

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 ***independent 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 $y_1, y_2, \dots, y_i \dots y_n$
- 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:

Residual = Observed value - Predicted value

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

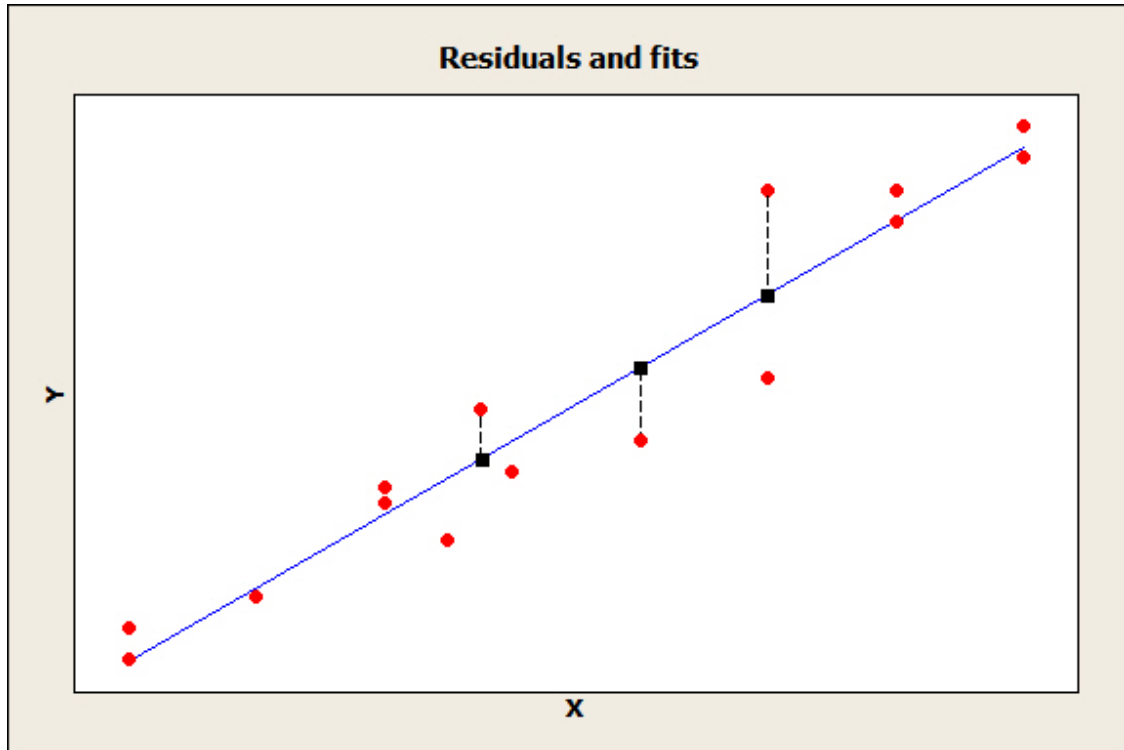


Figure 1:

- 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.
 $E(e) = 0$
- 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)
```

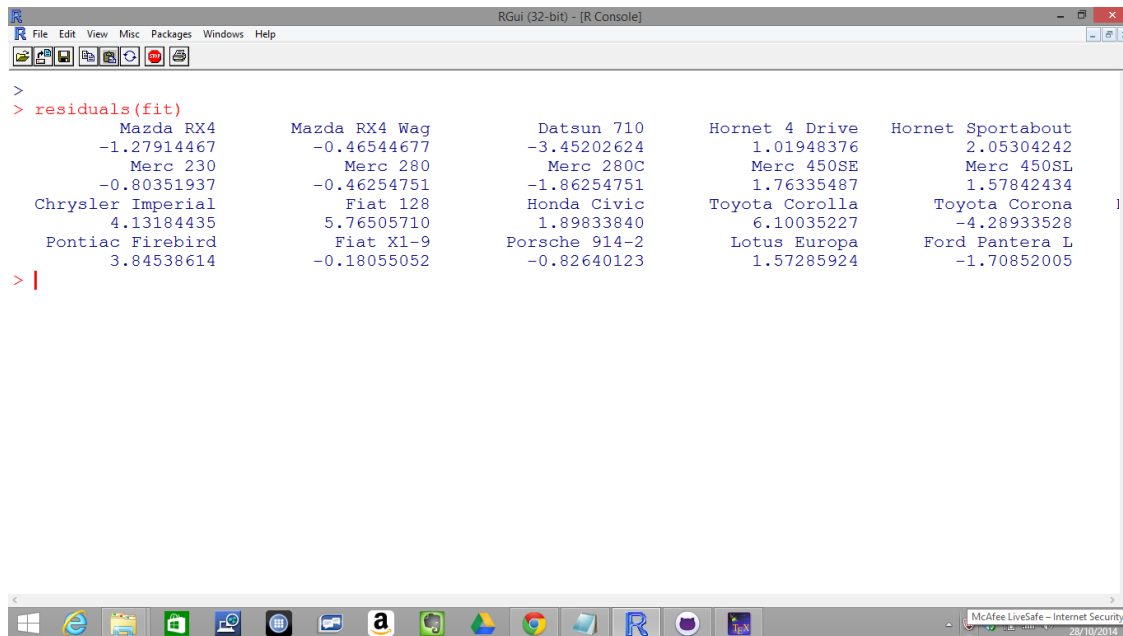


Figure 2:

```
> 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
```


2.1 resid - Extracting Model Residuals

- **residuals** is a generic function which extracts model residuals from objects returned by modeling functions.
- The abbreviated form **resid** is an alias for residuals. It is intended to encourage users to access object components through an accessor function rather than by directly referencing an object slot.
- All object classes which are returned by model fitting functions should provide a residuals method. (Note that the method is for **residuals** and not **resid**.)
- Methods can make use of **naresid** methods to compensate for the omission of missing values. The default, **nls** and **smooth.spline** methods do.

