AIMS Cameroon - Statistical Modelling
Session 5: Checking model assumptions (and hypothesis testing)

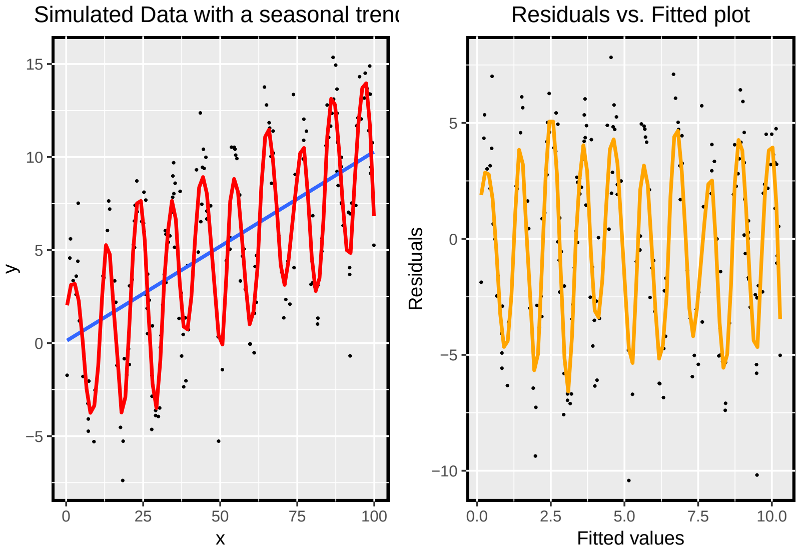
Contents

1 Applications of model checking

Before we check whether assumptions are met for specific linear models fit to a data set we will now look at illustrative examples where clearly at least one of assumptions of the linear model that we studied last week has not been met.

1.1 Examples of departures from assumptions

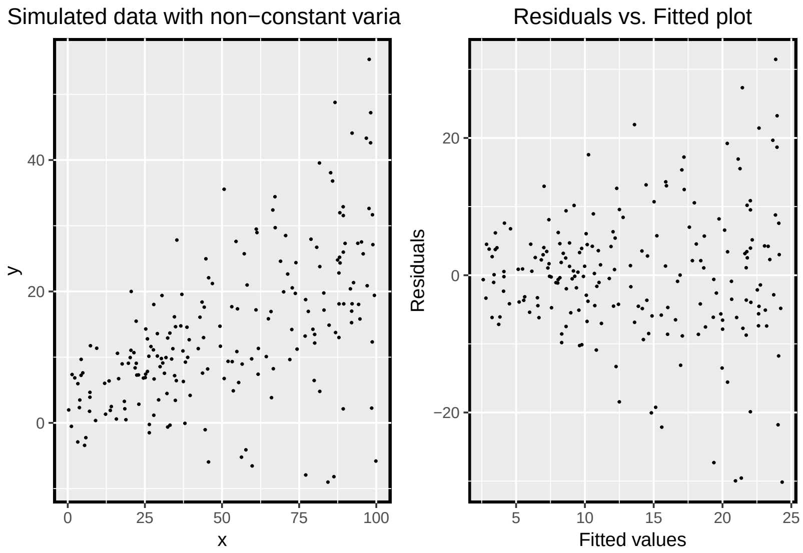
Here we will display some residual plots from simulated data to highlight departures from the assumptions of (a) a linear regression model being appropriate and (b) constant variance. The residual scale focuses attention on possible departures from the model.



The left hand plot shows the raw data with the typical seasonal "up and down" relationship, coupled with a linear trend. The blue line gives the linear regression fit to the data, which clearly is not adequate. In comparison if we used a non-parametric fit, we will get the red line as the fitted relationship. Now if we look at the residual plot (after fitting the linear regression) on the right we can clearly see that the residual plot retains some pattern (given by the orange line), which is a clear indication that the model linear model was not appropriate for this data set.

1.1.1 Non-constant variance

Now let us look at an example where we might experience non-constant variance.

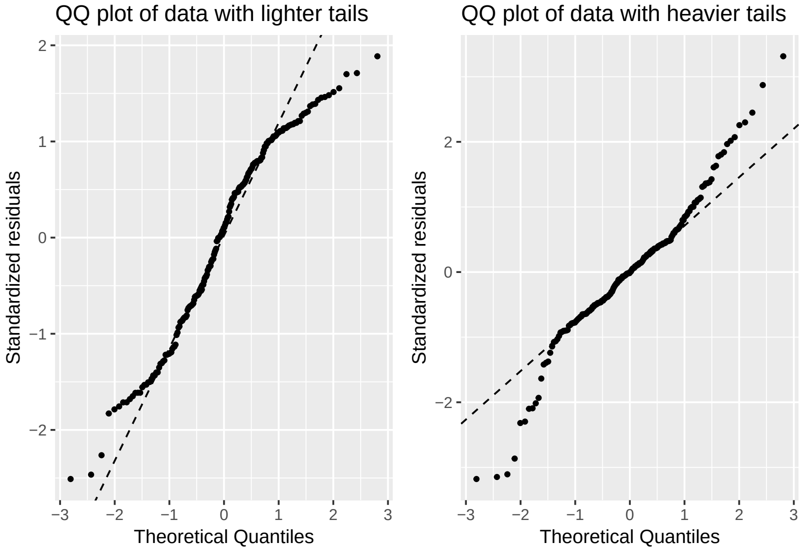


1.1.1.1 Lighter tails

In the next example, we see a QQ plot where the residuals deviate from the diagonal line in both the upper and lower tail. This plot indicated that the tails are 'lighter' (have smaller values) than what we would expect under the standard modelling assumptions. This is indicated by the points forming a "flatter" line than than the diagonal.

1.1.1.2 Heavier tails

In this final example, we see a QQ plot where the residuals deviate from the diagonal line in both the upper and lower tail. Unlike the previous plot, in this case we see that the tails are observed to be 'heavier' (have larger values) than what we would expect under the standard modeling assumptions. This is indicated by the points forming a "steeper" line than the diagonal.



Now we will apply the model checking plots to analyze model fit to a real data.

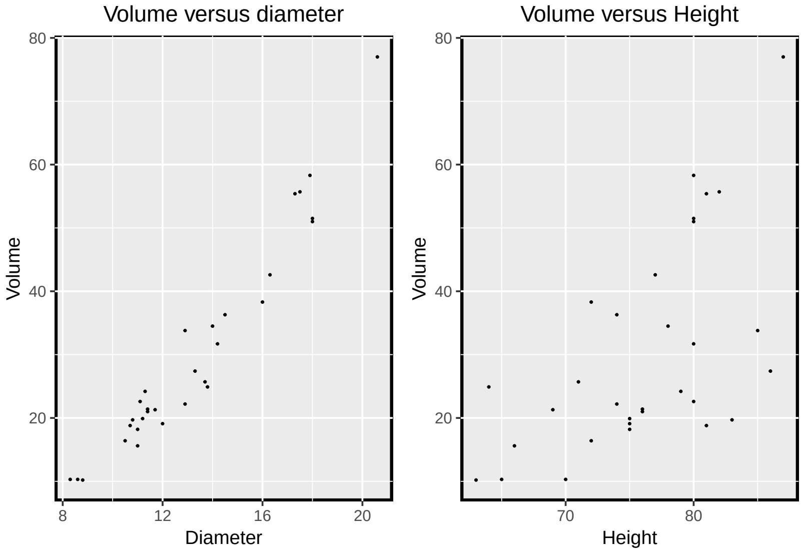
1.1.2 Example: Tree volume

The data set refers to the volume (cubic feet), diameter (inches) (at 54 inches above the ground) and height (feet) for a sample of 31 black cherry trees in the Allegheny National Forest Pennsylvania. The data were collected in order to find an estimate for the volume of a tree (and therefore for the timber yield), given its height and diameter. A starting point for estimating volume using these data is the geometric formula for a cylinder:

$$\text{volume} = \pi * \left(\frac{\text{diameter}}{2}\right)^2 * \text{height}$$

1.1.3 Exploratory Plots

We can start by exploring the relationship between the two predictors and the response in two separate plots.



