters,

- the estimates of the precision of the covariance parameters,

- fitted and predicted values.

3.12.4   Analyzing Influence in LME models

"*Influence* is defined by ? as "the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model". The goal of influence analysis is rather to identify influential cases and the manner in which they are important to the analysis. A consequence of this that cases may be to mark data points for deletion so that a better model fit can be achieved for the reduced data (?).

? considers several important aspects of the use and implementation of influence measures in LME models. *schabenberger* notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates.

? 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 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 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.

3.12.5 Influence in LME models (schab)

Likelihood based estimation methods, such as ML and REML, are sensitive to unusual observations. Influence diagnostics are formal techniques that assess the influence of observations on parameter estimates for β and θ. A common technique is to refit the model with an observation or group of observations omitted.West et al. (2007) examines a group of methods that examine various aspects of influence diagnostics for LME models. For overall influence, the most common approaches are the 'likelihood distance' and the 'restricted likelihood distance'.

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

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

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

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

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

determine whether the absence of observations changed the analysis.

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

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

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

3.13   Non-iterative Update Procedures

The change in the fixed-effects estimates following removal of the observations in U is

$$\hat{\beta} - \hat{\beta}_{(U)} = \boldsymbol{\Omega X V} \left(\boldsymbol{UPU} \right)$$

Residual variance

When σ^2 is profiled out of the marginal variance-covariance matrix, a closed-form estimate of σ^2 that is only based on only the remaining observation an be computed as follows, provided $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ [cite: Hurtado 1993]

Likelihood Distances

For noniterative methods the following computational devices are used to compute (restricted) likelihood distances provided that the residual variance σ^2 is profiled.

3.13.1 Miscellaneous

Mean Square Prediction Error

$$MSPR = \frac{\sum(y_i - \hat{y}_i)^2}{n^*} \tag{3.8}$$

Effects on parameter estimate

Cook's Distance. CD.

Effects on the fitted and predicted values

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.

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{3.9}$$

3.14 Christensen et al

Christensen, Pearson and Johnson (1992) (hereafter CPJ) studied case deletion diagnostics, in particular the analog of Cooks distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

?provides an overview of case deletion diagnostics for fixed effect models.

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

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

? develops these techniques in the context of REML

3.15   Deletion Diagnostics

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

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

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

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

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of 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.

In this section we introduce influence analysis and case deletion diagnosics. A full overview of the topic will be provided although there are specific tools that are particularly useful in the case of MCS problems: specifically the Cook's Distance and the DFBeta.

A discussion of how leave-k-out diagnostics would work in the context of MCS problems is required. There are several scenaros. 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 Subject) Out - All observations associated with a particular case or subject are omitted. (e.g. $\{x_1, x_2, x_3, y_1, y_2, y_3\}$)

Other metrics, such as the likelihood distance, will also be introduced, and revisited in a later section.

3.16   Case Deletion Diagnostics

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

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

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models. Linear models for uncorrelated data have well established measures to gauge the influence of one or more observations on the analysis. For such models, closed-form update expressions allow efficient computations without refitting the model.

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

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

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

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

3.17   Terminology for Case Deletion diagnostics

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

3.18   Terminology for Case Deletion diagnostics

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

Christensen et al. (1992)examines case deletion results for estimates of the variance components, proposing the use of one-step estimates of variance components for examining case influence. The method describes focuses on REML estimation, but can easily be adapted to ML or other methods. **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 $\boldsymbol{A}$.

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

3.19 DFBETAs

The measure that measures how much impact each observation has on a particular predictor is DFBETAs.

DFBETAS (standardized difference of the beta) is a measure that standardizes the absolute difference in parameter estimates between a (mixed effects) regression model based on a full set of data, and a model from which a (potentially influential) subset of data is removed. The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted. A value for DFBETAS is calculated for each covariate, and for each case, in the model separately.

$$
\begin{align}
DFBETA_a &= \hat{\beta} - \hat{\beta}_{(a)} \tag{3.10} \\
&= B(Y - Y_{\bar{a}} \tag{3.11}
\end{align}
$$

In the case of method comparison studies, there are two covariates, and one can construct scatterplots of the pairs of dfbeta values accordingly, both for LOO and LSO calculations. Furthermore 95% confidence ellipse can be constructed around these scatterplots. Note that with k covariates, there will be $k+1$ dfbetas (the intercept,β_0, and one β for each covariate). When the model is specified without an intercept term, as in the last chapter, there is a set of DFBETAs corresponding to each measurement method.

