1 Simple Linear Regression

In simple linear regression, we predict values on one variable from the values of a second variable.

- The variable we are predicting is called the ***dependent variable*** (or response variable) and is referred to as Y.
- The variable we are basing our predictions on is called the **i*ndependent variable*** (or predictor variable) and is referred to as X.

Remark: When there is only one predictor variable, the prediction method is called simple regression. Linear regression can have more than one predictor variable, i.e. Multiple Linear Regression.

In simple linear regression, the predicted values of Y when plotted as a function of X form a straight line on the scatter plot. This line is known as the ***regression line***.

- Suppose we construct our model using n observed values of the response variable $y_1, y_2, \ldots y_i \ldots y_n$
- For the original data set, there is a predicted value of each case of Y that corresponds to an observed value of Y.
- The difference between an observed value of the dependent variable (y_i) and the corresponding predicted value ($\hat{y}$) is called the residual (e_i). Each data point from the data set has one residual.
- Simply put, the values of the residuals are derived as follows:

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

$$e_i = y_i - \hat{y}_i$$

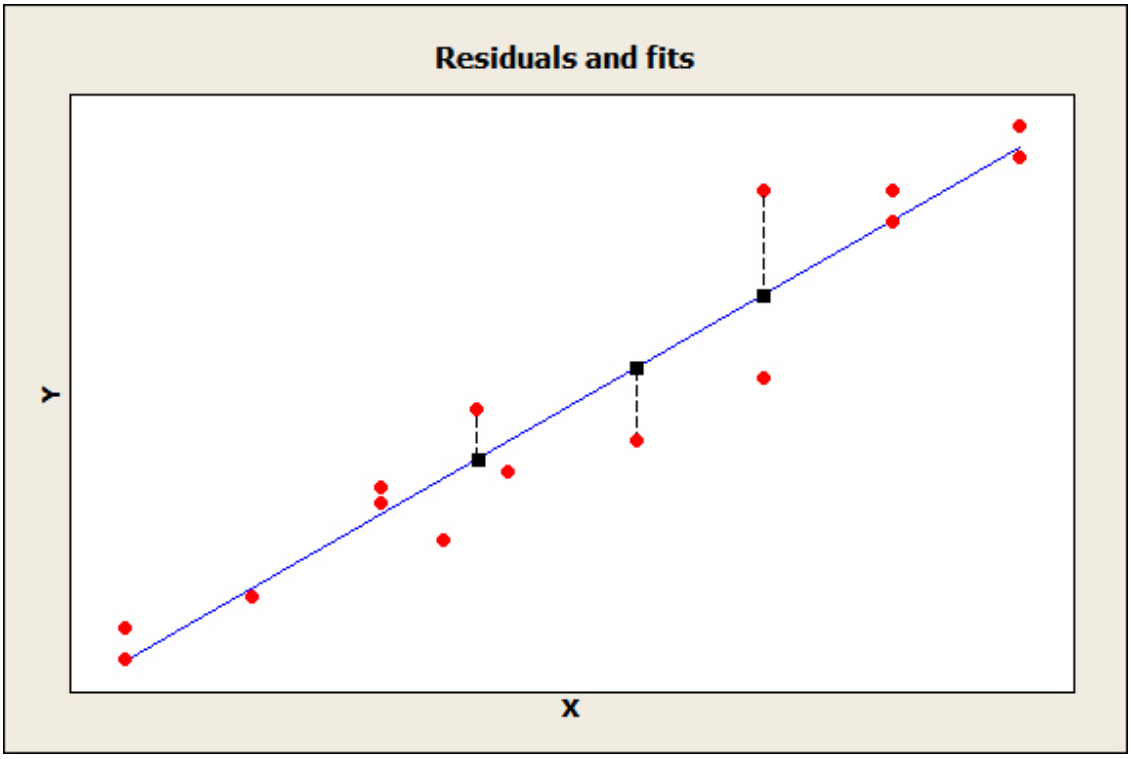


Figure 1:

- For three cases in the graphic above, the observed value (red dot) is linked to its corresponding predicted value (black dot) on the regression line (blue line). The difference (i.e. residual) is depicted using a dashed line. The magnitude of these residuals is of interest.
- The second of the three residuals will have a negative value.
- ***Ordinary Least Squares*** is a method of fitting a model, such that the total residual values are minimised.
- Important theoretical assumption underlying the OLS model: the sum of the residuals should equal to zero.

$$\sum e_i = 0$$

- An extension of this is that the expected value of the residuals is 0. $\mathrm{E}(e) = 0$
- Another Important Theoretical Assumption - The residuals are normally distributed. (more on that later)

1.1 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.

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.

