Residual Analysis for Linear and LME Models with R

Dublin R

June 28, 2014

Contents

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 cannot provide good answers to the underlying engineering or scientific questions under investigation.

Model diagnostic techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations.

1.1 Why Use Residuals?

If the model fit to the data were correct, the residuals would approximate the random errors that make the relationship between the explanatory variables and the response variable a statistical relationship. Therefore, if the residuals appear to behave randomly, it suggests that the model fits the data well. On the other hand, if non-random structure is evident in the residuals, it is a clear sign that the model fits the data poorly.

The subsections listed below detail the types of plots to use to test different aspects of a model and give guidance on the correct interpretations of different results that could be observed for each type of plot.

2 Introduction to Residuals

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero.

2.1 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.

Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

In the next lesson, we will work on a problem, where the residual plot shows a non-random pattern. And we will show how to "transform" the data to use a linear model with nonlinear data.

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

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

Possibilities include:

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

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

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

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

2.1.1 Autocorrelation

Adjacent residuals should not be correlated with each other (**autocorrelation**). If you can use one residual to predict the next residual, there is some predictive information present that is not captured by the predictors. Typically, this situation involves time-ordered observations.

For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated. You can include a variable that captures the relevant time-related information, or use a time series analysis.

2.1.2 Durbin-Watson Test for Autocorrelated Errors

The ***Durbin-Watson*** procedure is commonly used to to test for autocorrelationof residuals.

```
attach(mtcars)
 FitMod <- lm(mpg~wt+cyl)


# library(car)
durbinWatsonTest(FitMod)
```

```
> durbinWatsonTest(FitMod)
 lag Autocorrelation D-W Statistic p-value
   1       0.1302185      1.671096   0.252
 Alternative hypothesis: rho != 0
```

3 Standardization and Studentization

3.1 Standardization

A random variable is said to be standardized if the difference from its mean is scaled by its standard deviation. The residuals above have mean zero but their variance is unknown, it depends on the true values of θ. Standardization is thus not possible in practice.

3.2 Studentization

Instead, you can compute studentized residuals by dividing a residual by an estimate of its standard deviation.

3.3 Internal and External Studentization

If that estimate is independent of the $i-$th observation, the process is termed *external studentization*'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 studentization*internally studentized.

Externally *studentized residual* studentized residual require iterative influence analysis or a profiled residuals variance.

3.4 Computation

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

3.5 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.

4 Leverage and Influence

4.1 Influence

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

Cook's D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

4.2 Interpreting Cook's Distance

A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

4.3 Leverage

The leverage of an observation is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation.

For example, an observation with a value equal to the mean on the predictor variable has no influence on the slope of the regression line regardless of its value on the criterion variable. On the other hand, an observation that is extreme on the predictor variable has the potential to affect the slope greatly.

4.3.1 Calculation of Leverage (h)

The first step is to standardize the predictor variable so that it has a mean of 0 and a standard deviation of 1. Then, the leverage (h) is computed by squaring the observation's value on the standardized predictor variable, adding 1, and dividing by the number of observations.

4.4 Summary of Influence Statistics

- **Studentized Residuals** Residuals divided by their estimated standard errors (like t-statistics). Observations with values larger than 3 in absolute value are considered outliers.

- **Leverage Values (Hat Diag)** Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than $2(k+1)/n$ are considered to be potentially highly influential, where k is the number of predictors and n is the sample size.

- **DFFITS** Measure of how much an observation has effected its fitted value from the regression model. Values larger than $2\sqrt{(k+1)/n}$ in absolute value are considered highly influential.

- **DFBETAS** Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential.

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

- **Cooks D** Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than 4/n are considered highly influential.

4.5 Influential Observations : DFBeta and DFBetas

Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. dfbeta refers to how much a parameter estimate changes if the observation in question is dropped from the data set. Note that with k covariates, there will be k+1 dfbetas (the intercept,β_0, and 1 β for each covariate). Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

4.6 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.

4.7 Leverage

leverage is a term used in connection with regression analysis and, in particular, in analyses aimed at identifying those observations that are far away from corresponding average predictor values. Leverage points do not necessarily have a large effect on the outcome of fitting regression models.

Leverage points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.[1]

Modern computer packages for statistical analysis include, as part of their facilities for regression analysis, various quantitative measures for identifying influential observations: among these measures is partial leverage, a measure of how a variable contributes to the leverage of a datum.

