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.

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.

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

diagnostics do not often accompany LME model analyses. **?** discusses the state of LME diagnostics tools, providing a useful summary of established measures.

? describes the examination of model-data agreement as comprising several elements; residual analysis, goodness of fit, collinearity diagnostics and influence analysis.

1.1.1   Introduction to Residual Analysis

Residual analysis is a widely used model validation technique. A residual is simply the difference between an observed value and the corresponding fitted value, as predicted by the model. The rationale is that, if the model is properly fitted to the model, then the residuals would approximate the random errors that one should expect; if the residuals behave randomly, with no discernible trend. If some sort of non-random 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.

1.1.2   Residual Plots

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the

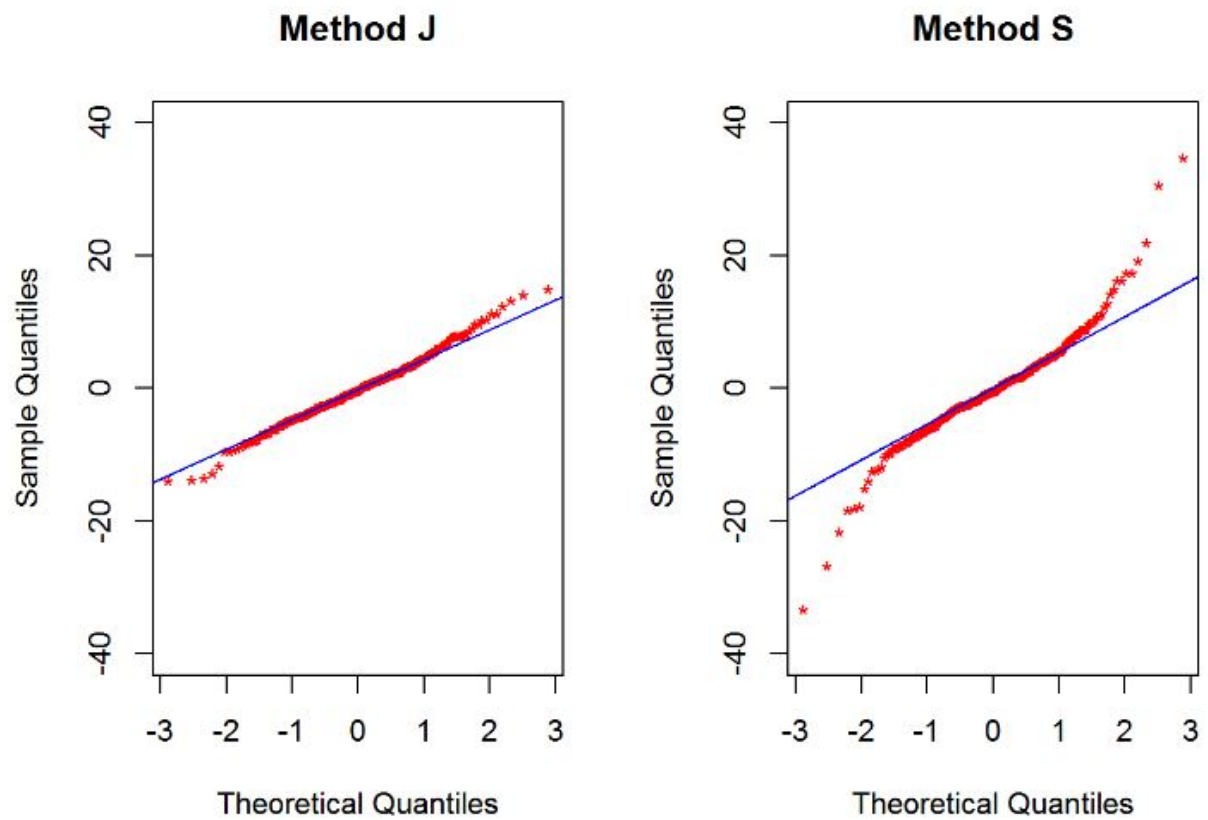data; otherwise, a non-linear model is more appropriate.
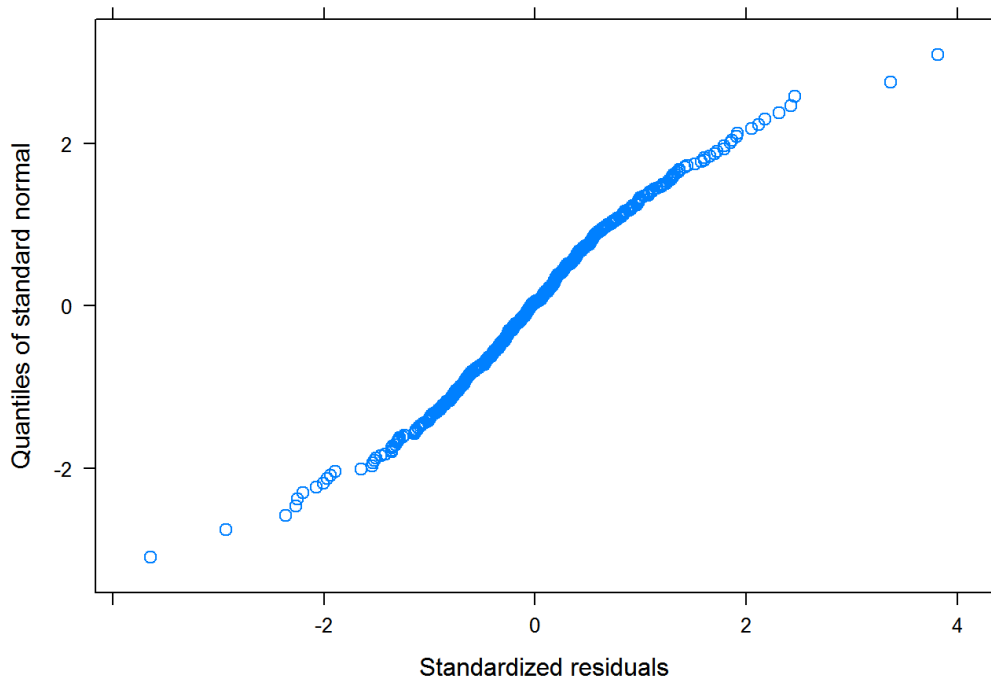