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".

Using DFBETAs to Assess Agreement

Suppose an LME model was formulated to model agreement for various (i.e. 2 or more) methods of measurement, 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 points align along the line of equality, then both methods can be said to be in agreement.**

For the model fitted to the blood data with the lme4 R package, the results tabulated below can be produced. All 85 subjects are ranked by Cook's Distance (with only the top 6 being presented here). The remaining columns are the DFBeta for each of the fixed effects, for each of the 85 subject.

Subject	Cook's D	methodJ	methodR	methodS
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
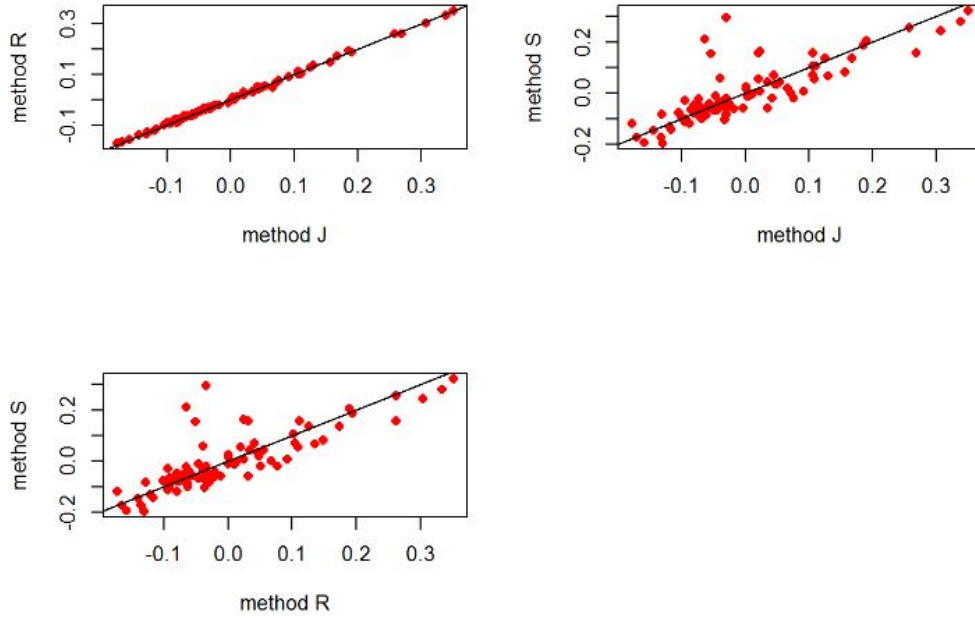
In the first of the three plots (*Top Right*), strong agreement between method J and method R is indicated. The other plots indicate lack of agreement of methods J and R with method S.

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 (see next section).

The Pearson Correlation coefficient of the DFBetas can be used in conjection with this analysis. A high correlation confirms good agreement. No threshold value for agreement is suggested, and analysts are advised to perform model diagnostics regardless of the correlation coeffient.

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. These diagnostic procedures are implementable using the **car** R package.

Deming Regression can be used to verify the line of equality. Significance test for Deming regression estimates are not available, but 95% bootstrap confidence intervals for the slope estimate and intercept estimates can be computed.

Additionally a mean difference plot can be used to identify outliers. 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.

If lack of agreement is indicated between methods of measurement, use of Roy's Testing is advised (This is the subject of the next section).