5 Regression Diagnostics with R

An excellent review of regression diagnostics is provided in John Fox's aptly named *Overview of Regression Diagnostics*. Dr. Fox's car package provides advanced utilities for regression modeling.

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

```
# Assume that we are fitting a multiple linear regression
# on the MTCARS data
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
```

5.1 Outliers

Assessment of Outliers can be carried out using the `outlierTest` function.

```
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```

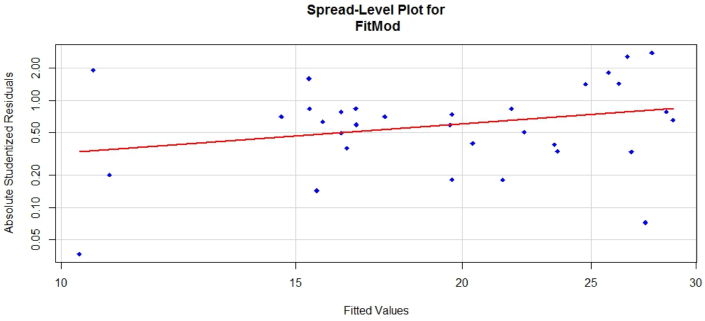
5.2 Added Variable Plots

```
# added variable plots
av.Plots(fit)
```

5.3 Non-constant Error Variance

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(FitMod)
# plot studentized residuals vs. fitted values
spreadLevelPlot(FitMod)
```

```
> ncvTest(FitMod)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.330027    Df = 1     p = 0.06802577
```



```
Suggested power transformation:  0.08866484
```

5.4 Influential Observations

```
# Influential Observations

# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle siz
```

6 Diagnostic Plots for Linear Models with `R`

Plot Diagnostics for an `lm` Object

6.1 Description

Six plots (selectable by `which`) are currently available:

1. a plot of residuals against fitted values,
2. a Scale-Location plot of *sqrt(— residuals —)* against fitted values,
3. a Normal Q-Q plot,
4. a plot of Cook's distances versus row labels,
5. a plot of residuals against leverages,
6. a plot of Cook's distances against leverage/(1-leverage).

By default, the first three and 5 are provided.

I explained the assumption of homoscedasticity and the plots that can help you assess it (including scale-location plots [2]) on CV here: What does having constant variance in a linear regression model mean? I have discussed qq-plots [3] on CV here: QQ plot does not match histogram. So, what's left is primarily just understanding [5], the residual-leverage plot.

To understand this, we need to understand three things:

- leverage,
- standardized residuals, and
- Cook's distance.

6.1.1 Leverage

To understand leverage, recognize that *Ordinary Least Squares* regression fits a line that will pass through the centre of your data, $(\bar{x}, \bar{y})$. The line can be shallowly or steeply sloped, but it will pivot around that point like a lever on a fulcrum. We can take this analogy fairly literally: because OLS seeks to minimize the vertical distances between the data and the line, the data points that are further out towards the extremes of X will push / pull harder on the lever (i.e., the regression line); they have more leverage. One result of this could be that the results you get are driven by a few data points; that's what this plot is intended to help you determine.

Another result of the fact that points further out on X have more leverage is that they tend to be closer to the regression line (or more accurately: the regression line is fit so as to be closer to them) than points that are near $\bar{x}$. In other words, the residual standard deviation can differ at different points on X (even if the error standard deviation is constant). To correct for this, residuals are often standardized so that they have constant variance (assuming the underlying data generating process is homoscedastic, of course).

One way to think about whether or not the results you have were driven by a given data point is to calculate how far the predicted values for your data would move if your model were fit without the data point in question. This calculated total distance is called **Cook's distance**. Fortunately, you don't have to rerun your regression model N times to find out how far the predicted values will move, Cook's D is a function of the leverage and standardized residual associated with each data point.