Figure 1.1.1:

This code will allow you to make QQ plots 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. Depending on the model, you can vary the level from 0, 1, 2 and so on

```
qqnorm(JS.roy1, ~ranef(.))


#  qqnorm(JS.roy1, ~ranef(.,levels=1)
```

1.2 Residual diagnostics

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 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.
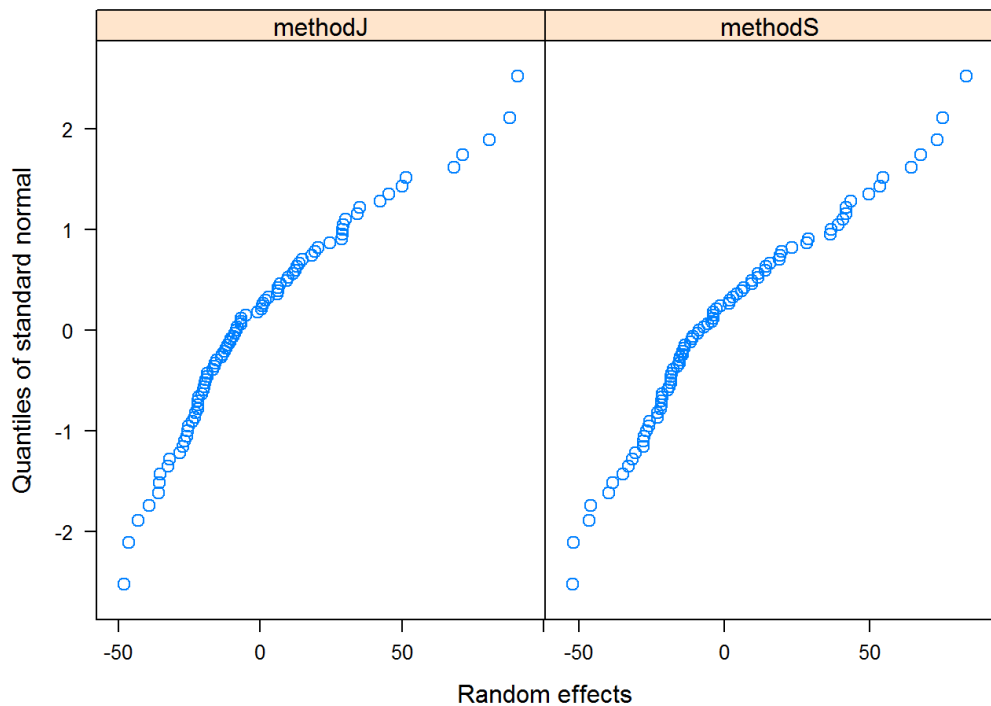
Figure 1.1.2:

Residuals

residuals.lme The residuals at level i are obtained by subtracting the fitted levels at that level from the response vector (and dividing by the estimated within-group standard error, if `type="pearson"`). The fitted values at level i are obtained by adding together 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.