Tukey Mean Difference Plots

44
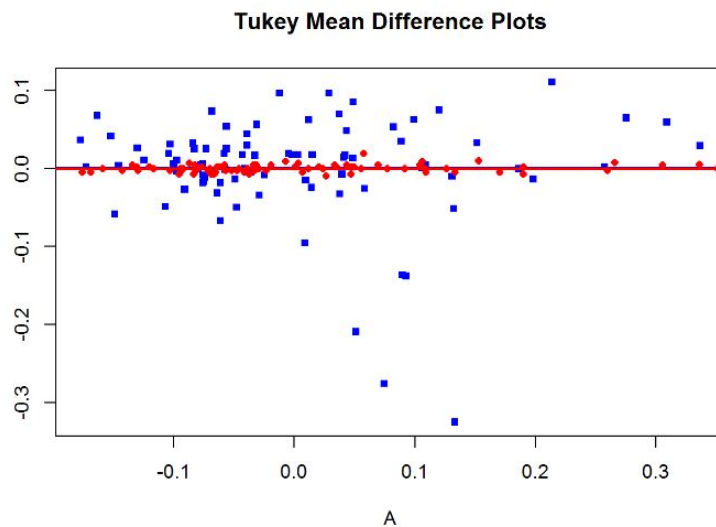
3.20 Computing DFBETAs with `R`

- This function computes the DFBETAS based on the information returned by the estex() function.

- The dfbeta refers to how much a parameter estimate changes if the observation or case in question is dropped from the data set.

- Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

- The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the th observation:

$$\text{Missing Formula}$$

  where is the th element of . In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter.

- **Belsley, Kuh, and Welsch (1980)** recommend 2 as a general cutoff value to indicate influential observations and as a size-adjusted cutoff.

3.21 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. DFFITS is a diagnostic meant to show how influential a point is in a statistical regression. It is defined as the change ("DFFIT"), in the predicted value for a point, obtained when that point is left out of the regression, "Studentized" by dividing by the estimated standard deviation of the fit at that point:

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

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

It is closely related to the studentized residual. For the sake of brevity, we will concentrate on the Studentized Residuals.

3.22 Predicted Values, PRESS Residual and the PRESS Statistic

The PRESS statistic is the sum of the squared PRESS residuals $\text{PRESS} = \sum \hat{\varepsilon}_{i(U)}^2$

The Prediction residual sum of squares (PRESS) is an value associated with this calculation.

When fitting linear models, PRESS can be used as a criterion for model selection, with smaller values indicating better model fits.

$$PRESS = \sum (y - y^{(k)})^2 \tag{3.12}$$

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.

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

3.22.1 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 $\boldsymbol{H}$.

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

3.22.2   DFFITs and MDFFITs

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

3.23   Model Diagnostics for Roy's Models

Further to previous work, this section revisits case-deletion and residual diagnostics, and explores how approaches devised by Galecki & Burzykowski (2013) can be used to appraise Roy's model. These authors specifically look at Cook's Distances and Likelihood Distances. For the Roy Model, Cook's Distances may also be generated using the **predictmeans**

As the model is structurally different from the models discussed in the earlier sections, Residual analysis will be briefly revisited.
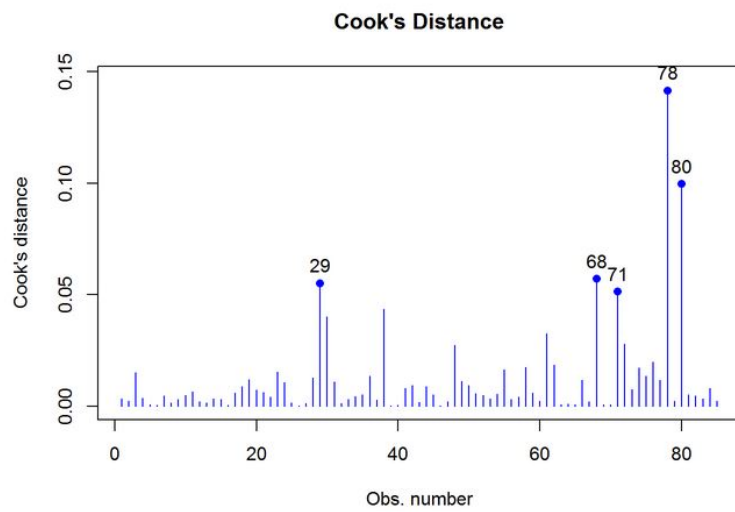
Figure 3.23.1:
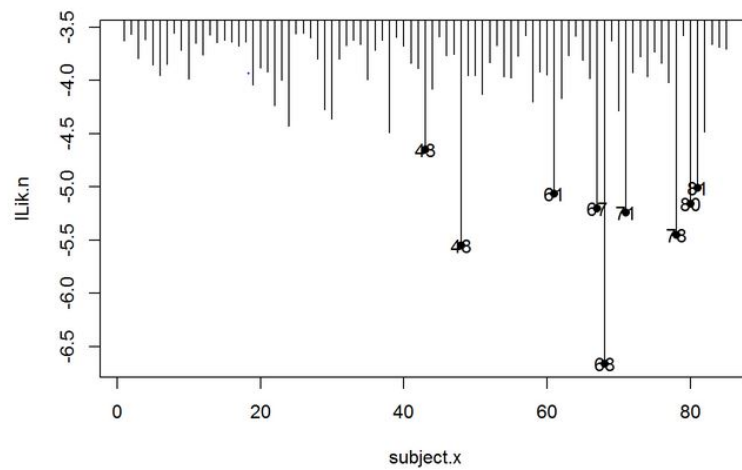
Figure 3.23.2:

8. Case Deletion Diagnostics for the Variance Ratios

Schabenberger advises on the use of deletion diagnostics for variance components of an LME model. 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 model., specifically the ratio of between subject variances and the within subject covariances respecitvely.

48
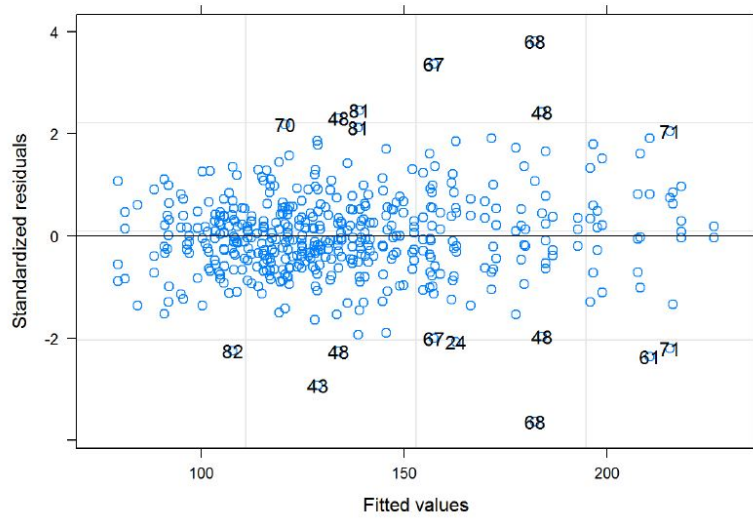
Figure 3.23.3:

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

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. As there may be several outliers (i.e. influential cases) present, the Grubbs test is not practical. However outlier detection using to Tukey's specification for boxplots (i.e. greater than $Q_3 + 1.5IQR$ or less than $Q_1 - 1.5IQR$), will suffice. Ranking the asbolute values of the standardizaed scores can also can be used to identify influential cases, even if the data is not normally distributed.

Bivariate Analyses may be applied jointly to the both sets of data sets, e.g Mahalanobis distances. The Mahalanobis distance, while not an intuitive measure in the context of the data, can be used to rank highly influential cases.

3.24   Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{3.13}$$

3.25 Case Deletion Diagnostics for Mixed 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.

An iterative analysis may seem computationally expensive. computing iterative influence diagnostics for n observations requires $n + 1$ mixed models to be fitted iteratively.

3.25.1 Extending deletion diagnostics to LMEs

after fitting a mixed, 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.

$$X = \begin{bmatrix} x'_i \\ X(i) \end{bmatrix}, Z = \begin{bmatrix} z'_{ij} \\ Z_{j(i)} \end{bmatrix}, Z = \begin{bmatrix} z'_{ij} \\ Z_{j(i)} \end{bmatrix},$$

$$y = \begin{bmatrix} y'_{ij} \\ y_{j(i)} \end{bmatrix} \text{ and } H = \begin{bmatrix} h_{ii} & h \\ h_{j(i)} & h \end{bmatrix}$$

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

$\boldsymbol{a}_{(i)}$ denotes a vector $\boldsymbol{a}$ with the i-th element, a_i, removed.

$$\breve{a}_i = \boldsymbol{a}_i - \boldsymbol{A}_{(i)} \boldsymbol{H}_{[i]} \boldsymbol{h}_i \tag{3.14}$$

3.25.2 Influence on measure component ratios

The general diagnostic tools for variance component ratios are the analogues of the Cook's distance and the information ratio.