2 mtcars example

Several data sets , intended as learning tools, are automatically installed when `R` *is installed. Many more are installed within packages to complement learning to use those packages. One of these is the famous* ***mtcars*** *data set, which is used in many data mining exercises.*

```
> data(mtcars)
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

```
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Suppose we fit a model with *mpg* (miles per gallon) as the response variable and *cyl* and *wt* (number of cylinders and weight of the car) as the predictor variables. We will call this fitted model `fit`.

```
fit <- lm(mpg ~ cyl + wt, data=mtcars)
```

```
> summary(fit)

Call:
lm(formula = mpg ~ cyl + wt, data = mtcars)

Residuals:
Min      1Q  Median      3Q     Max
-4.2893 -1.5512 -0.4684  1.5743  6.1004

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
cyl          -1.5078     0.4147  -3.636 0.001064 **
wt           -3.1910     0.7569  -4.216 0.000222 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.568 on 29 degrees of freedom
Multiple R-squared:  0.8302,    Adjusted R-squared:  0.8185
F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

```
residuals(fit1)
```

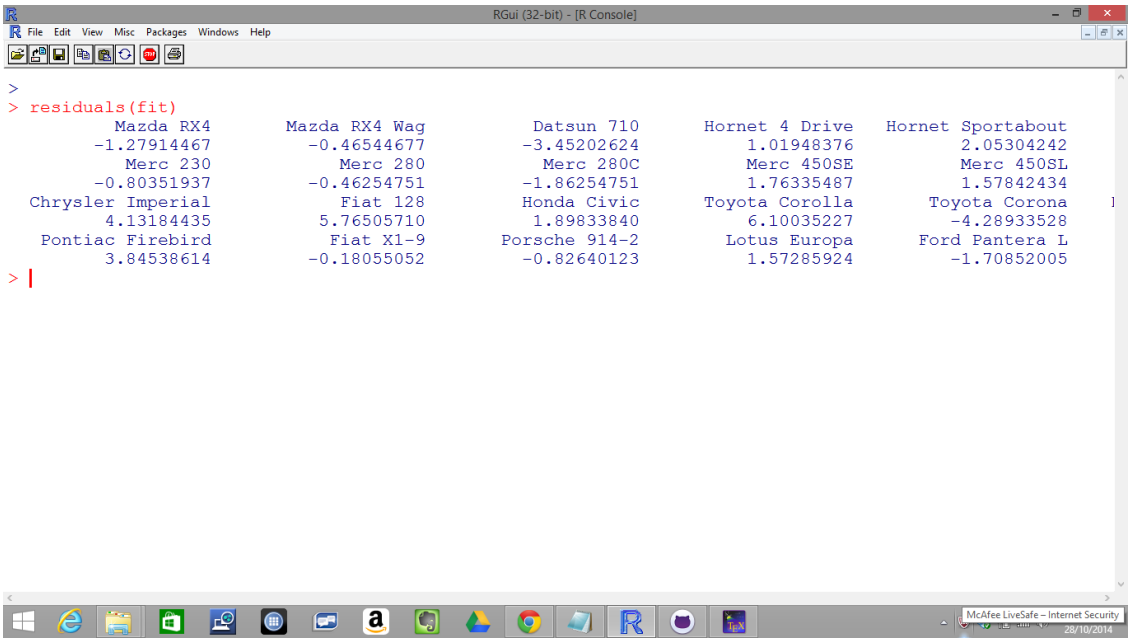


Figure 2:

```
> sum(residuals(fit))
[1] 1.096345e-15

> #Shapiro-Wilk Test for Normality
> shapiro.test(resid(fit))