```
residuals(fit)
```

```
resid(fit)
```

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.

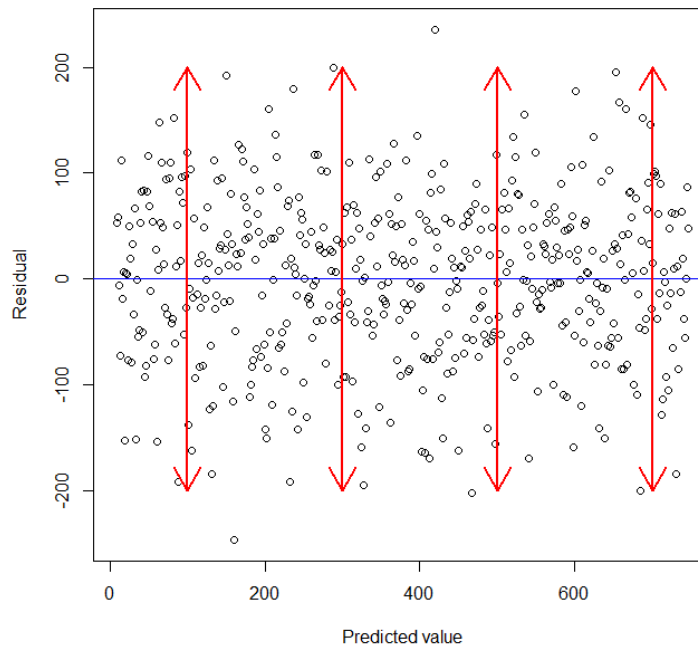


Figure 3:

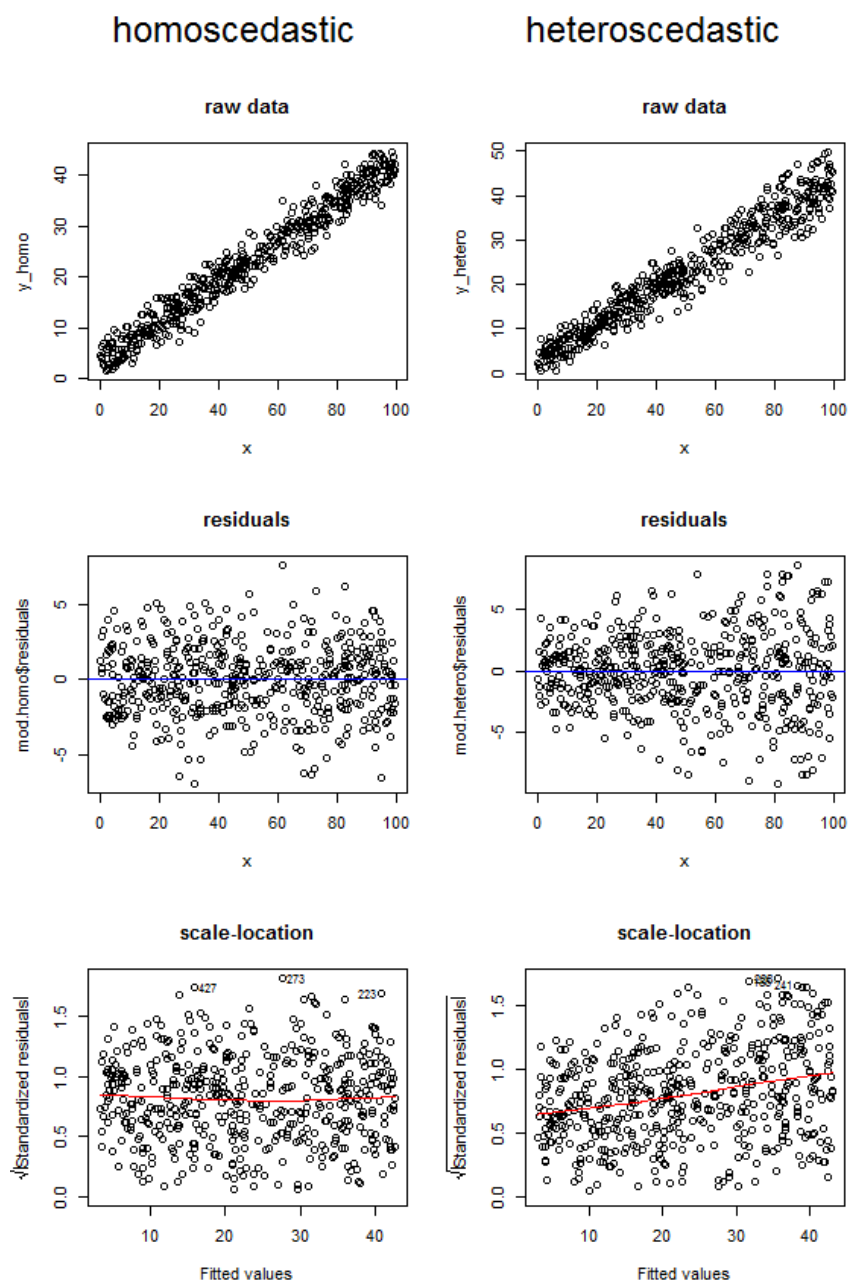


Figure 4:

3.2 More Definitions

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

- Leverage,
- Standardized 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

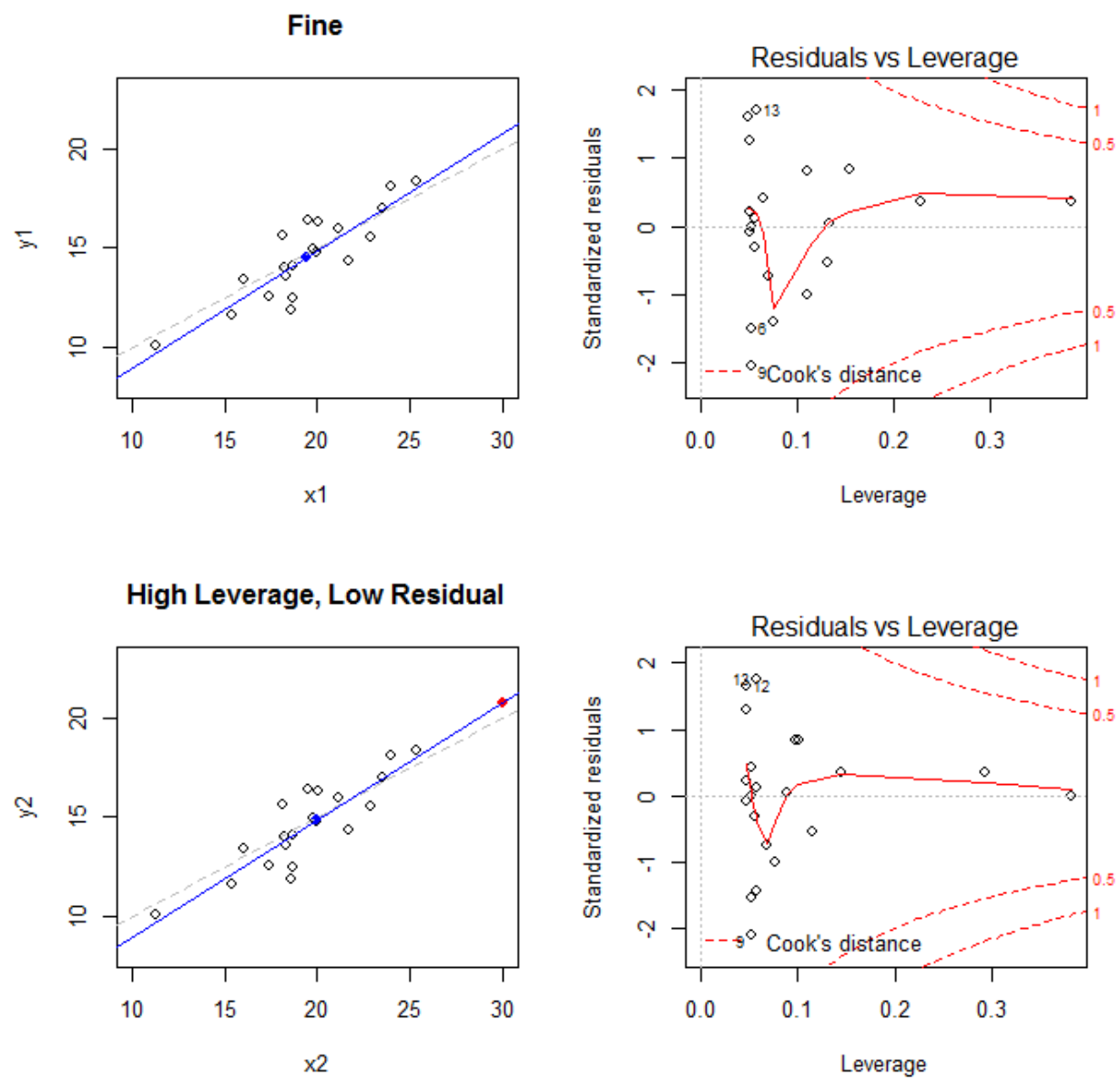


Figure 5:

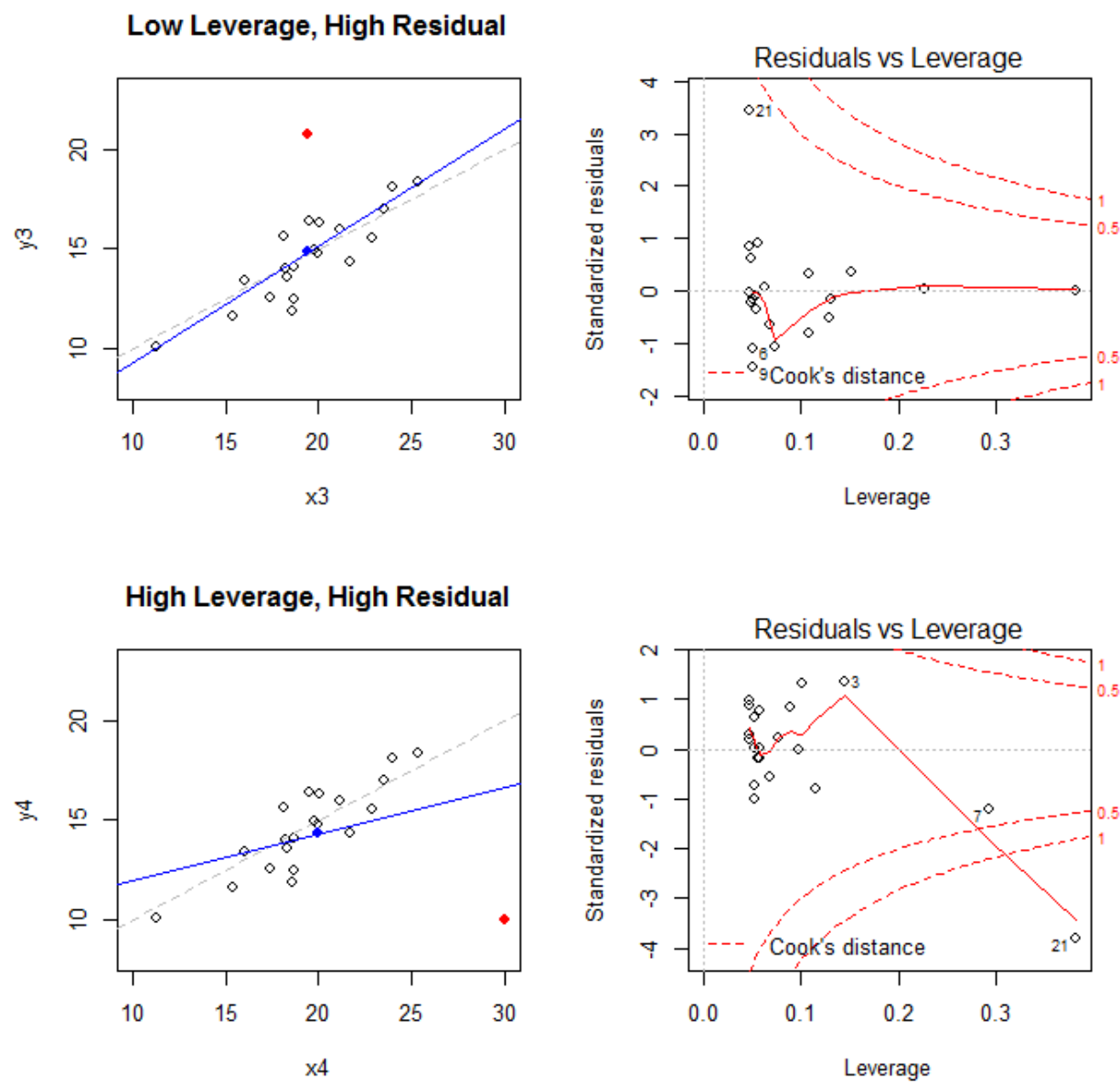


Figure 6:

- 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

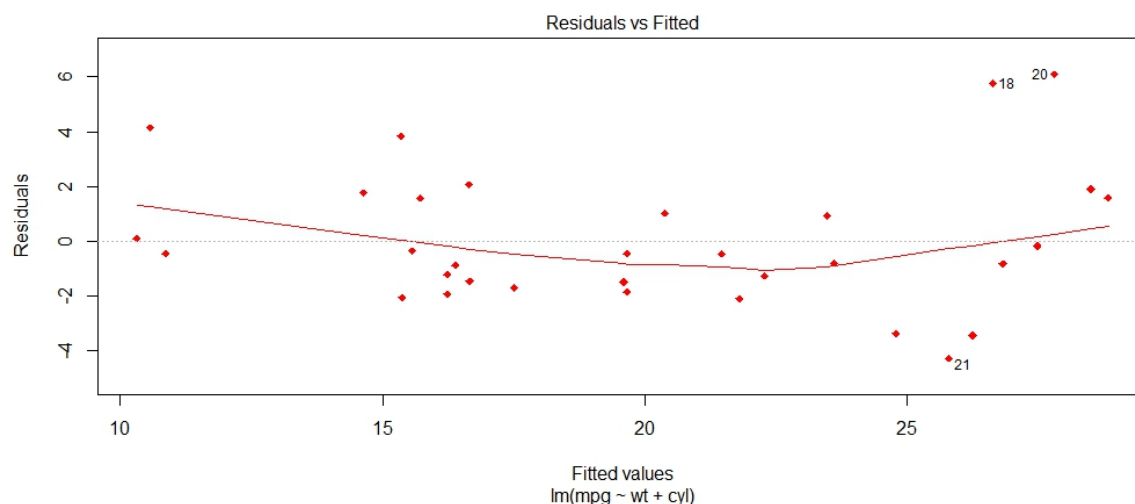
There are six plots (selectable by `which`) are currently available:

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 $leverage/(1-leverage)$.

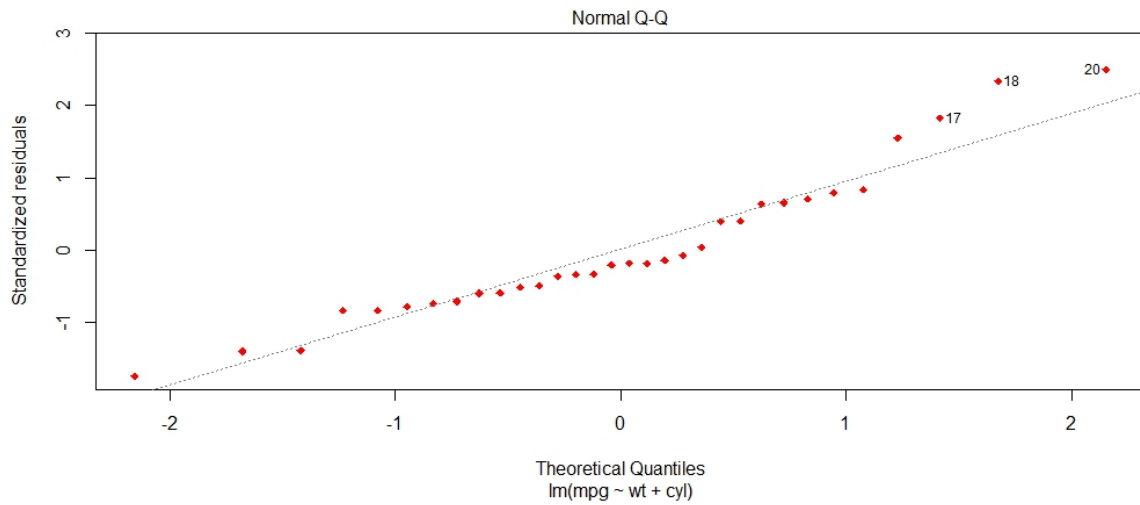
By default, the first three and 5 are provided, if you just type something like `plot(fit)`.