The analogue of Cooks distance measure for variance components γ is denoted $CD(\gamma)$.

$$CD_U(\gamma) = (\hat{\gamma}_{(U)} - \hat{\gamma})'[\text{var}(\hat{\gamma})]^{-1}(\hat{\gamma}_{(U)} - \hat{\gamma})$$

$$= -\boldsymbol{g}'_{(U)}(\boldsymbol{Q} - \boldsymbol{G})^{-1}\boldsymbol{Q}(\boldsymbol{Q} - \boldsymbol{G})\boldsymbol{g}_{(U)}$$

$$= \boldsymbol{g}'_{(U)}(\boldsymbol{I}_r + \text{var}(\hat{\gamma})\boldsymbol{G})^{-2}\text{var}(\hat{\gamma})\boldsymbol{g}_{(U)}$$

Large values of $CD(\gamma)$ highlight observation groups for closer attentions

The analogue of the information ratio measures the change in the determinant of the maximum likelihood estimates information matrix

$$IR\gamma = \frac{\det(\boldsymbol{Q} - \boldsymbol{G})}{\det(\boldsymbol{Q})}$$

Ideally when all observations have the same influence on the information matrix $IR\gamma$ is approximately one. Deviations from one indicate the group U is influential. Since $\text{var}(\hat{\gamma})$ and $\boldsymbol{I}_r$ are fixed for all observations, $IR\gamma$ is a function of $\boldsymbol{G}$, in turn a function of $\boldsymbol{C}_i$ and c_{ii}.

3.26 Case Deletion Diagnostics for Variance Ratios

In this section, case deletion diagnostics are used on the variance components of the model. Specifically the ratio of between subject variances and the within subject variances respecitvely.

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

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

Methods for Identifying 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. The first variant of Grubb's

test is used to detect if the sample dataset contains one outlier, statistically different than the other values. The test statistic is based by calculating score of this outlier G (outlier minus mean and divided by sd) and comparing it to appropriate critical values. Alternative method is calculating ratio of variances of two datasets - full dataset and dataset without outlier. The second variant is used to check if lowest and highest value are two outliers on opposite tails of sample. It is based on calculation of ratio of range to standard deviation of the sample. The third variant calculates ratio of variance of full sample and sample without two extreme observations. It is used to detect if dataset contains two outliers on the same tail.

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, (i.e. greater than $Q_3 + 1.5\text{IQR}$ or less than $Q_1 - 1.5\text{IQR}$), will suffice.

Mahalanbis Distance

Bivariate Analyses may be applied jointly to the both sets of data sets, e.g Mahalanobis distances.

The WSVR values are plotted against the corresponding BSVR values. Confidence Ellipses can be superimposed over the plot with minimal effort. Two ellipses are generated by this technique, a 50 % and 97.5% confidence ellipse respectively. Outlying cases are idenified by the plot. Subject 68 is evident

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.
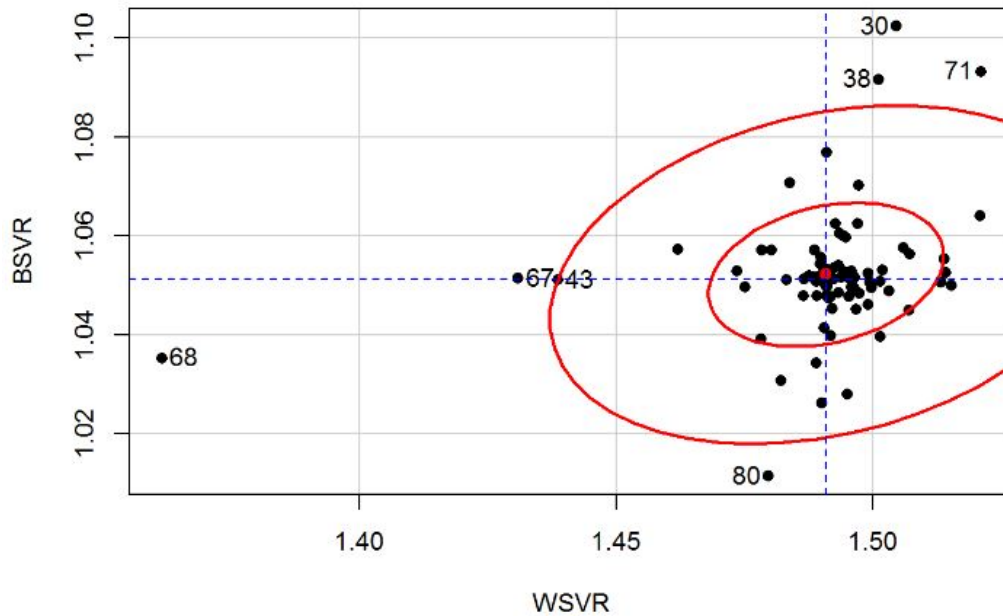