 Shapiro-Wilk normality test

 data:  resid(fit)
 W = 0.9375, p-value = 0.06341
```

3 Some Important Definitions

3.1 Homoscedasticity

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

Suppose you plot the individual residuals against the predicted value, the variance of the residuals predicted value should be constant.
Consider the red arrows in the picture below, intended to indicate the variance of the residuals at that part of the number line. For the OLS summption to be valid , the length of the red lines should be more or less the same.

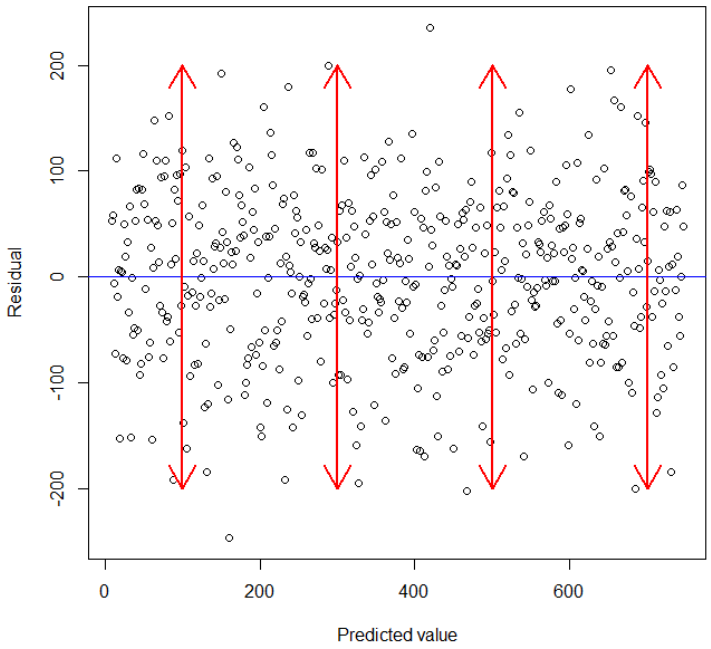


Figure 3:

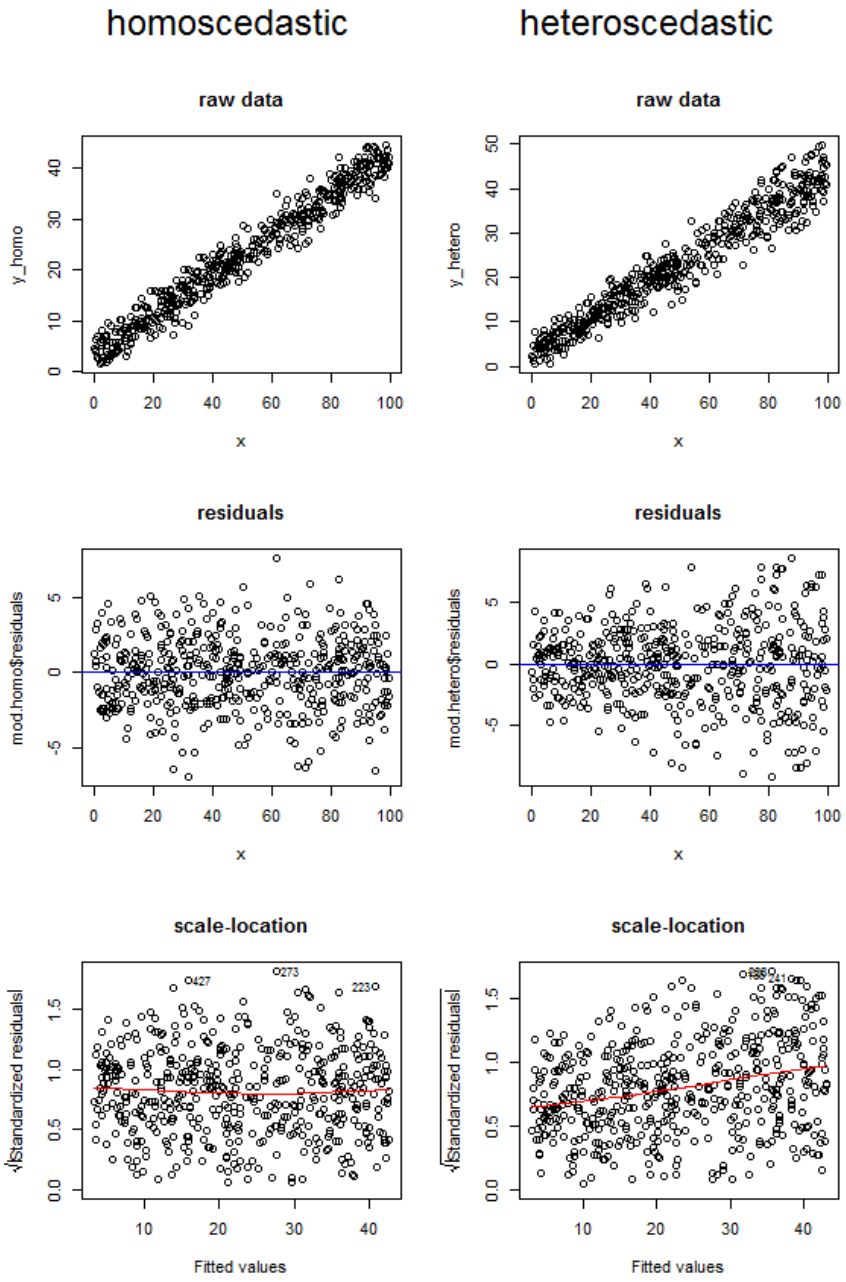

homoscedastic
heteroscedastic
raw data
raw data
y_homo
y_hetero
x
x
residuals
residuals
mod.homo$residuals
mod.hetero$residuals
x
x
scale-location
scale-location
√|Standardized residuals|
√|Standardized residuals|
427
273
223
Fitted values