```
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")
```

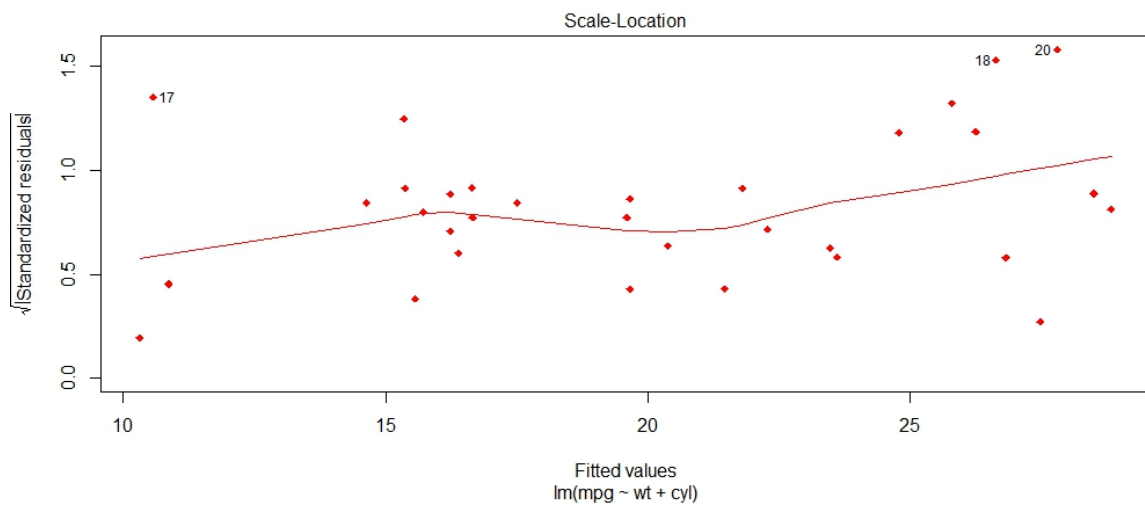
- The **Scale-Location** plot, also called Spread-Location (or S-L plot), takes the square root of the absolute residuals in order to diminish skewness ($\sqrt{|E|}$) is much less skewed than $|E|$ for Gaussian zero-mean E).
- **Plot 5** - The **Residual-Leverage** plot shows contours of equal Cook's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



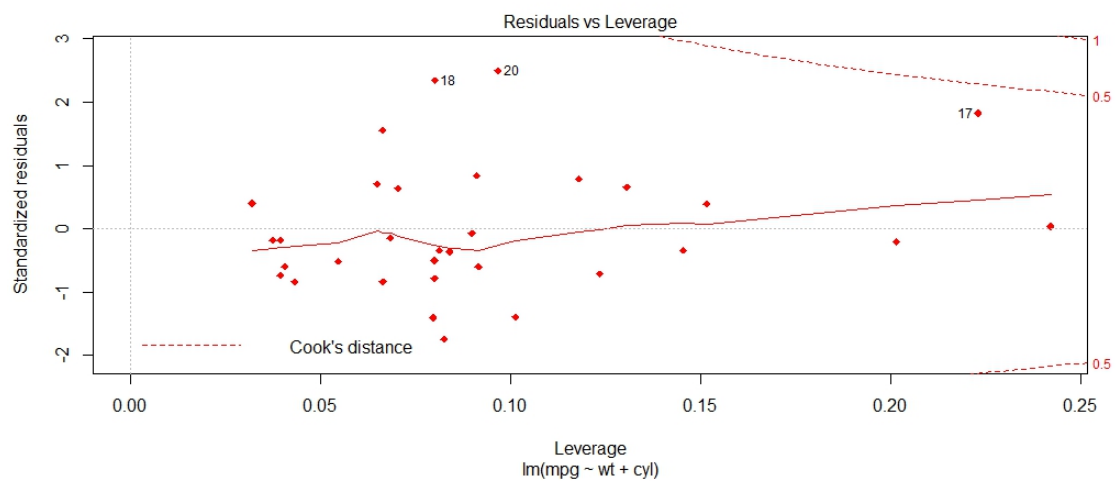
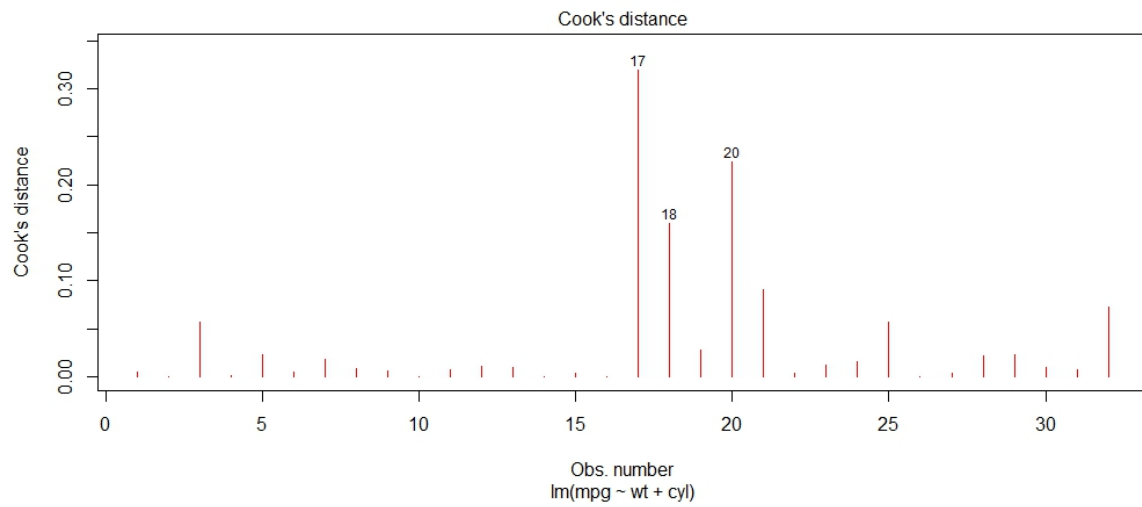
Plot 2 : Normal Probability Plot



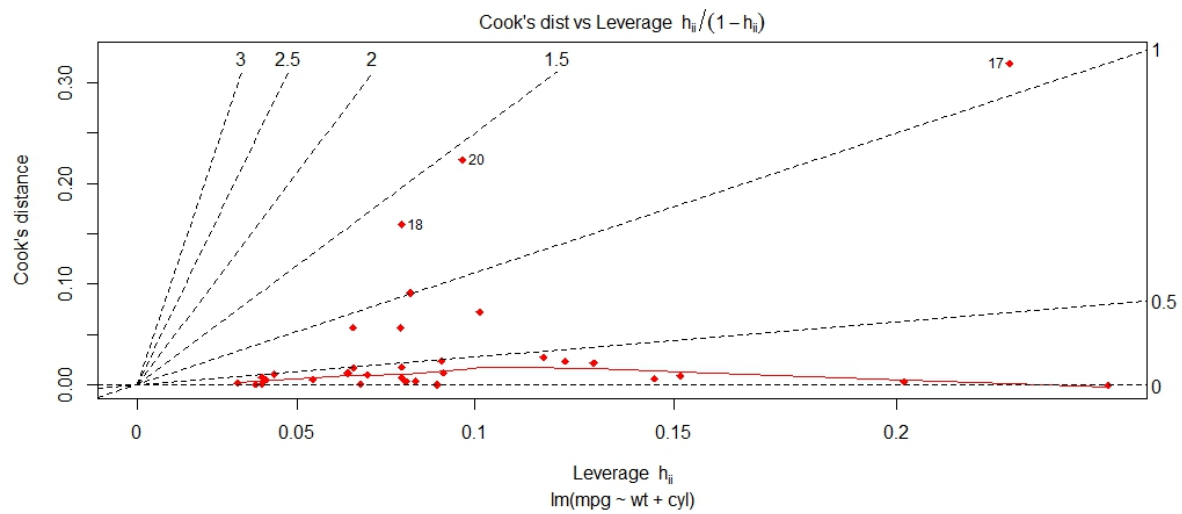
This plot is used to assess the validity of the normality of the residuals.



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

”Outliers are sample values that cause surprise in relation to the majority of the sample” (W.N. Venables and B.D. Ripley. 2002. Modern applied statistics with S. New York: Springer, p.119).

Crucially, surprise is in the mind of the beholder and is dependent on some tacit or explicit model of the data.

There may be another model under which the outlier is not surprising at all, say if the data really are lognormal or gamma rather than normal.

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.

6 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.
- 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.

- Cook's distance is the scaled change in fitted values. Each resulting element in a diagnostic calculation is the normalized change in the vector of coefficients due to the deletion of an observation.
- 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.
- 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 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.

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

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

- e_i is the residual (i.e., the difference between the observed value and the value fitted by the proposed model).

R code for computing Cook's Distance

