Statistical Modelling
Epiphany Term 2023
Lecture notes

Lecturer	Dr Tahani Coolen-Maturi
Email	tahani.maturi@durham.ac.uk
Office	MCS 3030
Office hour	See Blackboard Learn Ultra

Online information

Blackboard Learn Ultra (or Ultra for short) will be used to post lecture notes, problems sheets, tutorials and computer practicals.

Information on learning outcomes and assessment for the module may also be found in the Faculty Handbook:
`https://www.dur.ac.uk/faculty.handbook/module_description/?year=2022&module_code=MATH2697`.

Literature

(C) Crawley. *The R book* (2007): Wiley, ISBN 9786610900978. (electronic resource).

(K) Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*, (2000): Oxford, ISBN 0198507089.

(M) Mardia, Kent, & Bibby, *Multivariate Analysis*: Academic Press, 1979, ISBN 0124712509.

(MC) Petersen & Petersen (2008), *The Matrix Cookbook* (Version November 15, 2012), (PDF).

(N) Neter, Kutner, Nachtsheim, & Wasserman. *Applied Linear Statistical Models* (several editions with different combinations of authors 1974–2004): McGraw-Hill, ISBN 0256117365.

(R) Rice. *Mathematical Statistics and Data Analysis* (3rd edn., 2006): Brooks/Cole, ISBN 0495110892.

(RA) Raykov. *Basic Statistics — an introduction with R* (2013). Access via MyiLibrary.

(V) Venables, Smith, and the R Development Core Team. *An Introduction to R* (HTML) (PDF).

(W) Weisberg. *Applied Linear Regression* (3rd edn., 2005): Wiley-Interscience, ISBN 0471663794.

Based on lecture notes developed by Prof. Jochen Einbeck at Durham.

Contents

Chapter 1

Introduction

1.1 Supervised and unsupervised learning

Most statistical problems can be interpreted as a *learning* problem. One has observed some data sample, and wants to use this sample to learn something about the underlying population from which this data was sampled. We start directly with a simple illustrative environmental data example.

Example 1.1 *Scallop data.*

Scallops are small bivalves that live in deep waters and grow in shells, much the way oysters do. Fig. 1.1 (left) provides a graphical representation of $n = 127$ locations (represented by two variables, longitude `long` and latitude `lat`) at which scallops were collected in a 1990 survey cruise in the Atlantic continental shelf off Long Island, New York, USA.

The set of locations forms a bivariate data cloud of points $(\texttt{long}_i, \texttt{lat}_i)$, $i = 1, \dots n$. One can put these together to form a *data matrix*

$$\boldsymbol{Z} = \begin{bmatrix} \texttt{long}_1 & \texttt{lat}_1 \\ \vdots & \vdots \\ \texttt{long}_n & \texttt{lat}_n \end{bmatrix}.$$

One may be interested in using $\boldsymbol{Z}$ in order to gain some information on the scallop population that it represents. For instance, properties of interest could be

- the mean, $\bar{\boldsymbol{Z}} = (\overline{\texttt{long}}, \overline{\texttt{lat}})$;
- the variance-covariance structure of $\boldsymbol{Z}$ (Sec. 2.4);
- main directions of variability; here: south-west to north-east ('Principal component analysis', we will not cover this!)
- the existence of outliers (Sec. 2.5);
- when assuming a certain distributional shape: the parameter estimates and their properties ('Maximum Likelihood Estimation', Sec. 1.2.2);
- the existence of clusters or other features (we will not do this!) ...

We continue with the same data set but consider an additional variable. Beyond the locations, we have additional information available on the abundance of scallops caught by the cruiser. Specifically, we are given $\boldsymbol{Y} = (y_1, \dots, y_n)^T$, where y_i is the logarithm of the number of scallops found at location $(\texttt{long}_i, \texttt{lat}_i)$. We display this additional variable in vertical direction in a 3D scatterplot (Fig. 1.1 right). We observe from this plot that there is some roughly linear relationship between the variables (`long`, `lat`) contained in the original data matrix, $\boldsymbol{Z}$, and the *response vector*, $\boldsymbol{Y}$. The task is to quantify this relationship in some suitable form.

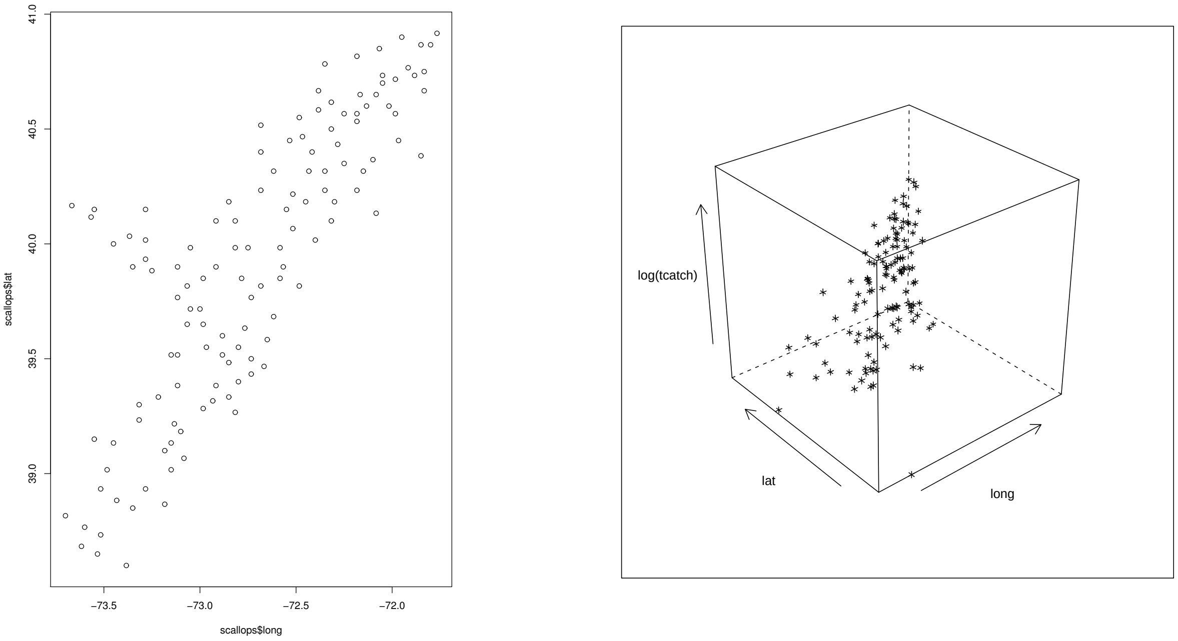


Figure 1.1: Longitudes and latitudes of scallop locations. Left: 2D scatterplot; right: with log–abundances.

Note that the view on the data has changed compared to the initial situation. We have now two types of variables: Those ones stored in $\boldsymbol{Z}$ play the role of an *input*, and that one stored in $\boldsymbol{Y}$ play the role of an *output*. We wish to gain some insight how the physical system considered transforms a given input into an output. We will achieve this through a *statistical model* which relates the elements of $\boldsymbol{Y}$ to the elements of $\boldsymbol{Z}$.

For example, a simple model that would spring into mind could be

$$y_i = \beta_1 + \beta_2 \texttt{long}_i + \beta_3 \texttt{lat}_i + \epsilon_i, \tag{1.1}$$

where the parameters β_1, β_2, and β_3 have to be estimated, and ϵ is some *noise* or *error*. In the example considered, such noise may stem from daily change of environmental conditions (weather, tides), unknown quantities related to the cruiser (speed, experience, etc.), measurement (counting) error, and other sources causing random variation.

The model (1.1) is an example for a *linear regression model*. We will deal with this model in detail in Chapter 3. There we will also study how to obtain estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ of the unknown parameters β_1, β_2, and β_3, thus enabling us to produce *fitted values*

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \texttt{long}_i + \hat{\beta}_3 \texttt{lat}_i.$$

How do we know whether these are good? Here, the observed (known) values y_i can play the role of the *teacher*: When all y_i are relatively close to the $\hat{y}_i$, then the model is good (in some sense), otherwise it is bad; i.e., one aims to minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{1.2}$$

Therefore this kind of statistical learning is often referred to as *supervised learning*. In contrast, learning problems involving just a data matrix $\boldsymbol{Z}$, as described earlier in this section, are referred to as *unsupervised learning* problems (Table 1.1).

Table 1.1: Supervised and unsupervised learning

data structure	Aim	Learning type
$\boldsymbol{Z}$	Investigate properties of $\boldsymbol{Z}$	**Unsupervised learning**
$[\boldsymbol{Z}, \boldsymbol{Y}]$	Learn how $\boldsymbol{Z}$ affects $\boldsymbol{Y}$	**Supervised learning**

Example 1.2 (R Code for Example 1.1)

```
#install.packages("remotes") #you need to do that once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> ?scallops
> dim(scallops)
 [1] 127   4
>  names(scallops)
 [1] "lat"    "long"   "tcatch" "y"
    # Note: y=log(tcatch)

> scallops[1:6,]
> head(scallops)
    # Both commands above do the same, namely to display the first 6 rows:
     lat      long tcatch       y
1 40.38333 -71.85000      1 0.00000
2 40.13333 -72.08333      2 0.69315
3 40.10000 -72.31667      7 1.94591
4 40.01667 -72.40000     13 2.56495
5 39.90000 -72.56667    530 6.27288
6 39.81667 -72.48333   2750 7.91936

# Find overall mean of positions
> c(mean(scallops$long), mean(scallops$lat)) # or, equivalently:
> colMeans(scallops[,c("long","lat")])
  long       lat
 -72.73215  39.91798

# Some graphical display in 2D:
> plot(scallops$long, scallops$lat)      # Fig. 1.1 (left)

# Now involve the response, y:
> library(lattice)
> cloud(y~long+lat, data=scallops)       # Fig. 1.1 (right)

# An additional tidbit
> plot(scallops$long, scallops$lat, cex=0.75)
> for (j in 1:126){
      segments(scallops$long[j],scallops$lat[j],scallops$long[j+1],
               scallops$lat[j+1], col=2)
  }   # gives the route presumably taken by the cruiser
```

1.2 Basics of multivariate analysis

Random vectors and densities

Let $X_j, j = 1, \ldots, q$ be a collection of real-valued random quantities (r.q.'s). Then

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_q \end{pmatrix}$$

forms a q- dimensional *random vector* (r.v.) and

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_q \end{pmatrix}$$

a particular realization of X. The probabilistic behavior of X is entirely determined by the distribution function of X,

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_q) = P(X_1 \leq x_1, \ldots, X_q \leq x_q).$$

An overview of important operations with density functions is provided in Table 1.2 for general reference.

Table 1.2: Common operations with density functions. Let $g : \mathbb{R}^q \longrightarrow \mathbb{R}$ be any arbitrary (integrable) function, $h : \mathbb{R}^q \longrightarrow \mathbb{R}^p$ a bijective and differentiable function, p a positive integer, and $\boldsymbol{y} = (y_1, \ldots, y_p)^T$.

(D1)	Marginalization ($p < q$)	$f(x_1, \ldots, x_p) = \int_{x_q} \cdots \int_{x_{p+1}} f(x_1, \ldots, x_q)\, dx_{p+1} \ldots dx_q$
(D2)	Conditioning ($p < q$)	$f(x_{p+1}, \ldots, x_q \vert x_1, \ldots, x_p) = f(x_1, \ldots, x_q)/f(x_1, \ldots, x_p)$
(D3)	Independence	$X_1, \ldots, X_q$ independent $\longleftrightarrow f(x_1, \ldots, x_q) = f(x_1) \cdot \ldots \cdot f(x_q)$
(D4)	Expectation	$E(g(X)) = \int g(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$
(D5)	Change of variables	The density of $Y = h(X)$ is $f_Y(\boldsymbol{y}) = f(h^{-1}(\boldsymbol{y})) \vert d\boldsymbol{x}/d\boldsymbol{y} \vert$.

Expectation and variance

The *expectation* of a r.v. $X = (X_1, \ldots, X_q)^T$ is given by

$$\begin{pmatrix} m_1 \\ \vdots \\ m_q \end{pmatrix} = \boldsymbol{m} = E(X) = \int \boldsymbol{x} f(\boldsymbol{x})\, d\boldsymbol{x} = \int \ldots \int \begin{pmatrix} x_1 \\ \vdots \\ x_q \end{pmatrix} f(x_1, \ldots, x_q) dx_1 \ldots dx_q = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_q) \end{pmatrix}$$

that is, the j–th component of $\boldsymbol{m}$ is just the expectation of the j–th component of X.

The *variance* is given by

$$\text{Var}(X) = E((X - \boldsymbol{m})(X - \boldsymbol{m})^T) = E(XX^T) - \boldsymbol{m}\boldsymbol{m}^T = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1q} \\ \vdots & \ddots & \vdots \\ \Sigma_{q1} & \cdots & \Sigma_{qq} \end{pmatrix} = \boldsymbol{\Sigma}, \qquad (1.3)$$

where

$$\begin{aligned} \Sigma_{ij} &= \text{Cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j) \qquad i \neq j \\ \Sigma_{jj} &\equiv \sigma_j^2 = \text{Var}(X_j) \end{aligned}$$

In short, we write

$$X \sim (\boldsymbol{m}, \boldsymbol{\Sigma})$$

meaning that X has some unspecified distribution with mean (expectation) $\boldsymbol{m}$ and variance $\boldsymbol{\Sigma}$. Any variance matrix $\boldsymbol{\Sigma}$ found via (1.3) has the following properties:

(i) $\boldsymbol{\Sigma}$ is symmetric,

(ii) $\boldsymbol{\Sigma}$ is positive semi-definite.

If a given matrix fulfils (i) and (ii) we call it a *valid* variance matrix, otherwise *it is invalid.*

Important related concepts

- The *correlation matrix* is defined as

$$\boldsymbol{R} = (R_{ij})_{1 \leq i \leq q, 1 \leq j \leq q}$$

with pairwise correlation coefficients

$$R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}},$$

where $R_{ij} = R_{ji} \in [-1, 1]$ for $i \neq j$ and $R_{ii} = 1$ for $i = 1, \ldots, q$. In matrix notation the correlation matrix can be expressed as

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{-1/2},$$

where $\boldsymbol{D} = \text{diag}(\Sigma_{11}, \ldots, \Sigma_{qq})$. A very useful feature of the correlation matrix $\boldsymbol{R}$ is that it is scale–invariant.

- We call the random variables X_i and X_j *uncorrelated* if $R_{ij} = 0$. If X_i and X_j are independent, then $\Sigma_{ij} = 0$ and, consequently, X_i and X_j are uncorrelated. Hence, under (D3), all X_i, X_j, for $i \neq j$, are uncorrelated so $\boldsymbol{R}$ becomes the identity matrix.

- Let $Y \in \mathbb{R}^p$ be a further random vector. We define the *covariance* between r.v.'s $X \in \mathbb{R}^q$ and $Y \in \mathbb{R}^p$ as

$$\text{Cov}(X, Y) = (\text{Cov}(X_i, Y_j))_{1 \leq i \leq q, 1 \leq j \leq p} = \text{Cov}(Y, X)^T.$$

- Sums of random vectors (of the same dimensionality): For $X \sim (\boldsymbol{m}, \boldsymbol{\Sigma})$, $Y \sim (\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\Sigma}})$, one has

$$X \pm Y \sim \Big(\boldsymbol{m} \pm \tilde{\boldsymbol{m}}, \boldsymbol{\Sigma} \pm \text{Cov}(X, Y) \pm \text{Cov}(Y, X) + \tilde{\boldsymbol{\Sigma}}\Big) = \Big(\boldsymbol{m} \pm \tilde{\boldsymbol{m}}, \boldsymbol{\Sigma} \pm 2\text{Cov}(X, Y) + \tilde{\boldsymbol{\Sigma}}\Big) \quad (1.4)$$

Note that the two covariance terms vanish if X and Y are independent.

1.2.1 Multivariate normal distribution

We say that the random vector $X = (X_1, \ldots, X_q)^T \in \mathbb{R}^q$ is *multivariate normal* (MVN) with parameters $\boldsymbol{m} \in \mathbb{R}^q$, $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ (pos. def.) if its density is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{m})\right\}, \quad (1.5)$$

where $|\boldsymbol{\Sigma}| \equiv \det(\boldsymbol{\Sigma})$. In short we write

$$X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma}).$$

In general the opposite is not true; i.e. $\Sigma_{ij} = 0 \not\Rightarrow f(x_i, x_j) = f(x_i)f(x_j)$. For instance, consider a random variable X_1 with $E(X_1) = 0$ and $E(X_1^3) = 0$ and the random variable $X_2 = X_1^2$. Clearly, X_1 and X_2 are dependent; however, $Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = E(X_3) = 0$.

Linear transformations and special cases

A very useful identity related to linear transformations of multivariate normal r.v. is the following: if $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$, $\boldsymbol{A} \in \mathbb{R}^{r\times q}$ and $\boldsymbol{b} \in \mathbb{R}^r$, then

$$\boldsymbol{A}X + \boldsymbol{b} \sim N_r(\boldsymbol{A}\boldsymbol{m} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T). \tag{1.6}$$

Recall that

$$E(\boldsymbol{A}X + \boldsymbol{b}) = E(\boldsymbol{A}X) + E(\boldsymbol{b}) = \boldsymbol{A}E(x) + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{m} + \boldsymbol{b}$$

$$\begin{aligned} Var(\boldsymbol{A}X + \boldsymbol{b}) &= E((\boldsymbol{A}X + \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{m} + \boldsymbol{b}))(\boldsymbol{A}X + \boldsymbol{b} - (\boldsymbol{A}m + \boldsymbol{b}))^T) \\ &= E((\boldsymbol{A}(X - \boldsymbol{m}))(\boldsymbol{A}(X - \boldsymbol{m}))^T) = E(\boldsymbol{A}(X - \boldsymbol{m})(X - \boldsymbol{m})^T\boldsymbol{A}^T) \\ &= \boldsymbol{A}E((X - \boldsymbol{m})(X - \boldsymbol{m})^T)\boldsymbol{A}^T \\ &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T. \end{aligned}$$

Eq. (1.6) is very useful as important properties may be derived in a very simple manner. Let look how we can use (1.6) for two important special cases.

- Marginalization:
  The general way of finding marginal distributions involves multiple integration; see (D1) in Table 1.2. For instance, we can find the marginal of X_1 by doing the following integration

  $$f(x_1) = \int_{x_2} \dots \int_{x_q} f(x_1, x_2, \dots, x_q)dx_2 \dots dx_q,$$

  which is a rather tedious process! Here we can use (1.6) in order to avoid this. Specifically, if we have $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$ then if want to find the marginal of X_j we simply define the row vector $\boldsymbol{A} = (0 \dots 0\ 1\ 0 \dots 0) \in \mathbb{R}^{1\times q}$ which is zero everywhere except at the j-th position where it equals 1. We also set $b = 0$. Then

  $$X_j = \boldsymbol{A}X + b \sim N(m_j, \Sigma_{jj}).$$

  So if the r.v. X is MVN, then each univariate component X_j is normally distributed.

- Standardization
  Again we have that $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite and therefore its square root matrix $\boldsymbol{\Sigma}^{1/2}$ exists. Then, we can consider the following transformation

  $$\boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{m}) = \boldsymbol{\Sigma}^{-1/2}X - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}.$$

  In the context of Eq. (1.6) we have $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{b} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}$; therefore,

  $$\begin{aligned} \boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{m}) &\sim N_q(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{m} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}) \\ &\sim N_q(\boldsymbol{0}, \boldsymbol{I}_q). \end{aligned}$$

  The transformed r.v. has zero means and covariance matrix equal to the identity matrix (zero covariances/corellations and variances equal to one).

Example 1.3 Bivariate normal distribution (BVN) with uncorrelated components.
Let

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Then

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{1}{2}(x_1 - m_1, x_2 - m_2) \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - m_1 \\ x_2 - m_2 \end{pmatrix} \right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{1}{2}(x_1 - m_1, x_2 - m_2) \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix} \begin{pmatrix} x_1 - m_1 \\ x_2 - m_2 \end{pmatrix} \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{1}{2}\frac{(x_1 - m_1)^2}{\sigma_1^2} \right\} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{1}{2}\frac{(x_2 - m_2)^2}{\sigma_2^2} \right\} \\ &= f(x_1) \cdot f(x_2) \end{aligned}$$

Hence, when X_1 and X_2 are *bivariate normal* and uncorrelated, then they are also independent. Obviously, this also holds the other way round: When X_1 and X_2 are normally distributed and independent, then they are BVN.

Example 1.4 *Visualizing bivariate normal distributions*

The following code visualizes the density $f(x, y)$ of a bivariate normal distribution generated by two independent random quantities $X \sim N(6, 3^2)$ and $Y \sim N(3, 1^2)$. The resulting picture is shown in Fig. 1.2.

```
> x<- seq(-8,20, length=51) # defines the range of x-values plotted
> y<- seq(-2,8, length=51)

> dens <- matrix(0,51,51) # creates an "empty" matrix of appropriate
                          # dimension which will be used for values of f(x,y)
> for (i in 1:51){
>  for (j in 1:51){
>    dens[i,j]<- dnorm(x[i],6,3)*dnorm(y[j],3,1) # uses independence (D3)
                                                 # to generate the joint density
>   }
> }

> persp(x, y, dens, theta=40, phi=20) # plots the  density in 3D
```

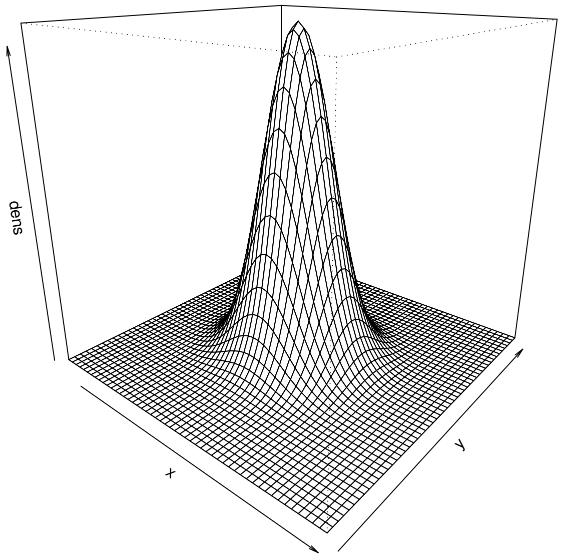


Figure 1.2: Plot of bivariate density function $f(x_1, x_2)$ with uncorrelated components.

Another way of visualizing multivariate densities are through "contours". Contours are defined as curves of equal density. For the MVN, these curves are ellipsoids

$$(\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{m}) = c^2,$$

for some constant c. For contours in R we use the following command.

```
> contour(x, y, dens) # contour plot of density
```

The corresponding graph is shown in Fig. 1.3.

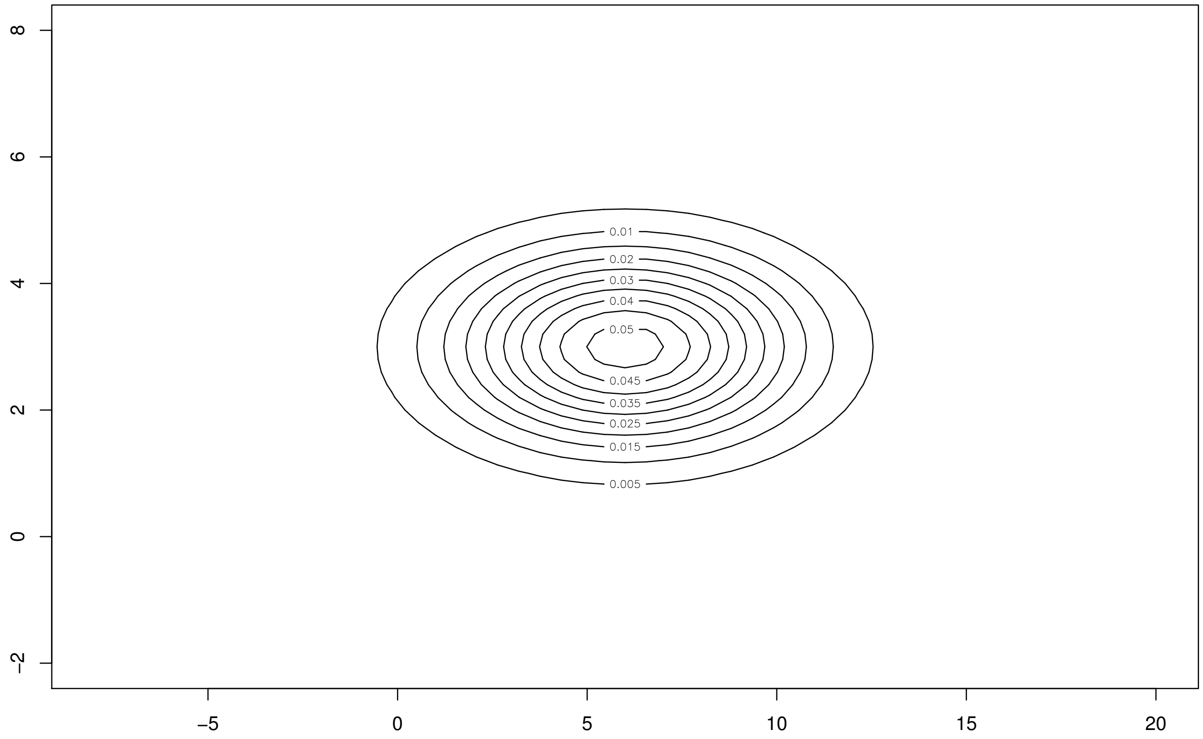


Figure 1.3: Contour plot of bivariate density function $f(x_1, x_2)$ with uncorrelated components.

Normality, independence and correlation

Simple results of the type above have let to the widespread misconception that, when X_1 and X_2 are each normally distributed and uncorrelated, then they are independent. This is **not** true; an example is provided below. Table 1.3 summarizes the relationship between multivariate independence/correlation/normality in diagrammatic form.

Table 1.3: Relationship between normality, multivariate normality, independence, and correlation

$X_1, \ldots, X_q$ are...	MVN		normal
independent		$\longleftrightarrow$	
	$\updownarrow$		$\downarrow$
uncorrelated		$\longrightarrow$	
otherwise		$\longrightarrow$	

Example 1.5 *Uncorrelated normal variables which are not independent.*

Let $X \sim N(0,1)$ and $Y = WX$, where the random variable W is independent of X and has probability mass

$$W = \begin{cases} -1, \text{ with probability } 1/2, \\ 1, \text{ with probability } 1/2. \end{cases}$$

First, let us examine the covariance of X and Y:

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) && (E(X) = 0) \\ &= E(X^2W) && (X^2, W \text{ independent}) \\ &= E(X^2)E(W) && (E(W) = 0) \\ &= 0. \end{aligned}$$

Thus, X and Y are uncorrelated. Next, we will prove that Y is also normally distributed by showing that $\Pr(Y \le x) = \Pr(X \le x)$.

$$\begin{aligned} \Pr(Y \le x) &= E(\Pr(Y \le x|W)) \\ &= \Pr(X \le x)\Pr(W = 1) + \Pr(-X \le x)\Pr(W = -1) \\ &= \Pr(X \le x)\Pr(W = 1) + \Pr(X \ge -x)\Pr(W = -1) \\ &= \Pr(X \le x)\frac{1}{2} + \Pr(X \le x)\frac{1}{2} && \text{(symmetry of normal)} \\ &= \Pr(X \le x). \end{aligned}$$

So, X and Y are both normally distributed and uncorrelated. However, they are not independent. For instance, $|Y| = |X|$.

Conditional normality

Any r.v. $X = (X_1 \ldots X_q)^T$ can be represented in terms of random subvectors. For instance, for $(p < q)$ we may write $X = (Y\ Z)^T$, where $Y = (X_1 \ldots X_p)^T \in \mathbb{R}^p$ and $Z = (X_{p+1} \ldots X_q)^T \in \mathbb{R}^{q-p}$. Then, for $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$ we have

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N_q\left(\begin{pmatrix} \boldsymbol{m}_Y \\ \boldsymbol{m}_Z \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_Z \end{pmatrix}\right),$$

where $\boldsymbol{m}_Y \in \mathbb{R}^p$, $\boldsymbol{m}_Z \in \mathbb{R}^{q-p}$, $\boldsymbol{\Sigma}_Y \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Sigma}_{YZ} \in \mathbb{R}^{p\times(q-p)}$, $\boldsymbol{\Sigma}_{ZY} \in \mathbb{R}^{(q-p)\times p}$ and $\boldsymbol{\Sigma}_Z \in \mathbb{R}^{(q-p)\times(q-p)}$.