Fitted values

Figure 4:

3.2 More Definitions

To understand a diagnostic plot called the residual-leverage plot, we need to understand three things:

- Leverage,
- Sstandardized residuals, and
- Cook's distance.

3.2.1 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.

This analogy can be taken 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 these diagnostic plots are intended to identify.

3.2.2 Standardization

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).

3.2.3 Cook's Distance

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

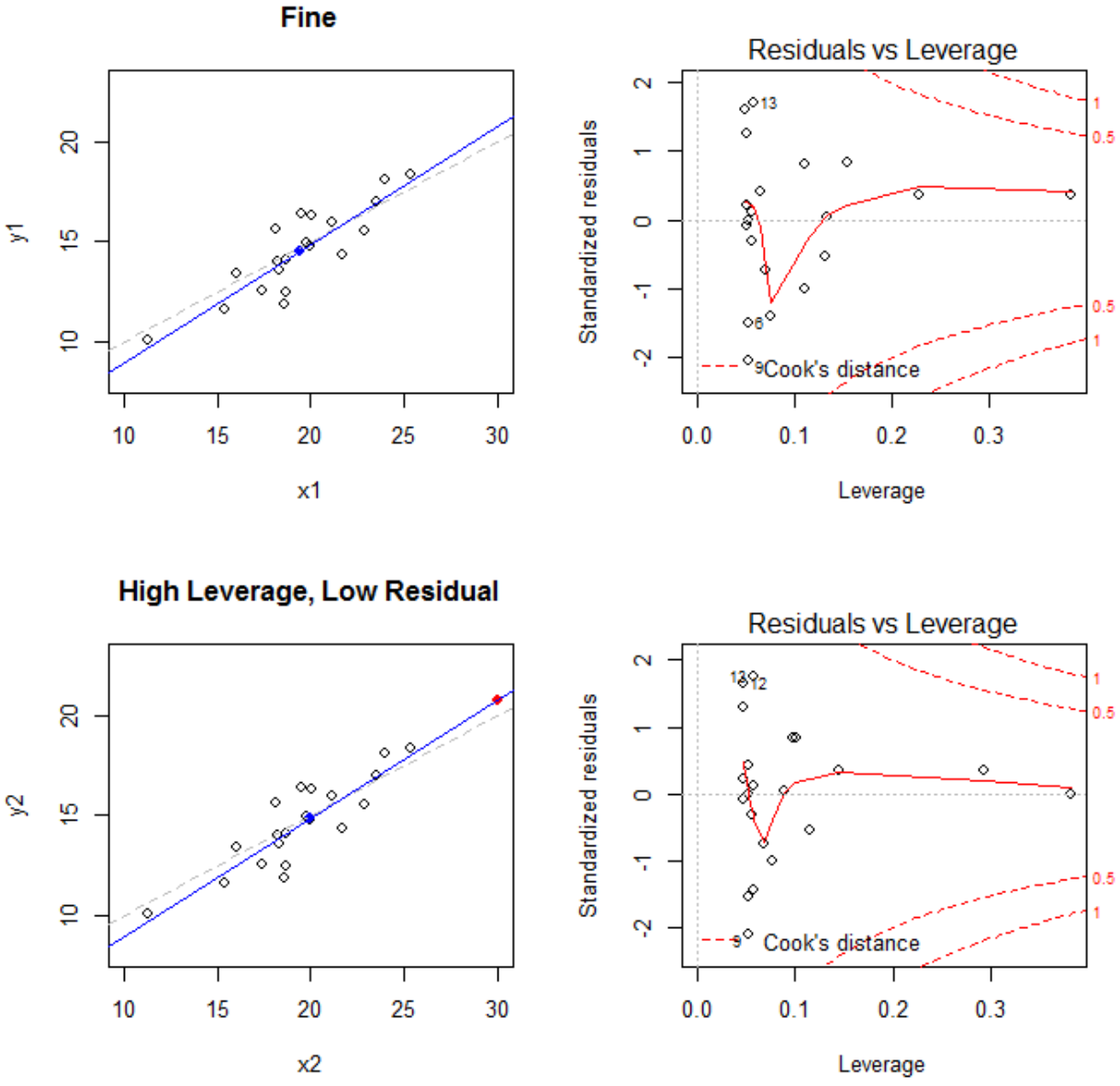

Fine
y1
x1
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
High Leverage, Low Residual
y2
x2
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Figure 5:

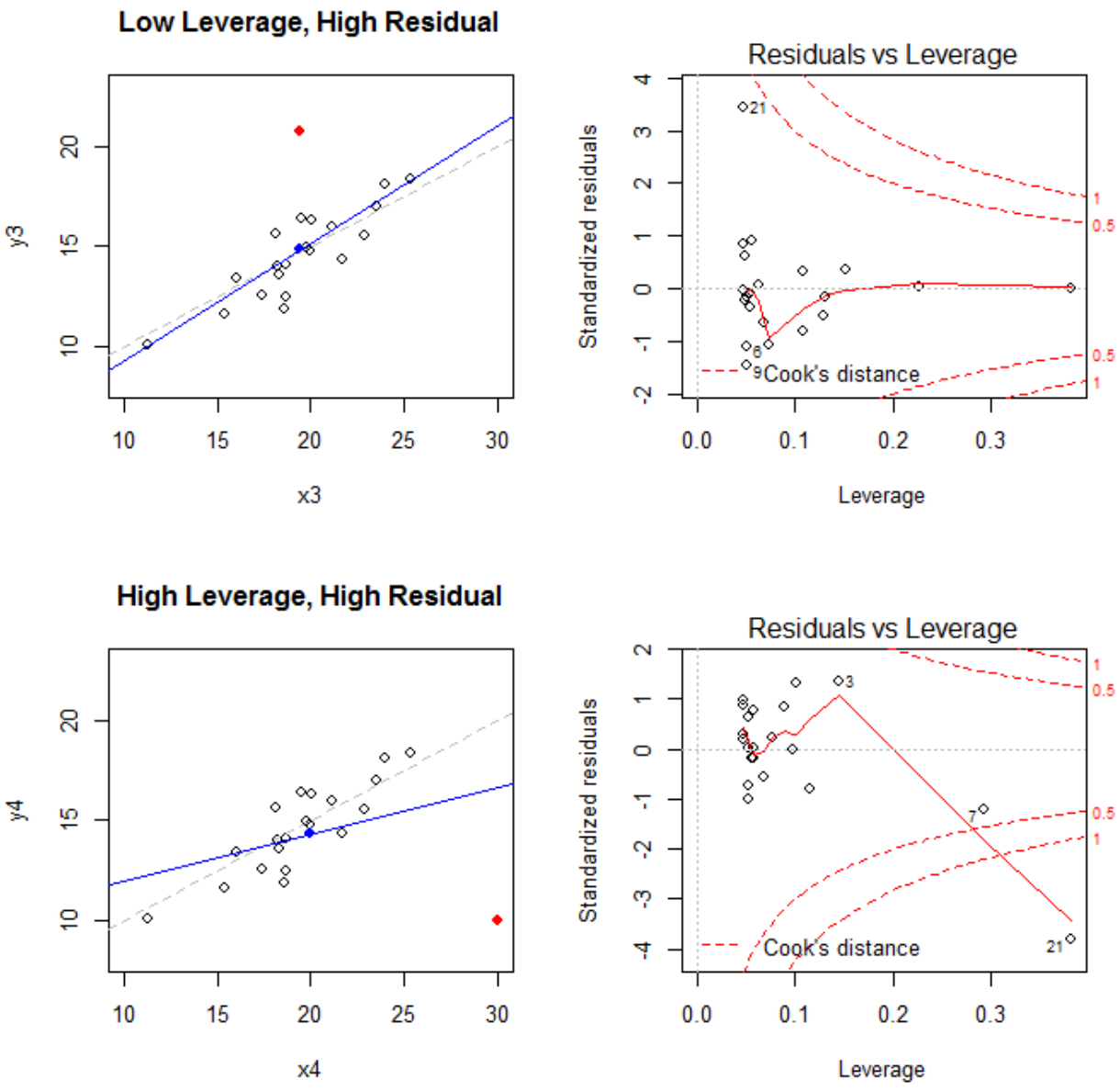

Low Leverage, High Residual
y3
x3
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