1.1.4 Suggested Model

Apart from the suggestion of a slight curvature in the plot of volume versus diameter, the scatterplots indicate that a multiple linear regression model with volume as a response and diameter and height as explanatory variables may be appropriate. **Linear model:**

$$\text{volume}_i = \beta_0 + \beta * \text{diameter}_i + \gamma * \text{height}_i + \epsilon_i$$

This model is shown below (along with the residual plots produced after fitting the model).

```
trees.lm=lm(Volume~Girth+Height,data=trees)
par(mfrow=c(1,2))
plot(trees.lm,2,,pch=16,cex=0.3)
plot(fitted(trees.lm),rstandard(trees.lm),
     xlab="Fitted Values",ylab="Standardized residuals",pch=16,cex=0.3)
```

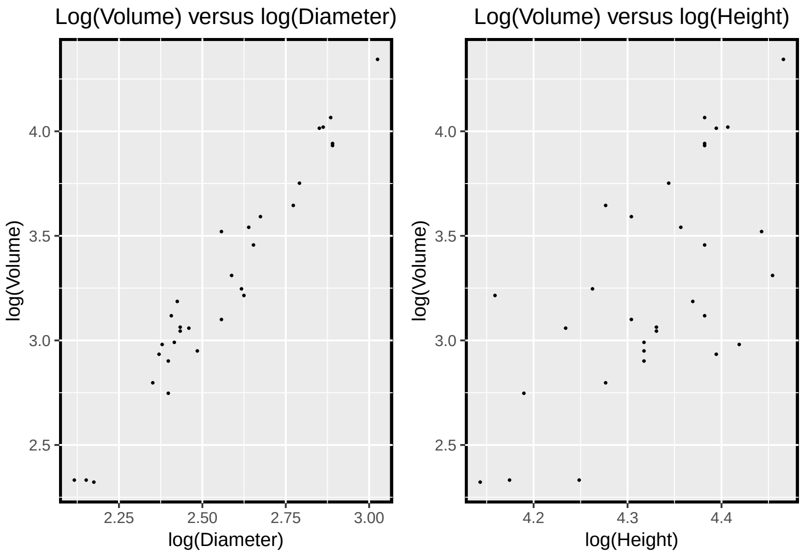
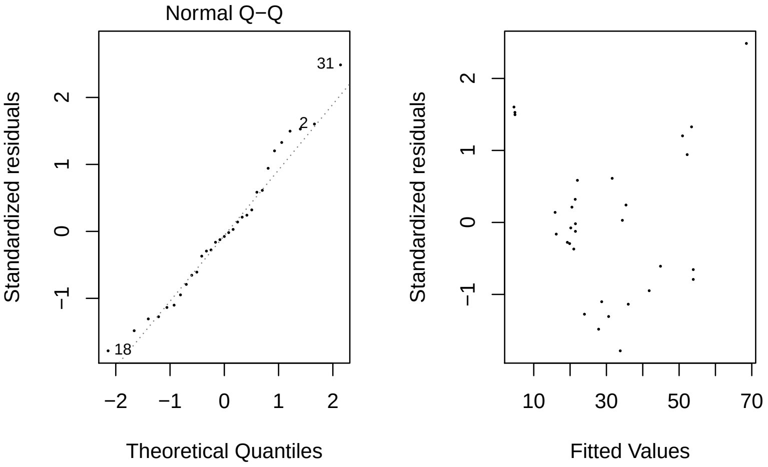


Figure 1: Log transformation of volume and diameter (left) and volume and height (right) of trees



While the initial scatterplots looked reasonable, the residuals versus fitted values plot highlights some evidence of curvature. This effect is not very marked, but there is a suggestion that the residuals tend to be positive, negative and then positive again, as we move from left to right in this plot.

This curvature (and the underlying geometric model) suggest that using a log transformation is appropriate for these data. (The log transform will produce, an additive, linear model from a multiplicative one.)

Linear model with a natural log transformation:

$$\log(\text{volume}_i) = \beta_0 + \beta_1 \log(\text{diameter}_i) + \gamma \log(\text{height}_i) + \epsilon.$$

Exploratory plots: We now take a log transformation of all the variables and again plot the response against the two predictors.

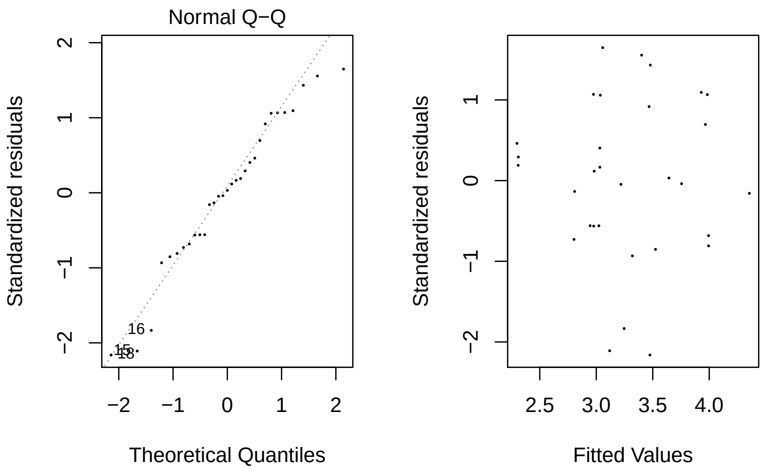


Figure 2: Normal Q-Q plot (left) and standardised residuals versus fitted values plot(right) of log transform model of trees data

1.2 Residual plots from linear model with log transformation:

When we move to the log scale, evidence of curvature in the residual plot disappears. However, the strongest argument for the use of the log transformation in this example is the underlying geometric model outlined earlier. The principal issue with these data is how volume should be predicted from diameter and height.