```
attach(mtcars)
fit = lm(mpg ~ cyl + wt )
cooks.distance(fit)
plot(cooks.distance(fit),type="b",pch=18,col="red")
```

$N = 32$

$k = 2$

$\text{cutoff} = 4 / (N - k - 1)$

$\text{abline}(h=\text{cutoff},lty=2)$

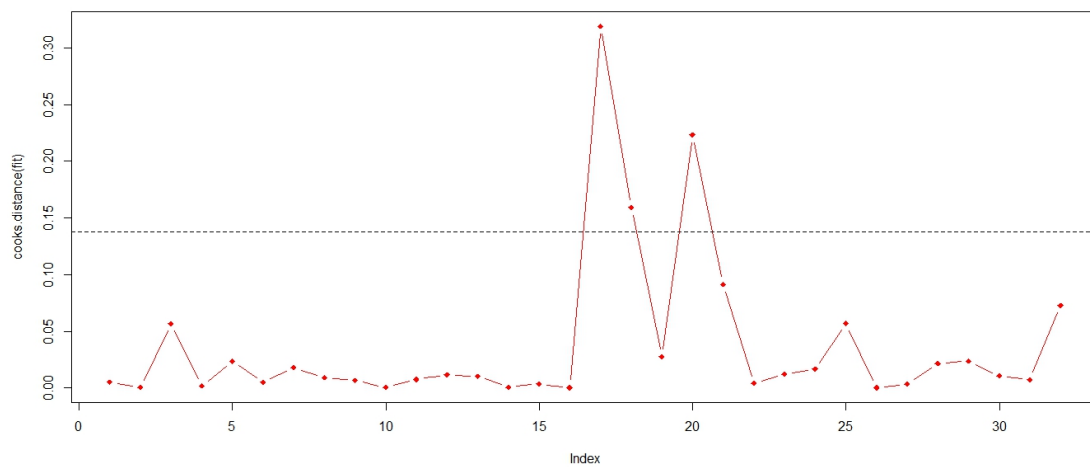


Figure 7:

6.1 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. (John Fox) (1991). *Regression Diagnostics: An Introduction*. Sage Publications.

- 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

$$\frac{4}{(N - k - 1)},$$

where N is the number of observations and k the number of explanatory variables.

- The R help file advises that an observation with Cook's distance larger than three times the mean Cook's distance might be an outlier. .
- John Fox (mentioned above), 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.

Remark

- 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** : *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 β_1 for each covariate).*
- Cook's distance is arguably more important if you are doing predictive modeling, whereas *DFBETA* is more important in explanatory modeling.
- **DFFITS**: Although the raw values resulting from the equations are different, Cook's distance and *DFFITS* are conceptually identical and there is a closed-form formula to convert one value to the other.

Regression Diagnostics

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.

6.2 Alcohol and Tobacco Data

This example is for exposition only. We will ignore the fact that this may not be a great way of modeling the this particular set of data!