High Leverage, High Residual
y4
x4
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Figure 6:

- The plots on the left show the data, the center of the data with a blue dot, the underlying data generating process with a dashed gray line, the model fit with a blue line, and the special point with a red dot.
- On the right are the corresponding residual-leverage plots; the special point is 21.
- The model is badly distorted primarily in the fourth case where there is a point with high leverage and a large (negative) standardized residual.

4 Diagnostic Plots for Linear Models with R

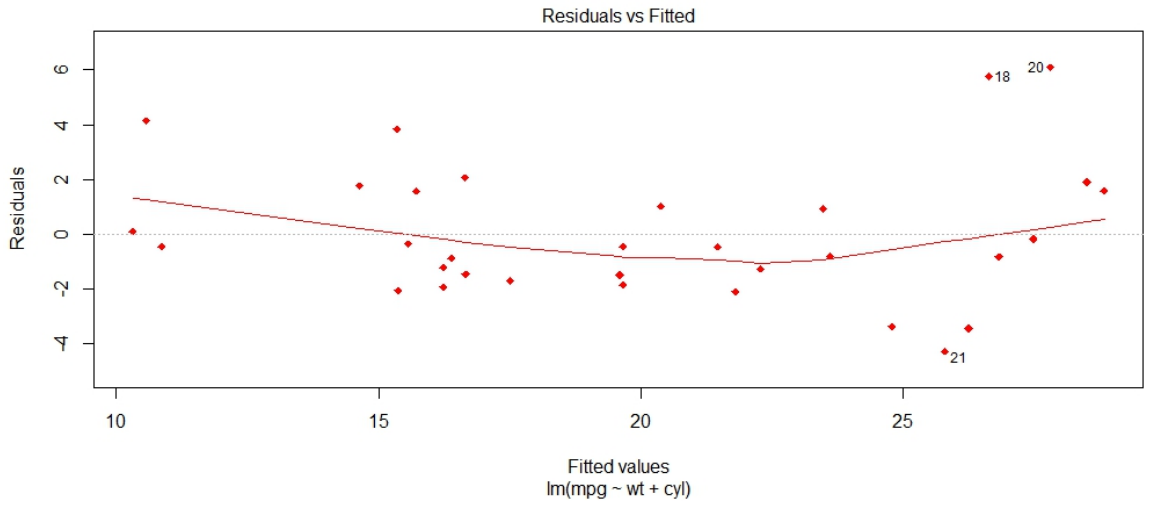
There are six plots (selectable by `which`) are currently available:

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)*.

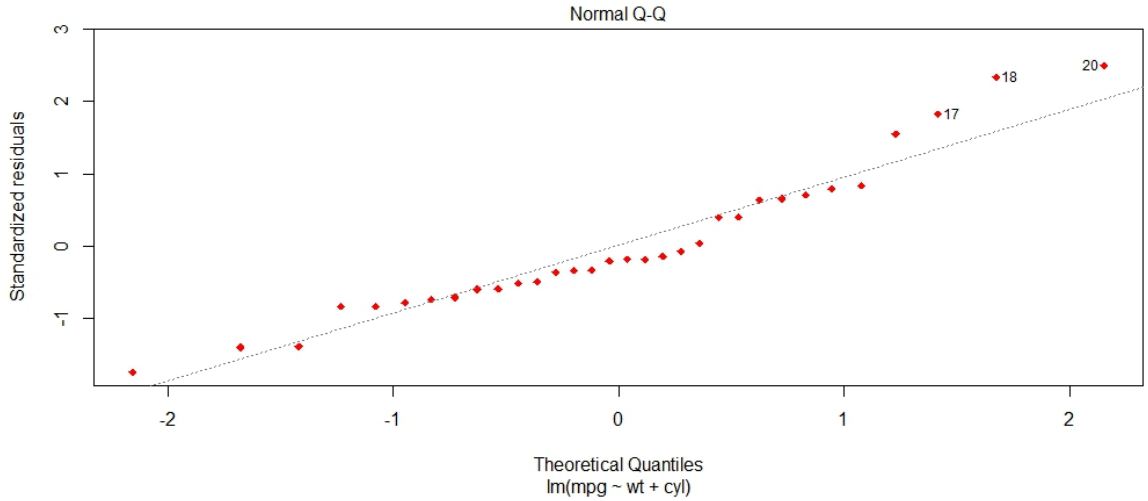
By default, the first three and 5 are provided, if you just type something like `plot(fit)`.