R output for Trees Data

```
tree.lm<- lm(formula = log(Volume) ~ log(Height) + log(Girth),data=trees)
summary(tree.lm)
```

```

Call:
lm(formula = log(Volume) ~ log(Height) + log(Girth), data = trees)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.168561 -0.048488  0.002431  0.063637  0.129223 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,	Adjusted R-squared:  0.9761 
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
anova(tree.lm)
```

```
Analysis of Variance Table
```

```
Response: log(Volume)
             Df Sum Sq Mean Sq F value    Pr(>F)    
log(Height)   1 3.4957  3.4957  527.76 < 2.2e-16 ***
log(Girth)    1 4.6275  4.6275  698.63 < 2.2e-16 ***
Residuals    28 0.1855  0.0066                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A useful summary is provided by R^2 and R^2(adj). For the model which incorporates both explanatory variables logged, R^2(adj) = 97.6%. Therefore, 97.6% of the variability in log volume can be explained by its dependence on log diameter and log height. For every one unit increase in log diameter, log volume increases by 1.98 on average, assuming that height remains the same. Similarly for every one unit increase in log height, log volume increases by 1.12 on average, assuming that the diameter remains the same.

1.3 Transformations and influential observations

We have already seen with the Trees data how a well chosen transformation can be very effective in harmonising the assumptions of a linear model. The most effective way of doing this is to consider the science of the process which has generated the data, to see if a natural transformation emerges.

When scientific guidance is not available, we can simply seek a transformation which makes the assumptions of the model more appropriate. The choice of transformation is one which involves experience and judgement.

1.3.1 Example: Mass and speed of quadrupedal rodents

Consider the Rodent data, where a plot on the original scale shows a rather odd pattern, with most of the data bunched up at the left hand side. Mass is presumably approximately proportional to volume. Speed is a linear measurement (roughly related to the length of the animal's stride) and so we might expect a geometrical transformation of some kind to apply again. The log transformation will produce an additive, linear model from a multiplicative one. The plots show that transforming both mass and speed in this way is very effective.

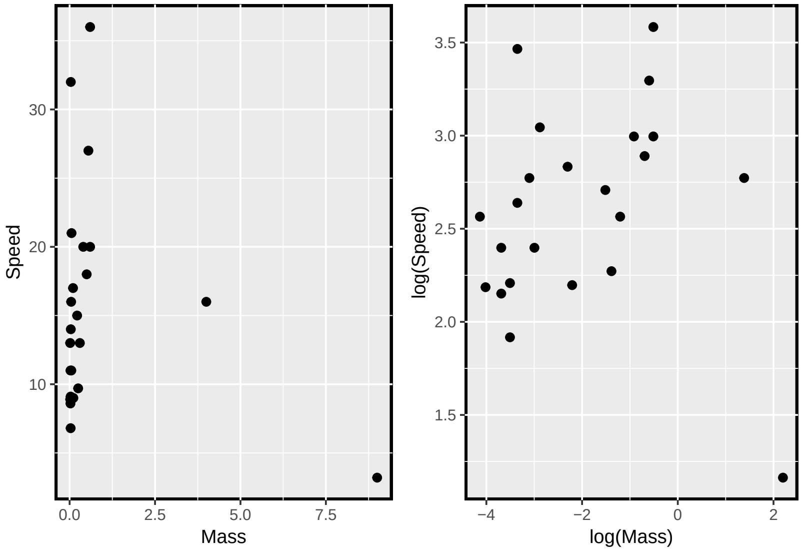
In an investigation of the relationship between mass and speed in animals, Garland (1983) collected information from published articles on these two variables for a large number of different species. These measurements are given below for a variety of four-footed rodents. (The common names of the species are taken from Corbet & Hill (1986)) Notice that the measurements are not all recorded to the same level of accuracy since the results have been collated from the work of a number of different scientists.

Table 1: North American rodent data

	Mass (kg)	Speed (ms^{-1})
North American Porcupine	9	3.2
Woodchuck	4	16
Long-clawed ground squirrel	0.6	36
Long-tailed souslik	0.6	20
Eastern grey squirrel	0.55	27
European souslik	0.5	18
European red squirrel and Persian squirrel	0.4	20
Belding's ground squirrel	0.3	13
Rat	0.25	9.7
American red squirrel	0.22	15
Golden Hamster	0.11	9
Eastern American chipmunk	0.1	17
Chisel-toothed kangaroo rat	0.05600	21
Meadow vole	0.05000	11

	Mass (kg)	Speed (ms^{-1})
Least chipmunk	0.04500	16
Merriman's kangaroo rat	0.03500	32
Fawn hopping mouse	0.03500	14
Pine mouse	0.030000	6.8
Deer mouse	0.030000	9.1
White footed mouse	0.02500	11
Woodland jumping mouse	0.02500	8.6
North American meadow jumping mouse	0.01800	8.9
House mouse	0.01600	13

```r
library(rpanel)
data(rodent)
a<-qplot(Mass,Speed, data=rodent)
b<-qplot(log(Mass),log(Speed),data=rodent)
grid.arrange(a,b,ncol=2)
```



```r
coef.lm1<-as.numeric(coef(lm(log(Speed)~log(Mass),data=rodent)))
coef.lm2<-coef(lm(log(Speed)~log(Mass),data=rodent[-1,]))
a<-qplot(log(rodent$Mass),log(rodent$Speed)) +
      geom_text( aes(x=-0.3, y=1.1, label="North American Porcupine",
                     color="red"),
                 show.legend = FALSE) +
     geom_point( aes(x=log(rodent[1,1]), y=log(rodent[1,2]),color="red"),
                 shape=21,size=5,show.legend = FALSE,alpha=1)+
     scale_shape(solid = FALSE) +
    stat_smooth(method = "lm", se = FALSE) +
   geom_abline(intercept=coef.lm2[1], slope=coef.lm2[2], color = "red")

### Need to replace this by lines excluding each point
b<-qplot(log(Mass),log(Speed),data=rodent)
grid.arrange(a,b,ncol=2)
```

Sometimes individual observations can exert a great deal of influence on our fitted model. One routine way of checking for this is to fit the model n times, missing out one of the n observation each time (i.e. 1st model

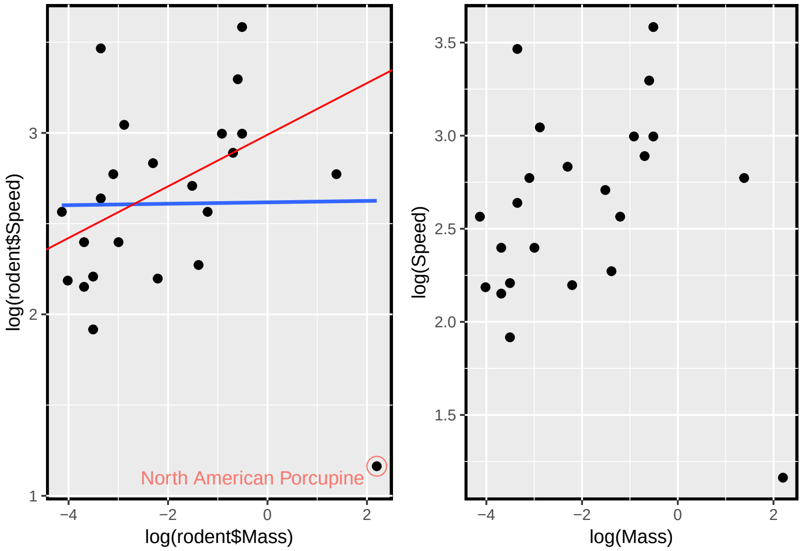


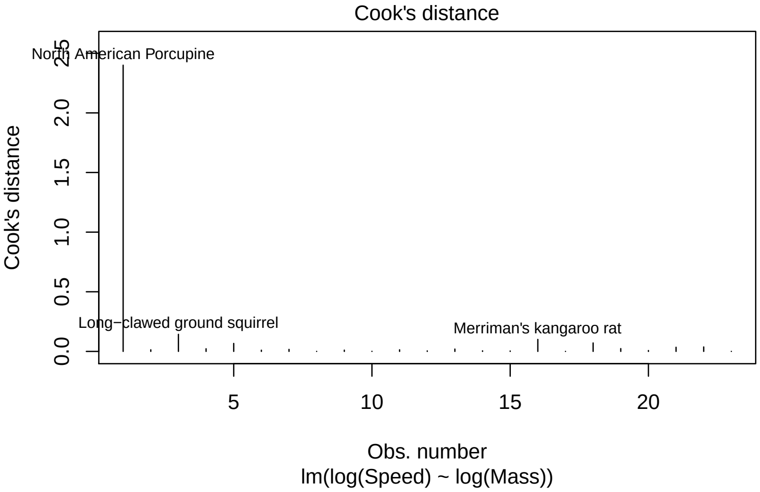
Figure 3: Linear model fit excluding individual points

includes all observations but without the 1st one, etc). A plot of all the lines for the Rodent data shows that the porcupine has a great deal of influence. When this point is omitted, the fitted line changes from horizontal (blue line) to one with a strong positive slope (red line). We can see that the porcupine exerts a great deal of influence on the model.

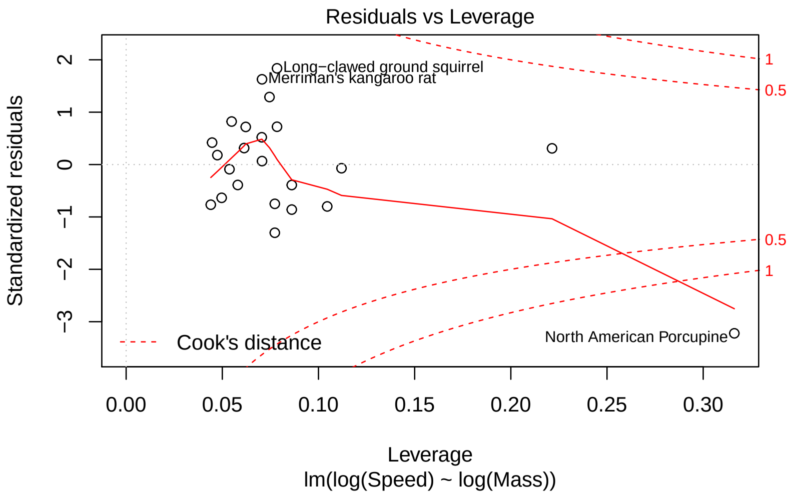
There is a good reason why porcupines do not follow relationship between mass and speed that we see in the other rodents. For most rodents, their speed is their defense against predators; whereas the porcupine is protected by its sharp spikes, so it doesn't need to move as quickly as other rodents.

Ideally we should fit multiple regression lines excluding each data point in turn. However, this isn't particularly feasible for very large numbers of data points. In R one can use Cook's distance which is available using the diagnostics of a linear model fit