```
alctob <- data.frame( cbind(  
Alcohol = c(6.47, 6.13, 6.19, 4.89, 5.63, 4.52,  
            5.89, 4.79, 5.27, 6.08, 4.02),  
Tobacco = c(4.03, 3.76, 3.77, 3.34, 3.47, 2.92,  
            3.20, 2.71, 3.53, 4.51, 4.56)),  
row.names = c("North", "Yorkshire", "Northeast",
```

```
"East Midlands", "West Midlands", "East Anglia",  
"Southeast", "Southwest", "Wales",  
"Scotland", "N. Ireland"))  
  
# Assume that we are fitting a multiple linear regression  
# on the MTCARS data  
library(car)  
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
```

```
alctobwo <- subset(alctob,rownames(alctob)!="N. Ireland")  
#without North Ireland  
  
plot(alctob$Tobacco, alctob$Alcohol,  
main="Weekly Household Spending on Alcohol vs. Tobacco",  
xlab="Tobacco Spending (GBP)",  
ylab="Alcohol Spending (GBP)",  
pch=16,col="red",cex=1.5,font.lab=2)  
#note N. Ireland in the bottom-right
```

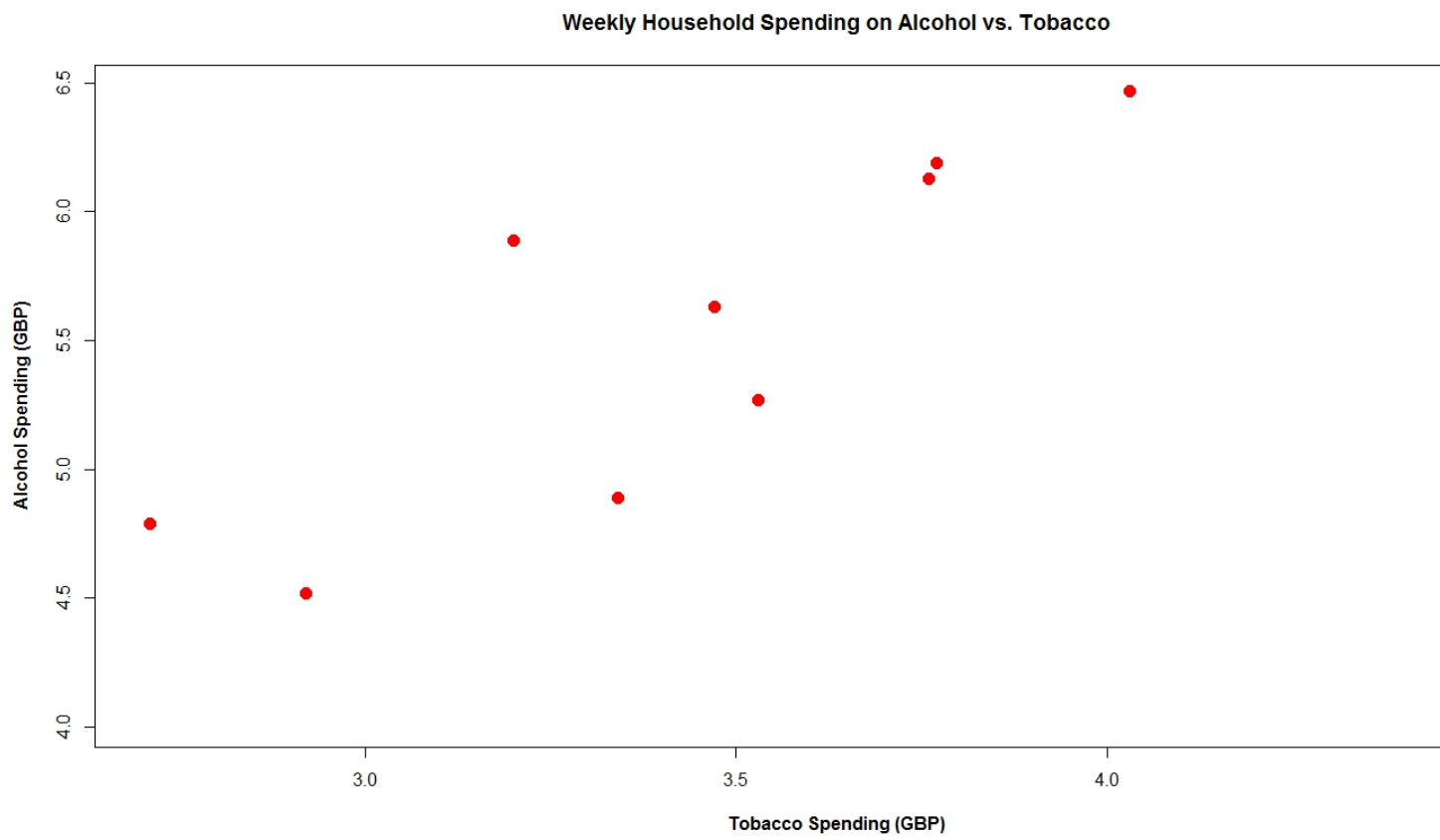


Figure 8:

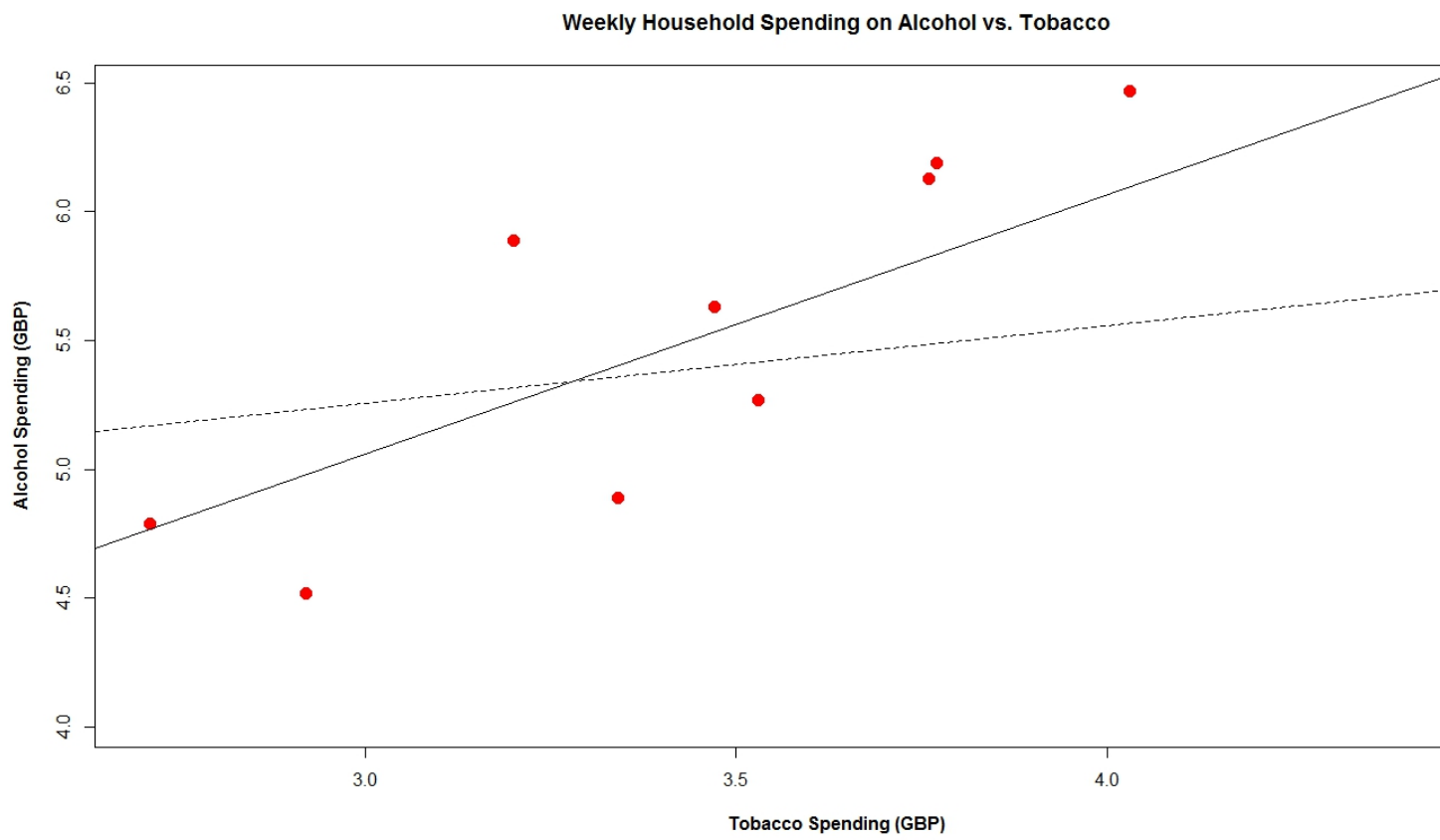


Figure 9:

All Observations

```
> summary(fit1)

Call:
lm(formula = Alcohol ~ Tobacco, data = alctob)

Residuals:
Min      1Q  Median      3Q      Max
-1.7080 -0.4245  0.2311  0.6081  0.9020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.3512     1.6067   2.708  0.0241 *
Tobacco        0.3019     0.4388   0.688  0.5087
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.8196 on 9 degrees of freedom
Multiple R-squared:  0.04998,    Adjusted R-squared:  -0.05557
F-statistic: 0.4735 on 1 and 9 DF,  p-value: 0.5087
```

Outlier Removed

```
> summary(fit2)

Call:
lm(formula = Alcohol ~ Tobacco, data = alctobwo)

Residuals:
Min      1Q  Median      3Q      Max
-0.51092 -0.42434  0.06056  0.34406  0.62991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0412     1.0014   2.038  0.07586 .
```



```
Tobacco      1.0059      0.2813    3.576  0.00723 **
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.446 on 8 degrees of freedom
```

```
Multiple R-squared:  0.6151,    Adjusted R-squared:  0.567
```

```
F-statistic: 12.78 on 1 and 8 DF,  p-value: 0.007234
```

Outliers

The conservative outlier test that we talked about in class uses the Bonferonni inequality to calculate the p-values we associate with the Student's-t test.

In R, we can use the `outlierTest` command to perform this test on our model. Remember that when we test for influence, we are testing the effect of an observation on model coefficients.

Therefore we need to give the `outlierTest` command a linear model as its input.

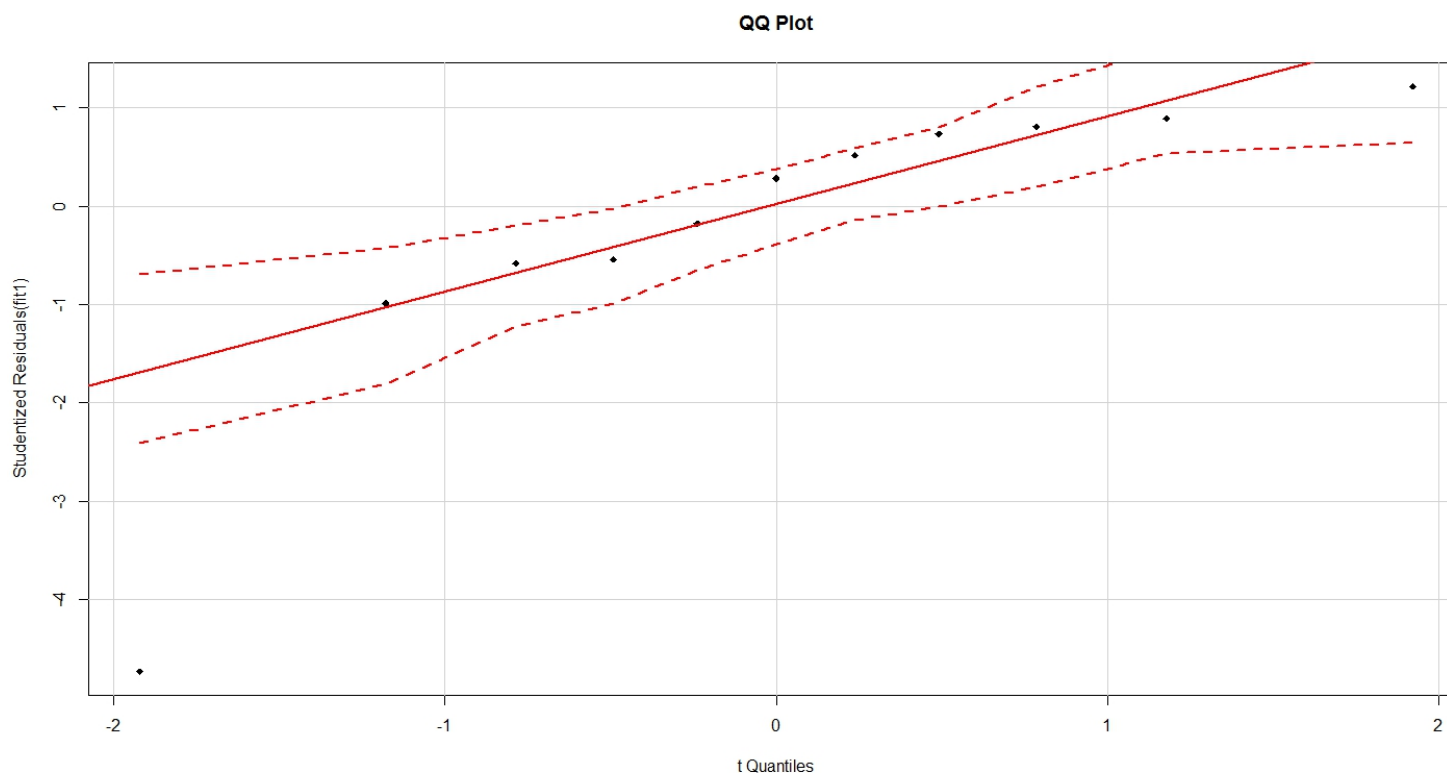
```
> # Assessing Outliers
> # syntax: outlierTest(fit)
> # Bonferonni p-value for most extreme obs
>
> outlierTest(fit1)
               rstudent  unadjusted p-value Bonferonni p
N. Ireland -4.732091          0.0014789      0.016268
>
```

We can also use R to calculate Cook's distance. Remember that generally we label any observation with Cook's distance greater than 1 as influential.