Figure 3.26.4:

3.27 Case Deletion Diagnostics for LME models

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

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

$$V_{[i]}^{-1} = \Lambda_{[i]} - \frac{\lambda\lambda\prime}{\nu ii} \tag{3.15}$$

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

$$X'_{[i]} V_{[i]}^{-1} X_{[i]} = X'V^{-1}X - \frac{\hat{x}_i \hat{x}'_i}{s_i} \tag{3.16}$$

$$(X'_{[i]} V_{[i]}^{-1} X_{[i]})^{-1} = (X'V^{-1}X)^{-1} + \frac{(X'V^{-1}X)^{-1}\hat{x}_i\hat{x}'_i(X'V^{-1}X)^{-1}}{s_i - \bar{h}_i} \tag{3.17}$$

$$X'_{[i]} V_{[i]}^{-1} Y_{[i]} = X\prime V^{-1}Y - \frac{\hat{x}_i \hat{y}'_i}{s_i} \tag{3.18}$$

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

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

3.27.1 Case Deletion Diagnostics

? develops case deletion diagnostics, in particular the equivalent of Cook's distance,

for diagnosing influential observations when estimating the fixed effect parameters and variance components.

3.27.2 Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{3.19}$$

3.27.3 Case Deletion Diagnostics for Mixed Models

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

? develops these techniques in the context of REML

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

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

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

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

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

Example:

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

3.27.4   Methods and Measures

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

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

- Cook's distance for LME models,

- likelihood distance,

- the variance (information) ration,

- the Cook-Weisberg statistic,

- the Andrews-Prebigon statistic.

3.27.5   Case Deletion Diagnostics

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

3.27.6   Case Deletion Diagnostics for Mixed Models

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

? develops these techniques in the context of REML

Chapter 4

Galecki

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.

Likelihood Displacement

4.0.7 Likelihood Distances

The likelihood distance is a global summary measure that expresses the joint influence of the subsets of observations, U, on all parameters in ϕ that were subject to updating. For classical linear models, the implementation of influence analysis is straightforward. ? points out the likelihood distance gives the amount by which the log-likelihood of the model fitted from the full data changes if one were to estimate the model from a reduced-data estimates. Importantly $LD(\psi_{(U)})$ is not the log-likelihood obtained by fitting the model to the reduced data set. It is obtained by evaluating the likelihood

function based on the full data set (containing all n observations) at the reduced-data estimates.

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

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

An overall influence statistic measures the change in the objective function being minimized. For example, in OLS regression, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance (**?**). In LME models, fitted by either ML or REML, an important overall influence measure is the likelihood distance (**?**). The procedure requires the calculation of the full data estimates $\hat{\psi}$ and estimates based on the reduced data set $\hat{\psi}_{(U)}$. The likelihood distance is given by determining

$$LD_{(U)} \;\; = \;\; 2\{l(\hat{\psi}) - l(\hat{\psi}_{(U)})\} \tag{4.1}$$

$$RLD_{(U)} \;\; = \;\; 2\{l_R(\hat{\psi}) - l_R(\hat{\psi}_{(U)})\} \tag{4.2}$$

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

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*.