```r
plot(lm(log(Speed)~log(Mass),data=rodent),4)
```



```r
plot(lm(log(Speed)~log(Mass),data=rodent),5)
```



The plot of Speed vs Mass shows an odd pattern. There is no obvious linear relationship between the two variables. Once we take the log transformation, the linear relationship is much more clear. However, there is still one point, located at the bottom right that looks odd.

Again we can see that the North American Porcupine stands out. The R library car also has a function called outlierTest which performs a formal test for detecting an outlier.

```r
library(car)
outlierTest(lm(log(Mass)~log(Speed),data=rodent))
```

```
                         rstudent unadjusted p-value Bonferonni p
North American Porcupine 4.026755         0.00066086       0.0152
```

Time to work on Task 1

2 Interval Estimation and Hypothesis Testing

Now we will focus on hypothesis testing of the regression parameters. In the previous session we considered diagnostics for and assumptions about linear regression models. Now we will consider inference for model parameters, model comparison and selection. We will construct interval estimates and hypothesis tests for various parameters of our models and use these along with R^2 for model selection. We will also consider basic approaches for model selection in the case of many explanatory variables.

Instead of solving several different types of inferential problems, e.g. involving a single parameter, involving two parameters or in some cases involving a linear combination of parameters we will develop a general theory for doing inference on linear combinations of parameters. Each of the cases described in the previous sentence can then be derived as a special case of the general theory. For example if we want to predict a future value, at $x = 5$ based on a simple linear model

$$y_i = \beta_0 + \beta_1 x + \epsilon_i \implies \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$

we are interested in the linear combination:

$$\beta_0 + 5\beta_1,$$

which can be written as:

$$(1 \quad 5)\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

We will now start by considering how to form linear functions of the parameters in a linear model.

2.1 The least-squares estimate of (linear) functions of the parameters in a (linear) model

Data: $(y_i, x_{1i}, x_{2i}, \ldots, x_{pi}); \quad i = 1, \ldots, n$

Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Suppose we want the least-squares estimate for a linear function of the parameters,

- Say $\mathbf{b}_1^T\boldsymbol{\beta}$ for some given vector $\mathbf{b}_1$, or
- possibly for a set of s linearly independent linear combinations $\mathbf{b}_1^T\boldsymbol{\beta}, \ldots, \mathbf{b}_s^T\boldsymbol{\beta}$, $s \leq p$ where the $\mathbf{b}_i$'s are given vectors.

It is always possible to create a non-singular transformation from $\boldsymbol{\beta} \leftrightarrow \boldsymbol{\phi}$ where

$$\boldsymbol{\phi} = \begin{pmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{b}_s^T \end{pmatrix} \boldsymbol{\beta} = \mathbf{B}\boldsymbol{\beta},$$

where $\mathbf{B}$ is a nonsingular matrix. So

$$\boldsymbol{\beta} = \mathbf{B}^{-1}\boldsymbol{\phi}$$

It is now possible to rewrite our model in terms of $\boldsymbol{\phi}$, where ϕ_1 or $\phi_1, \ldots, \phi_s$ are the parameters of interest.

Data: $(y_i, x_{1i}, x_{2i}, \ldots, x_{pi}); \quad i = 1, \ldots, n$

Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}\mathbf{B}^{-1})\boldsymbol{\phi} + \boldsymbol{\epsilon}$ where

$(\mathbf{X}\mathbf{B}^{-1})$ is an $n \times p$ matrix which is known and $\boldsymbol{\phi}$ is a p vector of unknown parameters.

The form of the model is mathematically equivalent to our original form, substituting $(\mathbf{X}\mathbf{B}^{-1})$ for the design matrix and $\boldsymbol{\phi}$ for the parameter vector.

Hence, we can write down the solution for the parameter estimates, based on least-squares, from our earlier results.

$$\begin{aligned}
\hat{\boldsymbol{\phi}} &= \{(\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})\}^{-1}(\mathbf{X}\mathbf{B}^{-1})^T\mathbf{Y} \\
&= \{(\mathbf{B}^{-1})^T(\mathbf{X}^T\mathbf{X})\mathbf{B}^{-1}\}^{-1}(\mathbf{B}^{-1})^T\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{B}^T(\mathbf{B}^T)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{B}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{B}\hat{\boldsymbol{\beta}}.
\end{aligned}$$

Hence the least-squares estimates of a set of linear functions of parameters is just the set of linear functions of the least-squares estimates.

A useful application of this result may sometimes simplify the calculation of least-squares estimates. The basic idea is that it may be possible to rewrite a model in terms of parameters whose estimates are "easier" to calculate and then we can transform back to the original parameters. This approach is often referred to as 'centering'. Centering in often useful to produce orthogonal columns which in turn gives us diagonal matrices to invert.

2.1.1 Example: Application of linear transformation of parameters for two parameter case

Data: $(y_i, x_i); \quad i = 1, \ldots, n$

Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}\mathbf{B}^{-1})\boldsymbol{\phi} + \boldsymbol{\epsilon}$

Suppose we transform

$y_i = \alpha + \beta x_i + \epsilon_i$ (Model 1) to

$y_i = \alpha' + \beta(x_i - \bar{x}) + \epsilon_i$ (Model 2)

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \leftrightarrow \boldsymbol{\phi} = \begin{pmatrix} \alpha' \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha + \beta\bar{x} \\ \beta \end{pmatrix}, \quad \bar{x} = (\sum_{i=1}^{n} x_i)/n.$$

$$E(\mathbf{Y}) = \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ . & . \\ . & . \\ . & . \\ . & . \\ 1 & (x_n - \bar{x}) \end{pmatrix} \begin{pmatrix} \alpha' \\ \beta \end{pmatrix} = \mathbf{X}\mathbf{B}^{-1}\boldsymbol{\phi}$$

where $\boldsymbol{\phi} = \mathbf{B}\boldsymbol{\beta}$ and $\mathbf{B} = \begin{pmatrix} 1 & \bar{x} \\ 0 & 1 \end{pmatrix}$

$$\begin{aligned} \hat{\boldsymbol{\phi}} &= \{(\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})\}^{-1}(\mathbf{X}\mathbf{B}^{-1})^T\mathbf{Y} \\ (\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1}) &= \begin{pmatrix} n & \sum_{i=1}^{n}(x_i - \bar{x}) \\ \sum_{i=1}^{n}(x_i - \bar{x}) & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix} \\ &= \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix} \end{aligned}$$

i.e. $(\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})$ is diagonal.

$$\begin{aligned} (\mathbf{X}\mathbf{B}^{-1})^T\mathbf{Y} &= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i(x_i - \bar{x}) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \end{pmatrix} \\ \{(\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})\}^{-1} &= \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix} \end{aligned}$$

i.e.

$$\hat{\boldsymbol{\phi}} = \begin{pmatrix} \sum_{i=1}^{n} y_i/n \\ \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix} = \begin{pmatrix} \hat{\alpha}' \\ \hat{\beta} \end{pmatrix}$$

Because of our choice of $\boldsymbol{\phi}$, $(\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})$ is easier to invert and hence the calculations are simpler.

From the nature of the transformation it is clear that