With these facts in mind, consider the plots associated with four different situations:

1. a dataset where everything is fine
2. a dataset with a high-leverage, but low-standardized residual point
3. a dataset with a low-leverage, but high-standardized residual point
4. a dataset with a high-leverage, high-standardized residual point

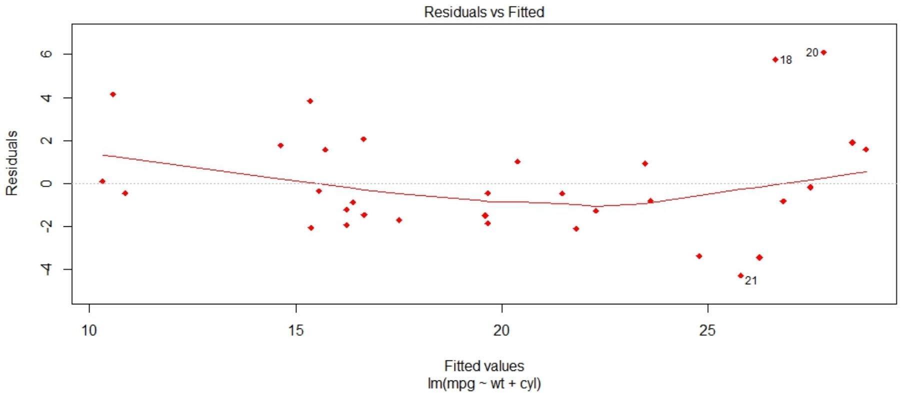
6.2 Diagnostic Plots for LMs

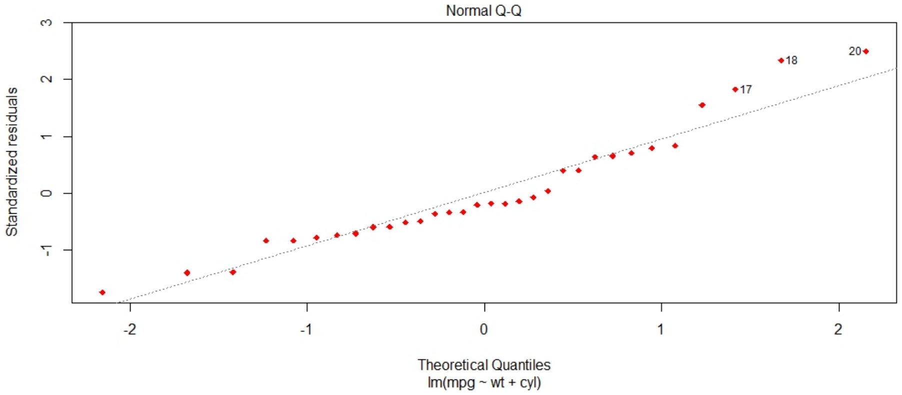
- The **Scale-Location** plot, also called Spread-Location (or S-L plot), takes the square root of the absolute residuals in order to diminish skewness (sqrt($|E|$)) is much less skewed than $|E|$ for Gaussian zero-mean E).

- The **Residual-Leverage** plot shows contours of equal Cook's distance, for values of `cook.levels` (by default 0.5 and 1) and omits cases with leverage one with a warning. If the leverages are constant (as is typically the case in a balanced aov situation) the plot uses factor level combinations instead of the leverages for the x-axis.

  (The factor levels are ordered by mean fitted value.)

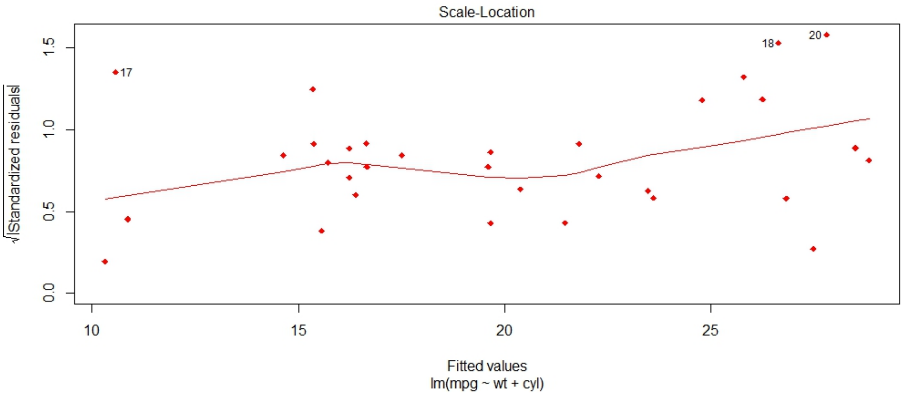
```
plot(lm(mpg~wt+cyl),which=c(1),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(2),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(3),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(4),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(5),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(6),pch=18,col="red")
```