```
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")
```

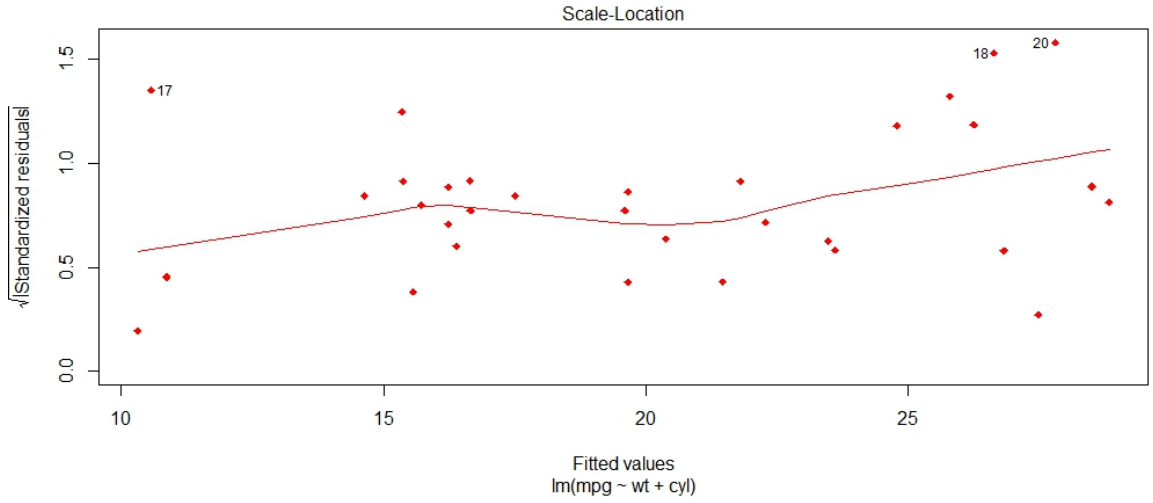
- The **Scale-Location** plot, also called Spread-Location (or S-L plot), takes the square root of the absolute residuals in order to diminish skewness (sqrt($|E|$)) is much less skewed than $|E|$ for Gaussian zero-mean E).
- **Plot 5** - The **Residual-Leverage** plot shows contours of equal Cook's 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.)



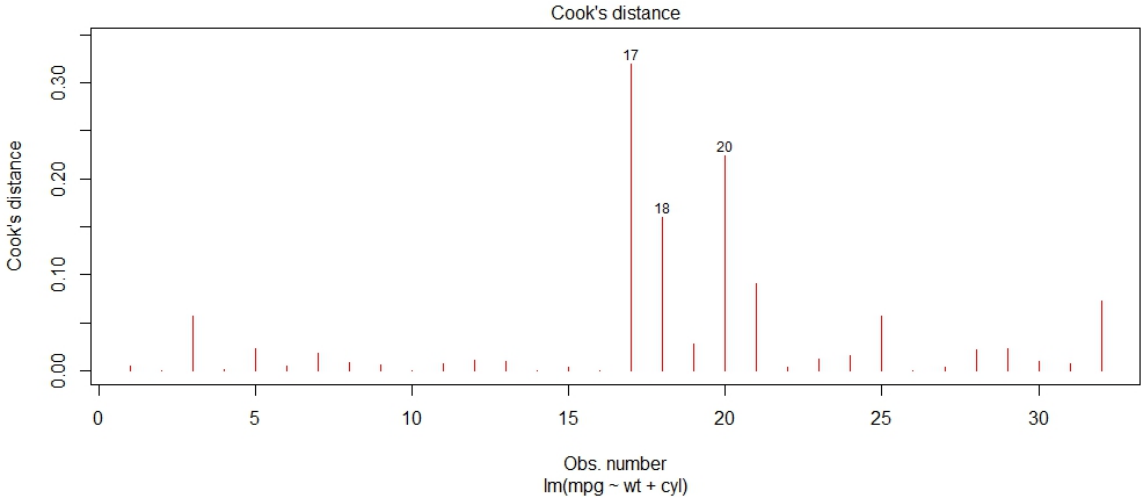
Plot 2 : Normal Probability Plot

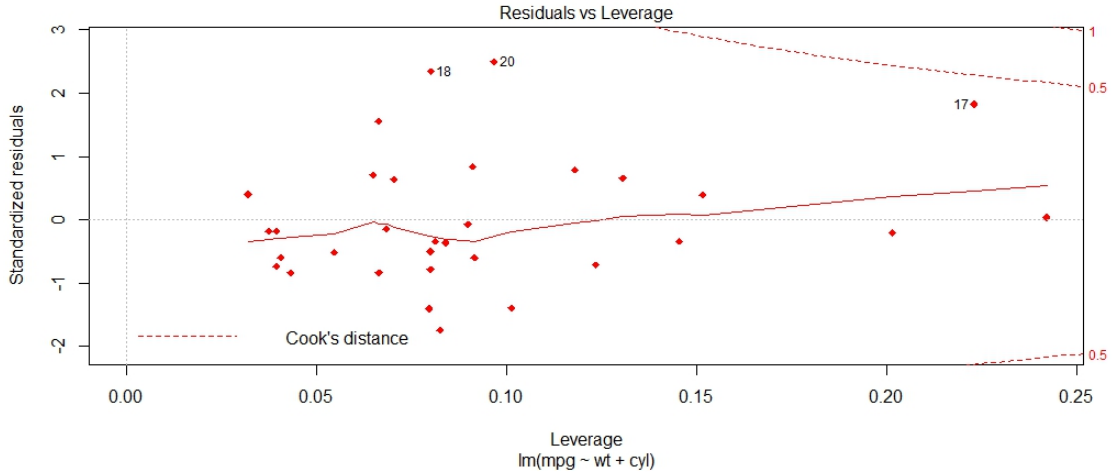


This plot is used to assess the validity of the normality of the residuals.

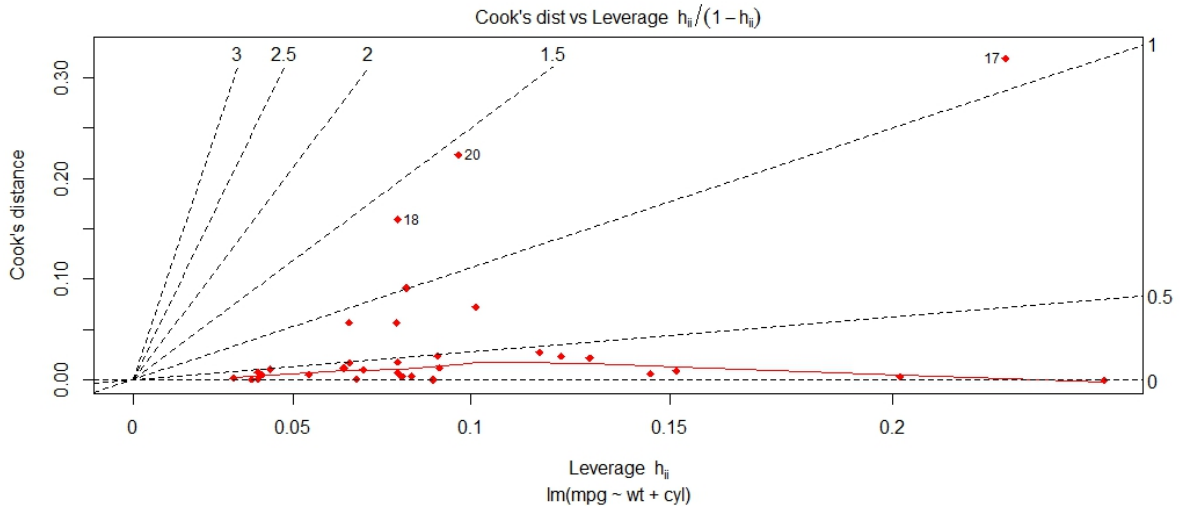


Plot 5 : Cook's Distance





Plot 6 : Cook's Distance vs Leverage



Plot the four default plots together:

```
par(mfrow=c(4,1))
plot(fittedmodel)
par(opar)
```

5 Outliers and Influential Observations

5.1 Outliers

Data points that diverge in a big way from the overall pattern are called outliers. There are four ways that a data point might be considered an outlier.

- It could have an extreme X value compared to other data points.
- It could have an extreme Y value compared to other data points.
- It could have extreme X and Y values.
- It might be distant from the rest of the data, even without extreme X or Y values.

After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line.

5.2 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.

5.3 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.

Purpose Cook's distance is useful for identifying outliers in the X values (observations for predictor variables). It also shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier.

Definition Cook's distance is the scaled change in fitted values. Each element in CooksDistance is the normalized change in the vector of coefficients due to the deletion of an observation.

Interpreting 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.

I would add that influential cases are not usually a problem when their removal from the dataset would leave the parameter estimates essentially unchanged: the ones we worry about are those whose presence really does change the results.

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, 0, and 1 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.

There is one other point worth making here. In observational research, it is often difficult to sample uniformly across the predictor space, and you might have just a few points in a given area. Such points can diverge from the rest. Having a few, distinct cases can be discomfiting, but merit considerable thought before being relegated outliers. There may legitimately be an interaction amongst the predictors, or the system may shift to behave differently when predictor values become extreme. In addition, they may be able to help you untangle the effects of colinear predictors. Influential points could be a blessing in disguise.

Cook's Distance Formula

It is calculated as:

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

where:

- $\hat{Y}_j$ 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;
- p is the number of fitted parameters in the model;
- MSE is the mean square error of the regression model.

For the case of simple linear regression, the following are the algebraically equivalent expressions

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

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

where: h_{ii} is the leverage, i.e., 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 residual (i.e., the difference between the observed value and the value fitted by the proposed model).

The Cook's distance, Di, of observation i is

is the jth fitted response value, where the fit does not include observation i. MSE is the mean squared error. p is the number of coefficients in the regression model. Cook's distance is algebraically equivalent to the following expression:

where ri is the ith residual, and hii is the ith leverage value.

CooksDistance is an n-by-1 column vector in the Diagnostics table of the LinearModel object.

How To After obtaining a fitted model, say, mdl, using fitlm or stepwiselm, you can:

Display the Cook's distance values by indexing into the property using dot notation, mdl.Diagnostics.CooksDistance Plot the Cook's distance values using plotDiagnostics(mdl,'cookd') For details, see the plotDiagnostics method of the LinearModel class. Determine Outliers Using Cook's Distance This example shows how to use Cook's Distance to determine the outliers in the data.

6 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 $\hat{V}$.

6.1 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.

6.1.1 Computation

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

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

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

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

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

6.1.2 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

6.1.3 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.

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

6.1.4 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.

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

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

6.1.5 DFBETA

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

6.2 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.

6.3 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.

7 Influential Points in Regression

Sometimes in regression analysis, a few data points have disproportionate effects on the slope of the regression equation. In this lesson, we describe how to identify those influential points.

8 Leverage and Influence

8.1 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.

8.2 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.

8.2.1 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.

8.3 Influential Points

An influential point is an outlier that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

This type of analysis is illustrated below. The scatter plots are identical, except that the plot on the right includes an outlier. The slope is flatter when the outlier is present (-3.32 vs. -4.10), so this outlier would be considered an influential point.

8.4 Without Outlier

- Regression equation: $\hat{y} = 104.78 - 4.10x$

- Coefficient of determination: $R^2 = 0.94$
- Regression equation: $\hat{y} = 97.51 - 3.32x$
- Coefficient of determination: $R^2 = 0.55$

The charts below compare regression statistics for another data set with and without an outlier. Here, the chart on the right has a single outlier, located at the high end of the X axis (where x = 24). As a result of that single outlier, the slope of the regression line changes greatly, from -2.5 to -1.6; so the outlier would be considered an influential point.

Sometimes, an influential point will cause the coefficient of determination to be bigger; sometimes, smaller. In the first example above, the coefficient of determination is smaller when the influential point is present (0.94 vs. 0.55). In the second example, it is bigger (0.46 vs. 0.52).

If your data set includes an influential point, here are some things to consider.

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

8.5 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.

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.
- 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.

It is calculated as:

$$Di = nj = 1(Y^j Y^j(i))2pMSE,$$

8.6 Leverage

- In statistics, 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.

9 Regression Deletion Diagnostics

This suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models discussed in Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), etc.