$$\hat{\alpha} = \hat{\alpha}' - \hat{\beta}\bar{x}$$

The benefits of this type of transformation are more obvious when applied to a 3 parameter example.

2.1.2 Example: Application of linear transformation of parameters for three parameter case

Data: $(y_i, x_{1i}, x_{2i}), \quad i = 1, \ldots, n$

Model: $E(Y_i) = \alpha + \beta x_{1i} + \gamma x_{2i}$

Reparametrise to

$$\text{Model}: E(Y_i) = \alpha' + \beta(x_{1i} - \bar{x}_{1.}) + \gamma(x_{2i} - \bar{x}_{2.})$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \leftrightarrow \boldsymbol{\phi} = \begin{pmatrix} \alpha' \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \alpha + \beta\bar{x}_{1.} + \gamma\bar{x}_{2.} \\ \beta \\ \gamma \end{pmatrix},$$

where $\bar{x}_{1.} = \sum_{i=1}^{n} x_{1i}/n, \quad \bar{x}_{2.} = \sum_{i=1}^{n} x_{2i}/n.$ i.e.

$$E(\mathbf{Y}) = \begin{pmatrix} 1 & (x_{11} - \bar{x}_{1.}) & (x_{21} - \bar{x}_{2.}) \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & (x_{1n} - \bar{x}_{1.}) & (x_{2n} - \bar{x}_{2.}) \end{pmatrix} \begin{pmatrix} \alpha' \\ \beta \\ \gamma \end{pmatrix} = \mathbf{X}\mathbf{B}^{-1}\boldsymbol{\phi}$$

$$\hat{\boldsymbol{\phi}} = ((\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1}))^{-1}(\mathbf{X}\mathbf{B}^{-1})^T\mathbf{Y}$$

$$((\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1})) = \begin{pmatrix} n & 0 & 0 \\ 0 & \sum_{i=1}^{n}(x_{1i} - \bar{x}_{1.})^2 & \sum_{i=1}^{n}(x_{1i} - \bar{x}_{1.})(x_{2i} - \bar{x}_{2.}) \\ 0 & \sum_{i=1}^{n}(x_{1i} - \bar{x}_{1.})(x_{2i} - \bar{x}_{2.}) & \sum_{i=1}^{n}(x_{2i} - \bar{x}_{2.})^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & \mathbf{0} \\ \mathbf{0}^T & \boldsymbol{\Psi} \end{pmatrix}$$

$$((\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1}))^{-1} = \begin{pmatrix} \frac{1}{n} & \mathbf{0} \\ \mathbf{0}^T & \boldsymbol{\Psi}^{-1} \end{pmatrix}$$

Hence inversion of $((\mathbf{X}\mathbf{B}^{-1})^T(\mathbf{X}\mathbf{B}^{-1}))$ is reduced to inversion of a (2×2) matrix, a great saving in calculation. In general a similar transformation will reduce the inversion of a $(p \times p)$ matrix to the inversion of a $(p-1) \times (p-1)$ matrix.

After calculation of $\hat{\boldsymbol{\phi}}, \hat{\alpha}$ can be obtained from

$$\hat{\alpha} = \hat{\alpha'} - \hat{\beta}\bar{x}_{1.} - \hat{\gamma}\bar{x}_{2.}$$

2.2 Inferences from regression equations

If we are interested in $\mathbf{b}^T\boldsymbol{\beta}$ (a linear function of the parameters), where $\mathbf{b}$ is a given vector of constants, we will use the concept of pivotal functions that you have learnt in the course Learning from Data.

2.2.1 Theorem: Pivotal function for a linear function of the parameters

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}}$$

is a pivotal function since

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}} \sim t(n-p),$$

where p is the number of parameters, n is the sample size and RSS is the residual sum-of-squares in a linear model.

This result is stated without proof. It is helpful to use the notation ***estimated standard error*** for the quantity

$$\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

(The word “estimated” is often omitted from the name).

The above result can be used to construct hypothesis tests and interval estimates for model parameters.

2.3 Hypothesis Testing

For example, if we were interested in making inferences about β in a simple linear regression model i.e. $y_i = \alpha + \beta x_i + \epsilon_i$, $\mathbf{b}^T\boldsymbol{\beta} = \beta$ i.e. $\mathbf{b}^T = (0 \quad 1)$ and this gives us:

$$\frac{\hat{\beta} - \beta}{\text{e.s.e}(\hat{\beta})} \sim t(n-p)$$

Under the null hypothesis:

H_0: $\beta = 0$ (where $\text{H}_1 : \beta \neq 0$)

$$\frac{\hat{\beta}}{\text{e.s.e}(\hat{\beta})} \sim t(n-p)$$

and $\frac{\hat{\beta}}{\text{e.s.e}(\hat{\beta})}$ is typically called the t-statistic. Therefore, the null hypothesis is rejected for large absolute values of the t-statistic, usually values > 2 i.e. for small p-values in R (where a p-value is the probability that we obtain a t-statistic value as extreme or more extreme if the null hypothesis is true). In general, we reject H_0 for p-values < 0.05 and this would indicate a significant relationship between a response and an explanatory variable in the model.

2.4 Interval estimate for $\mathbf{b}^T\boldsymbol{\beta}$

It is easy to show that an interval estimate for $\mathbf{b}^T\boldsymbol{\beta}$ with confidence c is

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}(\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}.$$

2.5 Prediction interval (PI) for Y given x

The quantity of interest here is a future observation of Y, Y_f say, when x takes the value x_f, which denotes the value of the explanatory variable at the position where a prediction of Y is required. The expected value $E(Y|x_f)$ can be written in the form $\mathbf{b}_f^T\boldsymbol{\beta}$.

For example, with a simple linear regression, $E(Y) = \alpha + \beta x$, we can write $E(Y|x_f) = \alpha + \beta x_f = \mathbf{b}^T\boldsymbol{\beta}$, where $\mathbf{b}^T = (1, x_f)$ and $\boldsymbol{\beta}^T = (\alpha, \beta)$.

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}(1 + \mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}}$$

is a prediction interval for y_f since

$$\frac{(\mathbf{b}^T\hat{\boldsymbol{\beta}} - \mathbf{b}^T\boldsymbol{\beta})}{\sqrt{\frac{RSS}{n-p}(1 + \mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}} \sim t(n-p)$$

Applying the results again, it is easy to show that a prediction interval for y_f with confidence c is

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}(1 + \mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}.$$

2.5.1 Examples

Now we will use the expressions for Hypothesis and confidence/prediction intervals to answer inferential questions for specific parameters and apply them to real data examples.

We might be interested in a 95% C.I. (confidence interval) for β for the model $y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \epsilon_i$. We can write this C.I. as

$$\hat{\beta} \pm t(n-p; 0.975)\text{s.e.}(\hat{\beta})$$

where

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and

$$\mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

`R` automatically prints the standard error (s.e.) of each individual parameter when it fits a regression model. This makes the construction of C.I.s for each parameter very easy. A parameter estimate is a random variable and in probability terms the estimated standard error of a parameter estimate is an estimate of its standard deviation.

2.5.2 Example: Hypothesis testing for Pregnancy Data

We have seen this example for parameter estimation.

Data: $(y_i, x_i) \quad i = 1, \ldots, 19$

Model: $E(Y_i) = \alpha + \beta x_i$

We will now test the slope parameter in the model β, build a confidence interval for β and finally predict a future observation y at $x = 27$.

To answer the questions we will need the following:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$$

$$\begin{aligned} (\mathbf{X}^T\mathbf{X}) &= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} 19 & 456 \\ 456 & 12164 \end{pmatrix} \\ (\mathbf{X}^T\mathbf{X})^{-1} &= \begin{pmatrix} 0.524763 & -0.019672 \\ -0.019672 & 0.000820 \end{pmatrix} \end{aligned} \tag{2}$$

and the R output