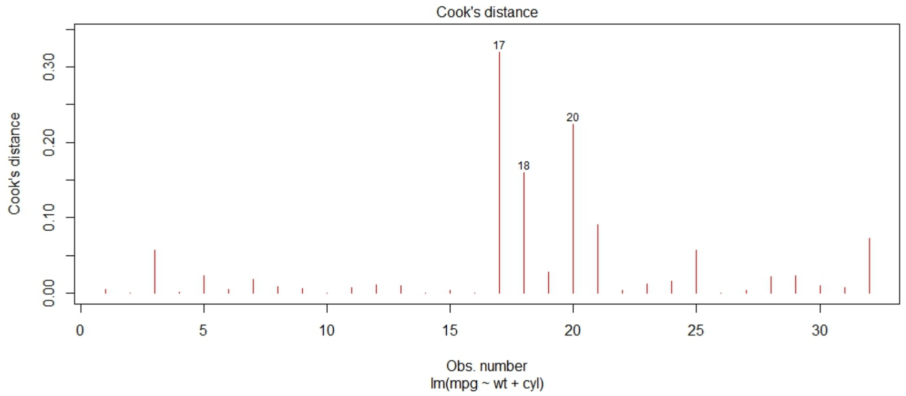


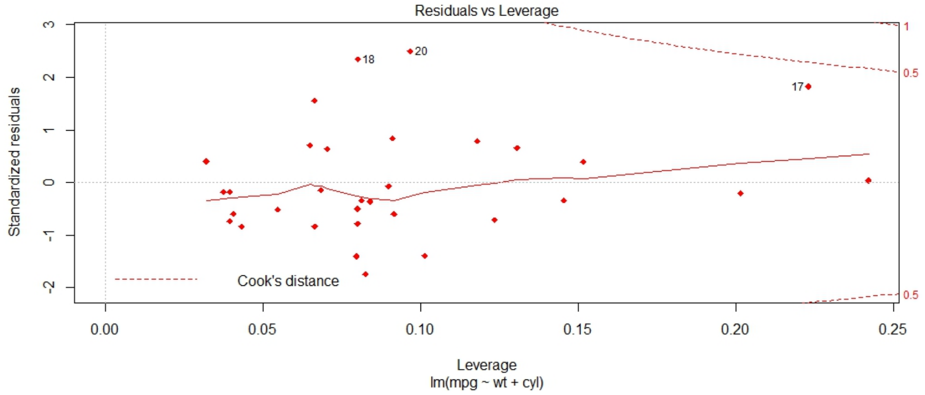
6.2.1 Plot 3 : Normal Probability Plot

This plot is used to assess the validity of the normality of the residuals.

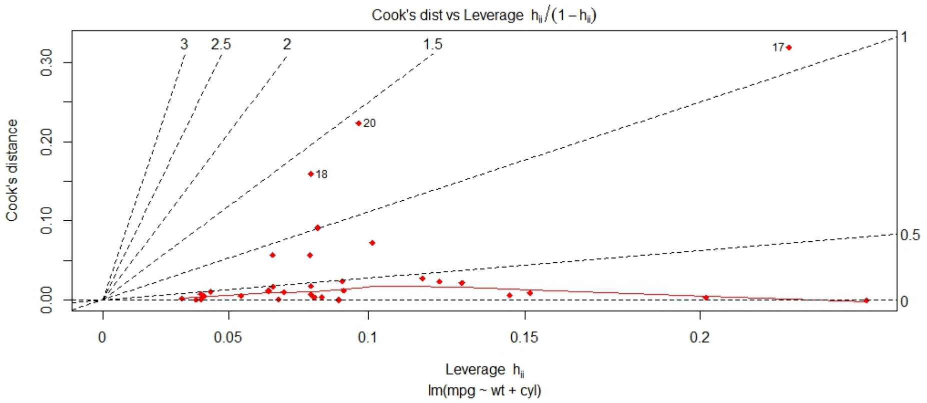


6.2.2 Plot 5 : Cook's Distance





6.2.3 Plot 6 : Cook's Distance vs Leverage



```
par(mfrow=c(4,1))

plot(fittedmodel)

par(opar)
```

7 Case Deletion

Case-deleted analysis is a popular method for evaluating the inuence of a subset of cases on inference.