```
> cooks.distance(fit1)
North      Yorkshire      Northeast      East Midlands
0.114101051 0.036517838 0.043728951 0.023600304
West Midlands East Anglia      Southeast      Southwest
0.004740759 0.147326647 0.046646563 0.077488350
Wales      Scotland      N. Ireland
0.001821694 0.068921892 1.747233521
```

Finally, one of the easier ways to evaluate our residuals and look for influential points is through plots.

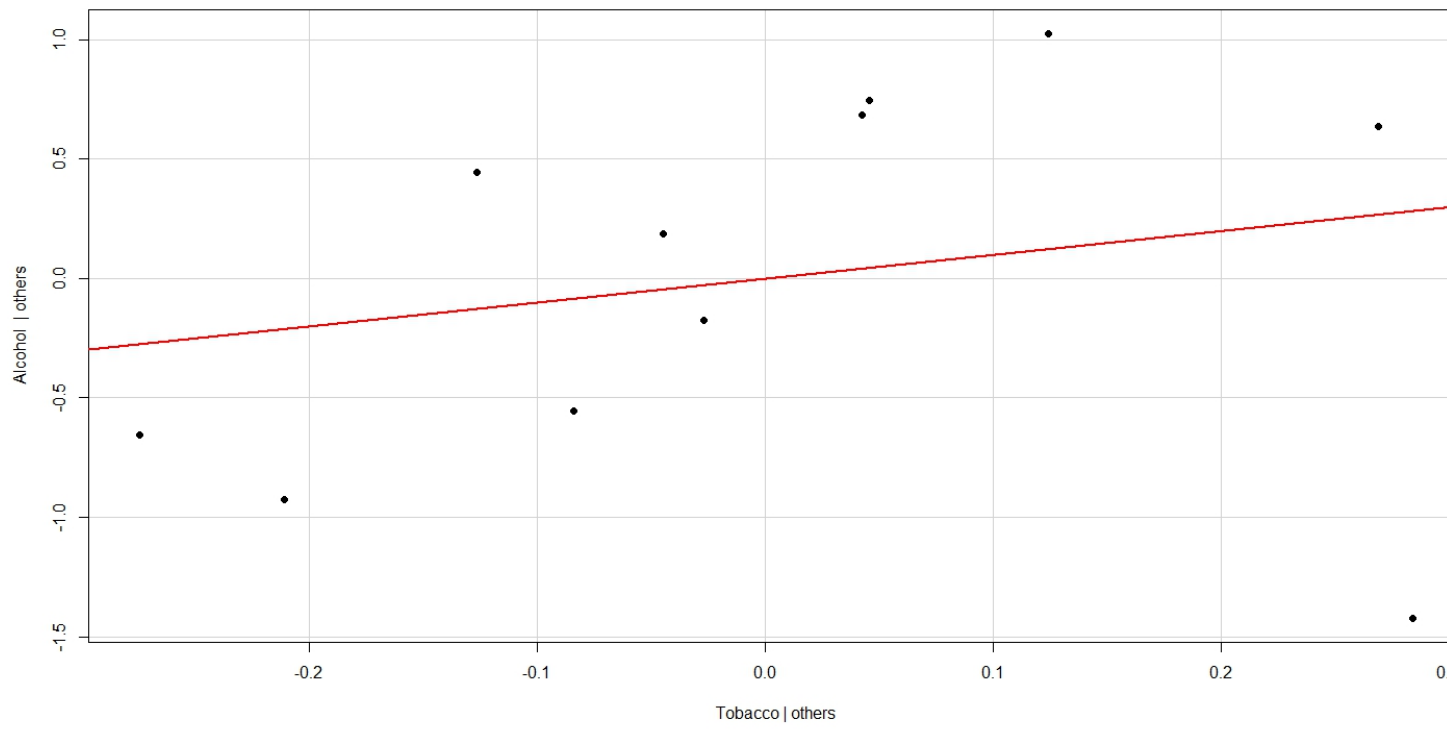
```
qqPlot(fit1, main="QQ Plot", pch=18, lim=c(-3,2)) #qq plot for studentized resid
leveragePlots(fit1) # leverage plots
```



In R, the `plot` command takes a special form when you pass it an `lm` object (see `?plot.lm` for all of the details).

Here we want to focus on three of the plots available through the command.

- The first one displays the residuals vs. the fitted values we use this to evaluate the mean, variance and correlation of residuals. If our assumptions of constant variance and uncorrelated residuals are violated we may be able to correct this with a variance-stabilizing transformation.
- The second plot helps us check the normality of the residuals. If the residuals are indeed normal, they should fall along the dashed line. Remember that the normality assumption for our errors allows us to determine the standard errors of our coefficients and predictions.
- The final plot will display our residuals vs. their leverage. The dashed red lines are level curves that denote a particular value of Cook's distance. We will pay attention to points lying beyond the distance of 1. Notice that when we have data with row labels, the points will be labeled with their names. Otherwise, the row number will be shown.



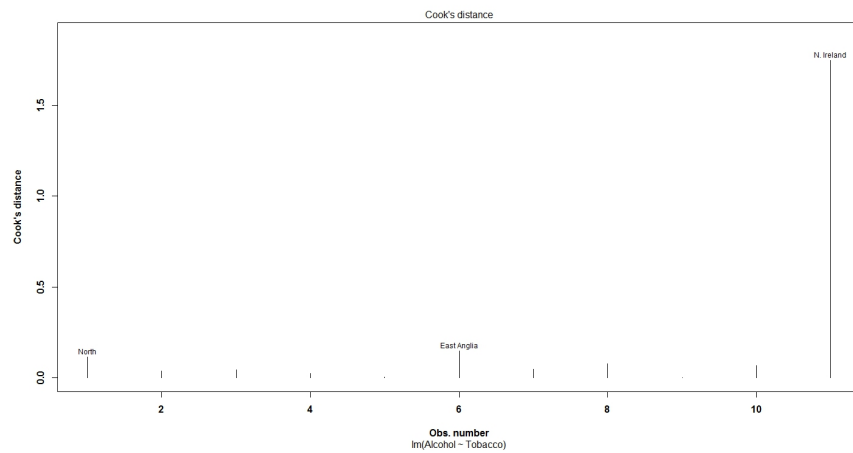


Figure 10:

```
plot(lmwith, which=c(1,2,5))
```

```
Influential Observations
# Influential Observations
# added variable plots
avPlots(fit1)

# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit1$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```

```
# Influence Plot
influencePlot(fit1,id.method="identify",
main="Influence Plot",
col="red"
sub="Circle size is proportional to Cook's Distance" )
```

Non-normality

```
# Normality of Residuals
# qq plot for studentized resid
qqPlot(fit1, main="QQ Plot")

# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

Non-constant Error Variance

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
spread vs. levels click to view
```

```
Multi-collinearity
# Evaluate Collinearity
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

6.3 Nonlinearity

```
# Evaluate Nonlinearity
# component + residual plot
crPlots(fit)
# Ceres plots
ceresPlots(fit)
component plus residual plot Ceres plots click to view
Non-independence of Errors
# Test for Autocorrelated Errors
durbinWatsonTest(fit)
```

6.4 Additional Diagnostic Help

The `gvlma()` function in the `gvlma` package, performs a global validation of linear model assumptions as well separate evaluations of skewness, kurtosis, and heteroscedasticity.

```
# Global test of model assumptions
library(gvlma)
gvmodel <- gvlma(fit)
summary(gvmodel)
```

6.5 Going Further

If you would like to delve deeper into regression diagnostics, two books written by John Fox can help: *Applied regression analysis and generalized linear models* (2nd ed) and *An R and S-Plus companion to applied regression*.

7 Standardization and Studentization

7.1 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.

7.2 Studentization

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

7.3 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.

7.4 Computation

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

8 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.

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.

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 = )
```

9 Standardized and Studentized Residuals

The standardized residual is the residual divided by its standard deviation.

Plot the standardized residual of the simple linear regression model of the data set `faithful` against the independent variable `waiting`.

We apply the `lm` function to a formula that describes the variable eruptions by the variable `waiting`, and save the linear regression model in a new variable `eruption.lm`. Then we compute the standardized residual with the `rstandard` function.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
> eruption.stdres = rstandard(eruption.lm)
```

We now plot the standardized residual against the observed values of the variable `waiting`.

```
> plot(faithful$waiting, eruption.stdres,
+      ylab="Standardized Residuals",
+      xlab="Waiting Time",
+      main="Old Faithful Eruptions")
> abline(0, 0) # the horizon
```

10 Leverage and Influence

10.1 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.

10.2 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.

10.3 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.

10.4 Summary of Influence Statistics

10.5 `rstudent`

The studentized residual RSTUDENT is estimated by $s(i)^2$ without the i th observation, not by s^2 . For example,

$$\text{RSTUDENT} = \frac{r_i}{s_{(i)} \sqrt{(1 - h_i)}}$$

Observations with RSTUDENT larger than 2 in absolute value may need some attention.

- **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.

11 Other 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} .

11.0.1 DFBETA

A group of measures that measures how much impact each observation has on a particular predictor are the 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.

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \quad (1)$$

$$= B(Y - Y_{\bar{a}}) \quad (2)$$

For k predictors in the model, there are $k + 1$ dfbetas.

11.0.2 DFFITS

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

$$DFFITS = \frac{\hat{y}_i - \widehat{y}_{i(k)}}{s_{(k)}\sqrt{h_{ii}}}$$

11.0.3 COVRATIO

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the i th observation:

$$COVRATIO = \frac{\det(s_{(i)}^2(X'_{(i)}X_{(i)})^{-1})}{\det(s^2(X'X)^{-1})}$$

Observations with

$$|COVRATIO - 1| \geq \frac{3k}{n}$$

where k is the number of parameters in the model and n is the number of observations used to fit the model, are worth further investigation.

11.1 Influential Observations : DFBeta and DFBetas

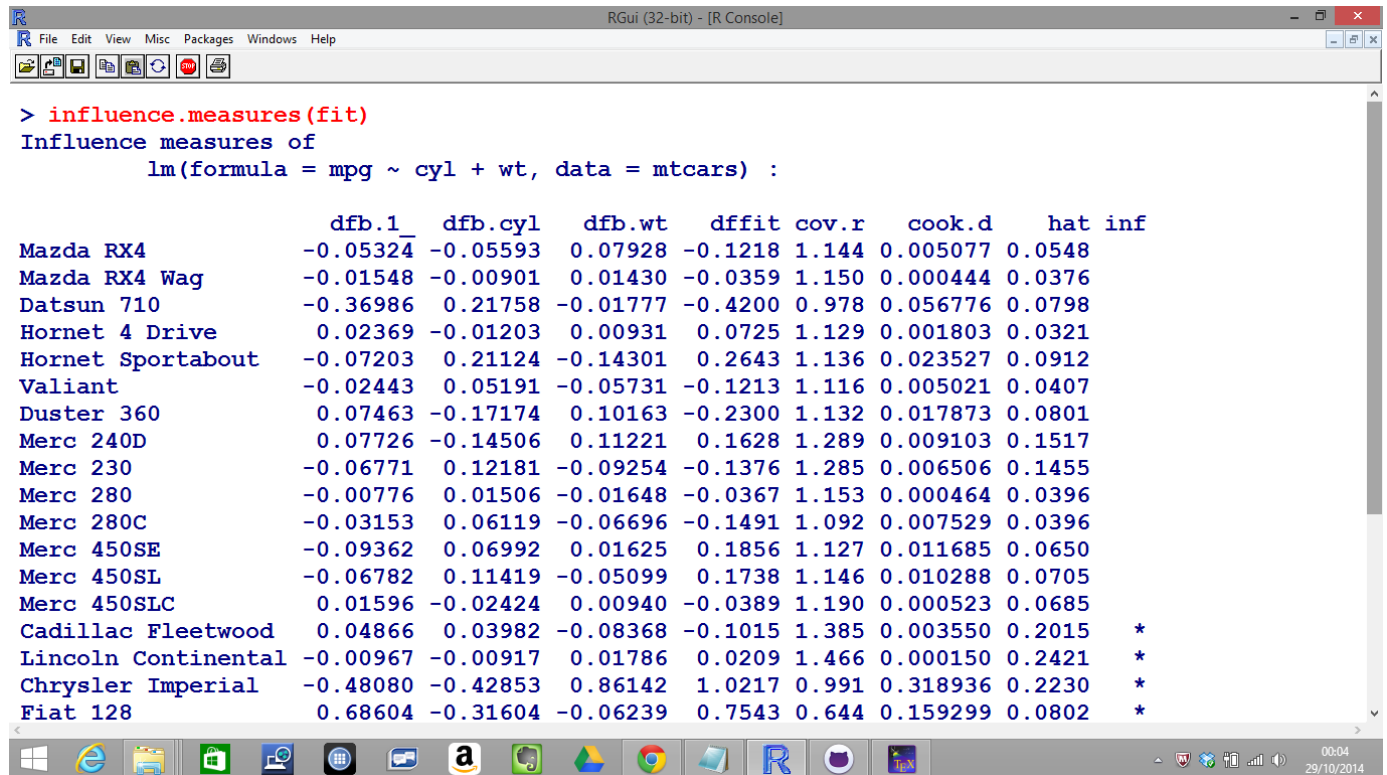


Figure 11:

12 Influential Points in Regression

```

inflm.fit <- influence.measures(fit)
which(apply(inflm.fit$is.inf, 1, any))

```

```
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

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.