The `logLik` Function

`logLik.lme` returns the log-likelihood value of the linear mixed-effects model represented by object evaluated at the estimated coefficients. It is also possible to determine the restricted log-likelihood, if relevant, using this function. For the Blood Data Example, the loglikelihood of the JS.roy1 model can be computed as follows.

```
> logLik(JS.roy1)
'log Lik.' -2030.736 (df=8)
```

4.1   Likelihood Distance

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

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

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

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

4.2   Likelihood Distance

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

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

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

The log-likelihood function l and restricted log-likelihood l_R ofthe LME model. $\boldsymbol{\psi}$ is the collection of all parameters (i.e. the fixed effects $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\theta}$).

Reduced data estimates $(\boldsymbol{\psi}_{(U)})$

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

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

Likelihood distance, known as likelihood displacements. The likelihood distance gives twice the amount by which the log likelihood of the full data changes if one were to use an estimate based on fewer data points. The likelihood distance is the a global summary measure of the influence of the observations in U jointly on all parameters.

An overall influence statistic measures the change in the objective function being minimized. In linear mixed models cook and Weisberg devised the Likelihood distance, known elsewhere as likelihood displacement.

The likelihood distance gives the amount of data by which the log-likelihood found when using the full data sets woudld change when the data set is reduced.

Importantly the value $l(\hat{\phi})_{(U)}$ is not the log-likelihood obtained from the reduced set, but the determining the likelihood function based on the full set at the reduced data estimates.

The approach can be applied to ML and REML models. Schabenberger (2004) uses in notation the subscript R to specify that REML models are under consideration.

The likelihood distance LD of observation group U is given by

$$LD_U = 2l(\hat{\phi}) - 2l(\hat{\phi}_{(U)}) \tag{4.3}$$

where l is the log likelihood function. If LD_U is large then the observation group U is influential on the likelihood function. The likelihood distance is a global summary measure, expressing the joint influence of the observation group U on all parameters ϕ subject to updating.

That U is influential is not grounds for deletion or changing the model. Should U be found to be influential, Schabenberger (2004) advises that the nature of that influence be determined. Estimates of the coefficients and precision of fixed effects, the coefficients and precision of covariance parameters, and fitted and predicted values should all be examined in light of determining that U is influential. Influence may be exerted by U on covariance parameters without affecting the fixed effects.

The likelihood distance has been widely used to detect outlying observations in data analysis. [cite: Cook and Weisberg] suggested that the likelihood distance may be compared to a χ^2 distribution for large samples.

Chapter 5

Cook's Distance

5.1 Cook's Distance for LMEs

Cook's Distance is a well known diagnostic technique used in classical linear models, extended to 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.

Diagnostic methods for fixed effects are generally analogues of methods used in classical linear models. Diagnostic methods for variance components are based on 'one-step' methods. *Cook (1986)* gives a completely general method for assessing the influence of local departures from assumptions in statistical models.

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

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

For random effect estimates, the Cook's distance is

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

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

5.1.1   Change in the precision of estimates

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.

Cook's distance

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

5.2   Cook's Distance

Cooks Distance (D_i) is an overall measure of the combined impact of the ith case of all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the kth case is deleted. $D_{(k)}$ can be calculated without fitting a new regression coefficient each time an observation is deleted. Cook's D statistics (i.e. colloquially Cook's Distance) is a measure of the influence of observations in subset U on a vector of parameter estimates (Cook, 1977).

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

If V is known, Cook's D can be calibrated according to a chi-square distribution with degrees of freedom equal to the rank of $\boldsymbol{X}$ (**?**).

63

5.3 Cook's Distance

- For variance components γ: $CD(\gamma)_i$,

- For fixed effect parameters β: $CD(\beta)_i$,

- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,

- For linear functions of $b\hat{e}ta$: $CD(\psi)_i$

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

Random Effects

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

linear functions

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

5.4 Exention of Cook's Distance methodology to LME models

Cook's Distance is extended to 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.

Diagnostic methods for variance components are based on 'one-step' methods. Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models. 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.

Cook's Distance was extended from classical linear models to LME models. For linear mixed effects models, Cook's distance can be extended to model influence diagnostics by definining.

$$CD_{\beta i} = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\beta} - \hat{\beta}_{[i]})}{p}$$

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

5.5   Cook's Distance for LMEs

Diagnostic methods for fixed effects are generally analogues of methods used in classical linear models. Diagnostic methods for variance components are based on 'one-step' methods. Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models.