A further property of the MVN relates to conditional distributions. Specifically, The conditional distribution of Y given a realization of $Z = z$ is again MVN; namely,

$$\begin{aligned} Y \mid Z = z &\sim N_p(\tilde{\boldsymbol{m}}_Y, \tilde{\boldsymbol{\Sigma}}_Y), \\ \tilde{\boldsymbol{m}}_Y &= \boldsymbol{m}_Y + \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{z} - \boldsymbol{m}_Z), \\ \tilde{\boldsymbol{\Sigma}}_Y &= \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZY}. \end{aligned}$$

When the covariance between Y and Z is zero the conditional distribution of Y is the marginal distribution $N_p(\boldsymbol{m}_Y, \boldsymbol{\Sigma}_Y)$.

Example 1.6 *The conditional distribution in the bivariate correlated case.*

If $X = (X_1\ X_2)^T$ BVN with non-zero correlation then the conditional distribution of X_1 is

$$X_1 \mid X_2 = x_2 \sim N\left(m_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - m_2), (1-\rho^2)\sigma_1^2\right),$$

where $\rho = \dfrac{\sigma_{12}}{\sigma_1\sigma_2}$ is the correlation.

1.2.2 Variance matrix estimation

The setup is as follows: We have data $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})^T$, $i = 1, \ldots, n$, which are n independent and identically distributed (iid) observations generated from a r.v.

$$X \sim (\boldsymbol{m}, \boldsymbol{\Sigma}) \in \mathbb{R}^q,$$

and which form together a data matrix

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1q} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{nq} \end{pmatrix}.$$

We denote by

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{iq} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{x}}_1 \\ \vdots \\ \bar{\boldsymbol{x}}_q \end{pmatrix}$$

the overall mean, which forms an unbiased estimate of the expectation $\boldsymbol{m}$ meaning that if we use the estimator $\hat{\boldsymbol{m}} = \bar{\boldsymbol{x}}$ then $E(\hat{\boldsymbol{m}}) = \boldsymbol{m}$. The goal of this subsection is to estimate $\boldsymbol{\Sigma}$.

Firstly, recall

$$\boldsymbol{\Sigma} = \mathrm{Var}(X) = E\left((X - \boldsymbol{m})(X - \boldsymbol{m})^T\right).$$

Replacing all expectations by means and $\boldsymbol{m}$ by $\bar{\boldsymbol{x}}$ (if the former is unknown), a natural candidate estimator for $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \overset{\text{Q. 2.1}}{=} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T \in \mathbb{R}^{q \times q}.$$

In fact, this turns out to be the Maximum Likelihood estimator for $\boldsymbol{\Sigma}$ (under the MVN assumption), therefore we give this estimator the suffix ML. It can be shown that $\hat{\boldsymbol{\Sigma}}_{ML}$ is not unbiased for $\boldsymbol{\Sigma}$, as

$$E(\hat{\boldsymbol{\Sigma}}_{ML}) = \left(1 - \frac{1}{n}\right)\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

However, since

$$\lim_{n \longrightarrow \infty} (\hat{\boldsymbol{\Sigma}}_{ML}) = \lim_{n \longrightarrow \infty} \left(1 - \frac{1}{n}\right)\boldsymbol{\Sigma} = \boldsymbol{\Sigma},$$

we say that $\hat{\boldsymbol{\Sigma}}_{ML}$ is *asymptotically unbiased.*

An unbiased estimator of $\boldsymbol{\Sigma}$ is obtained through the *sample variance matrix*,

$$\hat{\boldsymbol{\Sigma}}_{\text{sample}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{ML},$$

where essentially the fraction $n/(n-1)$ is a bias correction factor. We generally prefer $\hat{\boldsymbol{\Sigma}}_{\text{sample}}$ to $\hat{\boldsymbol{\Sigma}}_{ML}$.

Strictly, when talking of iid *observations* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we mean that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are realizations of n independent and identically distributed *random vectors*, which all have the same distribution as X. Note also, this does of course *not* imply that the components $X_1, \ldots, X_q$ need to be independent!

1.3 Mahalanobis distance

Given: Multivariate distribution

$$X \sim (\boldsymbol{m}, \boldsymbol{\Sigma}).$$

Question: What is (a fair measure of) the distance between a point $\boldsymbol{x} = (x_1, \dots, x_q)^T$ and the mean $\boldsymbol{m} = (m_1, \dots, m_q)^T$?

Example 1.7 (Continuation of Example 1.4).

Back to the BVN with uncorrelated components. Consider the contour plot shown in Fig. 1.2 in Example 1.4. One finds easily, by taking the length along the coordinate axes, that the (Euclidean) distance from, say, $(12, 3)^T$ to $\boldsymbol{m} = (6, 3)^T$ is 6, while the distance from $(6, 6)^T$ to $\boldsymbol{m}$ is only 3. But does this reflect accurately the effort to cover the distance? From the perspective of a hiker, the way from $(12, 3)^T$ to the summit is probably the easier one, as one starts from a higher basis already.

In more mathematical terms, taking the Euclidean distance

$$d_E = \sqrt{(x_1 - m_1)^2 + (x_2 - m_2)^2}$$

ignores the differing variability of the components of the random vector. To account for this, one can consider the *standardized Euclidean distance*

$$\begin{aligned} d_M &= \sqrt{\left(\frac{x_1 - m_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - m_2}{\sigma_2}\right)^2} = \\ &= \sqrt{(x_1 - m_1 \; x_2 - m_2) \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} x_1 - m_1 \\ x_2 - m_2 \end{pmatrix}} \end{aligned}$$

which gives, for the example above, the values 2 and 3, respectively, reflecting the perspective taken by the hiker.

—

One can generalize this notion to arbitrary random vectors: For any $X \sim (\boldsymbol{m}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ pos. def., one defines the *Mahalanobis distance* (to the mean) through

$$d_M(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{m})}.$$

If X is a MVN, i.e. $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$, then points with equal Mahalanobis distance to the mean lie on the same contour, and one has,

$$\begin{aligned} d_M^2(X, \boldsymbol{m}, \boldsymbol{\Sigma}) &= (X - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{m}) = \\ &= (X - \boldsymbol{m})^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} (X - \boldsymbol{m}) = \\ &= \underbrace{\left[\boldsymbol{\Sigma}^{-1/2} (X - \boldsymbol{m})\right]^T}_{\equiv Y \sim N_q(0, \boldsymbol{I})} \left[\boldsymbol{\Sigma}^{-1/2} (X - \boldsymbol{m})\right] = \\ &= Y^T Y = \sum_{j=1}^{q} \underbrace{Y_j^2}_{N(0,1)^2} \sim \chi_q^2 \end{aligned}$$

Note: The χ^2 property remains true (rule of thumb $n \geq 10q$) when $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are replaced by $\hat{\boldsymbol{m}} = \bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{ML}$ or $\hat{\boldsymbol{\Sigma}}_{\text{sample}}$).

In the following subsections we discuss important applications of the Mahalanobis distance. Firstly, we discuss how to verify the assumption of multivariate normality.

1.3.1 Checking multivariate normality

For given data $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \in \mathbb{R}^q$ sampled from a random vector $X \sim (\boldsymbol{m}, \boldsymbol{\Sigma})$, we wish to check whether or not X is a MVN. Let

$$d_i = d_M^2(\boldsymbol{x}_i, \boldsymbol{m}, \boldsymbol{\Sigma}), \tag{1.7}$$

where $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ may be replaced by appropriate estimates $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively.

We know that, if X is MVN, then $d_M^2(X, \boldsymbol{m}, \boldsymbol{\Sigma}) \sim \chi_q^2$. Hence, under multivariate normality, we would expect the values d_i to follow closely a χ^2 distribution with q degrees of freedom. This assumption can be easily checked in complete analogy to the well-known "normal probability plot" which is used to check a sample for univariate normality.

Specifically, the QQ plot for checking MVN is constructed as follows:

1. Sort the values of d_i, yielding ordered values $d_{(i)}$, $i = 1, \dots, n$.
2. Compute the quantiles $q_i = \chi^2_{q,1-(i-0.5)/n}$, i.e. the quantiles of the χ_q^2 distribution with a probability mass of $\frac{i-0.5}{n}$ to their left hand side.
3. Plot $d_{(i)}$ (vertical) versus q_i (horizontal).
4. Compare the plotted points to a straight line through the origin with slope equal to 1. Deviations from this line indicate deviations from MVN.

Example 1.8 *Fuel consumption data.* We consider data collected for the analysis of fuel consumption in 48 US states. There are nine variables in this data set but we are currently only interested in four of them: `TAX` (cents per gallon), `DLIC` (% population with driving licences), `INC` (average income in \$1000's), `ROAD` (1000's of miles).

```
#install.packages("remotes") #you need to do that once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> ?fuelcons
> head(fuelcons)
#  STATE  POP  TAX NLIC   INC  ROAD FUELC DLIC FUEL
#1    ME 1029  9.0  540 3.571 1.976   557 52.5  541
#2    NH  771  9.0  441 4.092 1.250   404 57.2  524
#3    VT  462  9.0  268 3.865 1.586   259 58.0  561
#4    MA 5787  7.5 3060 4.870 2.351  2396 52.9  414
#5    RI  968  8.0  527 4.399 0.431   397 54.4  410
#6    CN 3082 10.0 1760 5.342 1.333  1408 57.1  457
> fuel <- fuelcons[,c("TAX", "DLIC", "INC", "ROAD")]
> pairs(fuel)
```

We firstly estimate mean and variance:

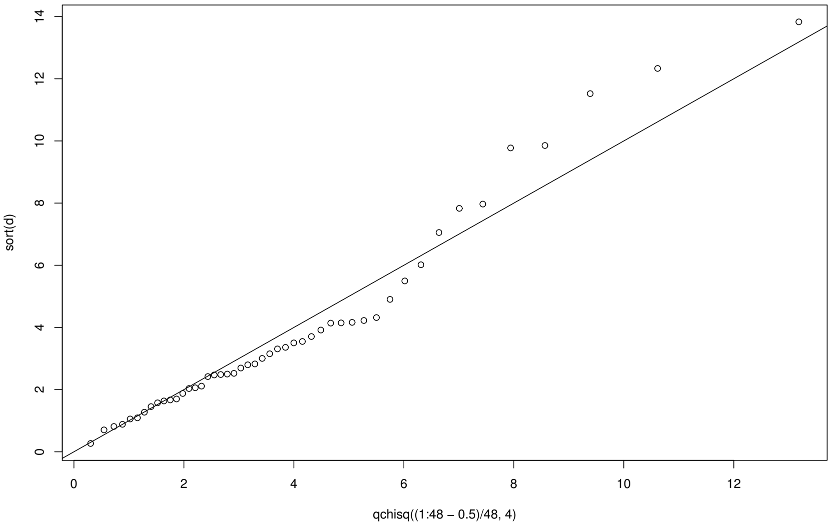
```
> m     <- colMeans(fuel)
> Sigma <- var(fuel)
```

A χ^2 probability plot for checking MVN via the Mahalanobis distances is obtained through

```
> d<- mahalanobis(fuel, m, Sigma)
> plot(qchisq( (1:48-0.5)/48,4),sort(d))
> abline(a=0,b=1)
```

The resulting plot is provided in Figure 1.4. We see some deviation from the straight line both in the middle part and in the right tail. Some deviations in the tail are hard to avoid even for relatively well–behaved data. The deviation in the middle part is potentially more relevant, as it concerns more data. We conclude that there is some moderate violation of multivariate normality in this data. Any further analysis based on the MVN 'working assumption' should be done with special care.

Figure 1.4: QQ plot $d_{(i)}$ versus q_i for Fuel data set.



We listed the steps required for producing the QQ plot but the question is what is the justification of this procedure as a valid check of multivariate normality. The justification is as follows. The ordered values can be used to evaluate the empirical distribution function (EDF) which depends on the sample and is given by

$$F_n(d_{(i)}) = \widehat{\Pr}(d \leq d_{(i)}) = \frac{\#\{i : d \leq d_{(i)}\}}{n} = \frac{i}{n},$$

where the random variable d is the square of Mahalanobis distance. This is the proportion of squared Mahalanobis distances that are less than or equal to $d_{(i)}$ for $i = 1, \dots, n$. On the other hand, the corresponding theoretical values of the χ^2_q distribution function are

$$\Pr\left(d \leq \chi^2_{q,1-\frac{i}{n}}\right) = \frac{i}{n}.$$

As a result the sets of values $\{\chi^2_{q,1-\frac{1}{n}}, \dots, \chi^2_{q,0}\}$ and $\{d_{(1)}, \dots, d_{(n)}\}$ should align approximately along a straight line passing through the origin with slope equal to 1 if the data are MVN. Note that in practice we use $1-(i-0.5)/n$ instead of $1-i/n$ (in step 2 above). This is due to symmetry reasons; for example, the expectation is that the observed value of $d_{(2)}$ to fall likely in the middle of the interval defined by $\chi^2_{q,1-\frac{1}{n}}$ and $\chi^2_{q,1-\frac{2}{n}}$.

1.3.2 Outlier detection

For a data sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^q$ and a specific data point $\boldsymbol{x} \in \mathbb{R}^q$ [which may or may not be one of the $\boldsymbol{x}_i$], we wish to test

$$H_0 : \boldsymbol{x} \quad \text{is not an outlier}$$

versus

$$H_1 : \boldsymbol{x} \quad \text{is an outlier}$$

at significance level α. The understanding of the word ‘outlier’ is here ‘an observation that deviates so strongly from the other sample values that it is unlikely to have been generated by the same mechanism’. We can equate the word ‘mechanism’ with ‘random vector’. A pragmatic assumption is required for the distribution of X; we choose $X \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$, where $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are usually unknown.

Under this assumption, we know that

$$d_M^2(X, \boldsymbol{m}, \boldsymbol{\Sigma}) \sim \chi_q^2$$

and this would then also be true for $\boldsymbol{x}$ under the null hypothesis. Hence, we reject H_0 if

$$d_M^2(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Sigma}) > \chi_{q,\alpha}^2. \tag{1.8}$$

where $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ need to be replaced by estimates if unknown. Typical choices for α are $\alpha = 0.05$ or $\alpha = 0.025$, with the latter being more common for this particular test.

Note that the test may give misleading results if the distribution of X is not MVN, or the sample size n is small. Since $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ need to be estimated from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, it would be conceptually preferable if $\boldsymbol{x}$ was **not** part of these data (as otherwise $\boldsymbol{x}$ may have a strong impact on $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{\Sigma}}$ and hence avoid its detection as an outlier, an effect known as *masking*).

Despite these concerns, the test is commonly carried out on *all* n observations simultaneously, that is $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{\Sigma}}$ are obtained from the complete data set, and then it is established for which of the n observations (1.8) holds. By construction, each of the n observations is then classified as outlier with $100 \times \alpha\%$ probability *even it is not outlying*, and so the *effective* type I error rate of the test, that is, *the probability that at least one observation will be incorrectly classified as outlier*, will be $1-(1-\alpha)^n \approx n\alpha >> \alpha$. This can be shown as follows

$$\begin{aligned}
\Pr(H_0 \text{ rejected for at least one } \boldsymbol{x}_i) &= 1 - \Pr(H_0 \text{ not rejected for all } \boldsymbol{x}_i) \\
&= 1 - (1-\alpha)^n \\
&= 1 - (1^n - n\alpha + \underbrace{\ldots\ldots\ldots\ldots\ldots}_{\text{very small}}) \\
&\approx n\alpha.
\end{aligned}$$

This is known as the 'multiple testing problem'. Several techniques have been proposed to address this issue; such as replacing $\chi_{q,\alpha}^2$ by $\chi_{q,\alpha/n}^2$ (*Bonferroni correction*). One will often find in practice that not any outliers are detected using this rule; one could say that the Bonferroni correction is 'overly conservative' in detecting outliers. The choice of $\alpha = 0.025$ (rather than 0.05) which is often used for this test is presumably driven by the attempt to mitigate somewhat for the multiple testing without needing to carry out a Bonferroni correction. Perhaps the most pragmatic point of view is to consider this methodology as a useful 'diagnostic device' rather than a formal statistical test.

Example 1.9 (Continuation of Example 1.8)

Let us initially investigate "by hand" whether there are potential outliers. Therefore, we use the function `identify()` as follows:

```
> plot(fuel$ROAD, fuel$INC)
> identify(fuel$ROAD, fuel$INC)
# [1]  7 12 37 ....

> plot(fuel$ROAD, fuel$TAX)
> identify(fuel$ROAD, fuel$TAX)
# [1]  7 12 37 ...
```

Now let's see whether formal outlier tests (using $\alpha = 0.05$ and $\alpha = 0.025$) based on the Mahalanobis distance confirm this result:

```
> which(d >   qchisq(0.95,4))
# [1]  6  7 12 37 45
> which(d >  qchisq(0.975,4) )
# [1]  7 37 45
```

Yes, this roughly confirms the result above (noting that outliers in multivariate space do not need to be outlying in any individual coordinate direction!).

Chapter 2

Linear models: Assumptions & Estimation

In Section 1.2, we were interested in studying the characteristics of a data matrix $\boldsymbol{Z}$ (or of the random vector from which it was generated). In this chapter, we are given data of type $[\boldsymbol{Z}, \boldsymbol{Y}]$, with $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$, and wish to gain insight into the character of the (statistical) dependence of $\boldsymbol{Y}$ on $\boldsymbol{Z}$.

2.1 Model specification

Example 2.1 *Scallop data – Cont. of Example 1.1.*

For the scallop data, we have given

$$\boldsymbol{Z} = \begin{pmatrix} \text{long}_1 & \text{lat}_1 \\ \vdots & \vdots \\ \text{long}_n & \text{lat}_n \end{pmatrix} \in \mathbb{R}^{n\times 2}, \quad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \log(\text{tcatch}_1) \\ \vdots \\ \log(\text{tcatch}_n) \end{pmatrix}.$$

A potentially useful model for these data was already suggested in (1.1), namely

$$y_i = \beta_1 + \beta_2 \text{long}_i + \beta_3 \text{lat}_i + \epsilon_i, \quad i = 1, \ldots, n$$

which, in matrix notation, takes the shape

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \text{long}_1 & \text{lat}_1 \\ \vdots & \vdots & \vdots \\ 1 & \text{long}_n & \text{lat}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = [\boldsymbol{1}, \boldsymbol{Z}]\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Generally, the *linear model* in matrix form is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where all notation is explained in Table 2.1. Taking the i-th row of (2.1), one can represent the linear model via

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{2.2}$$

which we will occasionally write without the index i,

$$y = \sum_{j=1}^{p} \beta_j x_j + \epsilon = \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon, \tag{2.3}$$

Note that the *predictor variables* $x_j, j = 1, \ldots, p$ (each of which corresponds to a column of $\boldsymbol{X}$) may be transformations or functions of the *covariate variables* which constitute the data matrix $\boldsymbol{Z}$. In other words, $\boldsymbol{X}$ is *not necessarily* equal to $[\boldsymbol{1}, \boldsymbol{Z}]$. An example for such a situation is provided in Example 2.2.

Table 2.1: Matrix notation used for the linear model.

Notations:

y : the response variable (also regressand, dependent variable, endogenous variable),

$x_1, \ldots, x_p$: the predictor variables (also regressors, independent or explanatory or exogenous variables)

$\beta_1, \ldots, \beta_p$: fixed unknown coefficients

ϵ : noise or error variable.

Model:

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \\ x_{n1} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where

$\boldsymbol{Y}^T = (y_1, \ldots, y_n)$ is the vector of responses,

$\boldsymbol{X}$ is the design matrix

$\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ is the $p-$ dimensional parameter vector,

$\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of errors.

Example 2.2 *Cement data.* The data below are from an experiment on the tensile strength of cement: Hald, A. (1952) *Statistical Theory with Engineering Applications*, New York: Wiley, 451.

curing time (days)	Tensile strength (kg/cm^2)
1	13.0 13.3 11.8
2	21.9 24.5 24.7
3	29.8 28.0 24.1 24.2 26.2
7	32.4 30.4 34.5 33.1 35.7
28	41.8 42.6 40.3 35.7 37.3

The plot in Figure 2.1 shows that the relationship between 'strength' and 'time' is (i) non-linear, (ii) appears to be increasing to a maximum (as should be expected) and (iii) the variability in strength increases with curing time.

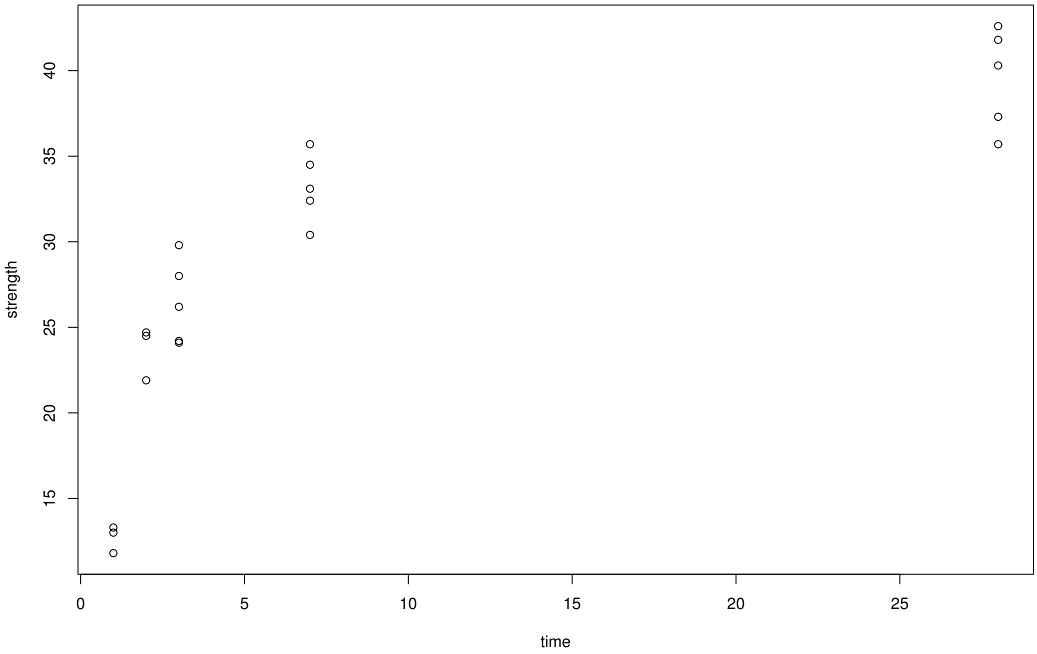


Figure 2.1: "tensile strength" of cement versus "curing time".

We want to be able to make reasonable predictions about tensile strength for any given value of curing time between the experiment's extremes. These predictions should minimally give (i) the typical strength, together with (ii) some measure of its uncertainty.

That there must be some uncertainty is clear from the data—different samples with the same curing time have different strengths. The fact that repeated trials were done for each curing time is very helpful as we can do a rough check of whether the variability is about the same at each curing time or whether it is more complicated than that. Note that the values of 'curing time' are *chosen in advance "by design"*. Look at the summaries

time	$\overline{\text{strength}}$	s_{strength}
1	12.70	0.79
2	23.70	1.56
3	26.46	2.46
7	33.22	2.03
28	39.54	2.95

where the $\overline{\text{strength}}$ are the group averages, while the s_{strength} values are the group standard deviations; for example, 39.54 is an *estimate* of $\text{E}[\text{strength} \mid \text{time} = 28]$ and $(2.95)^2$ is an estimate of $\text{Var}[\text{strength} \mid \text{time} = 28]$.

A plot of these standard deviations against the group means (not shown) reveals that things are complicated—the standard deviations are increasing (perhaps linearly) as the means increase. This is not desirable as one of the underlying assumptions of prevalent regression models is that the standard deviations are constant (as we will formalize later). When we see this happening a standard trick is *transform* the data: and here we transform to $y = \log(\text{tensile strength})$. Doing this we find:

time	$\bar{y}$	s_y
1	2.540	0.064
2	3.164	0.067
3	3.272	0.092
7	3.502	0.061
28	3.675	0.076

which has made the standard deviations roughly the same for each group of measurements. We see from the plot in Figure 2.2 that the relationship between time and log(strength) is still far from linear. Obviously cement cures to some maximum strength so we are looking for a mathematical function which tends to a finite maximum as time increases. Possibly the simplest such model is

$$E(y|\text{time}) = \beta_1 + \beta_2/\text{time}$$

for some constants $\beta_1 > 0$, $\beta_2 < 0$. Indeed, the plot in Figure 2.3 makes this look a plausible thing to do and the *least squares* line is shown. Translating all this to the notation introduced above, this means

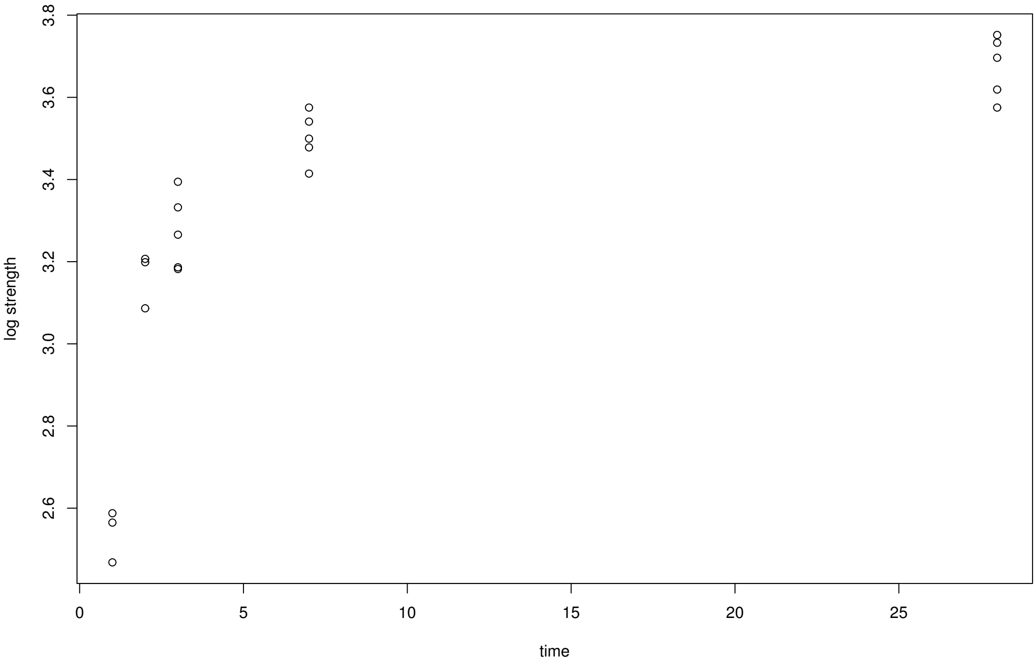


Figure 2.2: log "tensile strength" of cement versus "curing time".

$$\begin{aligned} y &= \log(\text{strength}) \\ x_1 &= 1, \\ x_2 &= 1/\text{time} \end{aligned}$$

and in matrix notation

$$\boldsymbol{Z} = \begin{pmatrix} \text{time}_1 \\ \vdots \\ \text{time}_n \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} \log(\text{strength}_1) \\ \vdots \\ \log(\text{strength}_n) \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1/\text{time}_1 \\ \vdots & \vdots \\ 1 & 1/\text{time}_n \end{pmatrix}.$$

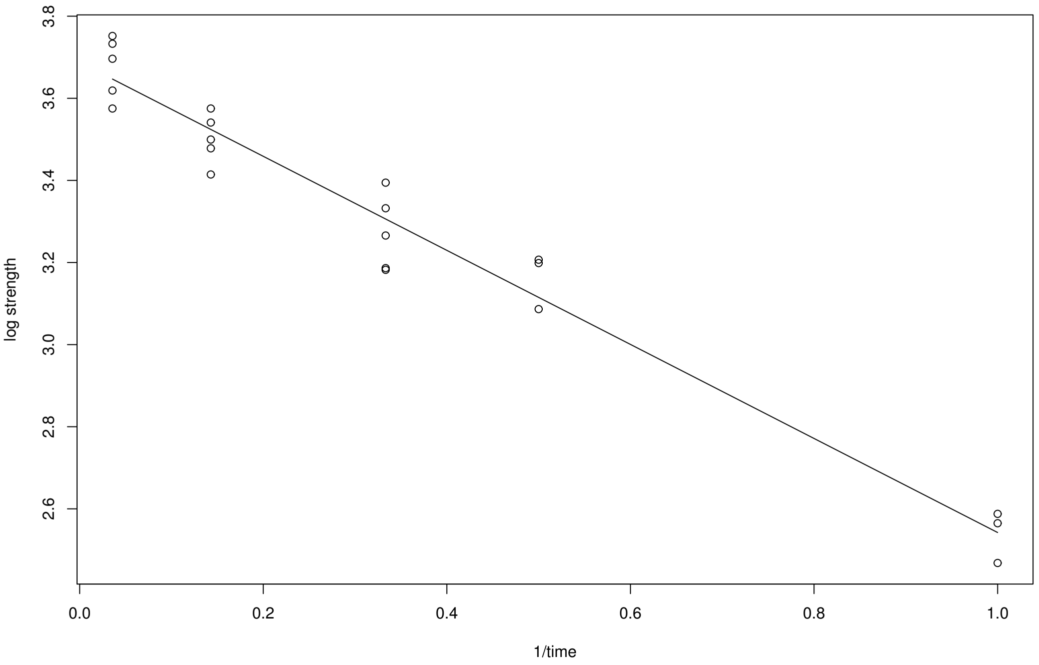


Figure 2.3: log “tensile strength” of cement versus “curing rate” with least squares line.

2.1.1 Linear model assumptions

Consider again the equation (2.2). This, in itself, does not imply any *assumptions*, it is just a *decomposition*. The term $\boldsymbol{x}_i^T\boldsymbol{\beta}$ expresses the tendency of the response to vary in some systematic fashion with $\boldsymbol{x}_i$, and the term ϵ_i expresses variation (“scattering”) around this line (curve, plane...) of statistical relationship. Of course, one has to ensure to do this in a meaningful way: If the fitted line (etc.) does not come close to the data then the decomposition does not make sense. Therefore, a linear model requires a couple of assumptions, which are listed in Table 2.2. Thereby the ϵ_i, $i = 1, \ldots n$ are considered as random variables, and the assumptions pertain to the distribution of the random variables. Note that then $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ takes the role of a *random vector*, and we could write (A1) and (A2) in condensed form as $\boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I})$. But, if $\boldsymbol{\epsilon}$ is random vector, then also $\boldsymbol{Y}$ is a random vector. Considering $\boldsymbol{X\beta}$ as a constant, we see that $\boldsymbol{Y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Including the assumption (A3), these distributions take the form $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and, via (1.6),

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}), \tag{2.4}$$

respectively. We call any model of type $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ which adheres to *at least* (A1) and (A2) a **linear model**. Normality (A3) is not strictly necessary for the pure parameter estimation/learning process, but it will be required for any further inference (confidence or prediction intervals, significance testing, etc.). It will be pointed out in each Section whether or not we actually require it.

Table 2.2: Assumptions of the linear model.

(A1)	Linearity	$E(\epsilon_i) = 0,$	i.e. $E(y_i\|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$
(A2)	Homoscedasticity Independence	$\text{Var}(\epsilon_i) = \sigma^2$ $\text{Cov}(\epsilon_i, \epsilon_j) = 0$	$\}\,\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$
(A3)	Normality	ϵ_i is normally distributed	

2.2 Estimation of model parameters

In this section, we assume (A1) and (A2) to hold for subsection 2.2.2, while (A3) is not strictly necessary. We wish to estimate the model parameters, i.e. $\boldsymbol{\beta}$ and σ^2.

2.2.1 Least-squares (LS) estimation of $\boldsymbol{\beta}$

We find estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ by minimizing

$$R(\boldsymbol{\beta}) = R(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The solution to this minimization problem is derived as follows.

$$\begin{aligned} R(\boldsymbol{\beta}) &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

Now using the identities $\frac{\partial \boldsymbol{a}^T\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{a}$ and $\frac{\partial \boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{A}\boldsymbol{\beta}$ (for $\boldsymbol{A}$ symmetric) we derive the gradient and set it equal to zero; specifically

$$\begin{aligned} \frac{\partial R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2(\boldsymbol{Y}^T\boldsymbol{X})^T + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \overset{!}{=} 0 \\ &\Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}. \end{aligned} \tag{2.5}$$

The above is a system of p linear equations, which are commonly referred to as "normal equations" within the context of statistical estimation.

Example 2.3 Consider the simple linear regression model

$$y_i = a + bx_i + \epsilon_i, i = 1, \ldots, n.$$

With $\boldsymbol{\beta}^T = (a, b)$, $\boldsymbol{X}^T = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$ and $\boldsymbol{Y} = (y_1 \ \ldots \ y_n)^T$, the normal equations take the familiar form

$$\begin{aligned} na + b\sum x_i &= \sum y_i \quad \text{(I)} \\ a\sum x_i + b\sum x_i^2 &= \sum x_i y_i \ \text{(II)} \end{aligned}$$

By multiplying (I) with $\frac{1}{n}\sum x_i$, subtracting it from (II), and after some algebra we find

$$\hat{b} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}, \tag{2.6}$$

and by plugging this solution into (I) we obtain

$$\hat{a} = \frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2 - n\bar{x}^2}. \tag{2.7}$$

In the general case of p predictors if $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular (invertible) the LS estimator is

$$\boxed{\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}} \tag{2.8}$$

which shows that each $\hat{\beta}_j$ is a known linear combination of the responses $y_1, \ldots, y_n$. We will show later that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

Remarks on Computation:

(1) The solution of (2.5) requires that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, which is the case if $\operatorname{rank}(\boldsymbol{X}^T\boldsymbol{X}) = p$. Since $\operatorname{rank}(\boldsymbol{X}^T\boldsymbol{X}) = \operatorname{rank}(\boldsymbol{X})$ ([R], Section 14.3, Lemma A), the normal equations have a formal solution if and only if the columns of $\boldsymbol{X}$ are linearly independent. Also, a necessary condition is in any case $n \geq p$.

(2) In practice, modern software never uses (2.8) for the computation of the LS estimator $\hat{\boldsymbol{\beta}}$, but uses decompositions of $\boldsymbol{X}^T\boldsymbol{X}$ (Cholesky decomposition [MC], Sec 5.5.1) or $\boldsymbol{X}$ (such as the QR decomposition [R], Sec 14.8) which allow for more efficient computation. The R software uses by default the QR decomposition.

Properties of $\boldsymbol{X}^T\boldsymbol{X}$, residuals and fitted values

- The matrix is symmetric: $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$.
- The matrix is positive semi-definite: for $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{a} \neq 0$ we have $\boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{a})^T(\boldsymbol{X}\boldsymbol{a}) = \sum_j b_j^2 \geq 0$, where $(b_1, \ldots, b_n)^T = \boldsymbol{b} = \boldsymbol{X}\boldsymbol{a}$.

The second property can be potentially used to verify that $\hat{\boldsymbol{\beta}}$ is indeed the vector which minimizes the objective function $R(\boldsymbol{\beta})$. Specifically we have to look at the matrix of the second partial derivatives (the *Hessian*). We have that

$$\frac{\partial R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^T}(-2\boldsymbol{Y}^T\boldsymbol{X} + 2\boldsymbol{X}^T\boldsymbol{X}) = 2\boldsymbol{X}^T\boldsymbol{X}.$$

This matrix is positive semi-definite (since $\boldsymbol{X}^T\boldsymbol{X}$ is positive semi-definite) and, therefore, its eigenvalues are all non-negative. We would need all eigenvalues to be strictly positive to conclude that we have a minimum. Without going into further detail one can show that $\hat{\boldsymbol{\beta}}$ is indeed a minimum. One can also reach this conclusion in a much simpler way; $R(\boldsymbol{\beta})$ is a convex function, therefore, the critical point is necessarily a minimum. So, $R(\boldsymbol{\beta})$ *is* minimized at $\hat{\boldsymbol{\beta}}$ with minimum

$$R(\hat{\boldsymbol{\beta}}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}}.$$

The *known* vector $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ of *"residuals"* $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ estimates the *unknown* vector of random errors ϵ_i; and the *known* vector

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \equiv \boldsymbol{H}\boldsymbol{Y}$$

of *"fitted values"* $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ estimates the *unknown* vector $\boldsymbol{X}\boldsymbol{\beta}$ of the (conditional) expectation of y_i: $E[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta}$ $(i = 1, \ldots, n)$. The *hat matrix* $\boldsymbol{H}$ will be considered in more detail later in this chapter.

Obviously, $y_i = \hat{y}_i + \hat{\epsilon}_i$, or in vector form, $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \hat{\boldsymbol{\epsilon}}$ (*observed= fitted plus residual*). The fitted values and residuals play an important role in prediction and model diagnostics.

Relationship with the ML estimator

The LS estimator of $\boldsymbol{\beta}$ is completely non-parametric in the sense that it does not require any of the assumtions listed in Table 2.2. If we were to use ML estimation on the other hand, we would require all of the assumptions in Table 2.2. However, the resulting estimate would be the same (see Q2.3 in sheet 2) and given by equation (2.8)! It should be noted that in general LS and ML estimation do not result in the same solutions.

For positive semi-definite matrices the second partial derivative test is generally inconclusive (meaning we may have a minimum or a saddle point).

2.2.2 Estimation of σ^2

Note that because $\mathrm{E}[\epsilon_i^2] = \mathrm{Var}[\epsilon_i] + (\mathrm{E}[\epsilon_i])^2 = \sigma^2 + 0^2 = \sigma^2$, it follows that

$$E\left[\frac{\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}}{n}\right] = \frac{1}{n}E\left[\sum_i \epsilon_i^2\right] = \frac{1}{n}\sum_i E\left[\epsilon_i^2\right] = \sigma^2. \tag{2.9}$$

But $\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}/n$ is unknown since the values of the random error ϵ_i are unknown. From (2.9) we could say that a natural candidate estimator is $n^{-1}\sum_i \hat{\epsilon}_i^2$. In fact, one can show that: (i) this is the ML estimator and (ii) this estimator is biased. A "bias-corrected" estimate of σ^2, based on the known residuals $\hat{\epsilon}_i$, is

$$\boxed{s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}{n-p} = \frac{\sum_1^n \hat{\epsilon}_i^2}{n-p} = \frac{R(\hat{\boldsymbol{\beta}})}{n-p} = \frac{\mathrm{RSS}}{n-p}} \tag{2.10}$$

where, as is customary, RSS denotes $R(\hat{\boldsymbol{\beta}})$ and is called the *"Residual Sum of Squares"* with $n-p$ degrees-of-freedom. In other words, s is an estimate of the magnitude of a "typical" random error ϵ. In R, s is referred to as "residual standard error". We will see later how dividing by $n-p$ rather than n makes s^2 an *unbiased* estimator of σ^2.

There is a related geometrical reason we divide by $n-p$ in equation (2.10); namely, that there are p constraints on the residuals so they only have $n-p$ *degrees-of-freedom*, not n, as is the case with the *unconstrained* unknown random errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ which can be anywhere in $\mathbb{R}^n$. The p constraints are given by

$$\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{0}.$$

This translates to

$$\underbrace{\begin{pmatrix} x_{11} & \ldots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \ldots & x_{np} \end{pmatrix}}_{\boldsymbol{X}^T} \underbrace{\begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix}}_{\hat{\boldsymbol{\epsilon}}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Multiplying the first row of $\boldsymbol{X}^T$ with $\hat{\boldsymbol{\epsilon}}$ yields $\sum_{i=1}^n x_{i1}\hat{\epsilon}_i = 0$. This means that as soon as the first $n-1$ residuals are known the last residual $\hat{\epsilon}_n$ is also known. In total, we have p such constraints, $\sum_{i=1}^n x_{ij}\hat{\epsilon}_i = 0$, for $j = 1, \ldots, p$, so $\hat{\boldsymbol{\epsilon}}$ has $n-p$ degrees of freedom. Also, note that if there is an intercept in the model, then $\sum_{i=1}^n \hat{\epsilon}_i = 0$; that is, the residuals sum to *zero*, because $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{n1})^T = (1, \ldots, 1)^T$. We will see later that these constraints are used as a basis for model checking.

Example 2.4 *Measurement model:*

The regression model with $p = 1$,

$$y_i = \mu + \epsilon_i, \quad (i = 1, \ldots, n)$$

is often referred to as the "measurement model", as μ may e.g. represent the value of an unknown physical constant which is to be estimated from $y_1, \ldots, y_n$. One has

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad , \quad \boldsymbol{\beta} = \mu$$

Hence, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = n$, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = y_1 + \cdots + y_n$ so that $n\hat{\mu} = y_1 + \cdots + y_n$; that is, $\hat{\mu} = \overline{y}$, which is just the *sample mean.* Also, $\hat{y}_i = \overline{y}$, $\hat{\epsilon}_i = y_i - \overline{y}$. Thus, $\mathrm{RSS} = \sum(y_i - \overline{y})^2$ and $\mathrm{df} = n-1$, and therefore the variance of the measurement errors is estimated as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2,$$

which corresponds to the well-known expression for *sample standard deviation.* This example highlights that common and simple statistical estimation problems can be expressed as special cases of linear regression and that is also the case for more complicated problems.

Example 2.5 (Continuation of Example 2.2)

We use the cement data to illustrate model parameter estimation using R.

```
#install.packages("remotes") #you need to do that once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> ?cement
```

Regard the row labels as case numbers i. The number of cases is $n = 21$. If we want to fit a simple linear regression model relating log(strength) to 1/time, as suggested in Example 2.2, we can use the R function `lm`: see its help file. To fit the model to the cement data and obtain estimates, residuals, fitted values and model matrix, we do as follows:

```
> cement.lm <- lm(log(strength) ~ I(1/time), data=cement)
> cement.lm

 Call:
lm(formula = log(strength) ~ I(1/time), data = cement)

Coefficients:
(Intercept)    I(1/time)
      3.688       -1.146

> coeffs <- cement.lm$coefficients   # LEAST SQUARES ESTIMATES OF INTERCEPT AND SLOPE
> coeffs
(Intercept)   I(1/time)
   3.687818   -1.145528
> cement.lm$residuals      # RESIDUALS FROM FITTED MODEL
          1           2           3           4           5           6
 0.02265927  0.04547395 -0.07419055 -0.02856765  0.08361883  0.09174896
.......................................................................
         19          20          21
 0.04944472 -0.07175606 -0.02791343
> cement.lm$fitted.values    # FITTED VALUES
       1        2        3        4        5        6        7        8
2.542290 2.542290 2.542290 3.115054 3.115054 3.115054 3.305976 3.305976
.......................................................................
      17       18       19       20       21
3.646907 3.646907 3.646907 3.646907 3.646907

> X<- model.matrix(cement.lm)
> X                      # DESIGN MATRIX
   (Intercept)  I(1/time)
1            1 1.00000000
2            1 1.00000000
3            1 1.00000000
4            1 0.50000000
5            1 0.50000000
6            1 0.50000000
7            1 0.33333333
8            1 0.33333333
.........................
.........................
20           1 0.03571429
21           1 0.03571429
```

Notice that the envelope `I()` is needed because we are employing a function of the original covariate `time`. If a covariate is directly used as predictor, then this envelope can be omitted.

As seen above, coefficients $\hat{\boldsymbol{\beta}}$, the residuals $\hat{\boldsymbol{\epsilon}}$, and the fitted values $\hat{\boldsymbol{Y}}$ are available from the *object* `cement.lm` for further processing. For example, we can obtain the residual sum of squares RSS, its degrees-of-freedom df and the residual standard error s as follows:

```
> RSS <- sum(cement.lm$residuals^2)
> RSS
[1] 0.1085086
> df <- cement.lm$df
> df
[1] 19
> s <- sqrt(RSS/df)
> s               # ESTIMATE OF SIGMA
[1] 0.07557102

# quicker, extract s directly from lm object:
> summary(cement.lm)$sigma
[1] 0.07557102

# Create Figure 3.3
plot(X[,2], log(cement$strength))
abline(coeffs)
```

2.3 Statistical properties of $\hat{\boldsymbol{\beta}}$ and s^2

2.3.1 Expectation and variance of $\hat{\boldsymbol{\beta}}$

Using assumption (A1), which implies $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$, one has

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = \mathrm{E}[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}] = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}[\boldsymbol{Y}] = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Under the variance assumption (A2), one gets

$$\begin{aligned}\mathrm{Var}[\hat{\boldsymbol{\beta}}] =& \mathrm{Var}[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}] = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{Var}[\boldsymbol{Y}]\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right]^{\mathrm{T}}\\ =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left[\sigma^2\boldsymbol{I}_n\right]\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2\end{aligned}$$

Example 2.6

(i) For the measurement model $y_i = \mu + \epsilon_i$, $i = 1, \dots, n$ (Example 2.4), we have

$$\begin{aligned}\hat{\mu} &= \bar{y}\\ \mathrm{Var}(\hat{\mu}) &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2/n\end{aligned}$$

(ii) For the simple linear regression model $y_i = a + bx_i + \epsilon_i$, $i = 1, \dots, n$, one has

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix}$$

with $\hat{a}$, $\hat{b}$ as in Example 2.3; that is

$$\hat{a} = \frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2 - n\bar{x}^2},$$

$$\hat{b} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

and

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2\begin{pmatrix} n & \sum x_i\\ \sum x_i & \sum x_i^2\end{pmatrix}^{-1}\\ &= \sigma^2\frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{pmatrix} \text{ or} \qquad (2.11)\\ &= \frac{\sigma^2}{n}\frac{1}{\sum x_i^2 - n\bar{\boldsymbol{x}}^2}\begin{pmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{pmatrix} \text{ or} \qquad (2.12)\\ &= \frac{\sigma^2}{n}\frac{1}{(n-1)\frac{\sum(x_i - \bar{\boldsymbol{x}})^2}{n-1}}\begin{pmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{pmatrix}\\ &= \frac{\sigma^2}{n}\frac{1}{(n-1)S_X^2}\begin{pmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{pmatrix}. \qquad (2.13)\end{aligned}$$

Note that the entries of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ tend to become small for $n \longrightarrow \infty$, meaning that the precision of estimator $\hat{\boldsymbol{\beta}}$ increases with sample size.

2.3.2 Variance of linear combinations of $\hat{\boldsymbol{\beta}}$

Generally, we are interested in a specified linear combination $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$; for example, for $\boldsymbol{c}^{\mathrm{T}} = (1, 0, \ldots, 0)$ we get $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta} = \beta_1$. Another important example is when we wish to draw inferences about $\mathrm{E}[y \,|\, \boldsymbol{x}_0] = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}$ at some (possibly previously unseen) value $\boldsymbol{x}_0 \in \mathbb{R}^p$ of the predictors. We estimate $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$ by the obvious unbiased estimator $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Clearly, one has $\mathrm{E}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}] = \boldsymbol{c}^{T}\boldsymbol{\beta}$, and because

$$\mathrm{Var}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}] = \sigma^2\, \boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}$$

the *standard deviation* of $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{SD}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}] = \sigma \sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}} \tag{2.14}$$

The standard deviation (SD) of $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ as a measure of the precision of the estimate $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is only useful if the value of σ is *known*. When it is not known we replace it by its estimate s and one obtains the *standard error* of $\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$

$$\mathrm{SE}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}] = s \sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}} \tag{2.15}$$

In the special case $\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots 0)^T$, with the 1 at $j-$th position, one gets for $j = 1, \ldots, p$

$$\mathrm{E}[\hat{\beta}_j] = \beta_j \qquad \mathrm{SD}[\hat{\beta}_j] = \sigma\sqrt{\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right]_{jj}} \tag{2.16}$$

and

$$\boxed{\mathrm{SE}[\hat{\beta}_j] = s\sqrt{\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right]_{jj}} \qquad (2.17)}$$

for the *standard error* (SE) of $\hat{\beta}_j$, the estimate of $\mathrm{SD}[\hat{\beta}_j]$.

Example 2.7 (Continuation of Example 2.5). R produces convenient summary objects of fitted linear model objects. For instance, for the cement data,

```
> summary(cement.lm)
```

produces a considerable amount of information (most of which will only be relevant later), which can either be read from the summary display or directly accessed from its components

```
> names(summary(cement.lm))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

For now, we are interested in the coefficient table which can be extracted via

```
> summary(cement.lm)$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  3.687818 0.02425278 152.05758 8.782730e-31
I(1/time)   -1.145528 0.05290007 -21.65457 7.472821e-15
```

Here, the column `Estimate` contains the $\hat{\beta}_j$, and the column `Std.Error` contains the $SE(\hat{\beta}_j)$. Let us try to verify these standard errors. According to our theory, these standard errors should be the same as

```
>  s * sqrt(diag(solve(t(X)%*%X)))
 (Intercept)   I(1/time)
 0.02425278  0.05290007
```

Note that we have made use of the residual standard error s as well as the design matrix $\boldsymbol{X}$ as obtained in Example 2.2. A slightly more convenient way to do this calculation is to make use of the summary component `cov.unscaled`, which is just $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$:

```
> XTXinv <- summary(cement.lm)$cov.unscaled
> s * sqrt(diag(XTXinv))
 (Intercept)  I(1/time)
  0.02425278  0.05290007
```

2.3.3 Expectation of s^2

To find the expectation of s^2 we need assumptions (A1) and (A2) to hold. Also, we will make use of three properties for traces and of three preliminary results.

The properties are the following:

(M1) $\mathrm{Tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{A}) + \mathrm{Tr}(\boldsymbol{B})$

(M2) $\mathrm{Tr}(\boldsymbol{AB}) = \mathrm{Tr}(\boldsymbol{BA})$ for symmetric $\boldsymbol{AB}$ and $\boldsymbol{BA}$

(M3) $\mathrm{Tr}(\boldsymbol{b}\boldsymbol{b}^T) = \boldsymbol{b}^T\boldsymbol{b}$

Next, the results that we will require.

(R1) We have that $\mathrm{E}[\hat{\boldsymbol{\epsilon}}] \overset{(A1)}{=} \mathrm{E}[\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] = \boldsymbol{X\beta} - \boldsymbol{X\beta} = \boldsymbol{0}$. So from this we obtain the first result

$$\mathrm{Var}[\hat{\boldsymbol{\epsilon}}] = \mathrm{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T] - \underbrace{\mathrm{E}[\hat{\boldsymbol{\epsilon}}]\mathrm{E}[\hat{\boldsymbol{\epsilon}}]^T}_{\boldsymbol{0}} = \mathrm{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T].$$

(R2) Now consider the hat matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. It is easy to show that $\boldsymbol{H}^T = \boldsymbol{H}$. Thus, we have that $\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{H}^2 = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{H}$. From this we obtain the second result

$$\begin{aligned}(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})^T &= \boldsymbol{I}_n - \boldsymbol{H}^T - \boldsymbol{H} + \boldsymbol{H}\boldsymbol{H}^T \\ &= \boldsymbol{I}_n - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H} \\ &= \boldsymbol{I}_n - \boldsymbol{H}.\end{aligned}$$

(R3) Finally, the third result is the following:

$$\mathrm{Tr}\left(\boldsymbol{H}\right) = \mathrm{Tr}\left(\left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right]\boldsymbol{X}^T\right) \overset{(M2)}{=} \mathrm{Tr}\left(\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) = \mathrm{Tr}\left(\boldsymbol{I}_p\right) = p$$

Now we can find the expectation of s^2. Recall from equation (2.10) that $s^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}/(n-p)$, so for simplicity we will start with $\mathrm{E}[\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}]$.

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}] &\overset{(M3)}{=} \mathrm{E}[\mathrm{Tr}\left(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T\right)] = \mathrm{Tr}\left(\mathrm{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T]\right) \overset{(R1)}{=} \mathrm{Tr}\left(\mathrm{Var}[\hat{\boldsymbol{\epsilon}}]\right) = \mathrm{Tr}\left(\mathrm{Var}[\boldsymbol{Y} - \hat{\boldsymbol{Y}}]\right) = \mathrm{Tr}\left(\mathrm{Var}[\boldsymbol{Y} - \boldsymbol{HY}]\right) \\ &= \mathrm{Tr}\left(\mathrm{Var}[(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]\right) = \mathrm{Tr}\left((\boldsymbol{I}_n - \boldsymbol{H})\underbrace{\mathrm{Var}[\boldsymbol{Y}]}_{\sigma^2\boldsymbol{I}_n\ \text{(A2)}}(\boldsymbol{I}_n - \boldsymbol{H})^T\right) = \sigma^2\mathrm{Tr}\left((\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})^T\right) \\ &\overset{(R2)}{=} \sigma^2\mathrm{Tr}\left((\boldsymbol{I}_n - \boldsymbol{H})\right) \overset{(M1)}{=} \sigma^2\left(\underbrace{\mathrm{Tr}(\boldsymbol{I}_n)}_{n} - \mathrm{Tr}(\boldsymbol{H})\right) \\ &\overset{(R3)}{=} \sigma^2(n-p).\end{aligned}$$

As a result we have that σ^2 is an unbiased estimator since

$$\mathrm{E}[s^2] = \frac{\mathrm{E}[\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}]}{n-p} = \sigma^2.$$

Chapter 3

Linear models: Inference & Prediction

3.1 Inference for linear model parameters

In this section we assume throughout assumptions (A1) to (A3), i.e. we further assume normality. We give some results without proof; some we prove.

3.1.1 Sampling distribution

The sampling distribution is the probability distribution of an estimator, when drawing repeatedly samples of the same size from the population. In our context this has to be understood as follows: For a given (fixed) set of predictor values $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n \in \mathbb{R}^p$, assume one repeatedly collects a set of responses $y_1, \ldots, y_n$. Each set of responses leads to a certain estimate $\hat{\boldsymbol{\beta}}$, and we are looking at the distribution of all those estimates. The sampling distribution is of crucial importance for deriving confidence intervals, critical values for hypothesis tests, etc.

Using the results from Subsection 2.3.1 and Equation (1.6), the sampling distribution of $\hat{\boldsymbol{\beta}}$ is given by

$$\boxed{\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \qquad (3.1)}$$

and that of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ is

$$\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta},\, \sigma^2\, \boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}\right). \tag{3.2}$$

Equations (3.1) and (3.2) are most times not particularly useful in practice because they involve the error variance σ^2 which is usually unknown. For this reason we will proceed to show that

$$(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}, \tag{3.3}$$

which will prove to be a very useful result. First let us re-express equation (2.10) as

$$s^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

so that

$$\begin{aligned} \frac{1}{\sigma^2}(n-p)s^2 &= \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &\overset{\text{see } Q2.4}{=} \underbrace{\frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}_{A} - \underbrace{\frac{1}{\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{B}. \end{aligned}$$

Upon careful examination both A and B are sum of squares of standard normal variates and, therefore, follow chi-square distributions. Specifically,

$$\begin{aligned} A &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y}])^T\mathrm{Var}[\boldsymbol{Y}]^{-1}(\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y}]) \sim \chi^2_n \end{aligned}$$

and

$$\begin{aligned} B &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &= (\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}])^T \mathrm{Var}[\hat{\boldsymbol{\beta}}]^{-1} (\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}]) \sim \chi^2_p. \end{aligned}$$

Thus, from the properties of the chi-square distribution we have that $A - B \sim \chi^2_{n-p}$ leading to the result in equation (3.3). Note that based on this result (under the further assumption (A3)), one can show that s^2 is an unbiased estimator in an alternative and simpler way (in comparison to the proof presented previously in Subsection 2.3.3); namely, we have that

$$\mathrm{E}[(n-p)s^2/\sigma^2] = \mathrm{E}[\chi^2_{n-p}] = n - p \Rightarrow \mathrm{E}[s^2] = \sigma^2.$$

Now, combining (3.2) and (3.3), one can arrive at a very useful result. We make use of the fact that s^2 and $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ are independent (not proven here). Hence,

$$\frac{\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}}{\mathrm{SE}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}]} = \frac{\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}}{s\sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}}} = \frac{\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}}{\underbrace{\sigma\sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}}}_{\mathrm{SD}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}]}} \frac{1}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} = t_{n-p} \quad (3.4)$$

is a Student-t distribution with $n - p$ degrees-of-freedom. In particular,

$$\boxed{\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}[\hat{\beta}_j]} \sim t_{n-p} \qquad (3.5)}$$

Example 3.1 As an illustrative toy example for (3.1), we simulate 500 data sets of sample size 50 from the true function

$$f(x) = 2x - 1$$

The predictors $x_1, \ldots, x_{50}$ are initially drawn from a uniform distribution on $[0, 1]$, but are then kept constant during the simulation process. The response is simulated as $y_i = f(x_i) + \epsilon_i$, where the $\epsilon_i, i = 1, \ldots, 50$ are drawn, in each of the 500 runs anew, from a normal distribution with $\sigma = 0.1$. One of the 500 simulated data sets is exemplarily shown in Figure 3.1 (top left).

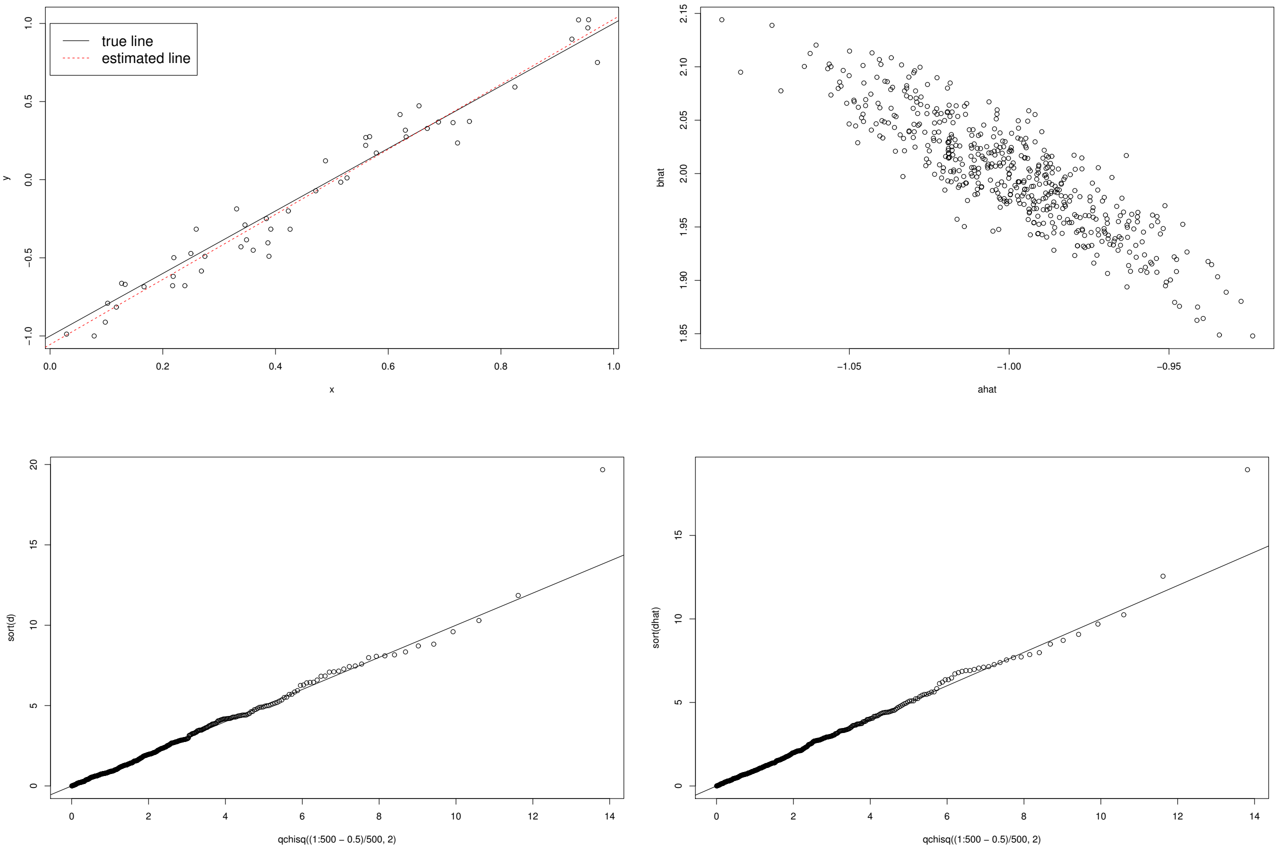
For each of the 500 runs, a linear model of type $y = a + bx + \epsilon$ is fitted to the data. The 500 estimates, say $\hat{\boldsymbol{\beta}}_j = (\hat{a}_j, \hat{b}_j)^T$, $j = 1, \ldots, 500$, of a and b are recorded. A scatterplot of all 500 estimates is provided in Figure 3.1 (top right). Do these form a bivariate normal distribution, as would be expected by result (3.1)? We check for bivariate normality through the χ^2 probability plots developed in Section 1.3. Construction of the QQ plots requires computation of Mahalanobis distances $d^2_M\left(\hat{\boldsymbol{\beta}}_j, E(\hat{\boldsymbol{\beta}}), \mathrm{Var}(\hat{\boldsymbol{\beta}})\right)$, $j = 1, \ldots, 500$. We consider two scenarios: on the left side in Figure 3.1 (bottom) we use the 'true' values $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} = (-1, 2)^T$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, with $\sigma = 0.1$, while the right plot uses estimates of mean and variance obtained directly from the scatterplot (see R code below). We see that the large bulk of the data follows very closely the straight line. Note that these are 500 observations, so a few deviating points at the upper end are tolerable! We also observe that the two plots look quite similar.

Overall, these plots do not give evidence that BVN is violated (this is, of course, as it should be!). The accompanying R code is provided below.

```
 # Define design and 'true' function
> x <- runif(50,0,1)
> fx<-  -1 + 2*x

 # Example data set
> y   <- fx+ rnorm(50,0,0.1)
> fit <- lm(y~x)
> plot(x,y)
```

Figure 3.1: Top left: Example data set simulated from the function $f(x) = 2x - 1$; right top: Plotted values of $\hat{b}$ versus $\hat{a}$ after 500 runs; bottom: QQ plot using true (left) and estimated (right) population parameters.



```
> abline(a=-1,b=2)
> abline(a=fit$coef[1], b=fit$coef[2], col=2, lty=2)
> leg.names<- c("true line", "estimated line")
> legend(0,1, leg.names, lty=c(1,2), col=c(1,2))

 # Simulation
> ahat <- rep(0,500)
> bhat <- rep(0,500)
> for (j in 1:500){
    y<- fx+ rnorm(50,0,0.1)
    fit<- lm(y~x)
    ahat[j] <- fit$coef[1]
    bhat[j] <- fit$coef[2]
}

> hat     <- cbind(ahat, bhat)
> plot(hat)

 # Is this a BVN? (If so, the squared Mahalanobis distances should be chi^2 with 2df)

 # Firstly, use ’true’ values of beta and Sigma
> beta<- c(-1,2)
> var <- 0.1^2 *summary(fit)$cov.unscaled # this is sigma^2* (X^TX)^{-1}
> d<- mahalanobis(hat, beta, var)
> plot(qchisq( (1:500-0.5)/500,2),sort(d))
> abline(a=0,b=1)

 # Secondly, estimate mean and variance from scatterplot
> betahat<- colMeans(hat)
> varhat <- var(hat)
> dhat <- mahalanobis(hat, betahat, varhat)
> plot(qchisq( (1:500-0.5)/500,2),sort(dhat))
> abline(a=0,b=1)
```

3.1.2 Significance tests

We reject a *null hypothesis* H_0 that $\beta_j = \beta_j^0$ at prescribed *significance level* $\alpha\,(0 < \alpha < 1)$ if

$$\boxed{T \equiv \left|\frac{\hat\beta_j - \beta_j^0}{\mathrm{SE}[\hat\beta_j]}\right| > t_{n-p,\alpha/2} \qquad (3.6)}$$

In particular, we reject $\beta_j = 0$ at level α if $\left|\hat\beta_j/\mathrm{SE}[\hat\beta_j]\right| > t_{n-p,\alpha/2}$. The hypothesis $H_0 : \beta_j = 0$ is the most common hypothesis to test in the regression setting, because it provides evidence whether covariate X_j has a statistically significant effect on Y.

Most statistical software packages do not carry out actual hypothesis tests, but compute, for an observed value of T, the *observed significance level* p^* ("p-value") as the probability

$$p^* = P(|t| \geq T), \tag{3.7}$$

where $t \sim t_{n-p}$ This can be seen as the probability of obtaining "by chance" (considering H_0 is true) a value of t at least as extreme as the observed one T. The smaller p^* is, the greater is the evidence (from the data) against the null hypothesis.

Note at this occasion, in R summary output, the column `t-value` contains the values $\hat\beta_j/\mathrm{SE}[\hat\beta_j]$ used for the test $H_0 : \beta_j = 0$, and the column `Pr(>|t|)` contains the corresponding $p-$values.

Example 3.2 (Continuation of Examples 2.5 and 2.6)

Test $H_0 : \beta_2 = -1$ vs $H_1 : \beta_2 \neq -1$, at level $\alpha = 0.05$. We use from Example 2.5 that $\hat\beta_2 = -1.146$ and $s = 0.0756$. Of course, the value $\mathrm{SE}(\hat\beta_2)$ could be read from the output of Example 2.7, but assuming that we have to do this manually, one finds from Example 2.6 (b)

$$\mathrm{SE}(\hat\beta_2) = s\frac{1}{\sqrt{\sum x_i^2 - n\bar x^2}} = ... = 0.0529.$$

(Note that $x_i = 1/\mathrm{time}_i$, so $\sum x_i^2 = 4.4140$ and $\bar x = 0.33617$.) Then

$$T = \left|\frac{-1.146+1}{0.0529}\right| = 2.760 > t_{21-2,0.025} = 2.093,$$

so we reject H_0 at the 5% level of significance.

3.1.3 Confidence intervals

From (3.4) we have

$$\Pr\left(\left|\frac{\boldsymbol{c}^{\mathrm T}\boldsymbol\beta - \boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}}{\mathrm{SE}[\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}]}\right| \leq t_{n-p,\alpha/2}\right) = 1-\alpha \iff$$

$$\Pr\left(-\mathrm{SE}[\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}]t_{n-p,\alpha/2} \leq \boldsymbol{c}^{\mathrm T}\boldsymbol\beta - \boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta} \leq \mathrm{SE}[\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}]t_{n-p,\alpha/2}\right) = 1-\alpha \iff$$

$$\Pr\left(\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta} - \mathrm{SE}[\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}]t_{n-p,\alpha/2} \leq \boldsymbol{c}^{\mathrm T}\boldsymbol\beta \leq \boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta} + \mathrm{SE}[\boldsymbol{c}^{\mathrm T}\hat{\boldsymbol\beta}]t_{n-p,\alpha/2}\right) = 1-\alpha.$$

Thus, a $100(1-\alpha)\%$ confidence interval (CI) for $\boldsymbol{c}^T\boldsymbol{\beta}$ has limits

$$\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \times \mathrm{SE}[\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}].$$

In particular, for $\boldsymbol{c} = (0 \dots 0\ 1\ 0 \dots 0)^T$, with 1 at $j-$th position, one gets

$$\boxed{\hat{\beta}_j \pm t_{n-p,\alpha/2} \times \mathrm{SE}[\hat{\beta}_j] \qquad (3.8)}$$

Interpretation of CIs

A confidence interval is the interval which covers with $(1-\alpha)$ probability the true (unknown) value of $\boldsymbol{\beta}_j$. This means that under repeated sampling we *expect* $100(1-\alpha)\%$ of the samples to produce confidence intervals which include the true value of $\boldsymbol{\beta}_j$. Note that this is not a probability statement for $\boldsymbol{\beta}_j$, which is considered a *fixed* yet unknown quantity.

Relationship between CIs and significance tests

Of course, there is a direct relationship between confidence intervals and significance tests. Specifically, we have that

$$\begin{aligned} \text{The } (1-\alpha) \text{ CI for } \hat{\beta}_j \text{ does not contain } \beta_j^0 &\iff \text{reject } H_0 : \beta_j = \beta_j^0 \text{ at significance level } \alpha \\ &\iff T > t_{n-p,\alpha/2} \\ &\iff p^* < \alpha, \end{aligned}$$

and *vice versa* when the CI contains β_j^0.

Example 3.3 (Continuation of Example 2.7)
R standard model output does not provide confidence intervals for parameters. We can do this ourselves:

```
>   s<- summary(cement.lm)$sigma
>   XTXinv <-  summary(cement.lm)$cov.unscaled
>   SE1<- s * sqrt(XTXinv[1,1]) # SE(\hat{\beta_1})
>   SE2<- s * sqrt(XTXinv[2,2]) # SE(\hat{\beta_2})
>   cement.lm$coef[1]  + c(-1,1) * qt(0.975,19)* SE1
[1] 3.637057 3.738580
>   cement.lm$coef[2] + c(-1,1) * qt(0.975,19)* SE2
[1] -1.256250 -1.034807
```

We observe that the latter interval does not contain the value -1, in conformity with the result from Example 3.2. Considerably simpler, we can do the same by invoking the built-in function `confint`:

```
>  confint(cement.lm,1,level=0.95)
               2.5 %  97.5 %
(Intercept) 3.637057 3.73858
>  confint(cement.lm,2,level=0.95)
             2.5 %    97.5 %
I(1/time) -1.25625 -1.034807
```

3.2 Prediction

Suppose we are contemplating a further observation $y_0 \equiv y|\boldsymbol{x}_0$ at predictor value $\boldsymbol{x}_0 = (x_{01} \dots x_{0p})^T$. The vector $\boldsymbol{x}_0$ may be either one of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ or a new value: fortunately, we do not need to distinguish between the two cases, as the subsequent theory is the same.

However, what we do need to distinguish carefully between are the following problems:

(i) Estimating the "population mean" $\mathrm{E}[y \,|\, \boldsymbol{x}_0] = \boldsymbol{x}_0^{\mathrm{T}} \boldsymbol{\beta}$.

(ii) Predicting the value $y_0 \equiv y|\boldsymbol{x}_0$.

Problem (i) has been covered: the estimate is $\widehat{\mathrm{E}[y \,|\, \boldsymbol{x}_0]} = \boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}}$ with standard error given by $s\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}$, and a $100(1-\alpha)\%$ *confidence interval* for $\mathrm{E}[y \,|\, \boldsymbol{x}_0]$ has limits

$$\boxed{\boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}} \pm\ t_{n-p,\alpha/2} \times \mathrm{SE}[\boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}} \pm\ t_{n-p,\alpha/2} \times s\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}}$$

Problem (ii) is about prediction: we are trying to predict $y_0 = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_0$ and the question in this case is what would be a reasonable estimate of y_0. A first rational observation is that the estimator should be of the from $y_0^* = \hat{y}_0 + \epsilon_0$ in order to account for both the systematic part $(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta})$ and the random part (ϵ_0) of the model. A natural candidate for $\hat{y}_0$ is, of course, the fitted value; i.e., $\hat{y}_0 = \widehat{\mathrm{E}[y \,|\, \boldsymbol{x}_0]}$, Given assumption (A1) we also know that $\mathrm{E}[\epsilon_0] = 0$, so 0 is a reasonable estimate of ϵ_0. Hence, our best estimate of y_0 must be $y_0^* = \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} + 0 = \hat{y}_0$, which is the same as the estimated population mean. However, things get different when looking at the variance. The variance of our prediction is

$$\mathrm{Var}[y_0^*] = \mathrm{Var}[\hat{y}_0 + \epsilon_0] = \mathrm{Var}[\hat{y}_0] + \mathrm{Var}[\epsilon_0] = \sigma^2 \boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \sigma^2,$$

so that the variance of the prediction $\hat{y}_0$ is larger than the variance of the estimator $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Note that $\hat{y}_0$ and ϵ_0 are independent so that their covariance is 0. One can further show that this is also the variance of the distance between y_0 (which is unknown) and $\hat{y}_0$; specifically,

$$\mathrm{Var}(y_0 - \hat{y}_0) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \sigma^2 \boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0,$$

where again $\hat{y}_0$, which is estimated based on the i.i.d observations $y_1, \dots, y_n$, is independent of y_0. Thus, the standard error of prediction (in short: *prediction error*) is

$$s\sqrt{1 + \boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

A so-called $100(1-\alpha)\%$ *prediction interval* for y_0 has limits

$$\boxed{\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \times s\sqrt{1 + \boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}}$$

Note carefully that the estimates in both problems (i) and (ii) are the *same*, but the confidence interval is *narrower* than the prediction interval. In fact, if n is "large", the estimate of $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is well-determined with "small" standard error, so that the prediction error is approximately s, as we might anticipate (because $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \approx \boldsymbol{0}$ for large n). In this case, for example, a 95% prediction interval has approximate limits

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p,0.025}\, s \approx \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm z_{0.025}\, s \approx \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm 2\, s,$$

since the t-distribution can be approximated by a normal distribution for a large number of degrees of freedom.

Finally, note, for predictions to be valid, we invoke the same assumptions about the distribution of $y|\boldsymbol{x}_0$ as for the previous n cases: y_0 is generated from the same linear model as $y_1, \dots, y_n$, and ϵ_0 and $\epsilon_1, \dots, \epsilon_n$ are independent $\mathrm{N}(0, \sigma^2)$ errors.

Example 3.4 *Measurement model:* (Continuation of Example 2.4)

There are no covariates and $y = \mu + \epsilon$, so we want a confidence interval for $\mathrm{E}[y] = \mu$ and a prediction interval for a further observation $y_0 = \mu + \epsilon_0$. Here, $p = 1$, $\hat{\mu} = \overline{y}$, $x_0 = 1$, $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = 1/n$ with $\mathrm{Var}[\hat{\mu}] = \sigma^2/n$, and so $\mathrm{SE}[\hat{\mu}] = s/\sqrt{n}$, where $s^2 = \sum(y_i - \overline{y})^2/(n-1)$. Hence, the confidence interval for μ has limits

$$\overline{y} \pm\ t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$$

and the corresponding prediction interval for y_0 has limits

$$\overline{y} \pm\ t_{n-1,\alpha/2} \times s\sqrt{1 + \frac{1}{n}} \simeq \overline{y} \pm\ z_{\alpha/2} \times s \qquad [\text{if}\, n\, \text{is ``large''}].$$

Example 3.5 (Continuation of Example 3.3)

Consider the tensile strength of cement at a given cure time with model

$$\mathrm{E}[\log(\text{strength}) \,|\, \text{time}] = \beta_1 + \beta_2/\text{time}$$

The following shows how both confidence intervals and prediction intervals can be computed with or without using the R function `predict`. Suppose we are interested in predicting tensile strength after a curing time of $t_0 = 10$ days.

```
>  cement.lm # least squares fit of model

 # predict  y=log(strength)  at time = t0 = 10 days
> t0    <- 10
> y0hat <-   as.numeric(cement.lm$coef %*% c(1,1/t0))
    # as.numeric(.) is required to transform the 1x1 matrix into a scalar
    # (would cause warning message at later stage otherwise)
> y0hat
  [1] 3.573266

 # CI for expected response:
> SE0 <- as.numeric(s* sqrt( c(1, 1/t0)%*%XTXinv %*% c(1,1/t0)))
> CI  <- y0hat + c(-1,1)* qt(0.975,19)* SE0

# PI for true response value:
> PE0 <-  as.numeric(s *sqrt(1+ c(1, 1/t0)%*%XTXinv %*% c(1,1/t0)))
> PI  <- y0hat +c(-1,1)* qt(0.975,19)* PE0


 # now, on original strength scale:
> exp(y0hat)
  [1] 35.63277
>  exp(CI)
  [1] 34.12271 37.20965
>  exp(PI)
  [1] 30.24326 41.98271

 # or, using the R function predict:
> t0 <- data.frame(time = 10)
> Ci <- predict(cement.lm, newdata = t0, interval = "confidence")
> Pi <- predict(cement.lm, newdata = t0, interval = "prediction")

 # back to original strength scale:
> exp(Ci)
> exp(Pi)
   # same as above

 # Finally, note that several predictions can be made in the same call,
 # and that the level of the interval can be changed. For instance,
> t0 <- data.frame(time=c(10,20,30))
> exp(predict(cement.lm, newdata = t0, interval = "prediction", level=0.99))
   # gives 99%PIs for all three times on original strength scale:
   #        fit      lwr      upr
   # 1 35.63277 28.47727 44.58624
   # 2 37.73326 30.11587 47.27738
   # 3 38.46059 30.68108 48.21268
```

Chapter 4

Factors

So far, we have mainly considered continuous covariates. Categorical covariates are called *factors.*

Example 4.1 The data below come from an experiment reported in Beall (1942), *Transformation of data from entomological field experiments*, Biometrika. Six different insect sprays were each applied to 12 plots and in each case the number of tobacco hornworms found in the plot is given.

		Number of insects											
	A	10	7	20	14	14	12	10	23	17	20	14	13
	B	11	17	21	11	16	14	17	17	19	21	7	13
Spray	C	0	1	7	2	3	1	2	1	3	0	1	4
	D	3	5	12	6	4	3	5	5	5	5	2	4
	E	3	5	3	5	3	6	1	1	3	2	6	4
	F	11	9	15	22	15	16	13	10	26	26	24	13

Spray is a *factor*, with *levels* $A, \ldots, F$ and $r = 12$ replicates per factor level.

4.1 Coding

For inclusion into a linear model, factors need to be coded: For a factor $\mathcal{A}$ with levels $1, \ldots, a$,

$$x_j^{\mathcal{A}} = 1_{\{\mathcal{A}=j\}},$$

i.e. an indicator taking the value 1 if the $j-$th factor level is attained, and 0 otherwise. In form of a coding matrix,

Spray-Ex.	$\mathcal{A}$	$x_1^{\mathcal{A}}$	$x_2^{\mathcal{A}}$	$\ldots$	$x_{a-1}^{\mathcal{A}}$	$x_a^{\mathcal{A}}$
A	1	1		$\ldots$		
B	2		1	$\ldots$		
$\vdots$	$\vdots$			$\ddots$		
E	$a-1$			$\ldots$	1	
F	a			$\ldots$		1

This type of coding is called *dummy-coding.* Other codings are possible as long as they enable a unique identification of factor levels. Including all a indicators into the LM, one gets the *unconstrained* model

$$E(y|\mathcal{A}) = \beta_0 + \beta_1 x_1^{\mathcal{A}} + \beta_2 x_2^{\mathcal{A}} + \ldots + \beta_{a-1} x_{a-1}^{\mathcal{A}} + \beta_a x_a^{\mathcal{A}}$$

i.e.

$$E(y|\mathcal{A} = j) = \beta_0 + \beta_j \qquad (j = 1, \ldots, a)$$

or

$$y_{jk} = \beta_0 + \beta_j + \epsilon_{jk} \tag{4.1}$$

for the $k-$th replicate at level j. In what follows, denote the number of replicates at level j by r_j. (In Example 4.1, $r_j \equiv r = 12$ for $j = 1, \dots, 6$). In matrix notation, (4.1) takes the form

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1r_1} \\ y_{21} \\ \vdots \\ y_{2r_2} \\ \vdots \\ \vdots \\ y_{a1} \\ \vdots \\ y_{ar_a} \end{pmatrix} = \left(\begin{array}{c:c:cccc} 1 & 1 & & & & \\ 1 & \vdots & & & & \\ \vdots & 1 & & & & \\ \vdots & & 1 & & & \\ \vdots & & \vdots & & & \\ \vdots & & 1 & & & \\ \vdots & & & \ddots & & \\ \vdots & & & & & 1 \\ \vdots & & & & & \vdots \\ 1 & & & & & 1 \end{array}\right) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{a-1} \\ \beta_a \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1r_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2r_2} \\ \vdots \\ \vdots \\ \epsilon_{a1} \\ \vdots \\ \epsilon_{ar_a} \end{pmatrix}$$

The design matrix $\boldsymbol{X}$ is an $n \times (a+1)$ matrix. Clearly, its columns are not linear independent as the sum over the latter a columns gives the first column. Hence, $\text{Rank}(\boldsymbol{X}) = a$. As $\text{Rank}(\boldsymbol{X}^T\boldsymbol{X}) = \text{Rank}(\boldsymbol{X})$, one has $\text{Rank}(\boldsymbol{X}^T\boldsymbol{X}) = a$ as well, so that $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{a+1 \times a+1}$ is not invertible. Thus, the unconstrained model is not feasible in practice. This is why we need *constraints* on the parameters. There is no unique way how to do this, but popular constraints are:

- The most common solution is a zero-constraint; e.g., to set $\beta_1 = 0$ (R does this too). In this case, level 1 takes the role of a *reference category.* Thus, the second column of the design matrix (corresponding to the first column of the coding matrix) can be cut off, as indicated through vertical dashed lines. This gives the *constrained model*

  $$E(y|\mathcal{A}) = \beta_0 + \beta_2 x_2^{\mathcal{A}} + \dots + \beta_a x_a^{\mathcal{A}}, \tag{4.2}$$

  i.e.

  $$E(y|\mathcal{A} = 1) = \beta_0$$

  and

  $$E(y|\mathcal{A} = j) = \beta_0 + \beta_j \qquad \text{for}\, j = 2, \dots a.$$

  This makes clear that the intercept represents the expected response for the reference category, and the parameters estimated for the other categories give the level effect relative to this reference category.

- Equivalently, one could set any other β_j, $j = 2, \dots, a$, equal to 0, in which case the $(j+1)$th column of the design matrix gets removed.

- A different type of constraint is the zero-sum constraint: $\sum_{j=1}^{a} \beta_j = 0$ (which arises naturally under *effect coding*).

Example 4.2 (Continuation of Example 4.1)
We firstly enter the data in a vector.

```
insects0 <-
+ c(
+   10,7,20,14,14,12,10,23,17,20,14,13,
+   11,17,21,11,16,14,17,17,19,21,7,13,
+   0,1,7,2,3,1,2,1,3,0,1,4,
+   3,5,12,6,4,3,5,5,5,5,2,4,
+   3,5,3,5,3,6,1,1,3,2,6,4,
+   11,9,15,22,15,16,13,10,26,26,24,13
+ )
```

For the inclusion into a linear model, we need one row for each observation:

```
> insects<- data.frame("spray"= c(rep("A",12), rep("B",12), rep("C",12),rep("D",12),
+   rep("E",12), rep("F",12)), "insects" = insects0)
> insects
   spray insects
1      A      10
2      A       7
3      A      20
4      A      14
5      A      14
6      A      12
7      A      10
8      A      23
9      A      17
10     A      20
11     A      14
12     A      13
13     B      11
14     B      17
15     B      21
.
.
.
68     F      10
69     F      26
70     F      26
71     F      24
72     F      13
```

Fortunately, we don't need to do the coding ourselves: R does this for us via the simple command `as.factor()`:

```
> insects$spray <- as.factor(insects$spray)
> insects$spray
 [1] A A A A A A A A A A A A B B B B B B B B B B B B C C C C C C C C C C C C D D
[39] D D D D D D D D D D E E E E E E E E E E E E F F F F F F F F F F F F
Levels: A B C D E F
```

Then we fit a linear model to the relationship between spray and insects, and display the design matrix and the fitted coefficients:

```
> fit <- lm(insects ~ spray, data= insects)
> model.matrix(fit)
...
> fit$coef
 (Intercept)       sprayB       sprayC       sprayD       sprayE       sprayF
    14.5000       0.8333     -12.4167      -9.5833     -11.0000       2.1667
```

Interpretation: For the reference group A we expect 14.5 insects per plot. For group B, we expect $14.5 + 0.8333$ insects per plot, and so on.

4.2 Experiments

In an *experiment*, the experimenter must identify at least one factor which (s)he manipulates, and at least one response variable to measure. Each combination of factor levels that an experimental unit receives is called a *treatment*. The goal of experiments is usually to provide stong evidence of cause-and-effect relationships, i.e. to decide whether (a substantial part of) the variation of the response can indeed be attributed to the different levels of the factor. For such conclusions to be valid, certain principles must be followed:

The three principles of experimental design:

(i) *Control:* The experimenter sets the values of the factors and tries to eliminate any other source of variation.

(ii) *Randomize:* Treatments have to be allocated at random, by the experimenter, to the individuals (i.e. individuals should not choose their treatment themselves, but also the experimenter should not allocate it just as he thinks best). This will

- Reduce selection bias
- Avoid confounding (i.e. reduce detrimental effects of minor violations of (i))

(iii) *Replicate.* Make sure to collect more than one observation for each administered treatment. One cannot conclude anything based on one observation!

Example 4.3

- The insects data are an example for a "randomized complete block design", where each two treatments from A to F are randomized within blocks of 12 neighboring plots.
- The cement data (Example 2.2) were also obtained from a designed experiment (the discrete hardening times play the role of the "factor", though we have not used it as a factor in the subsequent data analysis)

Experimental design finds application in many scientific fields. A prominent example in medicine, for instance, are *randomized control trials.* These are clinical trials designed to investigate the efficacy of new experimental drugs. In this setting carefully stratified random samples of individuals are randomly assigned either to the control group (which usually receives a *placebo* dose) or to the treatment group (which receives the experimental drug). Typically, these are *blinded* studies; that is, the individuals do not know whether they belong to the control group or to the treatment group. In such settings, we have $\mathrm{E}[y|\text{placebo}] = \beta_0$ and $\mathrm{E}[y|\text{drug}] = \beta_0 + \beta_1$ for an appropriate response variable y which accurately reflects the efficacy of the drug under consideration for the particular condition that needs treatment. Obviously, here the interest lies on testing hypotheses of the form $H_0 : \beta_1 \leq t$ *vs.* $H_1 : \beta_1 > t$ for some critical threshold value t (rejecting H_0 in this case implies that the drug is effective). Significance levels for this type of testing are far lower than the usual 5% level.

However, it is not possible to carry out all studies as experiments. Consider, for instance, a model as in Question 2.2 where we have

$$y_i = \mu + \delta x_i + \epsilon_i.$$

Here $x_i \in \{0, 1\}$ is a binary indicator/grouping variable. Now let us assume that the grouping variable distinguishes between non-smokers ($x_i = 0$) and smokers ($x_i = 1$) and that the response y_i is blood pressure. Obviously, we cannot allocate the treatment "smoking" at random, we can just observe the data, which are then likely to confound the treatment effect with other effects. For instance, smoking may be associated with other factors like, for instance, socio-economic status, which in turn influences other factors (e.g. dietary habits) that may affect the response of interest. Due to these confounding effects the resulting estimate $\hat{\delta}$ may be seriously inflated/deflated.

Investigative studies which do not follow the three principles above are called *observational studies.* Taken strictly, they cannot be used to infer about cause-and-effect relationships (though this is attempted all the time).

4.3 Factorial experiments

Experiments involving two or more factors are called *factorial experiments.* Certain technical terms in the language of factorial experiments are introduced and illustrated in the context of the example.

Example 4.4 (Animal survival times)

The data shown in Table 4.1 are the survival times (unit: 10 hours) of 48 animals (*experimental units*) with four animals randomly allocated to each of the 12 possible combinations of 3 *poisons* and 4 *antidote treatments.* The experiment was part of an investigation to combat the effects of certain toxic agents. Random allocation is intended to reduce bias by "scrambling up" any residual variation in the response (such as possibly resulting from unavoidable initial weight differences in the animals) not caused directly by the effects of the 12 combinations of poisons and antidote treatment. In the language of statistical experimental design, each combination of poison and antidote treatment is called a *treatment*, and the observations corresponding to each treatment form a *cell.* Thus, there are 12 treatments here corresponding to the 12 combinations of poison and antidote treatment. This arrangement is called a 3 × 4 *factorial design*, which is *complete* since no cell is empty and is *balanced* as we have the same number of *replicates* (4) per cell. Poisons and antidote treatments are *factors.* The factor *poison* has 3 *levels* and the factor *antidote treatment* has 4 *levels.*

	antidote treatment			
poison	A	B	C	D
I	0·31	0·82	0·43	0·45
	0·45	1·10	0·45	0·71
	0·46	0·88	0·63	0·66
	0·43	0·72	0·76	0·62
II	0·36	0·92	0·44	0·56
	0·29	0·61	0·35	1·02
	0·40	0·49	0·31	0·71
	0·23	1·24	0·40	0·38
III	0·22	0·30	0·23	0·30
	0·21	0·37	0·25	0·36
	0·18	0·38	0·24	0·31
	0·23	0·29	0·22	0·33

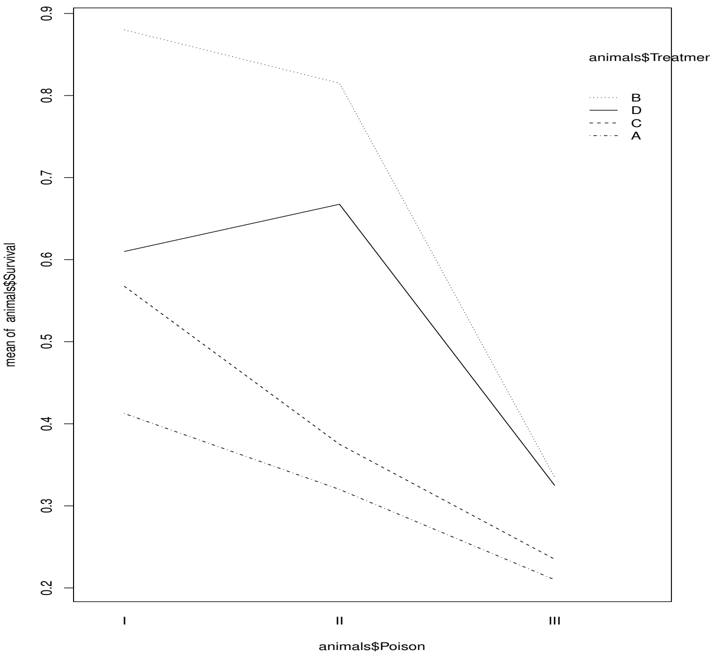


Table 4.1: Left: survival times (unit: 10 hours) of 48 animals comprising 4 replications of the 3 × 4 combinations of 3 poisons and 4 antidote treatments; right: interaction plot.

The data are stored in an R dataframe called `animals`, using one line for each observation, as necessary for the inclusion into a linear model. Code to read and display the data, as well as to produce the

There is an unfortunate double use of the word treatment here as it appears again in the name of one of the two factors, the *antidote treatment.*

interaction plot depicted in Table 4.1(right), is given below. Production of the "interaction plot"does not require fitting any model, it is just a visualization of the cell means.

```
#> install.packages("remotes") # you need to do this once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> data(animals)
> animals
    Poison Treatment Survival
 1       I         A     0.31
 2      II         A     0.36
 3     III         A     0.22
 4       I         B     0.82
 5      II         B     0.92
 6     III         B     0.30
 7       I         C     0.43
 8      II         C     0.44
 9     III         C     0.23
 10      I         D     0.45
 11     II         D     0.56
 12    III         D     0.30
 ...
 47     II         D     0.38
 48    III         D     0.33

> interaction.plot(animals$Poison,animals$Treatment, animals$Survival)
```

A linear model (without interaction) for a response y in a complete factorial design replicated r times with two factors, $\mathcal{A}$ with a levels and $\mathcal{B}$ with b levels, can be written as

$$y_{jk\ell} = \mu + \tau_j^{\mathcal{A}} + \tau_k^{\mathcal{B}} + \epsilon_{jk\ell} \tag{4.3}$$

with components

$\tau_j^{\mathcal{A}}$: *main effect* of level j of factor $\mathcal{A}$ ($j = 1, \dots, a$)

$\tau_k^{\mathcal{B}}$: *main effect* of level k of factor $\mathcal{B}$ ($k = 1, \dots, b$)

$\epsilon_{jk\ell}$: error term of replicate ℓ for the j, k factor combination ($\ell = 1, \dots, r$).

Similar as in Section 4.1, one finds that the unconstrained version of this model leads to a design matrix with $a + b + 1$ columns but rank

$$1 + (a - 1) + (b - 1) = a + b - 1,$$

resulting in the necessity to impose one constraint on each of the two parameter collections $\tau_1^{\mathcal{A}}, \dots, \tau_a^{\mathcal{A}}$ and $\tau_1^{\mathcal{B}}, \dots, \tau_b^{\mathcal{B}}$. The default constraints in R are $\tau_1^{\mathcal{A}} = \tau_1^{\mathcal{B}} = 0$.

Model (4.3) assumes that the two factors impact additively onto the response. If this is not the case, they are said to *interact*; i.e. the effect of the levels of factor $\mathcal{A}$ onto the expected response varies with the levels of $\mathcal{B}$, and vice versa. A full *interaction* model can be written in the form,

$$y_{jk\ell} = \mu + \tau_j^{\mathcal{A}} + \tau_k^{\mathcal{B}} + \tau_{jk}^{\mathcal{AB}} + \epsilon_{jk\ell} \tag{4.4}$$

where $\tau_j^{\mathcal{A}}$, $\tau_k^{\mathcal{B}}$, and $\epsilon_{jk\ell}$ are defined as above, and

$\tau_{jk}^{\mathcal{AB}}$: *interaction effect* of level j of factor $\mathcal{A}$ with level k of factor B.

The number of parameters in the unconstrained model appears to be $1 + a + b + ab = (a + 1)(b + 1)$. As there is no sense in having more than one parameter per cell, we will need appropriate constraints

For the animal survival data, $a = 3, b = 4, r = 4$.

For the animal survival data, $1 + (a - 1) + (b - 1) = 1 + 2 + 3 = 6$

which reduce this to effectively $p = ab$ parameters. One requires $1 + a + b$ constraints to restore the design matrix to full rank. The default constraints in R are

$$\tau_1^{\mathcal{A}} = \tau_1^{\mathcal{B}} = \tau_{1k}^{\mathcal{AB}} = \tau_{j1}^{\mathcal{AB}} = 0$$

for $j = 1, \ldots, a$ and $k = 1, \ldots, b$.

Even then $\boldsymbol{\beta}$ is a "big" vector compared with $n = abr$, and the residual degrees-of-freedom turn out to be $n - ab = ab(r-1)$, so we must have $r > 1$; otherwise, with just *one* replicate ($r = 1$) there are no degrees-of-freedom left for estimating σ^2— this means a "perfect fit" with all of the residuals equal to *zero*.

In analogy to the notation used in (4.2), we can re-express interaction model in (4.4) as

$$\mathrm{E}[y \,|\, \mathcal{A}, \mathcal{B}] = \mu + \sum_{j=2}^{a} \tau_j^{\mathcal{A}} x_j^{\mathcal{A}} + \sum_{k=2}^{b} \tau_k^{\mathcal{B}} x_k^{\mathcal{B}} + \sum_{j=2}^{a}\sum_{k=2}^{b} \tau_{jk}^{\mathcal{AB}} x_j^{\mathcal{A}} x_k^{\mathcal{B}} \tag{4.5}$$

where the $(a-1) + (b-1)$ predictors $x_j^{\mathcal{A}},\ j = 2, \ldots, a$ and $x_k^{\mathcal{B}},\ k = 2, \ldots b$ are such that $x_j^{\mathcal{A}} \in \{0, 1\}$ (absence or presence of level j of factor $\mathcal{A}$) with $\sum_{j=2}^{a} x_j^{\mathcal{A}} = 0$ for the observations belonging to the cell of the 1st reference-treatment level and $\sum_{j=2}^{a} x_j^{\mathcal{A}} = 1$ otherwise (similarly for factor $\mathcal{B}$).

If the terms $\tau_{jk}^{\mathcal{AB}}$ are all zero then the following statements are equivalent:

1. There is defined to be no interaction between $\mathcal{A}$ and $\mathcal{B}$.
2. The effects $\mathcal{A}$ and $\mathcal{B}$ are additive.
3. The difference in expected response between any two levels of $\mathcal{A}$ is the same for all levels of $\mathcal{B}$.
4. The difference in expected response between any two levels of $\mathcal{B}$ is the same for all levels of $\mathcal{A}$.

Of course, in data there will virtually always be nonzero estimates of the above quantities, so we must assess whether or not they are significantly different from zero.

Example 4.5 (Continuation of Example 4.4)

This is an edited R output to fit a model with main effects for poison and treatment and *no* interaction to the animals survival data:

```
> animals.addfit <- lm(Survival ~ Poison + Treatment, data = animals)
> summary(animals.addfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45229    0.05592   8.088 4.22e-10 ***
PoisonII    -0.07313    0.05592  -1.308  0.19813
PoisonIII   -0.34125    0.05592  -6.102 2.83e-07 ***
TreatmentB   0.36250    0.06458   5.614 1.43e-06 ***
TreatmentC   0.07833    0.06458   1.213  0.23189
TreatmentD   0.22000    0.06458   3.407  0.00146 **
```

The first 12 rows of the design matrix in R for the animal survival data for the model in (4.3) is shown below. In the full 48×6 model matrix, the first 12 rows are repeated 3 more times, corresponding to the 4 replicates of the 3×4 design.

```
> model.matrix(animals.addfit)
   (Intercept) PoisonII PoisonIII TreatmentB TreatmentC TreatmentD
1            1        0         0          0          0          0
2            1        1         0          0          0          0
3            1        0         1          0          0          0
```

```
4            1        0         0          1          0          0
5            1        1         0          1          0          0
6            1        0         1          1          0          0
7            1        0         0          0          1          0
8            1        1         0          0          1          0
9            1        0         1          0          1          0
10           1        0         0          0          0          1
11           1        1         0          0          0          1
12           1        0         1          0          0          1
.
.
.
```

The following is an edited R output to fit a model with main effects for poison, treatment *and* their interaction to the animals survival data:

```
> animals.interfit <- lm(Survival ~ Poison + Treatment + Poison:Treatment,
  data = animals)
> summary(animals.interfit)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.41250    0.07457   5.532 2.94e-06 ***
PoisonII             -0.09250    0.10546  -0.877   0.3862
PoisonIII            -0.20250    0.10546  -1.920   0.0628 .
TreatmentB            0.46750    0.10546   4.433 8.37e-05 ***
TreatmentC            0.15500    0.10546   1.470   0.1503
TreatmentD            0.19750    0.10546   1.873   0.0692 .
PoisonII:TreatmentB   0.02750    0.14914   0.184   0.8547
PoisonIII:TreatmentB -0.34250    0.14914  -2.297   0.0276 *
PoisonII:TreatmentC  -0.10000    0.14914  -0.671   0.5068
PoisonIII:TreatmentC -0.13000    0.14914  -0.872   0.3892
PoisonII:TreatmentD   0.15000    0.14914   1.006   0.3212
PoisonIII:TreatmentD -0.08250    0.14914  -0.553   0.5836

# Interpretation:
 # Expected response at Poison=III and Treatment=A:
> predict(animals.interfit, newdata=data.frame( Poison="III", Treatment="A"))
    # this is  0.41250-0.20250

 # Expected response at Poison=II and Treatment=B:
 > predict(animals.interfit, newdata=data.frame(Poison="II", Treatment="B"))
    # this is 0.41250 - 0.09250 + 0.46750 + 0.02750
```

Note, however, that the $p-$values give very little evidence to support the inclusion of the interaction terms, so the simpler additive model appears to be more adequate from this analysis.

Chapter 5

Analysis of variance

For the LM $E(y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ with intercept, with predictors $\boldsymbol{x} = (1, x_2, \ldots, x_p)$ under (A1-A3), we want to learn about the size of the contribution of different sources of variation (predictors, error) towards the total variation in the response y, and draw from this conclusions on the relative importance of these sources.

5.1 Explaining variation

The idea is to partition the total variation in the response

$$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

into two components

- the proportion of variation that is explained by the regression line (curve, hyperplane, etc). We shall call this component SSR, the *sum of squares for regression*, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$,
- the residual variation, i.e. proportion of variation that is unexplained. We shall call this SSE, *sum of squares for error*, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad [\ =\ RSS]$.

Indeed, it holds

$$SST = SSR + SSE. \tag{5.1}$$

This can be expressed in words as

$$\begin{aligned}&[\text{variation in } y \text{ ignoring } \boldsymbol{x}] = \\ &\quad = \ [\text{variation in } y \text{ due to regression on } \boldsymbol{x}] + [\text{residual variation in } y \text{ after regression on } \boldsymbol{x}]\end{aligned} \tag{5.2}$$

How can we use this decomposition in variation to assess whether or not the predictors $\boldsymbol{x} = (x_1, \ldots, x_p)$ are important in explaining the variation in y? The obvious approach would be to consider the ratio

$$R^2 \equiv \frac{SSR}{SST} = \frac{\text{variation in } y \text{ due to regression on } \boldsymbol{x}}{\text{variation in } y \text{ ignoring } \boldsymbol{x}}$$

which is the proportion of variation in y explained by $\boldsymbol{x}$. Equivalently, by virtue of (5.1), we could consider the ratio

$$\frac{SSR}{SSE} = \frac{\text{variation in } y \text{ due to regression on } \boldsymbol{x}}{\text{residual variation in } y \text{ after regression on } \boldsymbol{x}}$$

"large" values of the ratio indicating that $\boldsymbol{x}$ explains a substantial proportion of the original variation in y (when $\boldsymbol{x}$ is ignored). We do not know the sampling distribution of either of them, but a slightly modified version of the latter does the job:

5.1.1 The ANOVA table

It turns out that the following, so-called F–ratio, which is a slight adjustment to the previous ratio, obtained by dividing the numerator and denominator by their respective degrees-of-freedom, is the ratio to consider:

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{[\text{variation in } y \text{ due to regression on } \boldsymbol{x}]/(p-1)}{[\text{residual variation in } y \text{ after regression on } \boldsymbol{x}]/(n-p)}. \tag{5.3}$$

Provided our usual assumptions (A1-A3) hold, the sampling distribution of F under the *null hypothesis* that variation in y is not explained by $\boldsymbol{x}$ (equivalently, all β_j except the intercept are zero) can be shown to be an *F-distribution with numerator degrees-of-freedom $p-1$ and denominator degrees-of-freedom* $n-p$ (More detail on the F-distribution is provided in Table 5.1). Thus, “large” values of F, as evidence against the hypothesis, can be calibrated using the p-value

$$p^* = \mathrm{P}\left[F_{p-1,n-p} \geq F\right] \tag{5.4}$$

where the observed value F is given by (5.3). The smaller the value of p^*, the greater is the evidence against the *null hypothesis* that $E[y|\boldsymbol{x}] = \alpha$ when contrasted with the *alternative hypothesis* that $E[y|\boldsymbol{x}] = \boldsymbol{x}^T\boldsymbol{\beta}$. This procedure is known as the *overall F–test.*

An important, alternative way of displaying the F-ratio is

$$F = \frac{MSR}{MSE} \tag{5.5}$$

where MSR = SSR/$(p-1)$ and MSE = SSE/$(n-p)$ are called the “Mean Square Regression” and “Mean Square Error”, respectively. Note that MSE = s^2, the estimate of the error variance σ^2 in the full alternative model, while $SST/(n-1)$ is the estimate of σ^2 in the reduced null (“measurement”) model.

All this information is usually summarized in a ANOVA (ANalysis Of VAriance) table:

Source	df	SS	MS	F	p^*
Regression on $\boldsymbol{x}$	$p-1$	SSR	MSR	MSR/MSE	$\mathrm{P}\left[F_{p-1,n-p} > MSR/MSE\right]$
Error	$n-p$	SSE	MSE		
Total	$n-1$	SST			

Note that the last row of the table (“Total”) is not really necessary, and could be omitted without loss of information.

Example 5.1 *(Continuation of Example 4.5)*

Consider additive model for animals data:

```
> SSR <-  sum( (animals.addfit$fitted- mean(animals$Survival))^2)
> SSR
[1] 1.954219
> SSE  <- sum( (animals$Survival- animals.addfit$fitted)^2)
> SSE
[1] 1.050863
> F<-( SSR/(6-1) )/( SSE/(48-6))
[1] 15.62092
# p-value:
> 1-pf(15.62092, 5,42)
[1] 1.122701e-08
# Check with linear model summary:
> summary(animals.addfit)
# F-statistic: 15.62 on 5 and 42 DF,  p-value: 1.123e-08
```

Summarizing in the ANOVA table,

Source	df	SS	MS	F	p
Regression	5	1.954	0.3908	15.621	$1.12 \cdot 10^{-8}$
Error	42	1.051	0.0250		

We conclude that the predictors, as a whole, contribute significantly (and substantially) towards the variation in the response.

Table 5.1: A little bit more detail about the F-distribution.

- Let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$ be independent. Then the ratio

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \tag{5.6}$$

  is said to have an F-distribution with numerator degrees-of-freedom ν_1 and denominator degrees-of-freedom ν_2: we write $F \sim F_{\nu_1,\nu_2}$.

- Hence, for the F-test in (5.4) to be valid, we would need to establish that SSR $\sim \chi^2_{p-1}$, SSE $\sim \chi^2_{n-p}$ and that the two sums of squares are independent under the null hypothesis. To prove this is a little beyond the main aim of this course.

- For reference, the density of an F-distribution for any point $y > 0$ is

$$f(y|\nu_1, \nu_2) = K(\nu_1, \nu_2)\,\frac{y^{(\nu_1/2)-1}}{(\nu_2 + \nu_1 y)^{(\nu_2+\nu_1)/2}} \tag{5.7}$$

  and *zero* otherwise, and $K(\nu_1, \nu_2)$ normalises the p.d.f. to integrate to 1.

- It is worth noting that if $F \sim F_{\nu_1,\nu_2}$, then $\mathrm{E}[F] = \nu_2/(\nu_2 - 2)$. Hence, if the null hypothesis is true, we expect an F-ratio in the ANOVA table to be (slightly) in excess of 1 when ν_2 is not too small.

- The `R` functions `rf`, `pf`, `qf` and `df` have their usual meanings for the F-distribution.

- As usual, we denote the quantile corresponding to the probability mass α in the *right* tail of F_{ν_1,ν_2} by $F_{\nu_1,\nu_2,\alpha}$.

5.1.2 Sequential ANOVA

Can we decompose the 'Regression' component of the ANOVA table (as formulated in Section 5.1.1) further? That is, consider any sequence of m "nested" models $M_1 \subset \ldots \subset M_m$ with design matrices $\boldsymbol{X}_1 = \mathbf{1}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_j, \ldots, \boldsymbol{X}_m$, where $\boldsymbol{X}_j$ is $n \times p_j$ and $\boldsymbol{X}_{j+1}$ is obtained by adding $p_{j+1} - p_j$ columns to $\boldsymbol{X}_j$. In different but equivalent terms, model M_j is nested in M_{j+1} if they can be written as

$$M_j : \quad \mathrm{E}[y \,|\, \boldsymbol{x}] = \beta_1 + \beta_2 x_2 + \cdots + \beta_{p_j} x_{p_j} \tag{5.8}$$

$$M_{j+1} : \quad \mathrm{E}[y \,|\, \boldsymbol{x}] = \beta_1 + \beta_2 x_2 + \cdots + \beta_{p_j} x_{p_j} + \ldots + \beta_{p_{j+1}} x_{p_{j+1}} \tag{5.9}$$

where $p_{j+1} > p_j$, and $x_1, \ldots, x_{p_j}$ are the same in both models. Note carefully, there is *no* requirement that the values of $\beta_1, \ldots, \beta_{p_j}$ should be the same in the full and reduced models. To make this plain, we could have used (say) $\alpha_1, \ldots, \alpha_{p_j}$ instead of $\beta_1, \ldots, \beta_{p_j}$ to represent the coefficients in the second model, but we choose not to do this. Both models assume homogeneity of error variance, but again the value σ^2 in both models can be different. In what follows, we do not consider non–nested models.

The idea is now to assess how residual variation changes as we *add* successively columns to a design matrix; that is, *increase* the number of terms in the model for $\mathrm{E}[y \,|\, \boldsymbol{x}]$.

The corresponding residual sums of squares are such that $S_1 \geq S_2 \geq \cdots \geq S_m$ and

$$S_j - S_{j+1} \text{ is the "sum of squares explained by the extra columns" in } \boldsymbol{X}_{j+1}$$

with degrees-of-freedom $p_{j+1} - p_j$; equivalently, by the extra $p_{j+1} - p_j$ model terms. We can decompose S_1 as

$$S_1 = (S_1 - S_2) + (S_2 - S_3) + \cdots + (S_j - S_{j+1}) + \cdots + (S_{m-1} - S_m) + S_m$$

with corresponding degrees-of-freedom decomposition

$$n - p_1 = (p_2 - p_1) + (p_3 - p_2) + \cdots + (p_{j+1} - p_j) + \cdots + (p_m - p_{m-1}) + (n - p_m)$$

The *Sequential Analysis of Variance* table becomes

Source	df	SS	MS	F
$x_{p_1+1}, \ldots, x_{p_2}$	$p_2 - p_1$	$S_1 - S_2$	$(S_1 - S_2)/(p_2 - p_1)$	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{p_j+1}, \ldots, x_{p_{j+1}}$	$p_{j+1} - p_j$	$S_j - S_{j+1}$	$(S_j - S_{j+1})/(p_{j+1} - p_j)$	F_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{p_{m-1}+1}, \ldots, x_{p_m}$	$p_m - p_{m-1}$	$S_{m-1} - S_m$	$(S_{m-1} - S_m)/(p_m - p_{m-1})$	F_{m-1}
Error	$n - p_m$	S_m	$S_m/(n - p_m)$	
Total	$n - p_1$	S_1		

where under the null hypothesis $\mathrm{H}_0 : \beta_{p_j+1} = \cdots = \beta_{p_{j+1}} = 0$

$$F_j = \frac{(S_j - S_{j+1}) \,/\, (p_{j+1} - p_j)}{S_m / \,(n - p_m)} \sim F_{p_{j+1} - p_j, n - p_m} \qquad j = 1, \ldots, m-1 \tag{5.10}$$

and the corresponding $p-$value is obtained as usual.

With $\boldsymbol{X}_1$ being a column of 1's corresponding to an intercept term, there remain $(m-1)!$ different sequential ANOVA's depending on the order of the inclusion of the other columns. In general, all these ANOVA's will be different; that is, the values of SS, F, etc. for the same source of variation will be different if the order of inclusion of terms is altered. However, in the special case of a balanced factorial design, the ANOVA table does *not* depend on the order of inclusion of terms. In this case, we say that the sources of variation are *orthogonal* to each other.

Example 5.2 *(Student height/weight data)*
Personal characteristics were recorded for students taking Statistics in 1993/94. Suppose $y = \text{weight}, x_1 = 1, x_2 = \text{height}, x_3 = \text{gender}$ (with gender = 0 for male and gender = 1 for female). We determine the "full" model as

$$\mathrm{E}[y \mid \text{height}, \text{gender}] = \beta_1 + \beta_2 \,\text{height} + \beta_3 \,\text{gender} + \beta_4 \,\text{height} \cdot \text{gender}. \tag{5.11}$$

The model implies that we believe that weight is linearly related to height for both sexes, that is the straight line for men is

$$\mathrm{E}[y \mid \text{height}, \text{gender} = 0] = \beta_1 + \beta_2 \text{height} \tag{5.12}$$

and that for women is

$$\mathrm{E}[y \mid \text{height}, \text{gender} = 1] = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\text{height}. \tag{5.13}$$

Thus, β_3 is the difference in intercepts and β_4 is the difference in slopes. Hence, (i) $\beta_3 = 0$ corresponds to a common intercept β_1, (ii) $\beta_4 = 0$ corresponds to a common slope β_2, and (iii) $\beta_3 = \beta_4 = 0$ corresponds to a common straight line.

In order to produce a sequential ANOVA table, we first fit just the intercept, then *height* (a single straight line), then *height and gender* (two straight lines with the same slope but different intercepts) and finally *height and gender and height* × *gender* (two straight lines with different slopes and intercepts).

```
#install.packages("remotes") #you need to do that once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> ?student
> head(student)
   gender height stone lb   weight
 1      0     74    13  0 82.55381
 2      0     73    13 12 87.99692
 3      0     73     9  4 58.96701
 4      0     72    12  0 76.20352
 5      0     72    10  7 66.67808
 6      1     66     8  3 52.16312

> S1 <- sum(lm(weight ~ 1, data = student)$residuals^2)
[1] 6202.927
> S2 <- sum(lm(weight ~ 1 + height, data = student)$residuals^2)
[1] 1896.556
> S3 <- sum(lm(weight ~ 1 + height+gender, data = student)$residuals^2)
[1] 1868.036
> S4 <- sum(lm(weight ~ 1 + height+gender+height:gender, data = student)$residuals^2)
[1] 1850.411
```

Let S_j be the RSS of the $j-$th considered model and p_j be the number of parameters in that model. Hence, $S_1 = 6202.9$, $p_1 = 1$, $S_2 = 1896.6$, $p_2 = 2$, $S_3 = 1868.0$, $p_3 = 3$, $S_4 = 1850.4$, $p_4 = 4$, so that $S_1 - S_2 = 4306.3$ is the sum of squares for height, $S_2 - S_3 = 28.6$ is the sum of squares for gender, $S_3 - S_4 = 17.6$ is the sum of squares for height:gender, and $S_4 = 1850.4$ is the residual sum of squares. The necessary R code to carry out sequential ANOVA by hand is

```
> MSE <- S4/(43-4)
> MSE
[1] 47.44643
 # F-values:
> ((S1-S2)/1)/MSE
[1] 90.7628
```

```
> ((S2-S3)/1)/MSE
[1] 0.6011152
> ((S3-S4)/1)/MSE
[1] 0.3714676
 # compare each of these with quantile
> qf(0.95, 1, 39)
[1] 4.091279
```

which can be obtained much quicker using the function `anova()`,

```
 > anova(lm(weight ~ height*gender, data = student))
Analysis of Variance Table

Response: weight
              Df Sum Sq Mean Sq F value    Pr(>F)
height         1 4306.4  4306.4 90.7628 9.911e-12 ***
gender         1   28.5    28.5  0.6011    0.4428
height:gender  1   17.6    17.6  0.3715    0.5457
Residuals     39 1850.4    47.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[Note that the row denoted in R by `Residuals` is the same as what we call ‘Error’].

The table below shows how the estimates of the coefficients $\beta_1, \beta_2, \beta_3$ and β_4 change as we increase the number of terms in the model.

Intercept	height	gender	height:gender
66.49			
−103.211	2.458		
−87.339	2.242	−2.451	
−97.6607	2.3863	32.0842	−0.5149

When we fit `gender` before `height`, that is, fit the terms one-by-one from left-to-right in the model

$$\mathrm{E}[y \mid \text{gender}, \text{height}] = \beta_1 + \beta_2 \, \text{gender} + \beta_3 \, \text{height} + \beta_4 \, \text{gender} \cdot \text{height}$$

the ANOVA table below seems to tell a different story—two lines with the *same* slope but with *different* intercepts are required. The result is not surprising, as we expect a gender difference when we fit gender alone (ignoring height) at the first fit.

```
> anova(lm(weight ~ gender + height + gender:height, data = student))
Analysis of Variance Table

Response: weight
              Df  Sum Sq Mean Sq F value    Pr(>F)
gender         1 2679.56 2679.56 56.4755 4.251e-09 ***
height         1 1655.33 1655.33 34.8884 6.981e-07 ***
gender:height  1   17.62   17.62  0.3715    0.5457
Residuals     39 1850.41   47.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The table below shows how the estimates of the coefficients change as we increase the number of terms in the model in which `gender` enters before `height`, corresponding to the sequential ANOVA table in the `R` output above.

Intercept	gender	height	gender:height
66.49			
72.87	-16.15		
-87.339	-2.451	2.242	
-97.6607	32.0842	2.3863	$-$ 0.5149

Note, in particular, that the parameter estimates for the full model (last row) do not depend on the order of inclusion.

Important special cases:

- If $m = 2$, with M_1 corresponding to the measurement model, and M_2 to any other linear model with $p_2 \geq 2$, then the sequential ANOVA table constructed above corresponds to the ANOVA table as produced in Section 5.1.1, and the task of testing $H_0 : M_1$ versus $H_1 : M_2$ corresponds to the *overall F–test.*

- Assume $j + 1 = m$, that is, the larger of the two models corresponds to the "full" model. In this case, the test $H_0 : M_j$ versus $H_1 : M_m$ is also known as *partial F-test.* In the context of partial F–tests, the model $M_j = M_{m-1}$ is then often referred to as "reduced model". Note that, in order to test a reduced against a full model, one does not need to specify, or make any calculations, concerning the "lower" models $M_1, \ldots, M_{j-1}$, since all what is needed to calculate the test statistic F_j are the residual sums of squares of the two involved models.

Example 5.3 *(Continuation of Example 5.2)* We are now interested in the question of whether the same straight line suffices for both genders. This can be considered as a test problem where the 'reduced' model (5.12) serves as our null hypothesis, and the full model (5.11) as the alternative. There are two ways of approaching this problem:

Firstly, we can merge the information from two lines of the sequential ANOVA table:

```
> ((28.5+17.6)/2)/47.4   # merging two lines of
[1] 0.48634              # ANOVA table above
> 1-pf(0.48634, 2, 39)
[1] 0.6185519
```

Or secondly, we can answer this question through a partial F–test:

```
> student.fit = lm(weight ~ height*gender, data = student)
> RSS <- sum(student.fit$residuals^2)     # RSS[Full]
> RSS
[1] 1850.411
> df <- student.fit$df        # df[Full]
> df
[1] 39
> student.fit0 <- lm(weight ~ height, data = student)    # Reduced model
> RSS0 <- sum(student.fit0$residuals^2)   # RSS[Reduced]
> RSS0
[1]  1896.556
> df0 <- student.fit0$df      # df[Reduced]
> df0
[1] 41
> F <- ((RSS0-RSS)/(df0-df))/(RSS/df)     # F-value
> F
[1] 0.4862914
> p <- 1-pf(0.4862914,2,39)
   # p-value p = P[F(2,39) > 0.4862914]
> p
[1] 0.6185812
```

Of course, both approaches are equivalent and give the same result. All the above can be more conveniently carried out automatically by applying the R function anova to the two fits:

```
> anova(student.fit0, student.fit)
Analysis of Variance Table

Model 1: weight  ~ height
Model 2: weight  ~ height + gender + height:gender
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 1896.56
2     39 1850.41 2     224.3 0.4863 0.6186
```

Example 5.4 (Animal survival times)

```
> anova(animals.addfit)
Analysis of Variance Table

Response: Survival
          Df  Sum Sq Mean Sq F value    Pr(>F)
Poison     2 1.03301 0.51651  20.643 5.704e-07 ***
Treatment  3 0.92121 0.30707  12.273 6.697e-06 ***
Residuals 42 1.05086 0.02502
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This table indicates that variation in survival time for this “main effects only” model is significantly influenced by both factors, poison and treatment.

```
> anova(animals.interfit)
Analysis of Variance Table

Response: Survival
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Poison            2 1.03301 0.51651 23.2217 3.331e-07 ***
Treatment         3 0.92121 0.30707 13.8056 3.777e-06 ***
Poison:Treatment  6 0.25014 0.04169  1.8743    0.1123
Residuals        36 0.80072 0.022247
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This table indicates that there is little evidence for an interaction. Notice that the ANOVA lines for poison and treatment are the same for the both fits (with and without an interaction term in the model). Also, the residual sum of squares for the model without interaction is the sum of the sums of squares for residuals and the interaction for the model with interaction; and similarly for the degrees-of-freedom.

We change the order of inclusion of terms:

```
> anova(lm(Survival~Treatment+Poison+ Poison:Treatment, data=animals) )
Analysis of Variance Table

Response: Survival
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Treatment         3 0.92121 0.30707 13.8056 3.777e-06 ***
Poison            2 1.03301 0.51651 23.2217 3.331e-07 ***
Treatment:Poison  6 0.25014 0.04169  1.8743    0.1123
Residuals        36 0.80073 0.02224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If we fit treatment before poison, we see that the ANOVA lines are the same, except re-shuffled according to the order of fitting terms. Here the two sources of variation are orthogonal, a virtue of the "balanced" factorial design.

Notice that in an ANOVA table there is a row for each *source of variation* (i.e. for each factor), unlike in the `summary(lm)` output, where is a row for each (combination of) factor *level*(s).

Chapter 6

Model selection

6.1 Submodels

In this section we assume that there is a given *correctly specified* model

$$\mathrm{E}[y \mid \boldsymbol{x}] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} = x_1\beta_1 + \dots x_p\beta_p, \tag{6.1}$$

(featuring A1-A3), but we want to know if there are submodels with fewer terms which will be "almost as good".

Our goal is to find which terms, if any, can be deleted without important loss of information. Let $\mathcal{I}$ be the subset of indices of those terms to be included in a submodel and $\mathcal{D}$ the remainder to be deleted. The cardinality of these sets is denoted by $p_{\mathcal{I}}$ and $p_{\mathcal{D}}$, respectively, with $p_{\mathcal{I}} + p_{\mathcal{D}} = p$. The full mean function in (6.1) can be written

$$\mathrm{E}[y \mid \boldsymbol{x}] = \boldsymbol{x}_{\mathcal{I}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{I}} + \boldsymbol{x}_{\mathcal{D}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{D}}$$

The mean function for the submodel is

$$\mathrm{E}_{\mathcal{I}}[y|\boldsymbol{x}] = \boldsymbol{x}_{\mathcal{I}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{I}} \tag{6.2}$$

where $\mathrm{E}_{\mathcal{I}}$ indicates expectation with respect to an hypothesised mean function for the submodel with $p_{\mathcal{I}}$ terms which may or may not be correct: expectations or variances without such a subscript are with respect to the correct full model.

Dropping a term x_j can be particularly helpful when it is nearly a linear combination of other terms in the model, as dropping x_j is equivalent to replacing it in the full model by the linear combination.

"Good" submodels $\mathcal{I}$ will have $\mathrm{RSS}_{\mathcal{I}}$ close to the RSS of the full model, and $p_{\mathcal{I}}$ as small as possible. One will not achieve both goals arbitrarily well at the same time; one has to make a trade-off between goodness-of-fit and parsimony of the model.

One needs *selection criteria*, also called *information criteria*, to handle this tradeoff. There exist a large number of them; we introduce some important criteria below.

6.2 Selection criteria

(1) *Mallows' $C_{\mathcal{I}}$* is given by

$$C_{\mathcal{I}} = \frac{\mathrm{RSS}_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n \tag{6.3}$$

Equivalent expressions for $C_{\mathcal{I}}$ are

$$C_{\mathcal{I}} = \frac{\mathrm{RSS}_{\mathcal{I}} - \mathrm{RSS}}{s^2} + p_{\mathcal{I}} - p_{\mathcal{D}} = p_{\mathcal{D}}\,(F_{\mathcal{D}} - 1) + p_{\mathcal{I}} \tag{6.4}$$

where the subscript $\mathcal{I}$ refers to statistics computed from the submodel, while those without a subscript are computed from the full model. $F_{\mathcal{D}}$ is the F-statistic for testing the null hypothesis H_0 that $\boldsymbol{\beta}_{\mathcal{D}} = \mathbf{0}$.

Under H_0, one has $E(C_{\mathcal{I}}) = p_{\mathcal{D}}\,(E(F_{\mathcal{D}}) - 1) + p_{\mathcal{I}} \approx p_{\mathcal{D}}(1-1) + p_{\mathcal{I}} = p_{\mathcal{I}}$, indicating that good candidates for submodel mean functions will have $C_{\mathcal{I}} \leq p_{\mathcal{I}}$ (with $p_{\mathcal{I}}$ as small as possible). The last term in (6.4) shows that $C_{\mathcal{I}} \leq p_{\mathcal{I}}$ if and only $F_{\mathcal{D}} \leq 1$. Thus, we would tend to delete a set of terms $\mathcal{D}$ when the F-statistic $F_{\mathcal{D}}$ is less than 1.

(2) A criterion closely related to $C_{\mathcal{I}}$ is the AIC (Akaike Information Criterion), where one aims to minimize

$$AIC_{\mathcal{I}} = -2L_{\mathcal{I}} + 2p_{\mathcal{I}}.$$

Here $L_{\mathcal{I}}$ is the log-likelihood $L(\boldsymbol{\beta}_{\mathcal{I}}, \sigma)$ evaluated at $\boldsymbol{\beta}_{\mathcal{I}} = \hat{\boldsymbol{\beta}}_{\mathcal{I}}$ and a suitable estimate $\hat{\sigma}$ of σ. This can be seen as a maximum likelihood technique which penalizes large values of $p_{\mathcal{I}}$. One can show (Question 4.5) that $AIC_{\mathcal{I}}$ is equivalent to $C_{\mathcal{I}}$ under (A1)-(A3) if σ^2 in $AIC_{\mathcal{I}}$ is estimated by the common error variance estimator s^2 under the full model (6.1).

(3) A simpler idea is to minimize the residual mean square

$$s^2_{\mathcal{I}} = \frac{RSS_{\mathcal{I}}}{df_{\mathcal{I}}}$$

with $df_{\mathcal{I}} = n - p_{\mathcal{I}}$. This selects the submodel with the smallest estimated error variance σ^2.

(4) A variant of this is *Tukey's rule*: Minimize

$$s^2_{\mathcal{I}}/df_{\mathcal{I}} = \frac{RSS_{\mathcal{I}}}{(n - p_{\mathcal{I}})^2}$$

This criterion aims to simultaneously minimize the residual mean square and maximise its degrees-of-freedom, with stronger emphasis on the latter compared to method (3).

We will mainly consider methods (1) and (3) henceforth, since (2) and (4) can be considered as variants of these.

Example 6.1 The following data comprises a response,"heat evolved" $y =$ `heat` (in calories per gram of cement) during hardening of Portland cement considered as a function of the amounts of four covariates, chemical compounds contained in clinkers: "tricalcium aluminate" `aluminate`; "tricalcium silicate" `tri.silicate`; "tetracalcium alumino ferrite" `ferite`; and "dicalcium silicate" `di.silicate`, all as percentages of the weights of the clinkers.

```
> data(cement, package="MASS")
> cement <- cement[,c(5,1,2,3,4)]
> names(cement)<- c("heat", "aluminate", "tri.silicate", "ferite", "di.silicate")
> cement
    heat aluminate tri.silicate ferite di.silicate
1   78.5         7           26      6          60
2   74.3         1           29     15          52
3  104.3        11           56      8          20
4   87.6        11           31      8          47
5   95.9         7           52      6          33
6  109.2        11           55      9          22
7  102.7         3           71     17           6
8   72.5         1           31     22          44
9   93.1         2           54     18          22
10 115.9        21           47      4          26
11  83.8         1           40     23          34
12 113.3        11           66      9          12
13 109.4        10           68      8          12
```

We will assume the "largest model" has predictor terms x_1 = `aluminate`, x_2 = `tri.silicate`, x_3 = `ferite`, x_4 = `di.silicate` and $\mathrm{E}[y \mid x_1, x_2, x_3, x_4] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ is the full model, which we designate as "$1 + x_1 + x_2 + x_3 + x_4$", and $\mathrm{Var}[y \mid x_1, x_2, x_3, x_4] = \sigma^2$. The sequential ANOVA table below for this model, seems to indicate that the model $1+x_1+x_2$ might be a parsimonious model for these data, but we must be careful, as we know that an ANOVA can depend on the order of fitting terms.

```
> cement.fit.full <- lm(heat~ aluminate + tri.silicate + ferite + di.silicate,
   data = cement)
> anova(cement.fit.full)
Analysis of Variance Table

Response: heat
             Df  Sum Sq Mean Sq  F value    Pr(>F)
aluminate     1 1450.08 1450.08 242.3679 2.888e-07 ***
tri.silicate  1 1207.78 1207.78 201.8705 5.863e-07 ***
ferite        1    9.79    9.79   1.6370    0.2366
di.silicate   1    0.25    0.25   0.0413    0.8441
Residuals     8   47.86    5.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6.1 shows $C_{\mathcal{I}}$ and $s^2_{\mathcal{I}}$ applied to all $2^4 = 16$ possible submodels. Both criteria suggest the three models $1 + x_1 + x_2$, $1 + x_1 + x_2 + x_3$ and $1 + x_1 + x_2 + x_4$ as contenders. The submodel $1 + x_1 + x_2$ is preferred because it has fewer terms and also $C_{\mathcal{I}} = 2.68 < 3 = p_{\mathcal{I}}$.

Model contenders	$p_{\mathcal{I}}$	$df_{\mathcal{I}}$	$RSS_{\mathcal{I}}$	$s^2_{\mathcal{I}}$	$C_{\mathcal{I}}$
1	1	12	2715.76	226.31	442.99
$1+x_1$	2	11	1265.69	115.06	202.55
$1+x_2$	2	11	906.34	82.39	142.49
$1+x_3$	2	11	1939.40	176.31	315.15
$1+x_4$	2	11	883.87	80.35	138.73
$1+x_1+x_2$	3	10	57.90	5.79	2.68
$1+x_1+x_3$	3	10	1227.07	122.71	198.09
$1+x_1+x_4$	3	10	74.76	7.48	5.50
$1+x_2+x_3$	3	10	415.44	41.54	62.44
$1+x_2+x_4$	3	10	868.88	86.89	138.23
$1+x_3+x_4$	3	10	175.74	17.57	22.37
$1+x_1+x_2+x_3$	4	9	48.11	5.35	3.04
$1+x_1+x_3+x_4$	4	9	50.84	5.64	3.50
$1+x_1+x_2+x_4$	4	9	47.97	5.33	3.02
$1+x_2+x_3+x_4$	4	9	73.81	8.20	7.34
$1+x_1+x_2+x_3+x_4$	5	8	47.86	5.98	5.00

Table 6.1: All subset models for the cement data. For each submodel with $p_{\mathcal{I}}$ terms, $RSS_{\mathcal{I}}$ is the residual sum of squares; $s^2_{\mathcal{I}}$ is the estimate of the error variance σ^2; $C_{\mathcal{I}} = \frac{RSS_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n$ is Mallows' criterion ($s^2 = 5.98$ is estimate of σ^2 in full model).

6.3 Selection methods

Considering 2^k (for a model with an intercept: $k = p - 1$) submodels is manageable when $k \leq 10$: there are only $2^{10} = 1024$ submodels to consider, but $2^{20} \simeq 1,000,000$ submodels can be computationally time-consuming. A popular alternative is *stepwise regression* with two flavours: *forward selection*, in which terms are sequentially added to the mean function and *backward elimination*, in which terms are sequentially removed from the mean function. A third method alternates between the selection and elimination criteria.

6.3.1 Forward selection

Suppose at some stage there are $p_{\mathcal{I}}$ terms $\boldsymbol{x}_{\mathcal{I}}$ in the model with value $C_{\mathcal{I}}$ for Mallows' criterion.

(i) Add, in turn, each of the remaining $p - p_{\mathcal{I}}$ terms x_j to the current base set $\boldsymbol{x}_{\mathcal{I}}$ and for each of the resulting submodels with $p_{\mathcal{I}} + 1$ terms note the value of $C_{\mathcal{I} \cup \{j\}}$;

(ii) Find the term added in (i) which gives the *smallest* value of $C_{\mathcal{I} \cup \{j\}}$;

(iii) If the smallest value of $C_{\mathcal{I} \cup \{j\}}$ in (ii) is less than or equal to $C_{\mathcal{I}}$ then add the corresponding term to the base set $\boldsymbol{x}_{\mathcal{I}}$; otherwise *STOP*.

Example 6.2 (Continuation of Example 6.1)
Forward selection applied to the cement data with $C_{\mathcal{I}}$:

Submodel	$p_{\mathcal{I}}$	$C_{\mathcal{I}}$
1	1	442.99
$+x_1$	2	202.55
$+x_2$	2	142.49
$+x_3$	2	315.15
$+x_4$	2	*138.73
$1 + x_4$	2	138.73
$+x_1$	3	*5.50
$+x_2$	3	138.23
$+x_3$	3	22.37
$1 + x_4 + x_1$	3	5.50
$+x_2$	4	*3.02
$+x_3$	4	3.50
$1 + x_4 + x_1 + x_2$	4	STOP 3.02
$+x_3$	5	5.00

Table 6.2: Forward selection applied to the cement data with $C_{\mathcal{I}}$. A * indicates which submodel is selected at a particular stage. The forward selected submodel is $1 + x_4 + x_1 + x_2$; that is, 1 + `di.silicate` + `aluminate` + `tri.silicate`.

6.3.2 Backward elimination

As with forward selection of terms, suppose at some stage there are $p_{\mathcal{I}}$ terms $\boldsymbol{x}_{\mathcal{I}}$ in the model with value $C_{\mathcal{I}}$ for Mallows' criterion.

(i) Delete, in turn, each of the $p_{\mathcal{I}}$ terms x_j from the current base set $\boldsymbol{x}_{\mathcal{I}}$ and for each of the resulting submodels with $p_{\mathcal{I}} - 1$ terms note the value of $C_{\mathcal{I} \setminus \{j\}}$;

(ii) Find the term removed in (i) which gives the *smallest* value of $C_{\mathcal{I}\setminus\{j\}}$;

(iii) If the smallest value of $C_{\mathcal{I}\setminus\{j\}}$ in (ii) is less than or equal to $C_{\mathcal{I}}$ then remove the corresponding term from the base set $\boldsymbol{x}_{\mathcal{I}}$; otherwise *STOP*.

Example 6.3 (Continuation of Example 6.2) *Backward elimination applied to the Portland cement data with $C_{\mathcal{I}}$:*

Submodel	$p_{\mathcal{I}}$	$C_{\mathcal{I}}$
$1 + x_1 + x_2 + x_3 + x_4$	5	5.00
$-x_1$	4	7.34
$-x_2$	4	3.50
$-x_3$	4	*3.02
$-x_4$	4	3.04
$1 + x_1 + x_2 + x_4$	4	3.02
$-x_1$	3	138.23
$-x_2$	3	5.50
$-x_4$	3	*2.68
$1 + x_1 + x_2$	3	STOP 2.68
$-x_1$	2	142.49
$-x_2$	2	202.55

Table 6.3: Backward elimination applied to the cement data with $C_{\mathcal{I}}$. A * indicates which submodel is selected at a particular stage. The backward selected submodel is $1 + x_1 + x_2$; that is, 1+ `aluminate` + `tri.silicate`.

6.3.3 Stepwise selection

This combines the forward and backward procedures as follows. Proceed with forward selection, except at each stage test whether any term in the current submodel can be eliminated according to the backward elimination method. Stop when both the forward and backward criteria are satisfied.

Example 6.4 (Continuation of Example 6.3)

Here we apply the `R` functions `lm` and `step` to the Portland cement data to carry out backward elimination and forward selection — look also at the `help` for `step`. The results are the same as we got in the previous two tables. *Note that in what follows the notation C_p is used instead of $C_{\mathcal{I}}$ and p instead of $p_{\mathcal{I}}$ whereas in our development p is used exclusively for the number of terms in the full model.*

Backward Elimination: First we fit "largest" model using `lm`.

```
> cement.fit.full <- lm(heat ~ aluminate + tri.silicate + ferite + di.silicate,
data = cement)
```

Now we apply `step` to the full model fit `cement.fit.full` to do *backward elimination* (the default) using AIC $\equiv C_p$.

```
> step(cement.fit.full, scale = summary(cement.fit.full)$sigma^2)

Start:  AIC= 5
 heat ~ aluminate + tri.silicate + ferite + di.silicate


              Df Sum of Sq    RSS     Cp
```

```
- ferite        1     0.109 47.973 3.0182
- di.silicate   1     0.247 48.111 3.0413
- tri.silicate  1     2.972 50.836 3.4968
<none>                      47.864 5.0000
- aluminate     1    25.951 73.815 7.3375

Step:  AIC= 3.02
 heat ~ aluminate + tri.silicate + di.silicate

               Df Sum of Sq    RSS       Cp
- di.silicate   1      9.93  57.90   2.6782
<none>                       47.97   3.0182
- tri.silicate  1     26.79  74.76   5.4959
- aluminate     1    820.91 868.88 138.2259

Step:  AIC= 2.68
 heat ~ aluminate + tri.silicate

               Df Sum of Sq     RSS       Cp
<none>                        57.90   2.6782
- aluminate     1    848.43  906.34 142.4864
- tri.silicate  1   1207.78 1265.69 202.5488

Call:
lm(formula = heat ~ aluminate + tri.silicate, data = cement)

Coefficients:
 (Intercept)     aluminate  tri.silicate
     52.5773        1.4683        0.6623
```

Forward Selection: Here we start with the fit to the "smallest" model.

```
> cement.fit.intercept <- lm(heat ~ 1, data = cement)
```

Now we apply `step` to the intercept-only model fit `cement.fit.intercept` to do *forward selection* using AIC $\equiv C_p$. Note the use of `direction` and `scope`, which are not necessary for backward elimination.

```
> step(cement.fit.intercept, scale = summary(cement.fit.full)$sigma^2,
scope = list(lower = cement.fit.intercept, upper = cement.fit.full),
direction = "forward")

Start:  AIC= 442.92
 heat ~ 1

               Df Sum of Sq     RSS     Cp
+ di.silicate   1   1831.90  883.87 138.73
+ tri.silicate  1   1809.43  906.34 142.49
+ aluminate     1   1450.08 1265.69 202.55
+ ferite        1    776.36 1939.40 315.15
<none>                      2715.76 442.92

Step:  AIC= 138.73
 heat ~ di.silicate

               Df Sum of Sq    RSS       Cp
+ aluminate     1    809.10  74.76   5.4959
+ ferite        1    708.13 175.74  22.3731
+ tri.silicate  1     14.99 868.88 138.2259
<none>                      883.87 138.7308

Step:  AIC= 5.5
```

```
 heat ~ di.silicate + aluminate

               Df Sum of Sq    RSS     Cp
+ tri.silicate  1    26.789 47.973 3.0182
+ ferite        1    23.926 50.836 3.4968
<none>                      74.762 5.4959

Step:  AIC= 3.02
 heat ~ di.silicate + aluminate + tri.silicate

         Df Sum of Sq    RSS     Cp
<none>                47.973 3.0182
+ ferite  1     0.109 47.864 5.0000

Call:
lm(formula = heat ~ di.silicate + aluminate + tri.silicate, data = cement)

Coefficients:
 (Intercept)   di.silicate     aluminate  tri.silicate
   71.6483       -0.2365        1.4519        0.4161
```

Stepwise selection: Using `direction = ‘‘both’’` gives (for this data set!) the same result as backward elimination. Try yourself!

Chapter 7

Regression diagnostics & transformations

7.1 Regression diagnostics

There are two inter-related strands. *Residuals* are used to check the linear model assumptions, and *influential observations* (yet to define) are those cases featuring a particularly large impact on $\hat{\boldsymbol{\beta}}$ or s^2, or both. The key concept for the understanding of both is the *hat matrix*, from which important diagnostic tools are derived.

For a LM $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, recall that *fitted values* are given by

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY} = \boldsymbol{H}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = \boldsymbol{X\beta} + \boldsymbol{H\epsilon} \tag{7.1}$$

with $n \times n$ *hat matrix*

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \left(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_j\right)_{1\le i\le n, 1\le j\le n} \equiv (h_{ij})_{1\le i\le n, 1\le j\le n}$$

where $\boldsymbol{x}_i^T$ comprises the values of the predictors for case i. In (7.1) we have used that (H5) $\boldsymbol{HX} = \boldsymbol{X}$. The hat matrix has several other interesting properties which are summarized in Table 7.1. One also sees from (7.1) that $\hat{\boldsymbol{Y}}$ has the same mean function as $\boldsymbol{Y}$ but with error term $\boldsymbol{H\epsilon}$ in place of $\boldsymbol{\epsilon}$.

7.1.1 Leverage values and studentised residuals

Taking the $i-$th row of $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$, one finds that

$$\hat{y}_i = \ldots + h_{ii}y_i + \ldots$$

where the so-called *leverage values*

$$h_i \equiv h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

(also called: potential influence values, hat values) measure the impact of the $i-$th case on its own fitted value. It is always $0 \le h_i \le 1$, provided that the rows of $\boldsymbol{X}$ are different, and for models with an intercept, $\frac{1}{n} \le h_i \le 1$.

Residuals are given by

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$$

Thus the residuals $\hat{\boldsymbol{\epsilon}}$ are the same known linear functions of both the data $\boldsymbol{Y}$ and the random errors $\boldsymbol{\epsilon}$, and

$$\mathrm{Var}[\hat{\boldsymbol{\epsilon}}] = \mathrm{Var}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}] = (\boldsymbol{I} - \boldsymbol{H})\mathrm{Var}[\boldsymbol{\epsilon}](\boldsymbol{I} - \boldsymbol{H})^T = (\boldsymbol{I} - \boldsymbol{H})^2\sigma^2\boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{H})\sigma^2. \tag{7.2}$$

Table 7.2 summarizes properties of errors and residuals. Equation (7.2) implies that $\mathrm{Var}[\hat{\epsilon}_i] = (1 - h_i)\sigma^2$ and $\mathrm{Cov}[\hat{\epsilon}_i, \hat{\epsilon}_j] = -h_{ij}\sigma^2$.

Table 7.1: Properties of the hat matrix $\boldsymbol{H}$:

(H1)	$\boldsymbol{H}^2 = \boldsymbol{H}$ (idempotent)
(H2)	$\boldsymbol{H}^T = \boldsymbol{H}$ (symmetric)
(H3)	$\text{rank}\,[\boldsymbol{H}] = p$
(H4)	$\text{Tr}\,[\boldsymbol{H}] = p$
(H5)	$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and hence $\boldsymbol{X}^T\boldsymbol{H} = \boldsymbol{X}^T$
(H6)	If model has intercept: $\boldsymbol{H}\mathbf{1} = \mathbf{1}$; that is, $\sum_{i=1}^n h_{ij} = \sum_{i=1}^n h_{ji} = 1$ (Row and column sums of $\boldsymbol{H}$ are 1).
(H7)	$\boldsymbol{H}$ is positive semi–definite

Table 7.2: Properties of residuals $\hat{\boldsymbol{\epsilon}}$ and random error $\boldsymbol{\epsilon}$

Assumption	Errors	Residuals
–		$\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{\epsilon}} = 0$ (see Question 6.1).
–		$\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}} = 0$ (see Section 2.2.2).
–		If model has intercept term then $\mathbf{1}^T\hat{\boldsymbol{\epsilon}} = \sum \hat{\epsilon}_i = 0$.
(A1)	$\text{E}[\boldsymbol{\epsilon}] = \mathbf{0}$	$\text{E}[\hat{\boldsymbol{\epsilon}}] = \mathbf{0}$
(A2)	$\text{Var}[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}$	$\text{Var}[\hat{\boldsymbol{\epsilon}}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$
(A3)	$\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I})$	$\hat{\boldsymbol{\epsilon}} \sim N(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$

If $\epsilon_i \sim N(0, \sigma^2)$ then $\hat{\epsilon}_i \sim N(0, (1 - h_i)\sigma^2)$. Hence

$$\frac{\hat{\epsilon}_i}{\sigma\sqrt{1 - h_i}} \sim N(0, 1)$$

Diagnostic procedures are based on residuals $\hat{\boldsymbol{\epsilon}}$, which we would like to assume behave like the unobserved errors $\boldsymbol{\epsilon}$. Usefulness of such an assumption depends crucially on $\boldsymbol{H}$ since it relates $\hat{\boldsymbol{\epsilon}}$ to $\boldsymbol{\epsilon}$ via $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$

Replacing σ by its estimate s, we obtain the *studentised residuals*

$$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_i}}$$

It can be shown that $\text{E}[r_i] = 0$ and $\text{Var}[r_i] \approx 1$. Hence in a plot of r_i versus i, a case with a studentised residual greater in magnitude than 2 or 3 will suggest a possible outlier. We speak of 'internally' studentised residuals when all data, *including case* i, are used to estimate σ, otherwise one speaks of 'externally' studentised residuals.

The behaviour of $\hat{\epsilon}_i$, h_i, and r_i is illustrated in Figure 7.1. This plot shows a bivariate data set featuring some horizontal and vertical outlines, and a regression line (for y versus x) fitted to *all* data points. The $\triangle$ symbol within the bulk of the data is a data point where all of $\hat{\epsilon}_i$, h_i, and r_i are small. The vertical outlier (+) has $\hat{\epsilon}_i$ large and h_i small, and, hence, r_i moderately large. The data point in the top right corner has h_i large, but it is not outlying with respect to the regression model, so $\hat{\epsilon}_i$ is small and r_i still moderately sized. In contrast, the point in the bottom right corner ($\diamond$) has large $\hat{\epsilon}_i$, large h_i, and a very large value of r_i.

7.1.2 Influence analysis

While we use residuals to check the model, influence analysis assumes that the model is correct and studies robustness of conclusions to perturbations in the data. We define an observation to be *influential*

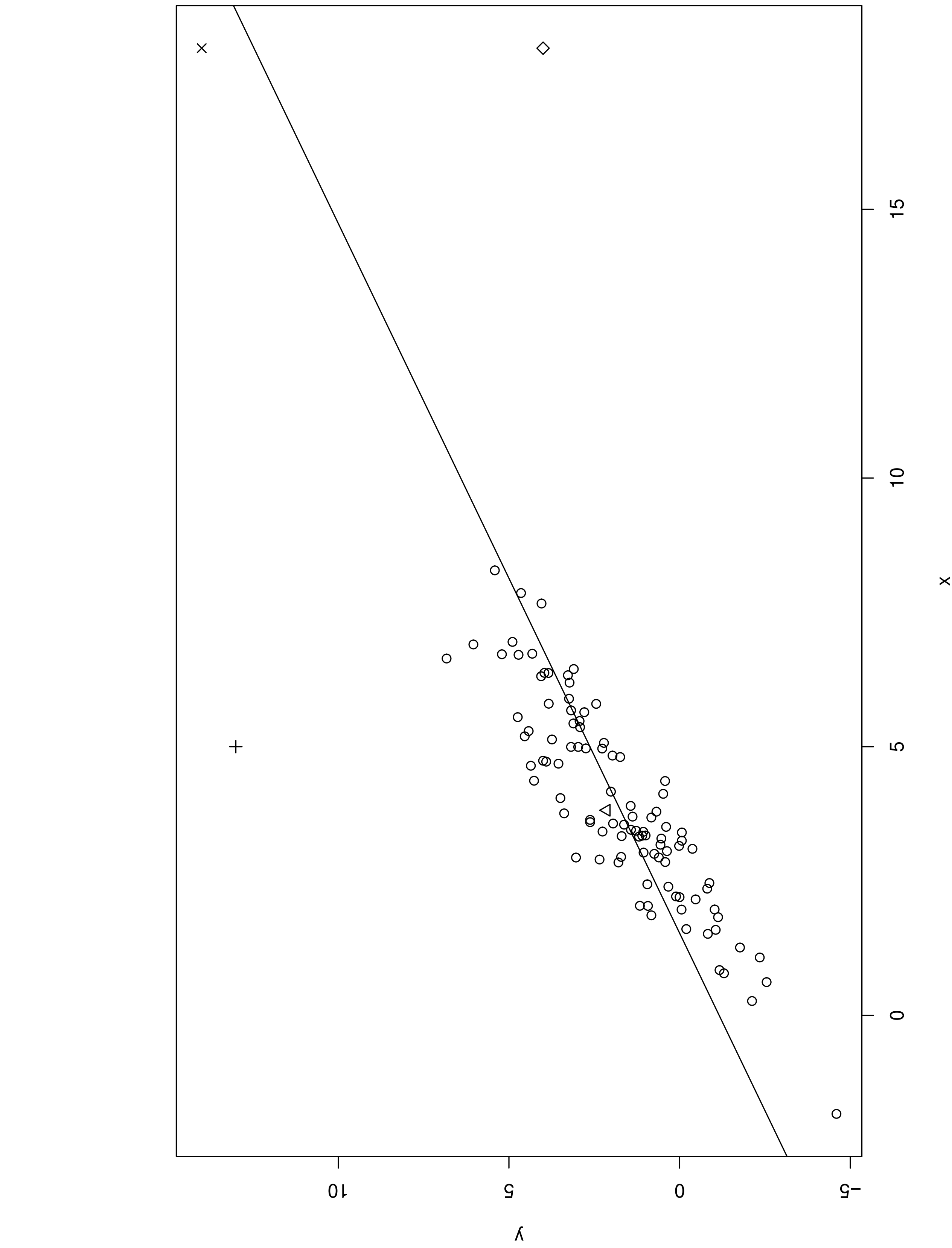


Figure 7.1: Illustration of $\hat{\epsilon}_i$, h_i, and r_i (see description in text and turn by 90 degrees!).

when $\hat{\boldsymbol{\beta}}$ or s^2 change substantially when deleting it (we will only investigate the behavior of $\hat{\boldsymbol{\beta}}$ henceforth). One idea is to look at the leverage values $h_i \equiv h_{ii}$.

Since $\text{Var}[\hat{\epsilon}_i] = (1 - h_i)\sigma^2 \searrow 0$ as $h_i \nearrow 1$ then essentially no matter what the value of y_i if $h_i \simeq 1$ then $\hat{\epsilon}_i \simeq 0$ so that the regression "plane" almost passes through y_i; that is $\hat{y}_i \simeq y_i$.

If case i has $h_i \simeq 1$ then it is *potentially* very influential in fitting the model. That is, without such a case i, the estimated regression coefficients and error variance *may* change substantially. However, a case can have large potential influence (leverage) but little actual influence. Thus, an overall influence measure should involve both $\boldsymbol{Y}$ and $\boldsymbol{X}$: the h_i only involve $\boldsymbol{X}$, not $\boldsymbol{Y}$.

The better idea is to delete cases from data, one at a time, and compare results with those from the fit to the *full* model. Those cases that cause major changes in analysis are called *influential.*
Notation: Subscript $_{(i)}$ means i-th case deleted; for example,

$$\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{X}_{(i)}^T \boldsymbol{Y}_{(i)}$$

In the above, $\hat{\boldsymbol{\beta}}_{(i)}, \boldsymbol{X}_{(i)}, \boldsymbol{Y}_{(i)}$ are respectively $p \times 1, (n-1) \times p, (n-1) \times 1$.

Cook's distance

We want to compare $\hat{\boldsymbol{\beta}}_{(i)}$ with $\hat{\boldsymbol{\beta}}$ using a single number. There are lots of possibilities but one popular choice is *Cook's distance*, D_i, which for case i is

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})^T (\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})}{ps^2} \overset{[W, p.200]}{=} \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

Note that

(a) $\hat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)}$ is $n \times 1$, *not* $(n-1) \times 1$.

(b) Since p is fixed, D_i is determined by two factors r_i^2 and h_i: it increases as r_i^2 increases and as h_i increases (because $h/(1-h)$ is an increasing function of h);

(c) the "studentised residual" r_i is a random quantity reflecting the "lack of fit" to the model of case i;

(d) the "potential influence" h_i is a non-random quantity reflecting the potential for case i to be influential, measuring the distance of $\boldsymbol{x}_i$ from the "centre of gravity" $\overline{\boldsymbol{x}}$ of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n$;

(e) Cases for which D_i is "large" have substantial influence on $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}$; and their *deletion* may result in substantial changes in conclusions;

(f) Cases for which $D_i \geq 1/2$ (rule of thumb!) should be considered carefully to see what effect their deletion would have.

Example 7.1 *Fuel consumption* We will carry out influence analysis for the fuel consumption data ([W], see also Example 1.8). The response is the fuel consumption `FUEL` ("gallons per person"). There are several covariates, of which we consider the following ones: `TAX` (cents per gallon), `DLIC` (% population with driving licences), `INC` (average income in $1000's), `ROAD` (1000's of miles). The data and a pairs plot are provided below:

```
#install.packages("remotes") #you need to do that once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> ?fuelcons
> fuelcons
   STATE   POP   TAX  NLIC   INC   ROAD FUELC DLIC FUEL
1     ME  1029  9.00   540 3.571  1.976   557 52.5  541
2     NH   771  9.00   441 4.092  1.250   404 57.2  524
3     VT   462  9.00   268 3.865  1.586   259 58.0  561
```

```
4     MA  5787  7.50  3060 4.870  2.351  2396 52.9  414
5     RI   968  8.00   527 4.399  0.431   397 54.4  410
6     CN  3082 10.00  1760 5.342  1.333  1408 57.1  457
7     NY 18366  8.00  8278 5.319 11.868  6312 45.1  344
8     NJ  7367  8.00  4074 5.126  2.138  3439 55.3  467
9     PA 11926  8.00  6312 4.447  8.577  5528 52.9  464
10    OH 10783  7.00  5948 4.512  8.507  5375 55.2  498
11    IN  5291  8.00  2804 4.391  5.939  3068 53.0  580
12    IL 11251  7.50  5903 5.126 14.186  5301 52.5  471
13    MI  9082  7.00  5213 4.817  6.930  4768 57.4  525
14    WI  4520  7.00  2465 4.207  6.580  2294 54.5  508
15    MN  3896  7.00  2368 4.332  8.159  2204 60.8  566
16    IA  2883  7.00  1689 4.318 10.340  1830 58.6  635
17    MO  4753  7.00  2719 4.206  8.508  2865 57.2  603
18    ND   632  7.00   341 3.718  4.725   451 54.0  714
19    SD   579  7.00   419 4.716  5.915   501 72.4  865
20    NE  1525  8.50  1033 4.341  6.010   976 67.7  640
21    KS  2258  7.00  1496 4.593  7.834  1466 66.3  649
22    DE   565  8.00   340 4.983  0.602   305 60.2  540
23    MD  4056  9.00  2073 4.897  2.449  1883 51.1  464
24    VA  4764  9.00  2463 4.258  4.686  2604 51.7  547
25    WY  1781  8.50   982 4.574  2.619   819 55.1  460
26    NC  5214  9.00  2835 3.721  4.746  2953 54.4  566
27    SC  2665  8.00  1460 3.448  5.399  1537 54.8  577
28    GA  4720  7.50  2731 3.846  9.061  2979 57.9  631
29    FL  7259  8.00  4084 4.188  5.975  4169 56.3  574
30    KY  3299  9.00  1626 3.601  4.650  1761 49.3  534
31    TN  4031  7.00  2088 3.640  6.905  2301 51.8  571
32    AL  3510  7.00  1801 3.333  6.594  1946 51.3  554
33    MS  2263  8.00  1309 3.063  6.524  1306 57.8  577
34    AR  1978  7.50  1081 3.357  4.121  1242 54.7  628
35    LA  3720  8.00  1813 3.528  3.495  1812 48.7  487
36    OK  2634  6.58  1657 3.802  7.834  1695 62.9  644
37    TX 11649  5.00  6595 4.045 17.782  7451 56.6  640
38    MT   719  7.00   421 3.897  6.385   506 58.6  704
39    ID   756  8.50   501 3.635  3.274   490 66.3  648
40    WY   345  7.00   232 4.345  3.905   334 67.2  968
41    CO  2357  7.00  1475 4.449  4.639  1384 62.6  587
42    NM  1065  7.00   600 3.656  3.985   744 56.3  699
43    AZ  1945  7.00  1173 4.300  3.635  1230 60.3  632
44    UT  1126  7.00   572 3.745  2.611   666 50.8  591
45    NV   527  6.00   354 5.215  2.302   412 67.2  782
46    WN  3443  9.00  1966 4.476  3.942  1757 57.1  510
47    OR  2182  7.00  1360 4.296  4.083  1331 62.3  610
48    CA 20468  7.00 12130 5.002  9.794 10730 59.3  524
```

```
> pairs(fuelcons[,c("FUEL", "TAX", "DLIC", "INC", "ROAD")])
```

We first fit a multiple linear regression model to the fuel consumption data and look at the residuals.

```
> fuel.lm <- lm(FUEL~ TAX+ DLIC+ INC+ROAD, data = fuelcons)
> summary(fuel.lm)                           # Part of the summary follows
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

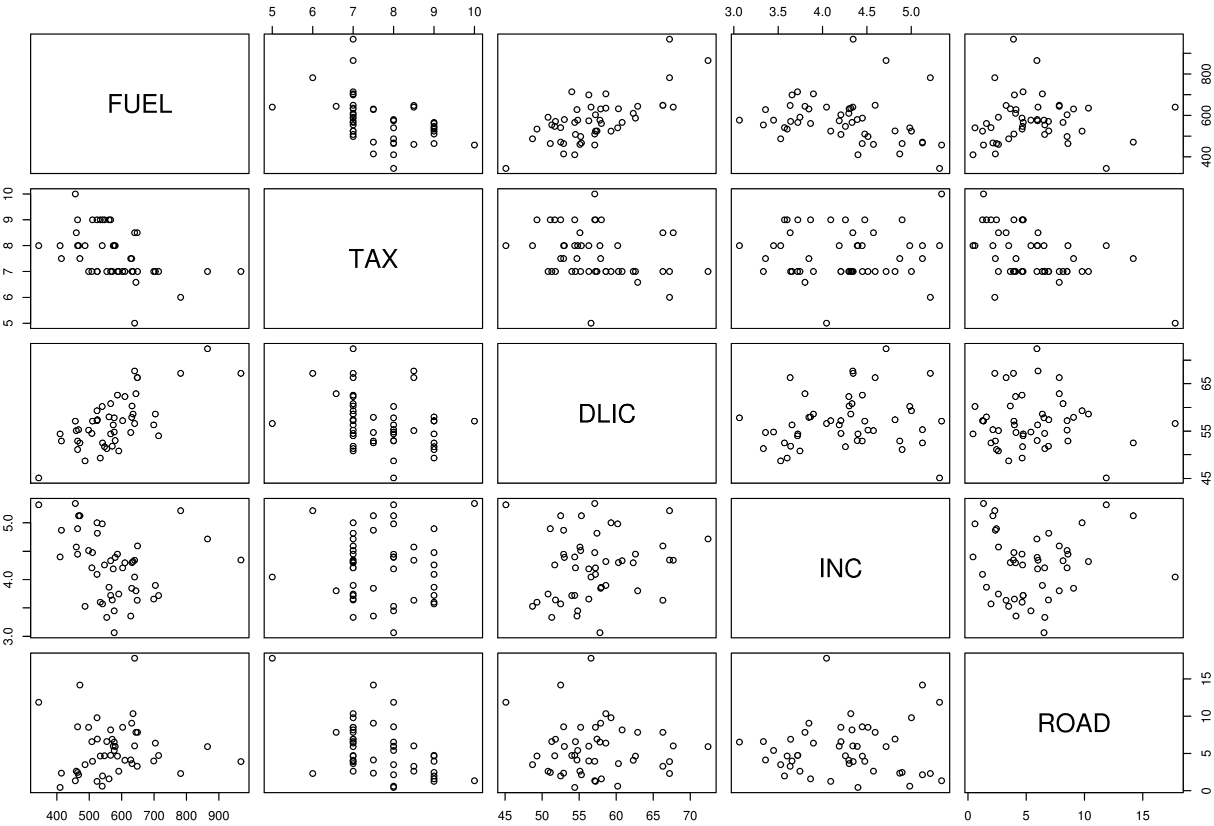


Figure 7.2: A matrix scatterplot of the fuel consumption data.

```
(Intercept)  377.291    185.541   2.033 0.048207 *
TAX          -34.790     12.970  -2.682 0.010332 *
DLIC          13.364      1.923   6.950 1.52e-08 ***
INC          -66.589     17.222  -3.867 0.000368 ***
ROAD          -2.426      3.389  -0.716 0.477999

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-Squared: 0.6787,     Adjusted R-squared: 0.6488

> e <- fuel.lm$res
> plot(e)                        # gives a plot of e_i against i
> identify(e)                    # case 40 outlying?
```

Externally and internally studentised residuals would be computed through

```
> r0 <- rstudent(fuel.lm)    # Externally studentised
> r  <- rstandard(fuel.lm)   # Internally studentised (let’s use these)
> plot(r)            # gives a plot of r_i against i.
> identify(r)        # also here, case 40 stands out -- clear outlier with r>3.
```

Next, we extract leverage values from the output of function `lm.influence`:

```
> fuel.inf <- lm.influence(fuel.lm)   # object containing influence information
                                        about regression object fuel.lm
> names(fuel.inf)
  # [1] "hat"          "coefficients" "sigma"        "wt.res"
```

```
> h <- fuel.inf$hat                   # hat: a vector containing h_i
> h
 [1] 0.09634480 0.07402210 0.08474641 0.12515339 0.09231812 0.22880802
 [7] 0.28322744 0.11069764 0.06076012 0.04790797 0.03582251 0.23047670
[13] 0.05563302 0.04330312 0.04408112 0.06413393 0.03814984 0.07341915
[19] 0.19037667 0.18745045 0.09536828 0.10941373 0.11270322 0.07370097
[25] 0.05175408 0.09124983 0.06577900 0.06465282 0.02652023 0.10906237
[31] 0.07816256 0.10890532 0.13778638 0.08036843 0.10415246 0.08097385
[37] 0.31510460 0.03963847 0.17091219 0.09972312 0.05434435 0.07451915
[43] 0.05698576 0.14884860 0.26597947 0.07231159 0.05632373 0.08792289

> plot(h)              # gives a plot of h_i against i
> identify(h)          # here, case 37 attains the top value
```

Cook's distances are easily obtained through

```
> d <- cooks.distance(fuel.lm)      # a vector of the Cooks distances D_i
> plot(d)                           # gives a plot of D_i against i. Notice again
> identify(d)                       # how case 40 stands out even though d[40] is
                                    # is only 0.3089626.
```

The produced plots of `e, r, h` and `d` are provided in Figure 7.3. Summarizing, case 40 (Wyoming) is clearly distinctive and likely to be an outlier, yet not actually influential in the sense of our rule of thumb (D_{40} is smaller than 1/2). Case 37 (Texas) has a relatively large leverage value and can be called potentially influential (several other observations could be classified as potentially influential too).

We can now consider fitting the full model without some of these possibly influential observations. For example, suppose we omit cases 37 (Texas) and 40 (Wyoming). We do this as

```
> newfuel.lm <- lm(FUEL~ TAX+ DLIC+ INC+ROAD, data = fuelcons, subset=c(-37,-40))
```

Part of the new summary is:

```
> summary(newfuel.lm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  433.744    157.866   2.748  0.00888 **
TAX          -31.663     11.155  -2.838  0.00702 **
DLIC          11.728      1.661   7.063 1.34e-08 ***
INC          -66.407     14.533  -4.569 4.43e-05 ***
ROAD          -1.192      3.078  -0.387  0.70061

Residual standard error: 55.8 on 41 degrees of freedom
Multiple R-Squared: 0.7009,     Adjusted R-squared: 0.6717
```

Comparing `summary(newfuel.lm)` with `summary(fuel.lm)`, we notice some improvements and changes, none particularly dramatic.

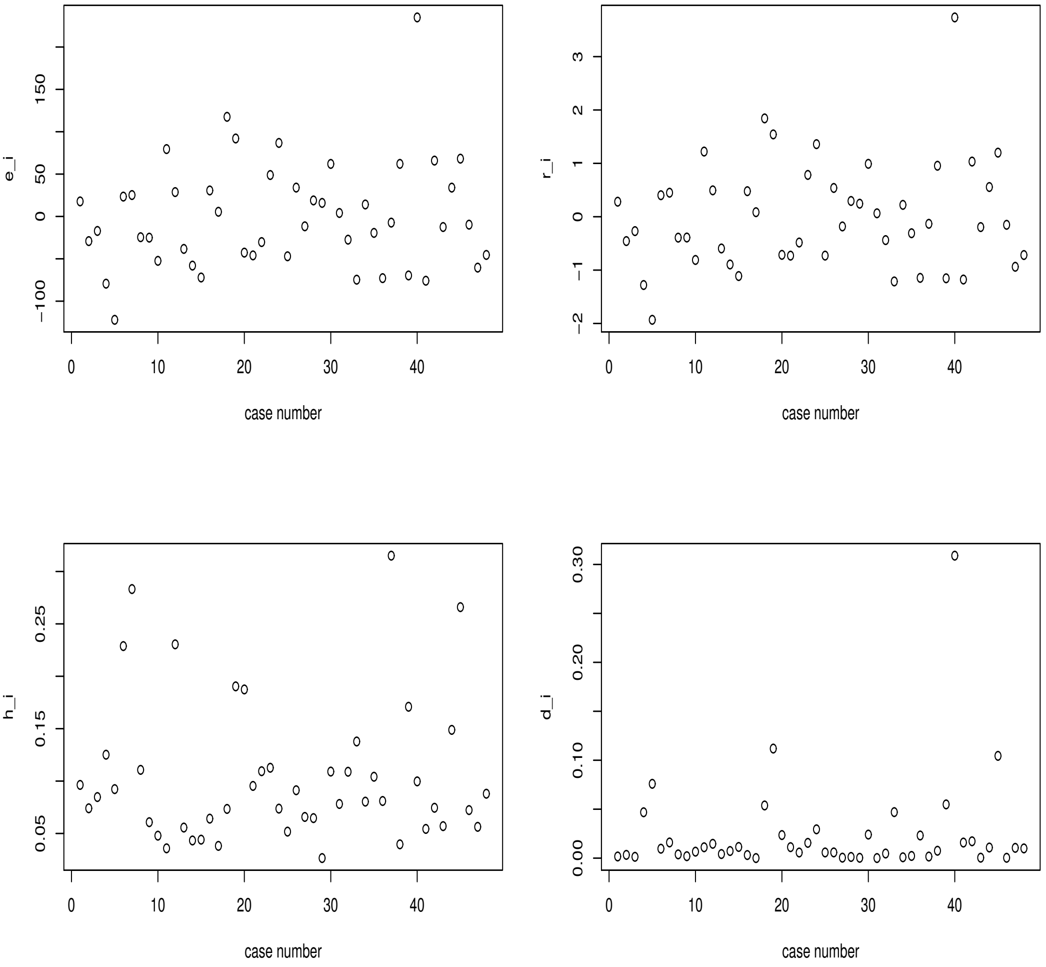


Figure 7.3: Residual and influence plots for fuel consumption data: Residuals $\hat{\epsilon}_i$; studentised residuals r_i; potential influences h_i; and Cook's distance D_i.

7.1.3 Model checking through residual diagnostics

We want to check

(A1) the form of $\mathrm{E}[y \mid \boldsymbol{x}]$ for all $\boldsymbol{x}$, equivalently $\mathrm{E}[\epsilon \mid \boldsymbol{x}] = 0$ for all $\boldsymbol{x}$;

(A2') $\mathrm{Var}[y \mid \boldsymbol{x}] = \sigma^2$ for all $\boldsymbol{x}$, equivalently $\mathrm{Var}[\epsilon \mid \boldsymbol{x}] = \sigma^2$ for all $\boldsymbol{x}$;

(A3) $y|\boldsymbol{x} \sim \mathrm{N}(\mathrm{E}[y \mid \boldsymbol{x}], \sigma^2)$ for all $\boldsymbol{x}$, equivalently, $\epsilon|\boldsymbol{x} \sim \mathrm{N}(0, \sigma^2)$ for all $\boldsymbol{x}$;

(A2") whether the errors $\epsilon_i \equiv \epsilon|\boldsymbol{x}_i, i = 1, \ldots, n$, are independent.

We will use residuals to check these four assumptions. In particular, we use the following previously established results to check items (A1) and (A2') above.

(a) $\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$

(b) $\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$

Equation (a) implies $0 = \sum_{i=1}^{n} x_{ij}\hat{\epsilon}_i$ for $j = 1, \ldots, p$, which corresponds exactly to the numerator of the correlation coefficient between the residuals and the j-th predictor (if there is an intercept in the model). Hence, one has

$$\mathrm{Corr}[\boldsymbol{x}_j, \hat{\boldsymbol{\epsilon}}] = 0 \qquad j = 1, \ldots, p$$

where $\boldsymbol{x}_j$ is the jth column of the design matrix $\boldsymbol{X}$.

Similarly (b) implies that the sample correlation

$$\mathrm{Corr}[\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\epsilon}}] = 0.$$

Thus, if the model assumptions (A1) and (A2') are correct, plots of $\hat{\epsilon}_i$ versus x_{ij} for $j = 1 \ldots, p$ and $\hat{\epsilon}_i$ versus $\hat{y}_i$ should be patternless. Otherwise,

- curvature suggests form of $\mathrm{E}[y \mid \boldsymbol{x}]$ is wrong
- a trumpet shaped plot suggests form of $\mathrm{Var}[y \mid \boldsymbol{x}] = \sigma^2$ for all $\boldsymbol{x}$ is wrong
- possibly both of the above!

A *Normal quantile plot* can be used to check assumption (A3). A plot

$$u_i = \Phi^{-1}\left(\frac{i - 0.5}{n}\right) \text{ versus } r_{(i)}$$

where $r_{(1)} < r_{(2)} < \cdots < r_{(n)}$ are the ordered studentised residuals $r_i = \hat{\epsilon}_i / s\sqrt{1 - h_i}$, should look like straight line $u = r$ through the origin with slope 45 degrees; otherwise, a curved plot suggests some form of non-normality.

A pattern in plot of $\hat{\epsilon}_i$ versus i or (if available) t_i, the *time* at which y_i was observed, may suggest non-independent errors, violating assumption (A2").

For standard graphical diagnostics in `R`, suppose `myfit.lm` results fitting a linear model using function `lm` in `R`. Then type `plot(myfit.lm)` and, as instructed, hit return to produce the following four plots:

(i) residuals versus fitted values plot;

(ii) Normal Q-Q plot of residuals;

(iii) scale-location plot: $\sqrt{|\text{internally studentised residuals}|}$ versus fitted values;

(iv) Either r_i vs h_i (continuous predictors) or r_i vs "Factor level combinations".

Example 7.2 (Continuation of Example 7.1)

Diagnostic plots for fuel consumption data. The corresponding plots are provided in Figure 7.4.

```
  # Residual plots (check for patterns)
> plot(fuelcons$TAX, fuel.lm$res)
> plot(fuelcons$DLIC, fuel.lm$res)
> plot(fuelcons$INC, fuel.lm$res)
> plot(fuelcons$ROAD, fuel.lm$res)
       # all ok.

> plot(fuel.lm$fitted, fuel.lm$res)
       # slight trumpet-shape?

  # check for normality
> qqnorm(fuel.lm$res)
> qqline(fuel.lm$res)
       # or
> qqnorm(rstandard(fuel.lm))
> qqline(rstandard(fuel.lm))
       # ok (small deviations in boundary region are acceptable).

  # check for residual autocorrelation and outliers
> plot(fuel.lm$res)
       # ok.

  # R standard diagnostics
  # plots not shown --- please produce yourself!
> par(mfrow=c(2,2))
> plot(fuel.lm)
   # top left: residual vs fitted
                     # no strong pattern here, but somewhat increasing spread
   # top right: r_i vs Gauss quantiles
                     # no strong indication against normality (boundaries are tolerable)
   # bottom left: sqrt(|r_i|) vs fitted
                     # slight tendency to heteroscedasticity  ("Location-scale-plot")
   # bottom right: r_i vs h_i
                     # No influential observations
```

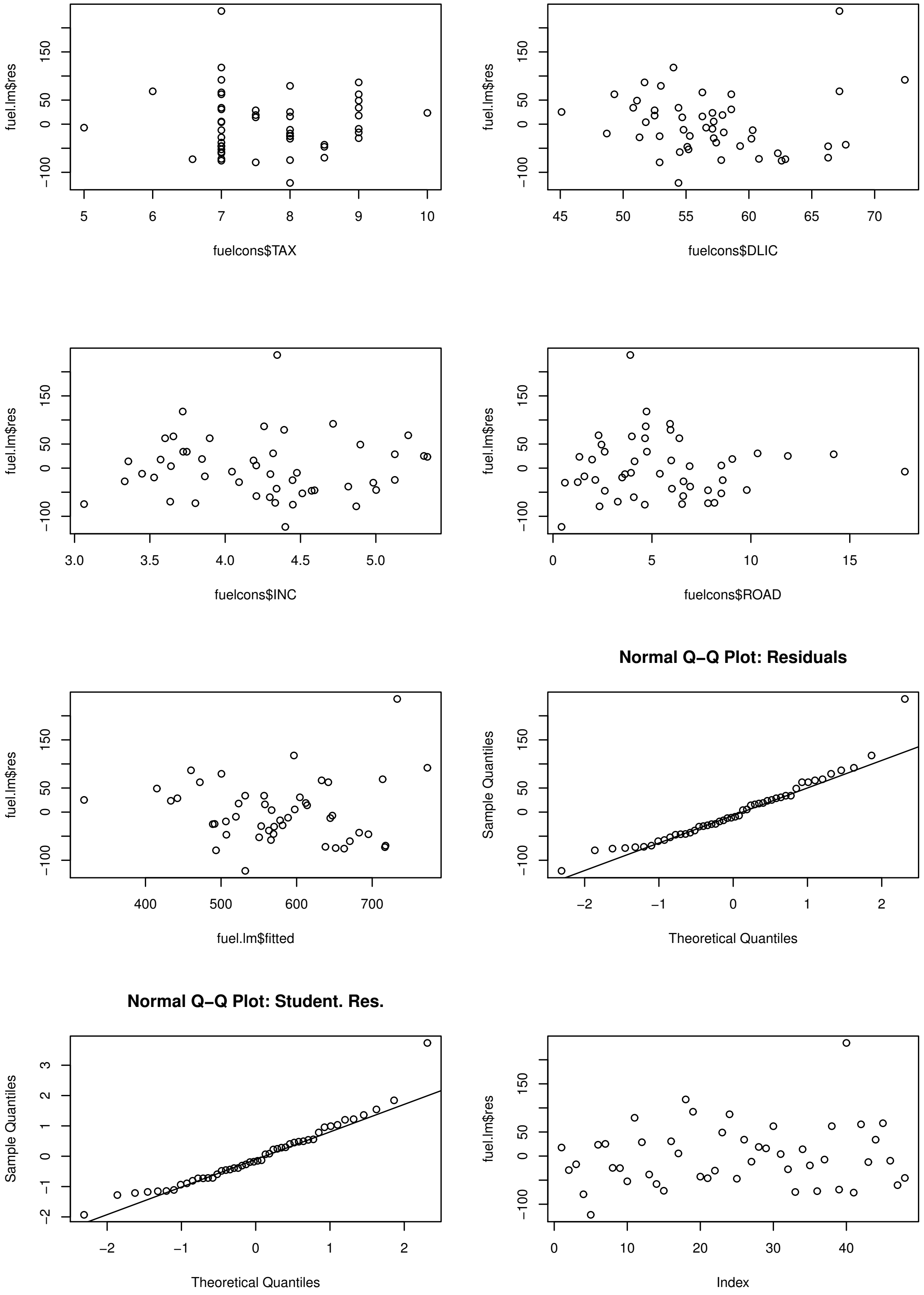


Figure 7.4: Diagnostic plots for fuel consumption data, in the order as produced in Example 7.2.

Example 7.3 (Animals data — Continuation of Example 5.4)

Figures 7.5 and 7.6 show the standard R diagnostic plots for the main effects-plus-interaction fit to both survival and reciprocal of survival in the animal factorial experiment involving the 12 combinations of the 4 treatments and 3 poisons investigated.

The residuals-versus-fitted values plot in Figure 7.5 indicates strongly that variance increases with mean, and this is supported in the scale-location plot. The Q-Q plot gives no support for the normality assumption. However, inspection of the corresponding plots in Figure 7.6 for the main effects-plus-interaction fit to the *reciprocal* of survival (rate of dying) suggest the standard four assumptions are valid on this scale of the response.

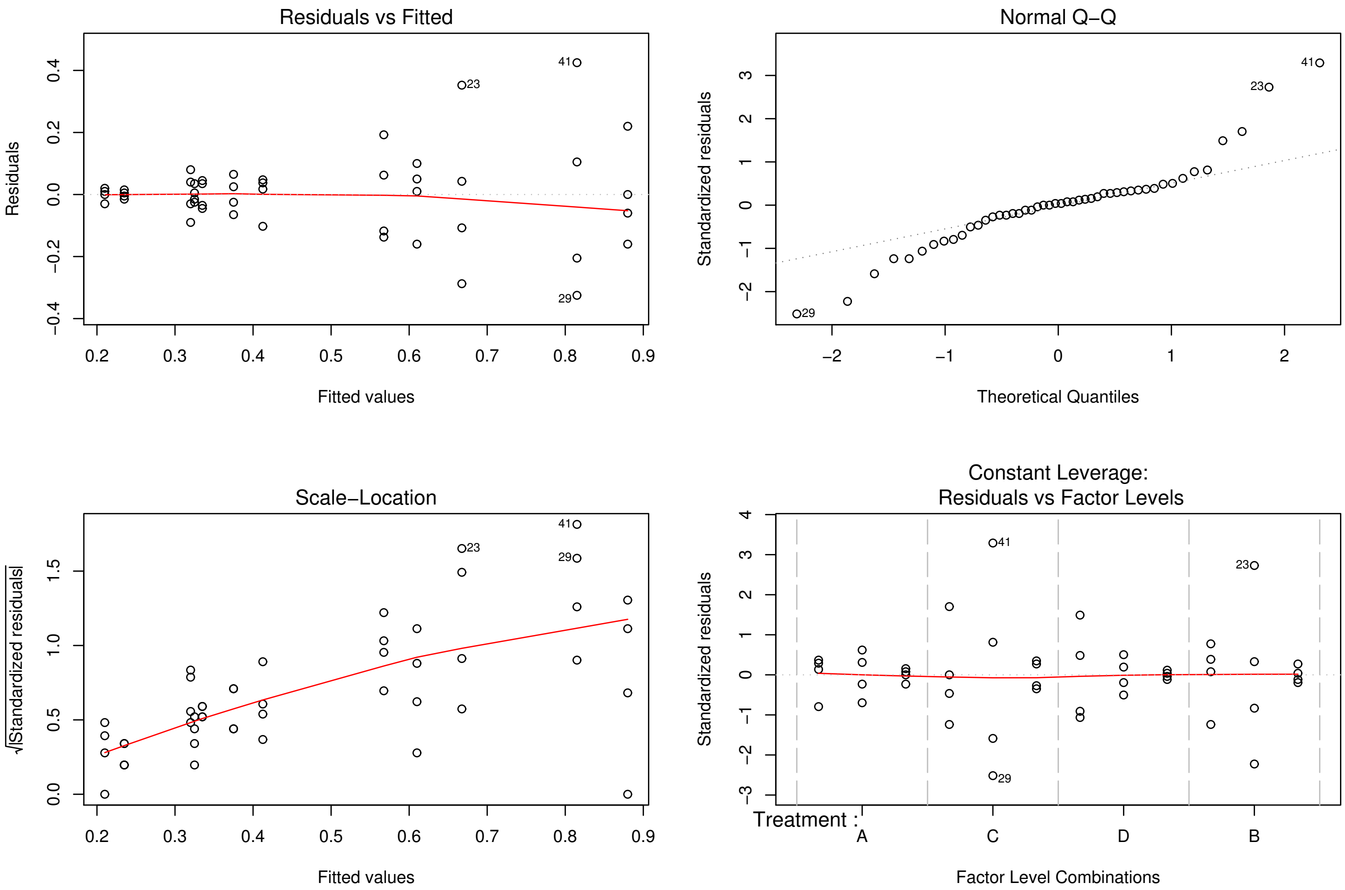


Figure 7.5: Diagnostic plots for main effects-plus-interaction fit to survival in animal factorial experiment (Example 4.4).

We also compare the ANOVA tables associated with the “main effects plus interaction fit for both suvival and its reciprocal, rate of dying.

```
# > install.packages("remotes") # you need to do this once
> remotes::install_github("tmaturi/sm2data")
> library(sm2data)
> data(animals)

> par(mfrow=c(2,2))
> plot(lm(Survival~ Poison * Treatment, data=animals))
> anova(lm(Survival ~ Poison * Treatment, data = animals))
                 Df  Sum Sq Mean Sq F value    Pr(>F)
 Poison           2 1.03301 0.51651 23.2217 3.331e-07 ***
 Treatment        3 0.92121 0.30707 13.8056 3.777e-06 ***
```

```
Poison:Treatment  6 0.25014 0.04169  1.8743    0.1123
Residuals        36 0.80073 0.02224
```

The corresponding ANOVA table for “rate of dying” fitted to the same model, given below, shows a increase in sensitivity compared with the previous analysis: the F-ratios for the main effects Poison and Treatment are larger (leading to smaller significance probabilities) and the F-value for interaction is smaller.

```
> plot(lm(1/Survival˜Poison* Treatment, data=animals))
> anova(lm(1/Survival ˜ Poison * Treatment, data = animals))
                  Df  Sum Sq  Mean Sq       F     Pr(>F)
Poison             2  34.877  17.4386  72.635  2.310e-13
Treatment          3  20.414   6.8048  28.343  1.376e-09
Poison:Treatment   6   1.571   0.2618   1.090     0.3867
Residual          36   8.643   0.2401
```

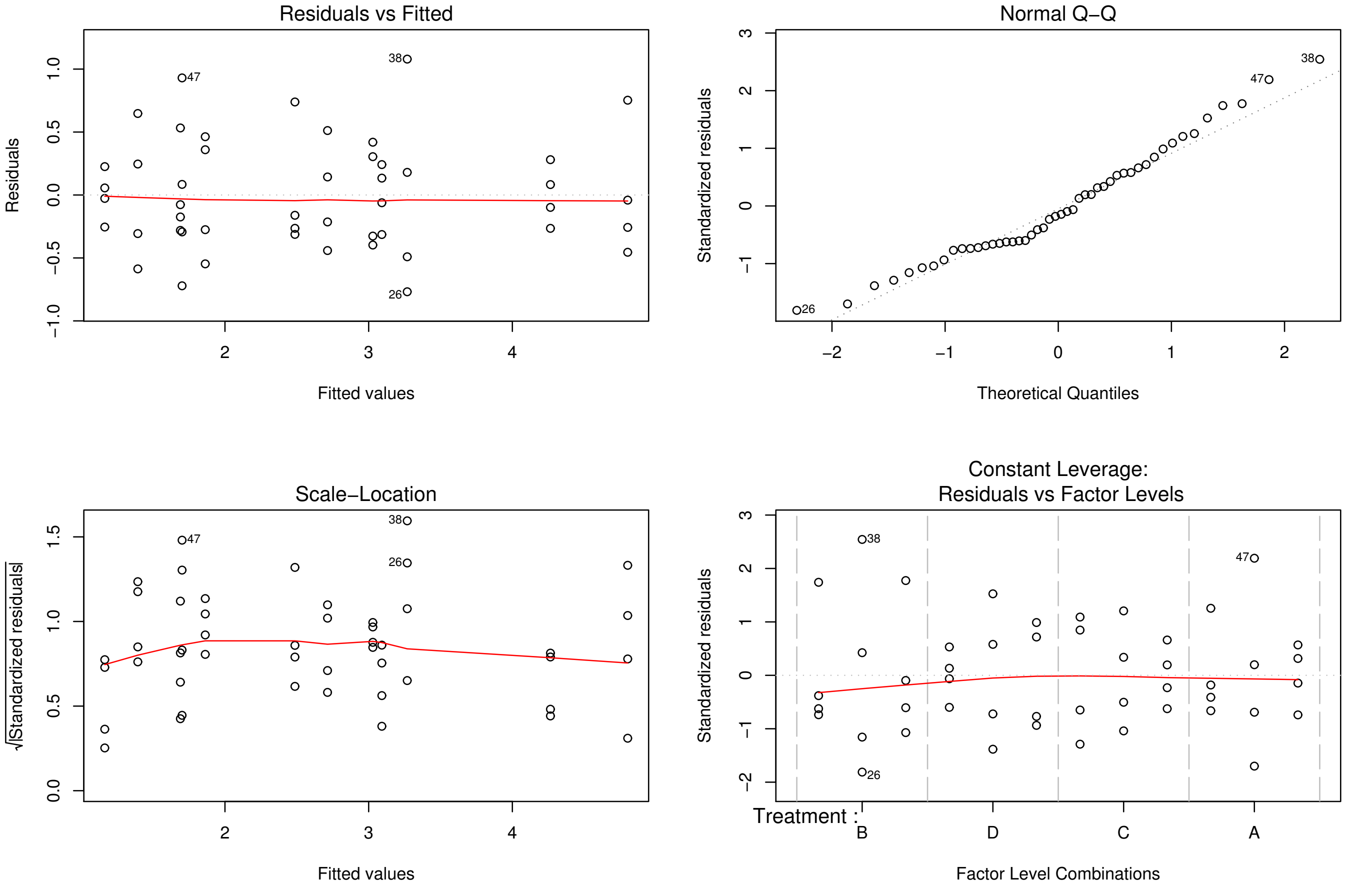


Figure 7.6: Diagnostic plots for main effects-plus-interaction fit to reciprocal survival in animal factorial experiment (Example 4.4).

This indicates that variation in survival time for the fit to “main effects plus interaction” model is significantly influenced by both factors, poison and treatment, while there is little evidence for interaction (even less than for the untransformed model).

The discussion so far indicates that the simple additive model `1/Survival` $\sim$ `Poison + Treatment` for rate of dying is well supported by the diagnostic plots and ANOVA tables. How we chose to consider rate of dying as the response will be considered in detail in Section 7.2.

Some remedies

There are many things we could do, but two common situations are:

(1) if a plot of $\hat{\epsilon}_i$ versus x_{ij} is curved, we may need to include an extra model term, or possibly transform x_j.

(2) if a plot of $\hat{\epsilon}_i$ versus $\hat{y}_i$ is trumpet shaped, we may need to transform the response y to stabilise its variance.

An interesting device to address (2) is the Box–Cox transformation.

7.2 Box-Cox transformations

Suppose we seek some power transformation of a *positive* response y of the form

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

such that $y^{(\lambda)}$ satisfies the standard model assumptions; namely,

(A1) $\mathrm{E}[y^{(\lambda)} \,|\, \boldsymbol{x}] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$

(A2') $\mathrm{Var}[y^{(\lambda)} \,|\, \boldsymbol{x}] = \sigma^2$ for all $\boldsymbol{x}$

(A3) $y^{(\lambda)}|\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)$ for all $\boldsymbol{x}$

(A2") $y_1^{(\lambda)}, \dots, y_n^{(\lambda)}$ are independent

As λ varies over $(-2, 2)$, $y^{(\lambda)}$ encompasses the reciprocal transformation ($\lambda = -1$), log ($\lambda = 0$), square root ($\lambda = \frac{1}{2}$), the original scale ($\lambda = 1$) and the square transformation ($\lambda = 2$). If the y_i are not positive we apply the transformation to $y_i + \gamma$, where γ is chosen to make all the $y_i + \gamma$ positive.

The Box-Cox method chooses λ via a likelihood argument as follows. Taking account of the Jacobian of the transformation from $y^{(\lambda)}$ to y, namely $y^{\lambda-1}$, the density of y_i is

$$f(y_i|\boldsymbol{x}_i) = \frac{y_i^{\lambda-1}}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i^{(\lambda)} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right]$$

Consequently, the log-likelihood for $\boldsymbol{\beta}, \sigma^2, \lambda$ based on independent $y_1, \dots, y_n$ is

$$L(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{1}{2}\left[n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^n \left(y_i^{(\lambda)} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] + (\lambda - 1)\sum_{i=1}^n \log y_i$$

If λ is regarded as fixed, the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}^{(\lambda)} \qquad \hat{\sigma}_\lambda^2 = \frac{RSS(\hat{\boldsymbol{\beta}}_\lambda)}{n}$$

where $RSS(\hat{\boldsymbol{\beta}}_\lambda)$ is the residual sum of squares for the regression of $\boldsymbol{Y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^{\mathrm{T}}$ on the columns of $\boldsymbol{X}$. Note when $\lambda = 1$ we get the ordinary least squares estimate of $\boldsymbol{\beta}$ and $\hat{\sigma}^2$ is a multiple of the usual estimate s^2 of σ^2. The *profile* log likelihood for λ, written here as $L_p(\lambda)$, is

$$L_p(\lambda) \equiv \max_{\boldsymbol{\beta}, \sigma^2} L(\boldsymbol{\beta}, \sigma^2, \lambda) = L(\hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}_\lambda^2, \lambda) = -\frac{n}{2}\log RSS_\lambda + (\lambda - 1)\sum \log y_i + \frac{n}{2}(\log n - 1),$$

where the latter term $\frac{n}{2}(\log n - 1)$ does not depend on λ and can therefore be omitted.

A plot of $L_p(\lambda)$ summarises the information concerning λ. An approximate $1 - \alpha$ confidence interval is the set of λ values for which

$$L_p(\lambda) \geq L_p(\hat{\lambda}) - \frac{1}{2}\chi_{1,\alpha}^2$$

where $\chi_{1,\alpha}^2$ is the quantile of the chi-squared distribution with 1 degree-of-freedom corresponding to the probability mass α in the right tail; for example, $\alpha = 0.05$ gives $\chi_{1,\,0.05}^2 = 3.84$ for a 95% interval [W, p. 289].

Example 7.4 (Continuation of Example 7.3)

Suppose we want to choose a value of λ such that a main effects only model (no interaction) plus the usual assumptions of independent Gaussian errors with constant variance for the response $\texttt{Survival}^\lambda$ is appropriate. We type

```
> library(MASS)
> boxcox(lm(Survival~Poison +Treatment, data=animals))
```

which generates Figure 7.7. The confidence interval contains −1, supporting the reciprocal transformation. The confidence interval in Figure 7.8, which shows the corresponding plot for the model using the reciprocal survival times contains 1. This gives further support to this model.

```
>   boxcox(lm((1/Survival) ~ Poison +  Treatment, data = animals))
```

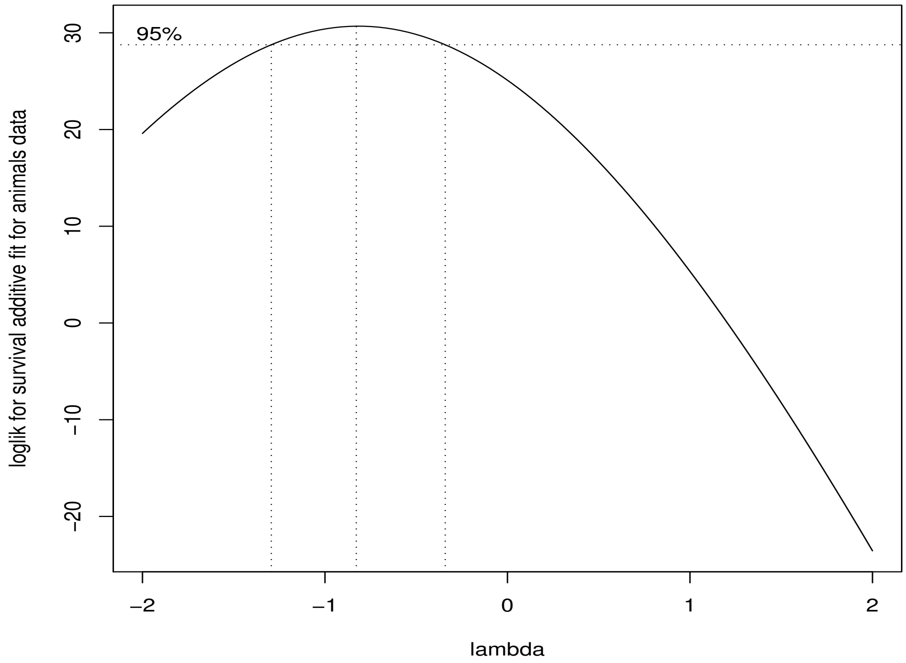


Figure 7.7: Box-Cox transformation log-likelihood plot for main effects fit to survival in animal factorial experiment.

.

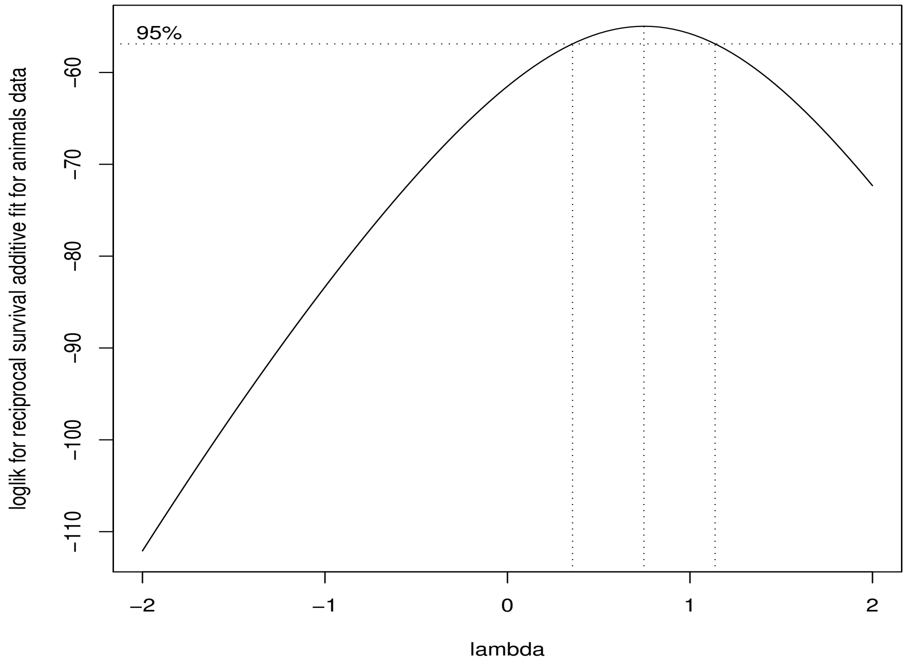


Figure 7.8: Box-Cox transformation log-likelihood plot for main effects fit to reciprocal survival in animal factorial experiment.

Appendix A

Some matrix algebra

Generally, we denote

- real-valued *matrices* by capital letters $\boldsymbol{A}, \boldsymbol{B}, \ldots$,
- real-valued *vectors* by small letters $\boldsymbol{a}, \boldsymbol{b}, \ldots$,
- *random vectors* by U, V, W, X, Y, Z, and their realizations by $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

Further,

- $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ denote *design matrices* (see Section 3), *response vectors*, and *data matrices*, respectively.

Below we provide a list of some important and useful terms and formulas from matrix algebra. Define therefore a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, a matrix $\boldsymbol{B} \in \mathbb{R}^{n \times m}$, and a vector $\boldsymbol{b} \in \mathbb{R}^n$.

- The *transpose* $\boldsymbol{B}^T$is the $m \times n$ matrix obtained from $\boldsymbol{B}$ by interchanging rows with columns;
- The *trace* $\mathrm{Tr}(\boldsymbol{A})$ is the sum of the n diagonal elements of $\boldsymbol{A}$;
- The *inverse* $\boldsymbol{A}^{-1}$ is the $n \times n$ matrix such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix having 1's along the diagonal and 0's otherwise.
- A matrix $\boldsymbol{A}$ is said to be *symmetric* if $\boldsymbol{A} = \boldsymbol{A}^T$.
- A matrix $\boldsymbol{A}$ is said to be *orthogonal* if $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, i.e. if $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.
- A symmetric matrix $\boldsymbol{A}$ is said to be *positive definite* if and only if $\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b} > 0$ for all vectors $\boldsymbol{b} \neq 0$, and *positive semi-definite* if the inequality is not strict (i.e., $\geq$). Similarly, $\boldsymbol{A}$ is said to be *negative definite* if and only if $\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b} < 0$ for all vectors $\boldsymbol{b} \neq 0$, and *negative semi-definite* if the inequality is not strict.
- The *eigenvectors* $\boldsymbol{v}_i$ and eigenvalues λ_i of a matrix $\boldsymbol{A}$ satisfy $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, $i = 1, \ldots, n$.
- The *principal minors* of a symmetric matrix $\boldsymbol{A}$ are the determinants of any $k \times k$ submatrix of $\boldsymbol{A}$ (where the submatrix is formed by removing any $q - k$ rows and the same $q - k$ columns).
- The *leading principal minors* of a symmetric matrix $\boldsymbol{A}$ are the determinants $\det(\boldsymbol{A}_{k \times k}), k = 1, \ldots, n$, where $\boldsymbol{A}_{k \times k}$ are obtained by taking only the first k rows and k columns of $\boldsymbol{A}$.
- If $\boldsymbol{A}$ is positive definite, then there exists precisely one positive definite matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ with $\boldsymbol{C}\boldsymbol{C} = \boldsymbol{A}$; we then define the *square root of a matrix* as $\boldsymbol{A}^{1/2} = \boldsymbol{C}$.

Of course, these are real-valued matrices resp. vectors too; they just have a special significance which is highlighted through an explicit denotation.

Table A.1: Properties of matrix operators. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$, $\boldsymbol{b} \in \mathbb{R}^n$.

(M1)	Let $\boldsymbol{C} \in \mathbb{R}^{n \times n}$. Then $\mathrm{Tr}(\boldsymbol{A}+\boldsymbol{C}) = \mathrm{Tr}(\boldsymbol{A}) + \mathrm{Tr}(\boldsymbol{C})$
(M2)	Let $\boldsymbol{C} \in \mathbb{R}^{m \times n}$. Then $\mathrm{Tr}(\boldsymbol{BC}) = \mathrm{Tr}(\boldsymbol{CB})$
(M3)	A special case of (M2) is: $\mathrm{Tr}(\boldsymbol{b}\boldsymbol{b}^T) = \boldsymbol{b}^T\boldsymbol{b}$
(M4)	$\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^n \lambda_i$
(M5)	$\det(\boldsymbol{A}) = \prod_{i=1}^n \lambda_i$
(M6)	$\boldsymbol{A}$ is positive (semi-) definite $\Leftrightarrow$ all eigenvalues of $\boldsymbol{A}$ are positive (non-negative)
(M7)	$\boldsymbol{A}$ is negative (semi-) definite $\Leftrightarrow$ all eigenvalues of $\boldsymbol{A}$ are negative (non-positive)
(M8)	$\boldsymbol{A}$ is positive definite $\Leftrightarrow$ $\boldsymbol{A}^{-1}$ is positive definite.
(M9)	$\boldsymbol{A}$ is invertible $\Leftrightarrow \det(\boldsymbol{A}) \neq 0$.
(M10)	$\boldsymbol{A}$ is positive definite $\Longrightarrow \det(\boldsymbol{A}) > 0$.
(M11)	Let $\boldsymbol{D} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $\det(\boldsymbol{D}) = ad - bc$
(M12)	With $\boldsymbol{D}$ as in (M11), one has $\boldsymbol{D}^{-1} = \frac{1}{\det(\boldsymbol{D})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.
(M13)	Let $\boldsymbol{a} \in \mathbb{R}^n$. Then $\frac{\partial \boldsymbol{a}^T\boldsymbol{b}}{\partial \boldsymbol{b}} = \frac{\partial \boldsymbol{b}^T\boldsymbol{a}}{\partial \boldsymbol{b}} = \boldsymbol{a}$.
(M14)	if $\boldsymbol{A}$ is symmetric, then $\frac{\partial \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}}{\partial \boldsymbol{b}} = 2\boldsymbol{A}\boldsymbol{b}$.

- The *determinant* $\det(\boldsymbol{A})$ is a rather complex function mapping a square matrix $\boldsymbol{A}$ to a scalar (geometrically, this is the volume of the parallelepiped spanned by the columns or rows of the matrix); see [K], Appendix A3.

- The *rank* of a matrix $\boldsymbol{B}$, denoted by $\mathrm{rank}(\boldsymbol{B})$, is defined as the largest number of linearly independent column vectors (or, equivalently, rows) in $\boldsymbol{B}$. This corresponds just to the dimension of the vector space spanned by the columns (or rows) of $\boldsymbol{B}$.

- The *derivative* (gradient) of a scalar function $g(\boldsymbol{b}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ w.r.t. $\boldsymbol{b}$ is defined as the vector $\frac{\partial g(\boldsymbol{b})}{\partial \boldsymbol{b}} = \left(\frac{\partial g(\boldsymbol{b})}{\partial b_1}, \ldots, \frac{\partial g(\boldsymbol{b})}{\partial b_n}\right)^T$.

Some important properties of these operators are summarized in Table A.1. A generally useful resource, where these (and many other) properties of matrices can be found, is the Matrix Cookbook [MC].