```
pregnancy<-read.csv("../Data/PROTEIN.CSV",header=T)
fit1<-lm(formula = Protein ~ Gestation,data=pregnancy)
summary(fit1)
```

```

Call:
lm(formula = Protein ~ Gestation, data = pregnancy)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16853 -0.08720 -0.01009  0.08578  0.20422 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.201738   0.083363   2.420    0.027 *  
Gestation   0.022844   0.003295   6.934 2.42e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1151 on 17 degrees of freedom
Multiple R-squared:  0.7388,	Adjusted R-squared:  0.7234 
F-statistic: 48.08 on 1 and 17 DF,  p-value: 2.416e-06
```

```
anova(fit1)
```

```
Analysis of Variance Table

Response: Protein
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Gestation  1 0.63667 0.63667  48.076 2.416e-06 ***
```

```
Residuals 17 0.22513 0.01324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing (1)

The hypotheses being tested for the coefficient of β are:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$

Since the p-value for gestation is < 0.001 (and hence < 0.05) the null hypothesis is rejected and we conclude that there is a statistically significant relationship between protein and gestation. The gestational age is a useful predictor of the protein level.

A confidence interval can be produced to provide a range of likely values for the coefficient of gestation.

Confidence Intervals (2)

So, a 95% C.I. for β is

$$0.0228 \pm t(17; 0.975)\sqrt{\frac{0.2251}{17}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and since $\mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ ie

$$\begin{aligned} 0.0228 \pm & \quad 2.11\sqrt{\frac{0.2251}{17}0.000820} \\ 0.0228 \pm & \quad 2.11(0.003295) \end{aligned}$$

i.e. 0.0228 ± 0.0070,

i.e. (0.016, 0.030)

The fact that this interval contains only positive values tells us that there is clear evidence that the average level of protein increases with gestation. The coefficient for β is highly likely to lie somewhere between 0.02 and 0.03.

Note: A confidence interval that includes zero indicates that there is insufficient evidence of a relationship between the response and the explanatory variable. In this situation we would expect the p-value for testing $\mathrm{H}_0 : \beta = 0$ to be > 0.05.

Prediction Intervals

In order to use this model in a clinical setting we need a means of telling what values of protein level are expected for a future healthy mother who attends this clinic. For example, if a woman who is 27 weeks pregnant has a protein level of 1.06, should this be regarded as unusual? A prediction interval helps us to answer this question.

A 95% P.I. for a future observation y at $x = 27$ is done in the following way:

Here $\mathbf{b} = \begin{pmatrix} 1 \\ 27 \end{pmatrix}$.

$$\begin{aligned}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} &= \begin{pmatrix}1 & 27\end{pmatrix}\begin{pmatrix}0.524763 & -0.019672\\ -0.019672 & 0.000820\end{pmatrix}\begin{pmatrix}1\\27\end{pmatrix}\\ &= \begin{pmatrix}0.524763-27\times 0.019672 & -0.019672+27\times 0.000820\end{pmatrix}\begin{pmatrix}1\\27\end{pmatrix}\\ &= \begin{pmatrix}-0.006381 & 0.002468\end{pmatrix}\begin{pmatrix}1\\27\end{pmatrix}\\ &= -0.006381+27\times 0.002468\\ &= 0.060255\end{aligned}$$

Thus the

Thus the prediction interval is

$$(0.02017 + 27 \times 0.0228) \pm 2.11\sqrt{\frac{0.2251}{17}(1 + 0.060255)}$$
$$(.57,1.07)$$

Since it lies within this interval (just) we have strong grounds for regarding a protein level of 1.06 as unusual. (Even if it did lie outside the interval, this only says that the result is unusual).

Note that Prediction intervals will always be wider than confidence intervals.

Analyzing the ANOVA table

The F statistic value: $\text{MS}_{\text{model}}/\text{MS}_{\text{residuals}}$ provides a test statistic that allows us to test whether there is any evidence that at least one of the model parameters is not zero.

The null hypothesis is H_0: all p parameters = 0, which will be tested against the alternative that at least one of the parameters is not zero.

If the null hypothesis is true, the statistic has an $\text{F}(\text{Df}_{\text{model}}, \text{Df}_{\text{residuals}})$ distribution. This implies that

$$F = \frac{MS_{\text{model}}}{MS_{\text{residuals}}} \sim F(Df_{\text{model}}, Df_{\text{residuals}}).$$

If H_0 is false, we would expect $\text{MS}_{\text{residuals}}$ to be smaller than MS_{model} and so large values of F should lead us to reject H_0. (i.e. for large values of F the p-value will be small)

In this example, the p-value is < 0.001 and hence the null hypothesis is rejected and we conclude that at least one of the parameters is not zero.

2.5.3 Example: Trees Data

Our full model for the trees data was

$$E(Y) = \alpha + \beta x_1 + \gamma x_2$$

where Y denotes log (volume), x_1 denotes log (diameter) and x_2 denotes log (height) of 31 trees.

The fitted model produces:

$$\hat{\beta} = \begin{pmatrix}-6.632\\1.983\\1.117\end{pmatrix}$$

$$RSS = 0.1855$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 96.5721 & 3.1393 & -24.1651 \\ 3.1393 & 0.8495 & -1.2275 \\ -24.1651 & -1.2275 & 6.3099 \end{pmatrix}$$

The matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ was obtained from `R`.

In order to check whether or not there is clear evidence of a relationship between each of the explanatory variables and the response, we can construct interval estimates for β and γ.

A 95% C.I. for β is given by

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-3;0.975)\sqrt{\frac{RSS}{n-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $\mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$$\hat{\beta} \pm t(28;0.975)\sqrt{\frac{0.1855}{28}0.8495}$$

$$1.983 \pm 0.15$$

$$(1.83, 2.13)$$

This interval does not contain zero. There is therefore clear evidence of a relationship between log(diameter) and log(volume), (i.e. log diameter is a significant predictor in addition to log height, and it is highly likely that the coefficient for log diameter lies between 1.83 and 2.13).

Note also that 2 is a plausible value for the coefficient of log(diameter). This is therefore consistent with the cylindrical model discussed in chapter 2, where $(\mathrm{V} = \pi(\frac{\mathrm{d}}{2})^2\mathrm{h};\ \log(\mathrm{V}) = (\pi/4){+}2\mathrm{logd}{+}\mathrm{logh})$.

A 95% C.I. for γ is given by

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t(n-3;0.975)\sqrt{\frac{RSS}{n-3}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $\mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$$\hat{\gamma} \pm t(28;0.975)\sqrt{\frac{0.1855}{28}6.3099}$$

$$1.12 \pm 0.42$$

$$(0.70, 1.54)$$

Again there is clear evidence of a relationship between log(height) and log(volume), since 0 does not lie in the interval estimate. Log height is a significant predictor in addition to log diameter. The results are again consistent with the cylindrical model since the value 1 lies in the interval. It is highly likely that the coefficient for log height lies between 0.70 and 1.54.