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.

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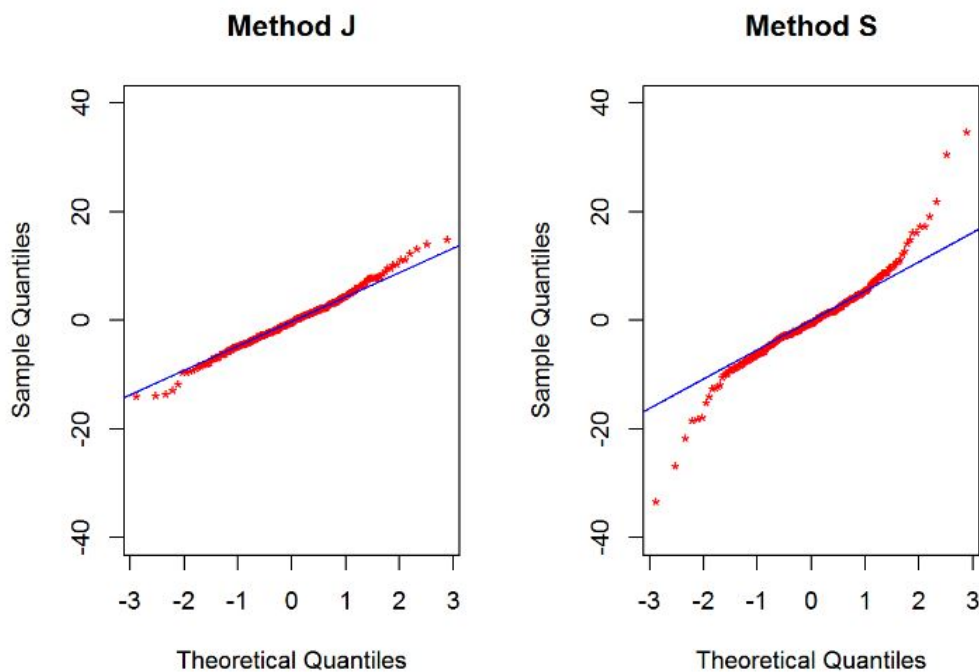
Normal probability plots can be rendered for each level of the random effects. LME models assume that not only the within-cluster residuals are normally distributed, but that each level of the random effects are as well.

Residuals can return three types of residuals:

- raw,

- Pearson, and

- standardized.