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.

Cook's Distance

Cooks Distance (D_i) is an overall measure of the combined impact of the ith case of all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the $i-$th case is deleted.

Importantly, $D_{(i)}$ can be calculated without fitting a new regression coefficient each time an observation is deleted.

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

Cook's Distance is a well known diagnostic technique used in classical linear models, extended to 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.

5.5.1   Cook's Distance

In statistics, Cook's Distance or Cook's D is a commonly used estimate of the influence of a data point when performing least squares regression analysis.In a practical ordinary least squares analysis, 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. It is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.

5.5.2   Change in the precision of estimates

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.

Interpretation

Specifically D_i can be interpreted as the distance one's estimates move within the confidence ellipsoid that represents a region of plausible values for the parameters.[clarification needed] This is shown by an alternative but equivalent representation of Cook's distance in terms of changes to the estimates of the regression parameters between the cases where the particular observation is either included or excluded from the regression analysis.

5.5.3   Cook's Distance

Some texts tell you that points for which Cook's distance is higher than 1 are to be considered as influential. Other texts give you a threshold of $4/N$ or $4/(Nk1)$, where N is the number of observations and k the number of explanatory variables. In your case the latter formula should yield a threshold around 0.1 .

John Fox (1), in his booklet on regression diagnostics is rather cautious when it comes to giving numerical thresholds. He advises the use of graphics and to examine in closer details the points with "values of D that are substantially larger than the rest". According to Fox, thresholds should just be used to enhance graphical displays.

In your case the observations 7 and 16 could be considered as influential. Well, I would at least have a closer look at them. The observation 29 is not substantially different from a couple of other observations.

(1) Fox, John. (1991). Regression Diagnostics: An Introduction. Sage Publications.

5.6 Cook's Distance for LMEs

Diagnostic methods for fixed effects are generally analogues of methods used in classical linear models. Diagnostic methods for variance components are based on 'one-step' methods. Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models.

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

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

For random effect estimates, the Cook's distance is

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

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

Random Effects

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

linear functions

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

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.

5.7 Cook's Distance for LMEs

Diagnostic methods for fixed effects are generally analogues of methods used in classical linear models. Diagnostic methods for variance components are based on 'one-step' methods. Cook (1986) gives a completely general method for assessing the influence of local departures from assumptions in statistical models.

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

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

For random effect estimates, the Cook's distance is

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

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

5.8 Cook's Distance

Cooks Distance (D_i) is an overall measure of the combined impact of the ith case of all estimated regression coefficients. It uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the kth case is deleted. $D_{(k)}$ can be calculated without fitting a new regression coefficient each time an observation is deleted.

5.8.1 Cook's Distance

- For variance components γ: $CD(\gamma)_i$,

- For fixed effect parameters β: $CD(\beta)_i$,

- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,

- For linear functions of $\hat{beta}$: $CD(\psi)_i$

Random Effects

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

linear functions

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

Cook (1977) greatly expanded the study of residuals and influence measures. Cook's key observation was the effects of deleting each observation in turn could be computed without undue additional computational expense. Consequently deletion diagnostics have become an integral part of assessing linear models.

Cook (1986) gave a completely general method for assessing influence of local departures from assumptions in statistical models.

5.8.2 Cook's Distance

In classical linear regression, a commonly used meausre of influence is Cook's distance. It is used as a measure of influence on the regression coefficients.

For linear mixed effects models, Cook's distance can be extended to model influence diagnostics by definining.

$$C_{\beta i} = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\beta} - \hat{\beta}_{[i]})}{p}$$

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

5.8.3 Cooks's Distance

Cook's D statistics (i.e. colloquially Cook's Distance) is a measure of the influence of observations in subset U on a vector of parameter estimates (Cook, 1977).

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

If V is known, Cook's D can be calibrated according to a chi-square distribution with degrees of freedom equal to the rank of $\boldsymbol{X}$ (?).

For LME models, Cook's distance can be extended to model influence diagnostics by defining.

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

Bibliography

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

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.

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

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

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

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

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.

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.