Cite: CPJ develops case-deletion diagnostics for detecting influential observations in mixed linear models. Diagnostics for both fixed effects and variance components are proposed. Computational formulas are given that make the procedures feasible. The methods are illustrated using examples.

7.1 Case Deletion Diagnostic Statistics

For ordinary generalized linear models, regression diagnostic statistics developed by Williams (1987) are commonly used in statistical platforms such as SS. These diagnostics measure the influence of an individual observation on model fit, and generalize the one-step diagnostics developed by Pregibon (1981) for the logistic regression model for binary data. Preisser and Qaqish (1996) further generalize regression diagnostics to apply to models for correlated data fit by generalized estimating equations (GEEs), where the influence of entire clusters of correlated observations is measured.

7.2 Matrix Notation for 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 $\boldsymbol{A}$.

7.3 Partitioning Matrices

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

7.4 CPJ's Three Propositions

7.4.1 Proposition 1

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \nu^{ii} & \lambda_i' \\ \lambda_i & \Lambda_{[i]} \end{bmatrix}$$

$$\boldsymbol{V}_{[i]}^{-1} = \boldsymbol{\Lambda}_{[i]} - \frac{\lambda_i \lambda_i'}{\lambda_i}$$

7.5 Proposition 2

(i) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{X}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}$

(ii) $= (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y})^{-1}$

(iii) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{Y}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}$

7.6 Proposition 3

This proposition is similar to the formula for the one-step Newtown Raphson estimate of the logistic regression coefficients given by Pregibon (1981) and discussed in Cook Weisberg.

8 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. The measure DFBETA is the studentized value of this difference.

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 $\hat{V}$.

8.1 Cook's Distance

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.

Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance are considered to merit closer examination in the analysis.

8.1.1 Computation

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j\ - \hat{Y}_{j(i)})^2}{p\ \mathrm{MSE}}.$$

The following are the algebraically equivalent expressions (in case of simple linear regression):

$$D_i = \frac{e_i^2}{p\ \mathrm{MSE}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right],$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^T (X^T X)(\hat{\beta} - \hat{\beta}^{(-i)})}{(1+p)s^2}.$$

In the above equations:

- $\hat{Y}_j$ is the prediction from the full regression model for observation j;
- $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted;
- h_{ii} is the i-th diagonal element of the hat matrix

  $$\mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T;$$

- e_i is the crude residual (i.e., the difference between the observed value and the value fitted by the proposed model);
- MSE is the mean square error of the regression model;
- p is the number of fitted parameters in the model

8.1.2 DFBETA

The DFBETA statistic for measuring the influence of the th observation is defined as the one-step approximation to the difference in the MLE of the regression parameter vector and the MLE of the regression parameter vector without the th observation. This one-step approximation assumes a Fisher scoring step, and is given by

8.1.3 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. It is closely related to the studentized residual.

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

8.1.4 PRESS

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.

$$PRESS = \sum (y - y^{(k)})^2 \tag{1}$$

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

8.1.5 DFBETA

$$\begin{aligned} DFBETA_a &= \hat{\beta} - \hat{\beta}_{(a)} && (2) \\ &= B(Y - Y_{\bar{a}} && (3) \end{aligned}$$

9 Robust Regression (Optional Section)

Robust regression is an alternative to ordinary least squares regression (OLS , the type of regression we have studied thus far) when data is contaminated with outliers or influential observations and it can also be used for the purpose of detecting influential observations.

Some important terms in linear regression.

Residual: The difference between the predicted value (based on the regression equation) and the actual, observed value.

Outlier: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Leverage: An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. Influence can be thought of as the product of leverage and outlierness.

Cook's distance (or Cook's D): A measure that combines the information of leverage and residual of the observation.

```
FitAll = lm(Taste ~ Acetic + H2S + Lactic)
plot(FitAll)
```

This will produce a set of four plots: residuals versus fitted values, a Q-Q plot of standardized residuals, a scale-location plot (square roots of standardized residuals versus fitted values, and a plot of residuals versus leverage that adds bands corresponding to Cook's distances of 0.5 and 1.

Using the plot to get a detailed interpretation of how well the model fits is beyond the scope of this module. However it is worth noting that plots identify particular observations that may warrant re-examination.