2.5.3.1 Task: Confidence interval for Autoanalyser data

Blood plasma concentrations are usually measured using a lengthy laboratory process. A simpler, cheaper method using an autoanalyser is often used. The autoanalyser is regularly tested to see if it is performing properly. On this occasion, 12 measurements have been made on samples of known concentration (3 replicates at each of 4 concentrations).

The following model has been fitted in `R` to estimate the autoanalyser concentration from the true concentration:

$$\text{autoanalyser}_i = \beta_0 + \beta_1 \text{true}_i + \epsilon_i, \quad i = 1, \dots, 12$$

Construct a 95% C.I. for the population mean autoanalyser concentration when the true concentration is 6 units and interpret the interval.

```
auto<-read.csv("../Data/autoanalyser.CSV")
auto.lm<-lm(autoanalyser~true,data=auto)
summary(auto.lm)
```

```

Call:
lm(formula = autoanalyser ~ true, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23333 -0.09583 -0.03333  0.10417  0.21667 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.68333    0.18993   3.598  0.00487 ** 
true         0.85000    0.04096  20.752  1.5e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1586 on 10 degrees of freedom
Multiple R-squared:  0.9773,    Adjusted R-squared:  0.975 
F-statistic: 430.6 on 1 and 10 DF,  p-value: 1.496e-09
```

```
anova(auto.lm)
```

```
Analysis of Variance Table

Response: autoanalyser
          Df  Sum Sq Mean Sq F value    Pr(>F)    
true       1 10.8375 10.8375  430.63 1.496e-09 ***
Residuals 10  0.2517  0.0252                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also have

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{180}\begin{pmatrix} 258 & -54 \\ -54 & 12 \end{pmatrix}$$

We require an interval estimate for

$$\alpha + 6\beta = \mathbf{b}^T\boldsymbol{\beta}$$

$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 6 \end{pmatrix}$

$$\begin{aligned} \mathbf{b}^T\hat{\boldsymbol{\beta}} &= \hat{\alpha} + 6 \times \hat{\beta} \\ &= 0.683 + 0.85 \times 6 \\ &= 5.783 \\ RSS &= 0.252 \\ n &= 12 \end{aligned}$$

t(10;0.975) $= 2.228$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{180}\begin{pmatrix} 258 & -54 \\ -54 & 12 \end{pmatrix}$$

Interval Estimate for $(\alpha + 6\beta)$: (5.61, 5.95)

The population mean autoanalyser concentration (when the true concentration is 6 units) is very likely to lie between 5.61 and 5.95 units.

Time to work on Task 2

2.5.4 Example: Confidence interval for the difference in two population means

Data: (y_{ij}); $\quad i = 1, 2; j = 1, \ldots, n_i$

Model: $E(Y_{ij}) = \mu_i$, $Y_{ij} \sim N(\mu_i, \sigma^2)$

Construct a confidence interval for $(\mu_1 - \mu_2)$

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ y_{12} \\ . \\ y_{1,n_1} \\ y_{21} \\ y_{22} \\ . \\ y_{2,n_2} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ . & . \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ . & . \\ 0 & 1 \end{pmatrix}$$

Interest is in $(\mu_1 - \mu_2)$, i.e. $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} = \hat{\boldsymbol{\beta}}$$

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} = (\bar{y}_1 - \bar{y}_2), \quad \mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = (\frac{1}{n_1} + \frac{1}{n_2})$$

What about the RSS (residual sum-of-squares)?

$$
\begin{aligned}
RSS &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} y_{ij}^2 - (n_1\bar{y_1}^2 + n_2\bar{y_2}^2) \\
&= RSS_1 + RSS_2
\end{aligned}
$$

where $RSS_i = \sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$

Interval for $\mu_1 - \mu_2$:

$$
(\bar{y_1} - \bar{y_2}) \pm t\left(n_1 + n_2 - 2; \frac{1+c}{2}\right)\sqrt{\frac{RSS_1 + RSS_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
$$

More examples, along the `R` codes can be found in the following reading materials:

- Sections 3.2, 3.5, 4.1, 4.2 and 4.4 from Linear Models with R by Julian J. Faraway
- Sections 3.9, 3.10 and 3.11 from Regression Analysis by Example - Samprit Chatterjee, Ali S. Hadi

Learning outcomes for Session 5

After studying this session's material you should be able to:

- check whether assumptions of linear models are satisfied (by interpreting residual plots),
- understand when transformations are needed,
- identify outliers from residual plots,
- be able to quote the general formulas for interval and prediction intervals,
- calculate interval estimates from summary statistics and/or R-output,
- perform hypothesis testing from summary statistics and/or R-output.

Tasks

1. Outlier detection for Scottish Hills data

The data set gives the record times in 1984 for 35 Scottish hill races. The variables are

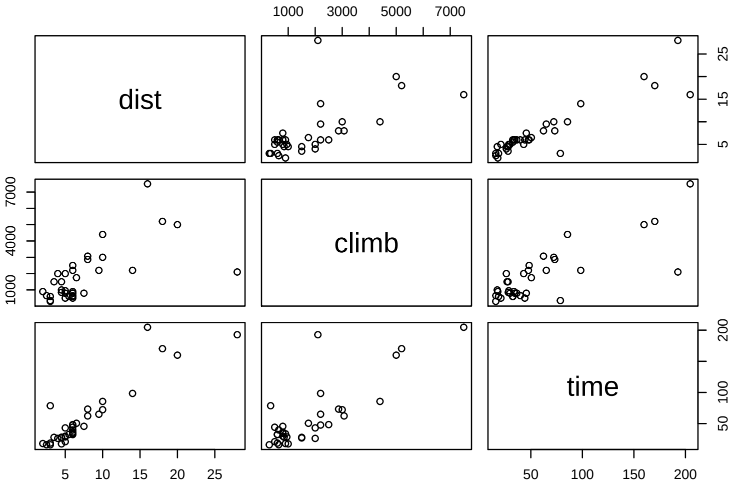
- `dist` distance in miles (on the map),
- `climb` total height gained during the route, in feet,
- `time` record time in minutes.

The goal is to predict `time` from the variables `dist` and `climb`.

```r
library(MASS)
data(hills)
head(hills)
```

```
               dist climb   time
Greenmantle     2.5   650 16.083
Carnethy        6.0  2500 48.350
Craig Dunain    6.0   900 33.650
Ben Rha         7.5   800 45.600
Ben Lomond      8.0  3070 62.267
Goatfell        8.0  2866 73.217
```

```r
pairs(hills)
```



Perform model diagnostics for the above data. In particular you can try doing the following:

- Check whether a `log` transformation provides better a prediction,
- Check whether there are any outliers,
- Check whether the assumption of the normality of errors is satisfied.

Answer to Task 1

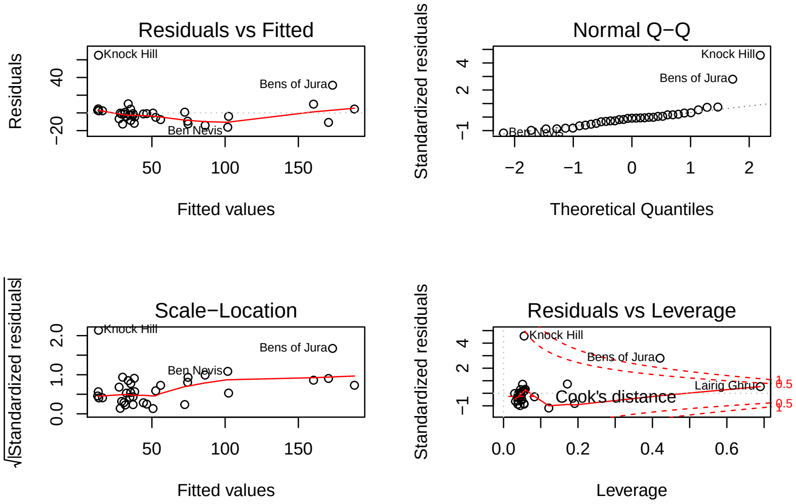
Without `log` transformation

```r
hills.lm=lm(time~.,data=hills)
summary(hills.lm)
```

```
Call:
```

```
lm(formula = time ~ ., data = hills)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.215  -7.129  -1.186   2.371  65.121 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.992039   4.302734  -2.090   0.0447 *  
dist         6.217956   0.601148  10.343 9.86e-12 ***
climb        0.011048   0.002051   5.387 6.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.68 on 32 degrees of freedom
Multiple R-squared:  0.9191,    Adjusted R-squared:  0.914 
F-statistic: 181.7 on 2 and 32 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(hills.lm)
```



```
outlierTest(hills.lm)
```

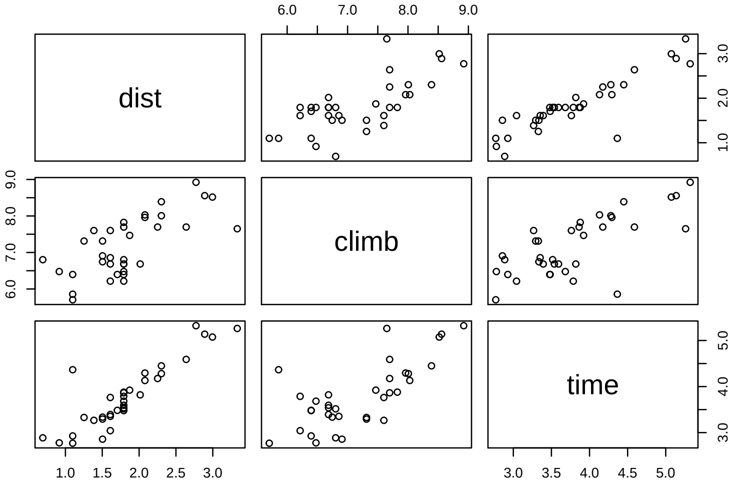
```
           rstudent unadjusted p-value Bonferonni p
Knock Hill 7.610845         1.3973e-08   4.8905e-07
```

We can see that the Knock Hill, Black Hill and Beans of Jura races stand out from the other races, the `outlierTest()` shows us that the Knock Hill race is an outlier.

Let us apply a `log` transformation to the variables and see if that provides a better fit.

Taking the `log` transformation

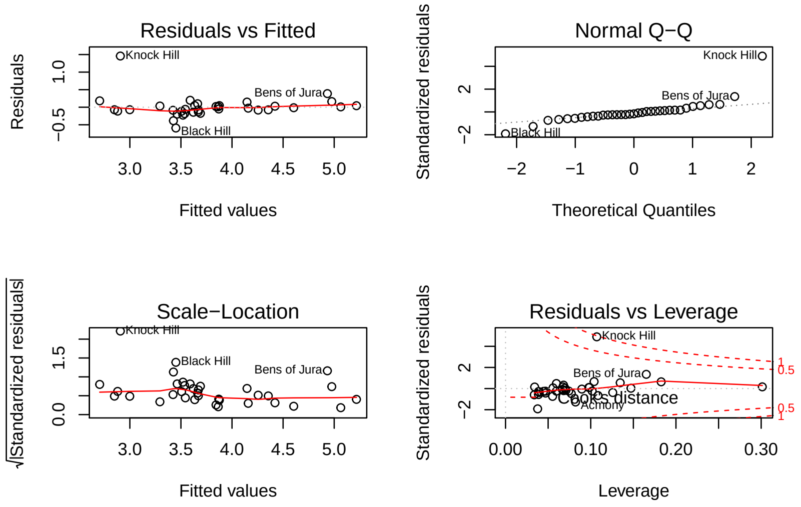
```
pairs(log(hills))
```



```
lhills=log(hills)
lhills.lm=lm(time~.,data=lhills)
summary(lhills.lm)
```

```
Call:
lm(formula = time ~ ., data = lhills)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59294 -0.11255 -0.05080  0.04439  1.45806 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.91921    0.53517   1.718   0.0955 .  
dist         0.89752    0.12803   7.011 6.04e-08 ***
climb        0.17100    0.09329   1.833   0.0761 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3149 on 32 degrees of freedom
Multiple R-squared:  0.8121,	Adjusted R-squared:  0.8003 
F-statistic: 69.15 on 2 and 32 DF,  p-value: 2.417e-12
```

```
par(mfrow=c(2,2))
plot(lhills.lm)
```



```
outlierTest(lhills.lm)
```

```
           rstudent unadjusted p-value Bonferonni p
Knock Hill 9.646392         7.5181e-11   2.6313e-09
```

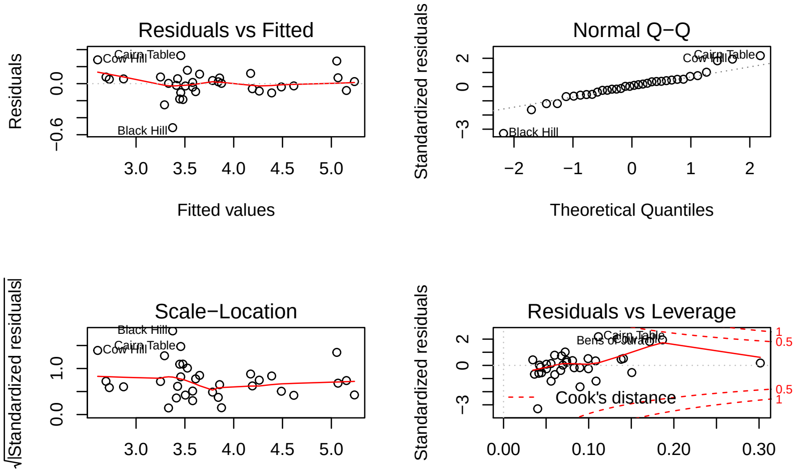
After taking a `log` transformation the plots show that the Beans of Jura and Black hill races fit in much better with the other races, however the Knock Hill race is still very influential; this is confirmed by the `outlierTest`.

We will remove the Knock Hill race and refit the data.

```
lhills <- lhills[rownames(lhills) != "Knock Hill",]
lhills.lm=lm(time~.,data=lhills)
summary(lhills.lm)
```

```

Call:
lm(formula = time ~ ., data = lhills)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51726 -0.07516  0.00873  0.06822  0.32955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.20937    0.28159   0.744    0.463    
dist         0.91328    0.06504  14.041 5.57e-15 ***
climb        0.25938    0.04826   5.375 7.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.16 on 31 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.949 
F-statistic: 307.9 on 2 and 31 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lhills.lm)
```



```
outlierTest(lhills.lm)
```

```
             rstudent unadjusted p-value Bonferonni p
Black Hill -4.031848         0.00034982     0.011894
```

After removing the Knock Hill race we can see that diagnostic plots all look much better, furthermore the adjusted R^2 value has increased so that now 95% of the variance in the data is explained by the model.

The `outlierTest` shows that Black Hill may also be an outlier, however you should be careful removing too many data points from the model — especially if you do not have a reason that explains why the data point is different from the other points, like we had in the rodent example.

Feel free to experiment and remove Black Hill, you will find that it only slightly improves adjusted R^2 and that `outlierTest` shows no more outliers.

2. Prediction interval for Autoanalyser data

Using the output from before, construct a 95% prediction interval for the autoanalyser concentration y when the true value x is 6 and interpret the interval

Answer to Task 2

$$5.783 \pm 2.228\sqrt{\frac{0.252}{10}(1+0.233)}$$

$$5.783 \pm 0.393$$

(5.390, 6.176)

A future observation for the autoanalyser concentration is highly likely to lie between 5.39 and 6.18 when the true value is 6.