1.3   Standardization and Studentization

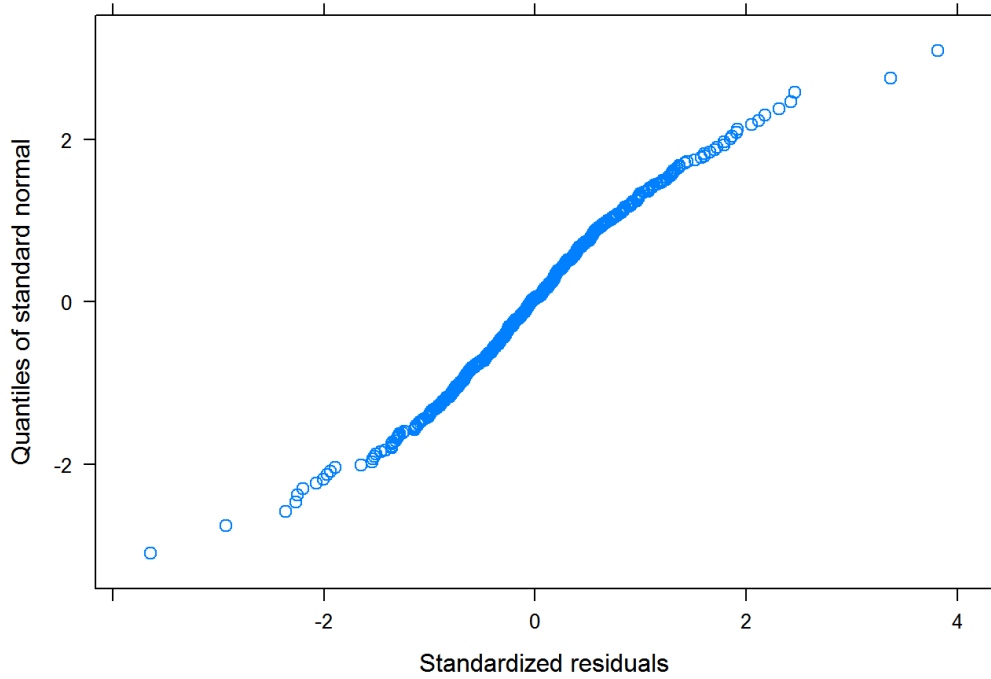To alleviate the problem caused by inconstant variance, the residuals can be scaled (i.e. divided) by their standard deviations. This results in a 'standardized residual'. A random variable is said to be standardized if the difference from its mean is scaled by its standard deviation. The residuals have mean zero but their variance is unknown, it depends on the true values of θ. Standardization is not possible in practice. Because 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
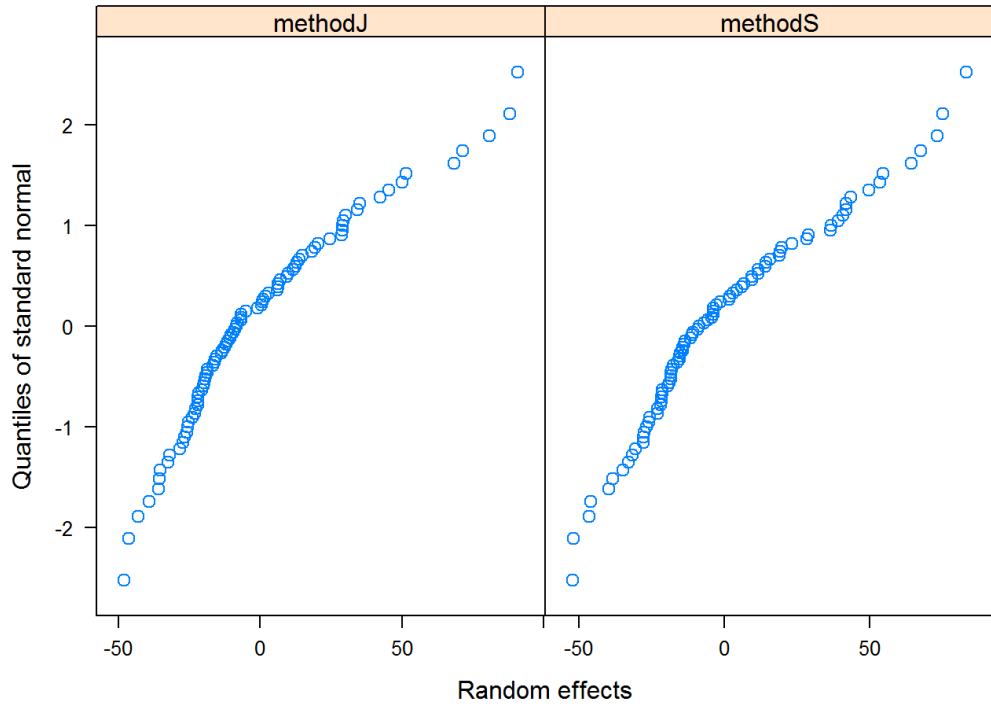
Figure 1.2.3:

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.

1.3.2 Internally and Externally Studentized Residuals

The computation of internally studentized residuals relies on the diagonal values of $V(\hat{\theta}) - Q(\hat{\theta})$ 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))

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

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. **?** 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 $E[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $E[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, 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 $\hat{y}_i$.

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

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.

2.1 Residual diagnostics

2.2 Taxonomy of LME Residuals

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

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.

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.

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.

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

- Conditional Residuals r_{ci}

- Marginal Residuals r_{mi}

? 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'.

2.3   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

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

2.3.1   Marginal Residuals

The marginal residuals are defined according to

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

2.3.2 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 (**?**).

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

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

2.4 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

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

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

2.5 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.

2.6 Residuals in the Blood Data Example

The fitted model used in the Blood data example, `JS.roy1`, was fitted using the `lme()` function from the nlme package, and as such, is stored as an `lme` object. The `residual` functions extracts residuals of a fitted LME model, depending on the type of residual required.