Remarks : The plots actually indicate the fitted model is actually quite good. The trend lines in the first and third plots demonstrate constant variance, and in the case of the first, the trend line through the centre of the plot follows the X=0 line quite well.

The Q-Q plot (i.e. the second plot) indicates that the assumption of normality is vindicated. The last plot indicates the no observations have unusually high Cooks Distances values. No observations are beyond the 0.5 (red line) threshold.

Robust regression is an alternative to least squares regression when data are contaminated with outliers or influential observations, and it can also be used for the purpose of detecting influential observation

Additionally I have added a line plot of the Cooks Distance values. Which observations have the highest values for Cooks Distance?

```
plot(cooks.distance(FitAll),type="b",pch=18,col="red")
```

9.1 The stackloss data set

Brownlee's Stack Loss Plant Data contains operational data of a plant for the oxidation of ammonia to nitric acid.

The variables are:

- Air Flow Flow of cooling air
- Water Temp Cooling Water Inlet Temperature
- Acid Conc. Concentration of acid [per 1000, minus 500]
- stack.loss Stack loss

9.2 Fitting a robust model (`rlm`

```
summary(rlm(stack.loss ~ ., data = stackloss))
```

```
> summary(rlm(stack.loss ~ ., stackloss))


Call: rlm(formula = stack.loss ~ ., data = stackloss)
Residuals:
     Min       1Q   Median       3Q      Max
-8.91753 -1.73127  0.06187  1.54306  6.50163
```

```
Coefficients:
            Value    Std. Error t value
(Intercept) -41.0265   9.8073    -4.1832
Air.Flow      0.8294   0.1112     7.4597
Water.Temp    0.9261   0.3034     3.0524
Acid.Conc.   -0.1278   0.1289    -0.9922

Residual standard error: 2.441 on 17 degrees of freedom
```

```
rlm(stack.loss ˜ ., stackloss, psi = psi.hampel, init = "lts")
```

```
> rlm(stack.loss ˜ ., stackloss, psi = psi.hampel, init = "lts")
Call:
rlm(formula = stack.loss ˜ ., data = stackloss, psi = psi.hampel,
    init = "lts")
Converged in 10 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-40.4748037   0.7410863   1.2250703  -0.1455242

Degrees of freedom: 21 total; 17 residual
Scale estimate: 3.09
```

9.3 Using Other *Psi* Operators

Fitting is done by ***iterated re-weighted least squares (IWLS).***

Psi functions are supplied for the Huber, Hampel and Tukey bisquare proposals as psi.huber, `psi.hampel` and **psi.bisquare**. Huber's corresponds to a convex optimization problem and gives a unique solution (up to collinearity). The other two will have multiple local minima, and a good starting point is desirable.

- huber
- bisquare
- hampel

```
rlm(stack.loss ~ ., stackloss, psi = psi.bisquare)
```

```
Call:
rlm(formula = stack.loss ~ ., data = stackloss, psi = psi.bisquare)
Converged in 11 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-42.2852537   0.9275471   0.6507322  -0.1123310

Degrees of freedom: 21 total; 17 residual
Scale estimate: 2.28
```

9.4 Implementation of Robust Regression

- When fitting a least squares regression, we might find some outliers or high leverage data points. We have decided that these data points are not data entry errors, neither they are from a different population than most of our data. So we have no proper reason to exclude them from the analysis.

- Robust regression might be a good strategy since it is a compromise between excluding these points entirely from the analysis and including all the data points and treating all them equally in OLS regression. The idea of robust regression is to weigh the observations differently based on how well behaved these observations are.

- The idea of robust regression is to weigh the observations differently based on how well behaved these observations are. Roughly speaking, it is a form of weighted and reweighted least squares regression (i.e. a two step process , first fitting a linear model, then a robust model to correct for the influence of outliers).

- Robust regression is done by iterated re-weighted least squares (IRLS). The rlm command in the MASS package command implements several versions of robust regression.

- There are several weighting functions that can be used for IRLS. We are going to first use the Huber weights in this example. We will then look at the final weights created by the IRLS process. This can be very useful.

- Also we will look at an alternative weighting approach to Hubers weighting called **Bisquare weighting**.

9.4.1 Huber Weighting

In Huber weighting, observations with small residuals get a weight of 1 and the larger the residual, the smaller the weight. This is defined by the weight function

$$w(e) = 1 for|e| \leq k \tag{4}$$

$$w(e) = \frac{k}{|e|} for|e| > k \tag{5}$$

The value k for the Huber method is called a ***tuning constant***; smaller values of k produce more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed.

The tuning constant is generally picked to give reasonably high efficiency in the normal case; in particular,$k = 1.345\sigma$ for the Hubers method, where σ is the standard deviation of the errors) produce 95-percent efficiency when the errors are normal, and still offer protection against outliers.

```
library(MASS)
FitAll.rr = rlm(Taste ~ Acetic + H2S + Lactic)
```

```
> summary(FitAll.rr)

Call: rlm(formula = Taste ~ Acetic + H2S + Lactic)
Residuals:
    Min      1Q  Median      3Q     Max
-16.163  -5.612  -1.153   5.487  27.106
```

```
Coefficients:
            Value    Std. Error t value
(Intercept) -20.7529  20.1001    -1.0325
Acetic       -1.5331   4.5422    -0.3375
H2S           4.0515   1.2715     3.1864
Lactic       20.1459   8.7885     2.2923

Residual standard error: 8.471 on 26 degrees of freedom
```

Regression Equation:

$$\hat{Taste} = -20.75 - 1.53Acetic + 4.05H2S + 20.14Lactic$$

From before, we noticed that observations 15 , 12 and 8 were influential in the determination of the coefficients. The following table indicates the weight given to each observation when using robust regression.

We can see that roughly, as the absolute residual goes down, the weight goes up. In other words, cases with a large residuals tend to be down-weighted.

```
> hweights <- data.frame(Taste = Taste, resid = FitAll.rr$resid,
+     weight = FitAll.rr$w)
> hweights2 <- hweights[order(FitAll.rr$w), ]
>
```

```
> hweights2[1:15, ]
   Taste      resid    weight
15  54.9  27.105636 0.4203556
```

```
12  57.2  17.518919 0.6504044
8   21.9 -16.162753 0.7049043
3   39.0  14.318512 0.7957592
18   6.4 -13.609277 0.8371707
28   0.7 -11.410452 0.9985018
1   12.3   9.990965 1.0000000
2   20.9  -1.692664 1.0000000
4   47.9  10.648009 1.0000000
5    5.6  -1.866642 1.0000000
6   25.9   2.632602 1.0000000
7   37.3   5.744433 1.0000000
9   18.1   4.775657 1.0000000
10  21.0   1.048052 1.0000000
11  34.9   5.723592 1.0000000
```

9.4.2 Implementation with Bisquare Weighting

Implementing with bisquare weighting simply requires the specification of the additional argument, as per the code below, highlighted in green)

```
> FitAll.rr.2 = rlm(Taste ~ Acetic + H2S + Lactic, psi = psi.bisquare)
```

```
> summary(FitAll.rr.2)

Call: rlm(formula = Taste ~ Acetic + H2S + Lactic, psi = psi.bisquare)
Residuals:
     Min       1Q   Median       3Q      Max
```

```
-15.7034  -5.1552  -0.9793   5.6933  27.7661

Coefficients:
            Value    Std. Error t value
(Intercept) -17.7730  20.7031    -0.8585
Acetic       -2.2650   4.6784    -0.4841
H2S           4.0569   1.3096     3.0977
Lactic       20.6885   9.0522     2.2855

Residual standard error: 7.878 on 26 degrees of freedom
```

Weights using Bisquare estimator.

```
> hweights2[1:15, ]
   Taste      resid    weight
15  54.9  27.766087 0.1884633
12  57.2  18.182810 0.5735669
8   21.9 -15.703388 0.6707319
3   39.0  14.384429 0.7193235
18   6.4 -13.462286 0.7516310
28   0.7 -11.190438 0.8246092
19  18.0 -11.112316 0.8269297
4   47.9  10.860173 0.8343637
1   12.3   9.852297 0.8625704
20  38.9  -8.952091 0.8858015
14  25.9   8.588121 0.8946576
30   5.5  -8.019522 0.9078077
7   37.3   6.329446 0.9420556
```

```
11  34.9   5.999726 0.9478611
13   0.7  -5.470990 0.9565447
```

9.4.3 Conclusion

We can see that the weight given to some observations is dramatically lower using the bisquare weighting function than the Huber weighting function and the coefficient estimates from these two different weighting methods differ. When comparing the results of a regular OLS regression and a robust regression, if the results are very different, you will most likely want to use the results from the robust regression. Large differences suggest that the model parameters are being highly influenced by outliers. Different functions have advantages and drawbacks. Huber weights can have difficulties with severe outliers, and bisquare weights can have difficulties converging or may yield multiple solutions.

10 Residual Analysis for GLMs (Optional Section)

10.1 Pearson and Deviance Residuals

Pearson Residuals

The Pearson residual is the raw residual divided by the square root of the variance function $V(\mu)$. The Pearson residual is the individual contribution to the Pearson χ^2 statistic.

For a binomial distribution with mi trials in the ith observation, it is defined as

$$r_{Pi} = \sqrt{m_i}\frac{r_i}{\sqrt{V(\hat{\mu_i})}}$$

For other distributions, the Pearson residual is defined as

$$r_{Pi} = \frac{r_i}{\sqrt{V(\hat{\mu_i})}}$$

The Pearson residuals can be used to check the model fit at each observation for generalized linear models. The standardized and studentized Pearson residuals are

$$r_{Psi} = \frac{r_{Pi}}{\sqrt{\hat{\phi}(1-h_i)}}$$

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

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

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

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

10.2 Diagnostics for Logistic Regression

10.3 Diagnostics for Poisson Regression

11 Residual Analysis for LME Models

11.1 Introduction

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.

Cite:Zewotir 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.

11.2 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

11.3 Cook's Distance applied to LMEs

- 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 $\hat{beta}$: $CD(\psi)_i$

11.4 Iterative and non-iterative influence analysis for LMEs

Cite: Schabenberger 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.

11.5 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.**Cite: Schabenberger** describes the choice between iterative influence analysis and non-iterative influence analysis.

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.

Information Ratio

12 Measures 2

12.1 Cook's Distance

- For variance components γ

Diagnostic tool for variance components

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

12.2 Variance Ratio

- For fixed effect parameters β.

12.3 Cook-Weisberg statistic

- For fixed effect parameters β.

12.4 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.

12.5 Likelihood Distance

?

12.6 Diagnostics for LMEs with R

influence.ME: Tools for detecting influential data in mixed effects models

influence.ME provides a collection of tools for detecting influential cases in generalized mixed effects models. It analyses models that were estimated using lme4. The basic rationale behind identifying influential data is that when iteratively single units are omitted from the data, models based on these data should not produce substantially different estimates. To standardize the assessment of how influential a (single group of) observation(s) is, several measures of influence are common practice, such as DFBETAS and Cook's Distance. In addition, we provide a measure of percentage change of the fixed point estimates and a simple procedure to detect changing levels of significance.

You should have a look at the R package ***influence.ME***. It allows you to compute measures of influential data for mixed effects models generated by lme4.

An example model:

```
library(lme4)
model <- lmer(mpg ~ disp + (1 | cyl), mtcars)
```

The function `influence` is the basis for all further steps:

```
library(influence.ME)
infl <- influence(model, obs = TRUE)
```

Calculate Cook's distance:

```
cooks.distance(infl)
```

Plot Cook's distance:

```
plot(infl, which = "cook")
```

13 Case-Deletion Diagnostics for LMEs The CPJ Paper

13.1 Case-Deletion results for Variance components

Cite: CPJ 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.

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

13.2 CPJ Notation

$$\boldsymbol{C} = \boldsymbol{H}^{-1} = \begin{bmatrix} c_{ii} & \boldsymbol{c}_i' \\ \boldsymbol{c}_i & \boldsymbol{C}_{[i]} \end{bmatrix}$$

Cite: CPJ noted the following identity:

$$\boldsymbol{H}^{-1}_{[i]} = \boldsymbol{C}_{[i]} - \frac{1}{c_{ii}}\boldsymbol{c}_{[i]}\boldsymbol{c}'_{[i]}$$

Cite: CPJ use the following as building blocks for case deletion statistics.

- $\breve{x}_i$
- $\breve{z}_i$
- $\breve{z}_ij$
- $\breve{y}_i$

- p_ii
- m_i

All of these terms are a function of a row (or column) of $\boldsymbol{H}$ and $\boldsymbol{H}_{[i]}^{-1}$