```
Usage
influence.measures(model)

rstandard(model, ...)
## S3 method for class 'lm'
rstandard(model, infl = lm.influence(model, do.coef = FALSE),
          sd = sqrt(deviance(model)/df.residual(model)), ...)
## S3 method for class 'glm'
rstandard(model, infl = influence(model, do.coef = FALSE),
          type = c("deviance", "pearson"), ...)
```

```
rstudent(model, ...)
## S3 method for class 'lm'
rstudent(model, infl = lm.influence(model, do.coef = FALSE),
         res = infl$wt.res, ...)
## S3 method for class 'glm'
rstudent(model, infl = influence(model, do.coef = FALSE), ...)

dffits(model, infl = , res = )
```

```
dfbeta(model, ...)
## S3 method for class 'lm'
dfbeta(model, infl = lm.influence(model, do.coef = TRUE), ...)

dfbetas(model, ...)
## S3 method for class 'lm'
dfbetas(model, infl = lm.influence(model, do.coef = TRUE), ...)
```

```
covratio(model, infl = lm.influence(model, do.coef = FALSE),
         res = weighted.residuals(model))
```

```
cooks.distance(model, ...)
## S3 method for class 'lm'
cooks.distance(model, infl = lm.influence(model, do.coef = FALSE),
               res = weighted.residuals(model),
               sd = sqrt(deviance(model)/df.residual(model)),
               hat = infl$hat, ...)
## S3 method for class 'glm'
cooks.distance(model, infl = influence(model, do.coef = FALSE),
               res = infl$pear.res,
               dispersion = summary(model)$dispersion,
               hat = infl$hat, ...)
```

```
hatvalues(model, ...)
## S3 method for class 'lm'
hatvalues(model, infl = lm.influence(model, do.coef = FALSE), ...)

hat(x, intercept = TRUE)
```

Arguments

model an R object, typically returned by lm or glm.

infl influence structure as returned by lm.influence or influence (the latter only for t

res (possibly weighted) residuals, with proper default.

sd standard deviation to use, see default.

dispersion dispersion (for glm objects) to use, see default.

hat hat values H[i,i], see default.

type type of residuals for glm method for rstandard.

x the X or design matrix.

intercept should an intercept column be prepended to x?

... further arguments passed to or from other methods.

Details

- The primary high-level function is `influence.measures` which produces a class "infl" object tabular display showing the DFBETAS for each model variable, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the hat matrix. Cases which are influential with respect to any of these measures are marked with an asterisk.
- The functions `dfbetas`, `dffits`, `covratio` and `cooks.distance` provide direct access to the corresponding diagnostic quantities.
- Functions `rstandard` and `rstudent` give the standardized and Studentized residuals respectively.
- (These functions re-normalize the residuals to have unit variance, using an overall and leave-one-out measure of the error variance respectively.)
- Values for generalized linear models are approximations, as described in Williams (1987) (except that Cook's distances are scaled as F rather than as chi-square values). The approximations can be poor when some cases have large influence.
- The optional `infl`, `res` and `sd` arguments are there to encourage the use of these direct access functions, in situations where, e.g., the underlying basic influence measures (from `lm.influence` or the generic influence) are already available.
- Note that cases with `weights == 0` are dropped from all these functions, but that if a linear model has been fitted with `na.action = na.exclude`, suitable values are filled in for the cases excluded during fitting.
- The function `hat()` exists mainly for S (version 2) compatibility; we recommend using `hatvalues()` instead.

10 LME Models

11 residuals.lme nlme- Extract lme Residuals

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.

```
fm1 <- lme(distance ~ age + Sex,
     data = Orthodont, random = ~ 1)
head(residuals(fm1, level = 0:1))
summary(residuals(fm1) /
        residuals(fm1, type = "p"))

# constant scaling factor 1.432
```

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 lmehas an n-by-p fixed-effects design matrix X and an n-by-q random-effects design matrix 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

$$r_{Cond} = yX\hat{\beta}Z\hat{b},$$

and the marginal raw residual is

$$r_{Mar} = yX\hat{\beta}.$$

For more information on other types of residuals, see the ResidualType name-value pair argument.