For an lme object, the residuals at level i are obtained by subtracting the fitted levels at that level from the response vector (and dividing by the estimated within-group standard error, if `type="pearson"`).The Pearson residual is the raw residual divided by the square root of the variance function (here, the Within-group standard error for both methods, 6.11 and 9.11 respectively). The fitted values at level i are obtained by adding together 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.

"`response`" : the raw residuals (*observed* - *fitted*) are used. This is the default option.

"`pearson`" : the standardized residuals (raw residuals divided by the corresponding standard errors) are used;

"`normalized`" : the normalized residuals (standardized residuals pre-multiplied by the inverse square-root factor of the estimated error correlation matrix) are used.

```
data.frame( response = resid(JS.roy1, type = "response"),
pearson  = resid(JS.roy1, type = "pearson"),
normalized = resid(JS.roy1, type = "normalized") )
```

```
response       pearson     normalized
1    -4.65805902 -0.761587227 -0.7615872269
2    -0.88701342 -0.145025661  0.0776238081
3    -5.16580898 -0.844603753 -0.8446037530
4     2.29041830  0.374480726  0.6450898404
5     7.87508366  1.287567009  1.2875670086
6    -6.57048659 -1.074266908 -1.5090772378

......................................
```

For the J observations, the variance is 6.116252 whereas for the S observations, the denominator is 9.118144. (with the expected ratio of 1.490806)

```
> pearson %>%
+   as.numeric %>%
+   matrix(nrow=85) %>%
+   round(4)
[,1]    [,2]    [,3]    [,4]    [,5]    [,6]
[1,] -0.7616  0.2194  0.3829 -0.2983  0.3597 -0.0790
[2,] -0.1450  0.1820 -0.1450 -0.5014  0.1567  0.2663
[3,] -0.8446  0.4634  0.1364 -0.1630 -0.2727  0.1660
[4,]  0.3745 -0.2795 -0.2795 -0.2658 -0.2658  0.6115
[5,]  1.2876 -0.6744 -0.6744  0.8935 -0.0935 -0.8612
[6,] -1.0743  1.8687 -0.7473 -0.0383  0.2908 -0.3673

......................................
```

We can plot the residuals against the fitted values, to assess the assumption of constant variance.

```
# standardized residuals versus fitted values

plot(JS.roy1, resid(., type = "pearson") ~ fitted(.) ,

abline = 0, id = 0.05)
```
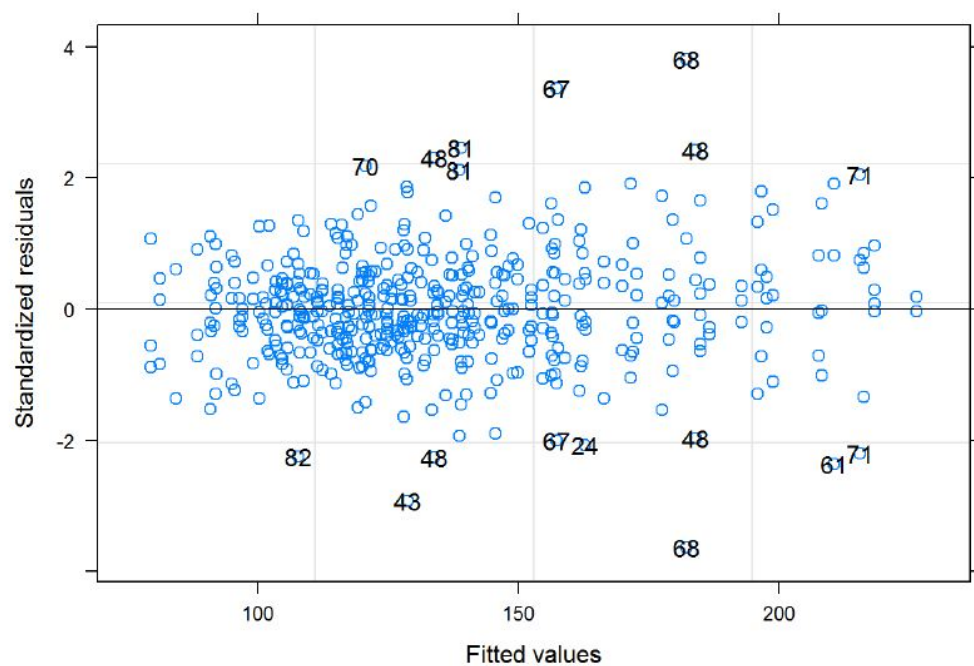

Figure 2.6.1: