KU LEUVEN kulak

Wetenschap & Technologie Kulak

# Statistische modellen en data-analyse

## Bachelor in de wiskunde, Bachelor in de chemie

Prof. dr. Mia Hubert & Prof. dr. Stefan Van Aelst

KU Leuven, Departement wiskunde

Academiejaar 2018-2019

# Contents

# Part I

# Multivariate Statistics

# Chapter 1

# Beschrijvende multivariate statistiek

## 1.1 Vectoriële notatie

Bij *univariate* statistiek zijn de gegevens van de vorm $x_1, x_2, \ldots, x_n$ met alle $x_i \in \mathbb{R}$.

Enkele veelgebruikte beschrijvende statistieken zijn in dat geval

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{het steekproefgemiddelde} \\
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{de empirische variantie} \\
s &= \sqrt{s^2} \quad \text{de empirische standaardafwijking}
\end{aligned}$$

Grafische voorstellingen van een univariate steekproef zijn o.a. het histogram en de boxplot.

Bij *multivariate* statistiek, hebben we $n$ objecten en metingen voor $p$ variabelen. Deze brengt men samen in een matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ waarbij de rijen de objecten voorstellen en elke kolom een variabele weergeeft.

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Bijgevolg is

$$x_{ij} = \text{meting voor variabele } j \text{ voor het } i\text{-de object.}$$

Dit schrijft men ook vaak als

$$\boldsymbol{X} = (\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_n)^{\tau}$$

waar de objecten voorgesteld worden als *kolomvectoren* $\in \mathbb{R}^{p\times 1}$:

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

1. Het steekproefgemiddelde

   Voor elke variabele kunnen we het univariate steekproefgemiddelde berekenen:

   $$\overline{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

   De combinatie geeft het multivariate steekproefgemiddelde

   $$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \begin{pmatrix} \overline{x}_{.1} \\ \overline{x}_{.2} \\ \vdots \\ \overline{x}_{.p} \end{pmatrix} \in \mathbb{R}^{p\times 1}$$

2. De empirische covariantiematrix

Voor elk van de variabelen bepalen we de univariate empirische variantie, en voor elk paar van variabelen de empirische covariantie:

$$\begin{aligned} s_{jj} &:= s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})^2 \\ s_{jk} &:= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})(x_{ik}-\overline{x}_{.k}) \end{aligned}$$

Merk op dat $s_{jj}$ ook als speciaal geval van $s_{jk}$ kan gezien worden (als $j = k$).

Dit levert de empirische covariantiematrix $\boldsymbol{S}$ die steeds symmetrisch is:

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix} \in \mathbb{R}^{p\times p}$$

Vaak gebruikt men ook de notatie

$$\boldsymbol{W} = (n-1)\boldsymbol{S} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & & & \\ w_{p1} & w_{p2} & \dots & w_{pp} \end{pmatrix}$$

met

$$\begin{aligned} w_{jj} &= \sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})^2 && \text{“kwadratensom”} \\ w_{jk} &= \sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})(x_{ik}-\overline{x}_{.k}) && \text{“som van kruisproducten”} \end{aligned}$$

Om deze reden wordt de matrix $\boldsymbol{W}$ ook wel *SSCP matrix* genoemd (“Sum of Squares and Cross Products”).

**Result.**

$$\boldsymbol{W} = \sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^{\tau}$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^{\tau}$$

3. De empirische correlatiematrix

De correlatie tusen twee variabelen is gedefinieerd als

$$r_{jk} := \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})(x_{ik}-\overline{x}_{.k})}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\overline{x}_{.j})^2}\sqrt{\sum_{i=1}^{n}(x_{ik}-\overline{x}_{.k})^2}}$$

Hierbij geldt dat $r_{jk} = r_{kj}$, $r_{jj} = 1$ en $-1 \leqslant r_{jk} \leqslant 1$. Zo verkrijgen we de empirische correlatiematrix

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix} \in \mathbb{R}^{p\times p}$$

Merk op dat

$$\boldsymbol{R} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$$

waarbij $\boldsymbol{D}^{-\frac{1}{2}} = diag(\frac{1}{\sqrt{s_{11}}}\dots,\frac{1}{\sqrt{s_{pp}}})$.

Een alternatieve manier om $\boldsymbol{R}$ te bekomen is via de gestandaardiseerde variabelen

$$z_{ij} = \frac{x_{ij}-\overline{x}_{.j}}{\sqrt{s_{jj}}} \qquad \text{“}z\text{-scores”}$$

Voor de gestandaardiseerde gegevens $\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_n\}$ wordt

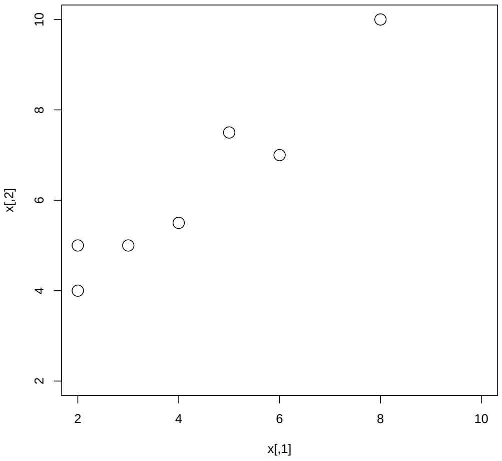
$$\begin{aligned} \overline{z}_{.j} &= 0 \\ (s_z)_{jj} &= \frac{s_{jj}}{s_{jj}} = 1 \\ (s_z)_{jk} &= \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = r_{jk} \end{aligned}$$

**Result.**

$$\boldsymbol{R}(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) = \boldsymbol{S}(\boldsymbol{z}_1,\dots,\boldsymbol{z}_n)$$

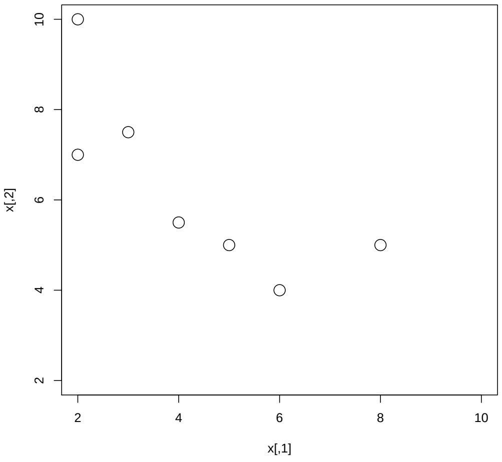
## 1.2 Grafische voorstellingen

Voor bivariate ($p = 2$) gegevens kunnen we een scatterplot maken.



$$\boldsymbol{x} = \begin{pmatrix} 3 & 5 \\ 4 & 5.5 \\ 2 & 4 \\ 6 & 7 \\ 8 & 10 \\ 2 & 5 \\ 5 & 7.5 \end{pmatrix}$$

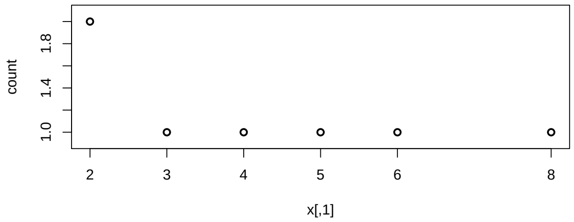
De correlatie tussen beide variabelen is duidelijk positief. Als we de eerste variabele ongewijzigd laten en de tweede variabelen herordenen, verandert de scatterplot:

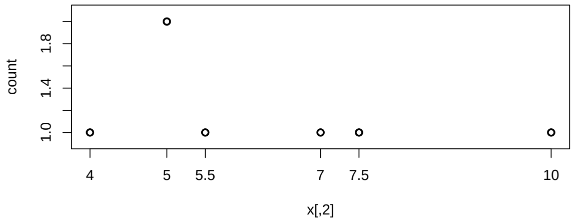


$$\boldsymbol{x} = \begin{pmatrix} 3 & 7.5 \\ 4 & 5.5 \\ 2 & 7 \\ 6 & 4 \\ 8 & 5 \\ 2 & 10 \\ 5 & 5 \end{pmatrix}$$

Nu is de correlatie duidelijk negatief.

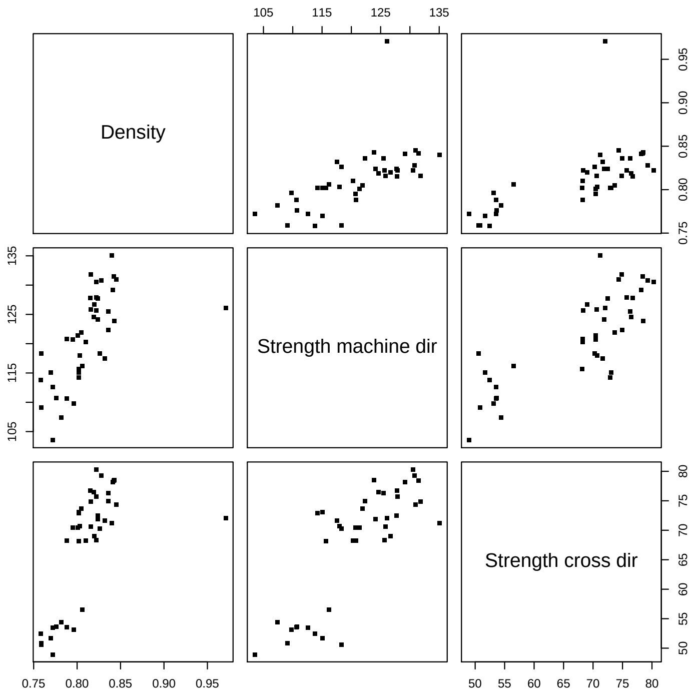
We zouden ook twee afzonderlijke grafieken (bv. dotplots) van de marginalen kunnen construeren.





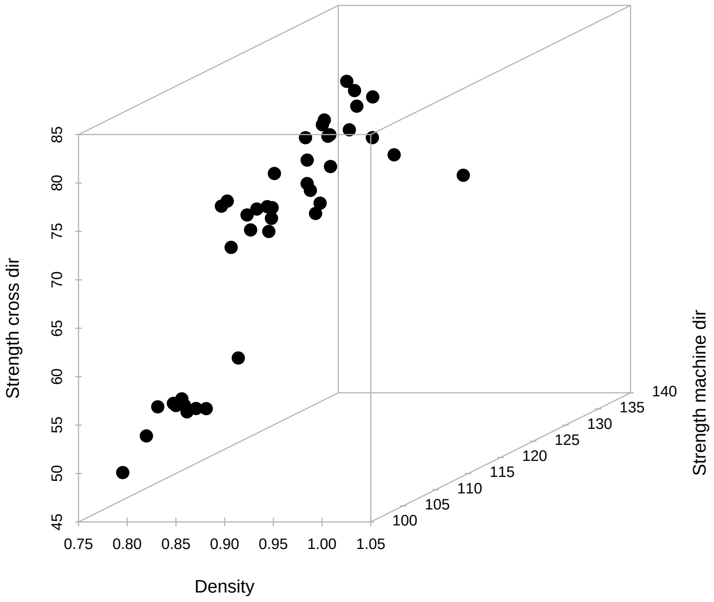
Voor beide bovenstaande scatterplots levert dat hetzelfde paar dotplots. Kennis van de marginalen is dus niet voldoende om de scatterplot te reconstrueren. Vertrekkende van de marginalen kunnen we wel $\overline{x}_{.1}, s_{11}, \overline{x}_{.2}, s_{22}$ berekenen. Voor de covariantie $s_{12}$ is meer informatie nodig (in beide bovenstaande scatterplots is het teken van de covariantie duidelijk verschillend).

**Voorbeeld:** papierkwaliteit ($p = 3$). Drie variabelen meten de dichtheid van het papier en de stevigheid in de machinerichting zowel als in de richting daar loodrecht op (bron: SONOCO Products, Inc.). We maken een matrix van scatterplots, geordend zoals in de covariantiematrix.



Merk op: er is 1 duidelijke “uitschieter”, een observatie waarvoor de eerste variabele een abnormaal hoge waarde heeft. Bovendien lijken de gegevens verdeeld over twee groepen volgens de derde variabele.

Een driedimensionale scatterplot geeft



# Chapter 2

# Cluster analysis

## 2.1 Data heterogeneity and cluster analysis

Statistical models assume that all data follow a particular model. However, when the data is generated from a heterogeneous population, it is likely that observations from different subpopulations behave differently and thus follow a different statistical model. Before applying a statistical model, it is thus important to investigate whether the dataset is homogeneous or not. If the data turns out to be heterogeneous, then the different subpopulations may need to be analyzed separately.

Cluster analysis is a statistical technique to explore the heterogeneity of a dataset. Its objective is to search for groups, or *clusters*, within a dataset, in such a way that objects within the same cluster resemble each other in some manner, and that objects in different clusters are *dissimilar*, do not resemble each other. If there are only two or three variables in the sample, then it is possible to detect the groups visually. When the number of variables increases however, we need an algorithm to perform the cluster analysis.

Broadly speaking, there are two types of clustering algorithms. The first type, the *partitioning* algorithms, divide the sample into $K$ clusters, where the number of clusters $K$ needs to be specified by the user. The second type, the *hierarchical* algorithms, result in a full hierarchy of clusters of the sample, and the user can decide the number of clusters to use based on this hierarchy.

## 2.2 Dissimilarity matrices

In cluster analysis, a sample can have two very distinct structures. First of all, the data can be represented in the usual manner, using a $n \times p$ data matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

where each of the $n$ rows corresponds to an observation and each of the $p$ columns corresponds to a variable. Alternatively, it can happen that the actual objects are unobserved, but that the "differences" or *dissimilarities* between each pair of objects are known. These dissimilarities can be represented in a $n \times n$ dissimilarity matrix

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & & & \ddots & \\ d(n,1) & & \cdots & & 0 \end{pmatrix}$$

where $d(i,i') = d(i',i)$ is the dissimilarity between objects $i$ and $i'$. For the remainder of the chapter, we will use the term observation for an object when all its variables are observed.

**Def.** The **dissimilarity** $d(i,i')$ measures how different objects $i$ and $i'$ are. This dissimilarity must satisfy the following three axioms:

1. $d(i,i) = 0$: the dissimilarity of an object to itself is zero.
2. $d(i,i') \geqslant 0$: dissimilarities are non-negative.
3. $d(i,i') = d(i',i)$: dissimilarities are symmetric.

Note that this does not imply that a dissimilarity is a metric, because the triangle inequality doesn't necessarily hold.

**Example: dissimilarity**

If the variables in the dataset are all observed quantitative variables, we can use an actual metric for the dissimilarity, such as

$$d(i,i') = \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|_2 = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2} \qquad \text{(Euclidean distance), or}$$

$$d(i,i') = \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|_1 = \sum_{j=1}^{p}|x_{ij} - x_{i'j}| \qquad \text{(Manhattan distance).}$$

Note that this makes the dissimilarities strongly dependent on the choice of measurement units of the variables, and that the variable with the largest variance will influence the clustering the most. As such, the data need to be *standardised* in advance if all variables are considered equally important:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad \text{($z$-scores)}$$

where $\bar{x}_j$ and $s_j^2$ are the sample mean and sample variance of the $j$-th variable. Examples of such variables are physical measurements like height, weight, etc.

If the variables are of another type, or if there are several types of variables present in the sample, we can still compute dissimilarities, but using different techniques. We will not go into detail about these methods in this course, but interested readers can consult Kaufman & Rousseeuw (1990), among others.

## 2.3 $K$-means clustering

The *$K$-means clustering* algorithm is one of the most popular, and intuitive, clustering algorithm. This algorithm is intended for situations where all the variables in the sample are observed quantitative variables, and where the dissimilarity measure is the Euclidean distance

$$d(i,i') = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2} = \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|_2.$$

Once again, we can standardise the variables to eliminate the method's sensitivity to changes in the variables' measurement units.

The goal of the $K$-means clustering algorithm is to minimize the sum of the squared Euclidean distances of all data points to the center of their cluster, which is given by the mean of all the observations of that cluster. We thus try to minimize the objective function

$$L(C_1,\ldots,C_K) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2$$

with $\boldsymbol{m}_k$ the mean of the observations of cluster $C_k$.

This idea can be translated to the following algorithm:

1. Choose $K$ initial points $\boldsymbol{m}_1,\ldots,\boldsymbol{m}_K$ in the sample space. These are the initial estimates for the means of the clusters and do not have to correspond to actual data points.

2. Given a current set of means $\{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_K\}$, assign each observation $\boldsymbol{x}_i$ to the cluster $C_k$ corresponding to the closest current cluster mean, so

$$k = \underset{1\leqslant k\leqslant K}{\operatorname{argmin}} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2,$$

3. Compute the mean $\boldsymbol{m}_k$ for each of the clusters $C_k$, because

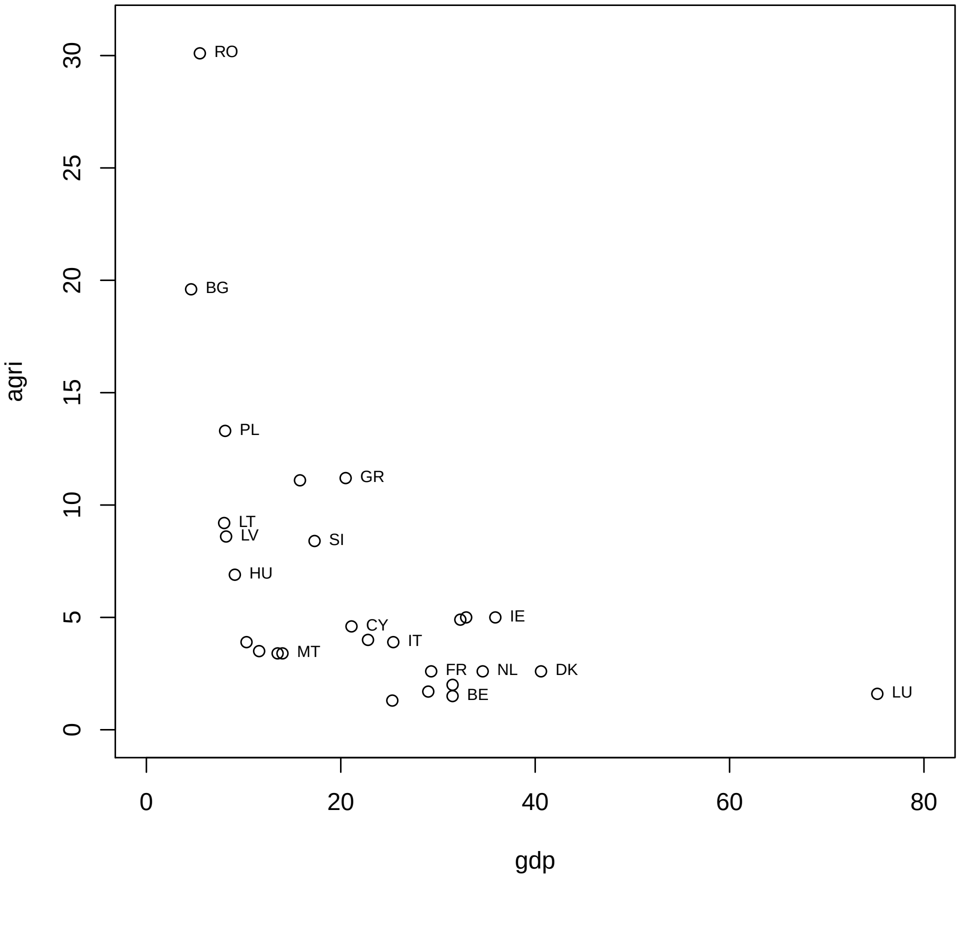
$$\boldsymbol{m}_k = \underset{\boldsymbol{m}}{\operatorname{argmin}} \sum_{\boldsymbol{x}_i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}\|_2^2.$$

4. Repeat steps 2 and 3 until the cluster assignments in step 2 do not change anymore.

This algorithm will always converge, since the value of the objective function will decrease every time steps 2 and 3 are performed, and the function is non-negative. However, the algorithm does not guarantee convergence to the optimal solution. As such, it is recommended to perform the algorithm several times with different initial values for $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ in step 1, and choose the solution which has the lowest resulting value for the objective function. Also note that the optimal solution is not unique, because a permutation of the clusters and the cluster means would not alter the value of the objective function.

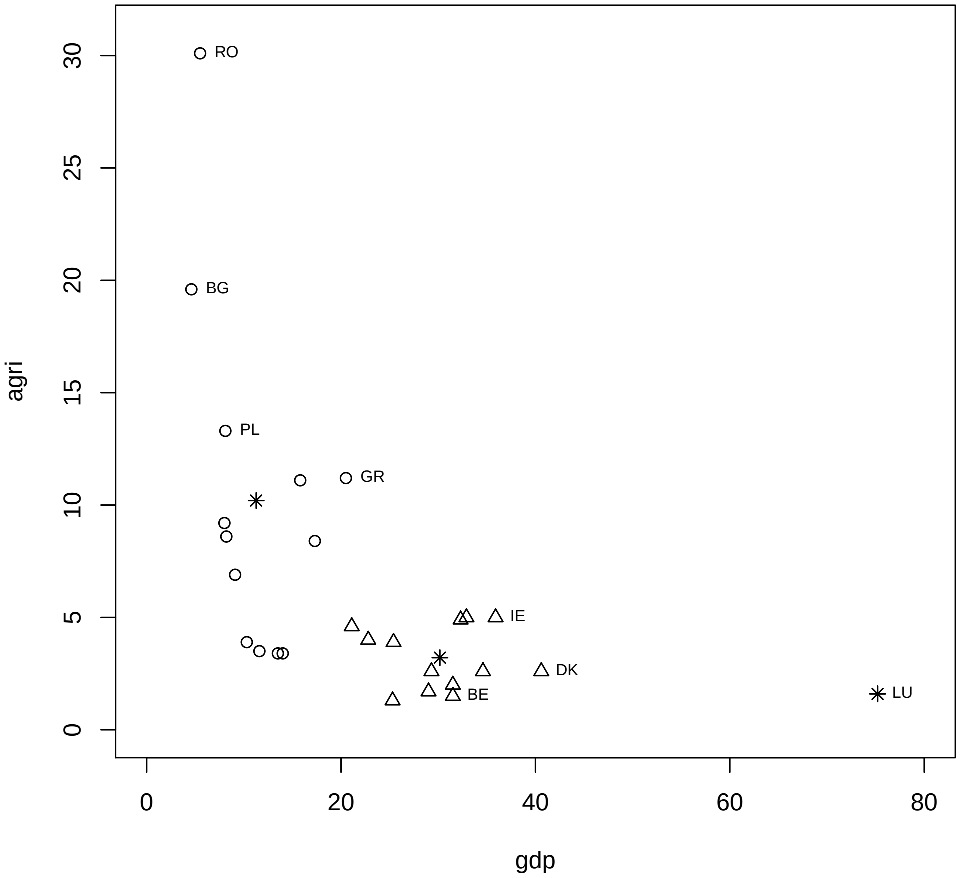
**Example: EU data**

We have data on the gross domestic product per capita (`gdp`, in 1000 euros) and the percentage of the working population employed in the agricultural sector for all the EU states in the year 2009, and we wish to see if there are clear groups in the data. Below is a scatterplot of these data.



Visually, we would say that there are three groups in the data: the country on the far right of the plot, Luxemburg, and the two or three countries in the top left of the plot, Romania, Bulgaria, and possibly Poland as well. Hence, we will try to find $K = 3$ clusters with the $K$-means algorithm.

```
clust.kmeans3 <- kmeans(agri09, 3, iter.max = 25, nstart = 50)
clust.kmeans3

K-means clustering with 3 clusters of sizes 13, 13, 1

Cluster means:
       gdp      agri
1 11.26923 10.200000
2 30.16915  3.207692
3 75.20000  1.600000

Clustering vector:
BE BG CZ DK DE EE IE GR ES FR IT CY LV LT LU HU MT NL AT PL PT RO
 2  1  1  2  2  1  2  1  2  2  2  2  1  1  3  1  1  2  2  1  1  1
SI SK FI SE UK
 1  1  2  2  2

Within cluster sum of squares by cluster:
[1] 958.7477 385.6727   0.0000
 (between_SS / total_SS =  80.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"

plot(agri09, pch = clust.kmeans3$cluster, xlim = c(0, 80), ylim = c(0, 31))
points(clust.kmeans3$centers, pch = 8)
```



We see that Luxemburg is assigned to its own cluster $C_1$, but that Romania, Bulgaria and Poland are not assigned to their own cluster. Instead, several other nations are grouped together with those three, such as Greece, Latvia, etc. The clustering result is visualized in the scatterplot. The stars in the plot represent the cluster means, and the circle, the triangles, and the plusses represent the three clusters. We see that the countries which have a low GDP per capita (below 20 000 euros) are in one cluster, the 'richer' countries except Luxemburg in the second, and Luxemburg has its own cluster.

We can also try the algorithm with $K = 4$ clusters, and obtain:

```
clust.kmeans4 <- kmeans(agri09, 4, iter.max = 25, nstart = 50)
clust.kmeans4
```

```
K-means clustering with 4 clusters of sizes 1, 2, 12, 12

Cluster means:
       gdp      agri
1 75.20000  1.600000
2  5.05000 24.850000
3 30.92492  3.091667
4 13.12500  7.291667

Clustering vector:
BE BG CZ DK DE EE IE GR ES FR IT CY LV LT LU HU MT NL AT PL PT RO
 3  2  4  3  3  4  3  4  3  3  3  4  4  4  1  4  4  3  3  4  4  2
SI SK FI SE UK
 4  4  3  3  3

Within cluster sum of squares by cluster:
[1]   0.0000  55.5300 294.4689 381.7917
 (between_SS / total_SS =  89.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Now, we observe that Luxemburg still has its own cluster, and that Romania and Bulgaria are in a cluster of their own as well, cluster $C_3$. Compared to the clustering with three clusters, Cyprus has now been assigned to the ‘poorer’ countries instead of the ‘richer’ countries.

## 2.4 Partitioning Around Medoids

### 2.4.1 Clustering algorithm

As illustrated in the previous section, $K$-means clustering is one of the most intuitive clustering algorithms, but it has three serious drawbacks:

1. the algorithm cannot be used if the data include non-quantitative variables, for which the Euclidean distance is not appropriate,
2. the algorithm cannot be used if the data are given in the form of a dissimilarity matrix instead of the actual observations,
3. as it is based on the squared Euclidean distances, it is very sensitive to outliers.

To overcome these issues, we can impose that the cluster centers are observations themselves. This overcomes the first two issues because we can either compute dissimilarities between observations even if not all variables in the dataset are quantitative or because we already have the dissimilarities available. The third issue is overcome by trying to maximize the sum of the dissimilarities, instead of the squared dissimilarities.

Basically, the idea behind this method is to choose $K$ representative objects $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ from the dataset, called the *medoids*, one for each cluster, such that the objective function

$$\sum_{i=1}^{n} \min_{j=1,\ldots,K} d(i, \boldsymbol{m}_j) \tag{2.4.1}$$

is minimised, where $d(i, \boldsymbol{m}_j)$ is the dissimilarity between the $i$-th observation in the sample and medoid $\boldsymbol{m}_j$. This amounts to minimising the sum of the dissimilarities of all observations to their nearest medoid. Once the medoids have been determined, assign each object $i$ to the cluster $C_j$ for which

$$d(i, \boldsymbol{m}_j) \leqslant d(i, \boldsymbol{m}_{j'}) \text{ for all } j' = 1, \ldots, K.$$

This clustering technique is called the *k-medoid* method.

The *partitioning around medoids* (PAM) algorithm that determines the medoids proceeds in the following steps:

1. Initialisation phase (or BUILD step):
   - Medoid $\boldsymbol{m}_1$ is the object for which $\sum_{i=1}^{n} d(i, \boldsymbol{m}_1)$ is smallest, and set $\boldsymbol{m}_2$ to $\boldsymbol{m}_K$ equal to $\boldsymbol{m}_1$.
   - Set medoid $\boldsymbol{m}_2$ equal to the object that decreases the objective function (2.4.1) the most…
   - Set medoid $\boldsymbol{m}_K$ equal to the object that decreases the objective function (2.4.1) the most.
2. For each pair of objects $(i, \boldsymbol{m}_j)$ with $i \notin \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K\}$ and $j = 1, \ldots, K$, determine how much the objective function would change if $\boldsymbol{m}_j$ were replaced with object $i$, and set medoid $\boldsymbol{m}_j$ equal to $i$ for the pair which decreases the objective function the most. If no replacement would decrease the objective function, do not perform any replacement. This is the SWAP step.
3. Repeat step 2 until no replacement takes place.

**Simple example**

Assume that we have a dataset consisting of five objects $a$, $b$, $c$, $d$, and $e$, and that the data is represented as a matrix of dissimilarities:

| | $a$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 0 | | | | |
| $b$ | 2 | 0 | | | |
| $c$ | 6 | 3 | 0 | | |
| $d$ | 8 | 7 | 5 | 0 | |
| $e$ | 9 | 6 | 5 | 4 | 0 |

We wish to perform clustering with $K = 2$ clusters, using the partitioning around medoids algorithm. First, for each object $i$, compute the sum of the dissimilarities of that object to all objects in the dataset. This gives

$$\begin{array}{lll} a : 2+6+8+9 = 25 & b : 2+3+7+6 = 18 & c : 6+3+5+5 = 19 \\ d : 8+7+5+4 = 24 & e : 9+6+5+4 = 24, & \end{array}$$

so we choose object $b$ as our first initial medoid $\boldsymbol{m}_1$, with $\sum_i d(i, \boldsymbol{m}_1) = 18$. Next, compute for each object $a$, $c$, $d$, $e$ how much the objective function (2.4.1) would change if a second cluster is added around that object.

| Medoids | Clusters | | (2.4.1) |
|---|---|---|---|
| $b, a$ | $\{b, c, d, e\}$ | $\{a\}$ | $3 + 7 + 6 = 16$ |
| $b, c$ | $\{a, b\}$ | $\{c, d, e\}$ | $2 + 5 + 5 = 12$ |
| $b, d$ | $\{a, b, c\}$ | $\{d, e\}$ | $2 + 3 + 4 = 9$ |
| $b, e$ | $\{a, b, c\}$ | $\{d, e\}$ | $2 + 3 + 4 = 9$ |

We observe that we achieve the largest decrease if we choose object $d$ or $e$ as the medoid of the second cluster, so let's take $\boldsymbol{m}_2 = d$. Now we have obtained our initial clusters $C_1 = \{a, b, c\}$ and $C_2 = \{d, e\}$.

In the next step, we replace medoids with objects not in the set of medoids until we cannot decrease the objective function (2.4.1) any further.

| Replacement | New medoids | New clusters | | (2.4.1) |
|---|---|---|---|---|
| $\boldsymbol{m}_1 \to a$ | $a, d$ | $\{a, b\}$ | $\{c, d, e\}$ | $2 + 5 + 4 = 11$ |
| $\boldsymbol{m}_1 \to c$ | $c, d$ | $\{a, b, c\}$ | $\{d, e\}$ | $6 + 3 + 4 = 13$ |
| $\boldsymbol{m}_1 \to e$ | $e, d$ | $\{b, c, e\}$ | $\{a, d\}$ | $6 + 5 + 8 = 19$ |
| | | $\{b, e\}$ | $\{a, c, d\}$ | $6 + 8 + 5 = 19$ |
| $\boldsymbol{m}_2 \to a$ | $b, a$ | $\{b, c, d, e\}$ | $\{a\}$ | $3 + 7 + 6 = 16$ |
| $\boldsymbol{m}_2 \to c$ | $b, c$ | $\{a, b\}$ | $\{c, d, e\}$ | $2 + 5 + 5 = 12$ |
| $\boldsymbol{m}_2 \to e$ | $b, e$ | $\{a, b, c\}$ | $\{d, e\}$ | $2 + 3 + 4 = 9$ |

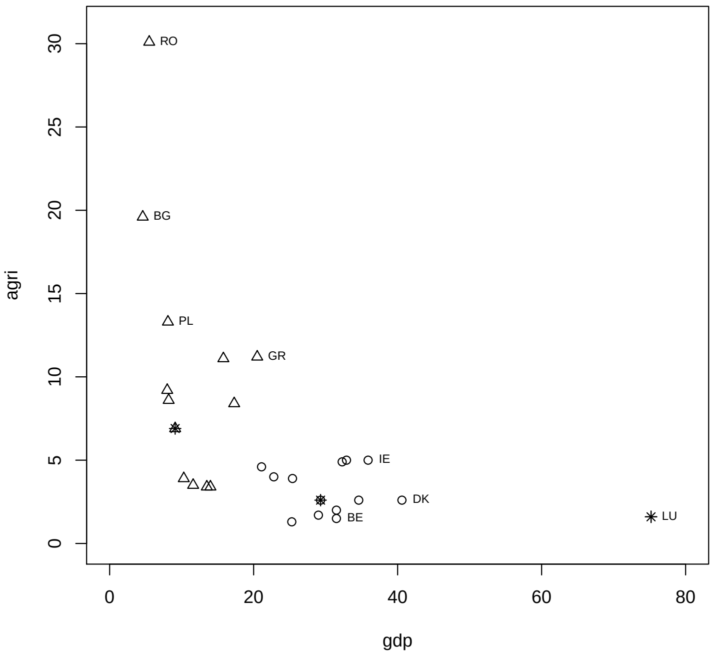
In this case, we see that there is one possible replacement which will not change the value of the objective function, but no replacement will decrease it below 9. Therefore, we terminate the algorithm and select $C_1 = \{a, b, c\}$ and $C_2 = \{d, e\}$, with medoids $\boldsymbol{m}_1 = b$ and $\boldsymbol{m}_2 = d$ respectively, as the clusters in the data.

### Example: EU data

We revisit the EU data, but this time, we will search for $K = 3$ clusters using the partitioning around medoids algorithm.

```
clust.pam3 <- pam(agri09, 3)
clust.pam3

Medoids:
   ID  gdp agri
```

```
FR 10 29.3  2.6
HU 16  9.1  6.9
LU 15 75.2  1.6
Clustering vector:
BE BG CZ DK DE EE IE GR ES FR IT CY LV LT LU HU MT NL AT PL PT RO
 1  2  2  1  1  2  1  2  1  1  1  1  2  2  3  2  2  1  1  2  2  2
SI SK FI SE UK
 2  2  1  1  1
Objective function:
   build     swap
6.326925 5.786994

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"
 [5] "isolation"  "clusinfo"   "silinfo"    "diss"
 [9] "call"       "data"

plot(agri09, pch = clust.pam3$cluster, xlim = c(0, 80), ylim = c(0, 31))
points(clust.pam3$medoids, pch = 8)
```



We find the same clusters as the one we obtained with the 3-means algorithm: the ‘poorer’ countries, centered around Hungary, the ‘richer’ countries, centered around France, and Luxemburg by itself. We also see that after completing the build step of the algorithm, the objective function was 6.33, while the swap step of the algorithm reduced it further to 5.79. The mediods are indicated by stars in the scatterplot.

## 2.5 Graphical representations

For datasets with two or three variables, and where all variables are observed, we can make a plot of the data and indicate, by various choices of markers, to which cluster we have assigned each observation. However, if the data are high-dimensional, or if the data are given in the form of a dissimilarity matrix, we can not use a classical or three-dimensional scatterplot to visualise the observations and the clusters to which they belong. However, a graphical representation of these data is still desirable, hence we illustrate two techniques for this purpose: the silhouette plot and the clusplot.

### 2.5.1 Silhouette plot

For the silhouette plot, we first have to compute the *silhouette value* $s(i)$ of each object $i$. This value shows how well the object belongs to the cluster $C_j$ to which it is assigned.

1. First, compute the average dissimilarity of the object $i$ to all other objects in its assigned cluster $C_j$:

$$a(i) = \frac{1}{|C_j| - 1} \sum_{i' \in C_j, i' \neq i} d(i, i'), \tag{2.5.1}$$

   where $|C_j|$ is the number of objects in cluster $C_j$. If the cluster $C_j$ contains only the object $i$, we set $a(i) = 0$.

2. Next, consider any cluster $C_{j'}$ with $j' \neq j$ and define the average dissimilarity of object $i$ to all objects in cluster $C_{j'}$:

$$d(i, C_{j'}) = \frac{1}{|C_{j'}|} \sum_{i' \in C_{j'}} d(i, i'). \tag{2.5.2}$$

3. Denote $b(i) = \min_{j' \neq j} d(i, C_{j'})$ and define the *neighbour* of object $i$ as the cluster $C_{j'}$ for which $d(i, C_{j'}) = b(i)$. This is the second-best cluster for object $i$, according to the clustering method used.

4. The silhouette value $s(i)$ of object $i$ can then be defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1].$$

If this value $s(i)$ is close to one, the object $i$ is well classified in cluster $C_j$; if the value $s(i)$ is close to zero, the object $i$ lies intermediate between its assigned cluster $C_j$ and its neighbour, and the assignment isn't clear cut. If the value $s(i)$ is close to $-1$ however, the object $i$ is badly classified.

The **silhouette** of a cluster $C_j$ is a plot of the silhouette values $s(i)$, ranked in decreasing order, of all objects $i$ in the cluster. If the silhouettes for all clusters are put next to each other, we obtain the **silhouette plot**. This allows us to easily compare the quality of the different clusters.

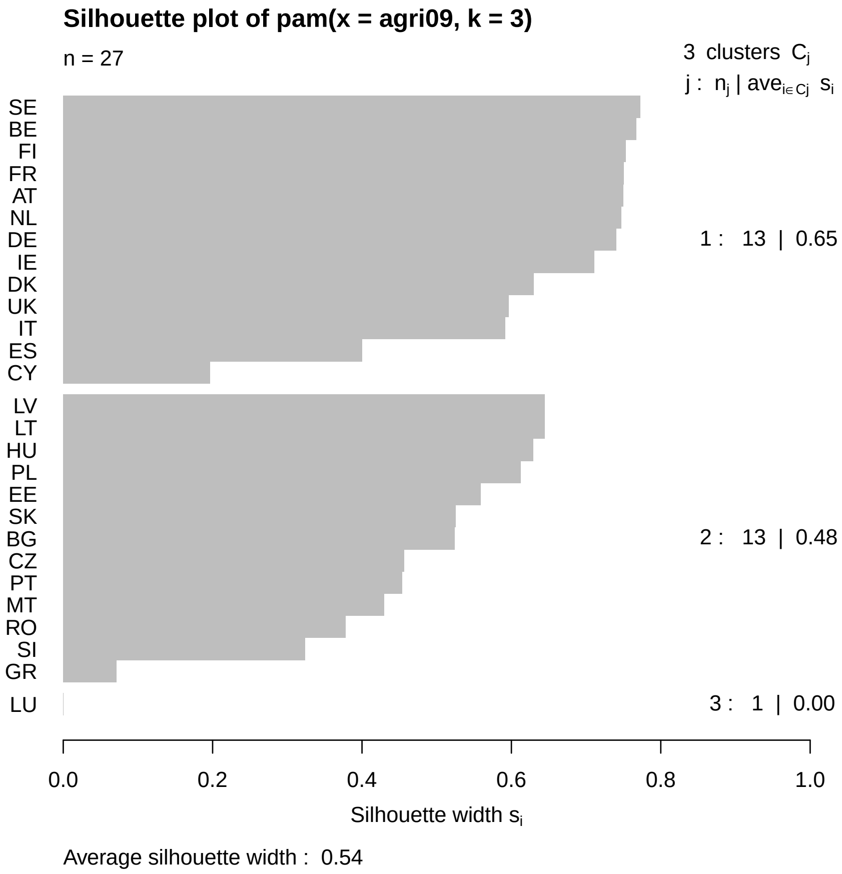
The *overall silhouette width* of the silhouette plot is the average of the silhouette values $s(i)$ of every object $i$ in the dataset. To obtain the *silhouette coefficient*, run the partitioning around medoids algorithm for several values of $K$, and compare the resulting silhouette width. The highest of these is the silhouette coefficient (SC), and indicates how much structure the data contains, as indicated in the table below. If a weak structure is detected, it is advisable to try additional clustering methods.

| SC | Interpretation |
|---|---|
| 0.71 – 1.00 | Strong structure |
| 0.51 – 0.70 | Reasonable structure |
| 0.26 – 0.50 | Weak or artificial structure |
| $\leqslant 0.25$ | No substantial structure |

### Example: EU data

We revisit the EU data example. For the 3-medoid PAM clustering we obtain the silhouette plot:

```
plot(silhouette(clust.pam3))
```



On the bottom of the plot, we see the average silhouette width of 0.54, indicating that we have found a reasonable structure in the data. We also observe that the assignment of Cyprus (CY) to the cluster of ‘richer’ countries is a bit of a borderline case, and that the assignment of Greece (GR) to the cluster of ‘poorer’ countries is also a borderline case, as indicated by their low silhouettes.

### 2.5.2 Clusplot

The clusplot is a technique that creates a two-dimensional graphical representation of the data, and the clusters that have been detected, by making an appropriate approximation of the data first.

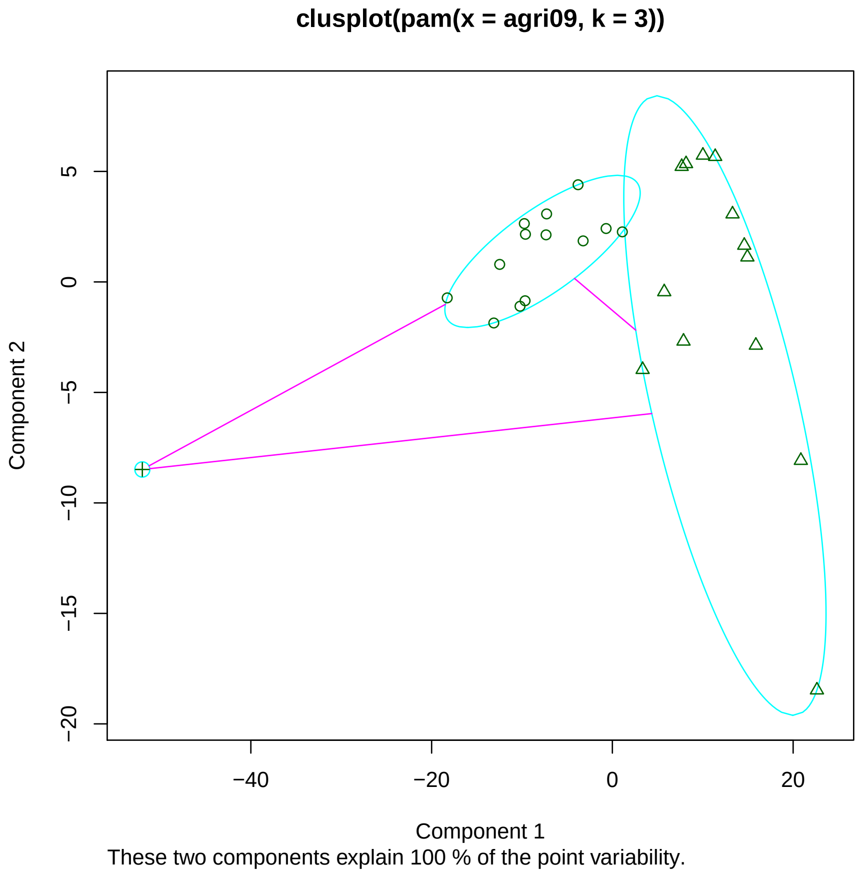
If the data is given in the form of a matrix of observations, we can first perform a principal component analysis, (see later) and make a scatterplot of the first two components. Although this method will lead to a loss of information, any significant structure in the data is often already visible in the first few principal components, meaning we will still be able to visualise the clusters that we have found using our clustering algorithm.

Alternatively, if the data are given in the form of a dissimilarity matrix, we can use the technique of *principal coordinates analysis* or *multidimensional scaling.* Going into detail about the mechanisms behind this technique is beyond the scope of this course, but in general terms the method returns a set of points $\boldsymbol{y}_i$, $i = 1, \ldots, n$, such that the pairwise distances between points $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i'}$ are approximately equal to the dissimilarity between objects $i$ and $i'$.

### Example: EU data

We revisit the EU data example. For the 3-medoid PAM clustering we obtain the clusplot:

```
clusplot(clust.pam3)
```



Note that each cluster is represented by the ellipse with smallest area that contains all the points in the cluster.

## 2.6 Fuzzy Analysis

Most clustering methods will allocate each object to exactly one cluster, and these methods are referred to as crisp clustering methods. On the other hand, sometimes it is not clear to which cluster an object should be assigned, for example, when the object is an outlying object, or when it lies between two clusters. Hence, it can be advantageous to spread each object over the various clusters and obtain a *fuzzy clustering*.

A fuzzy clustering method will compute, for each object $i$ and each cluster $C_j$, a *membership* $u_{ij}$ which indicates how strongly object $i$ belongs to the cluster $C_j$. These memberships have to satisfy the following conditions:

- $u_{ij} \geqslant 0$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, K$, and
- $\displaystyle\sum_{j=1}^{K} u_{ij} = 1$ for all $i = 1, \ldots, n$.

Summarizing, we can say that crisp memberships are either 0 or 1, and that fuzzy memberships lie between 0 and 1.

One possible way of performing this fuzzy clustering is through minimisation of the objective function

$$\sum_{j=1}^{K} \frac{\sum_{i,i'=1}^{n} u_{ij}^2 u_{i'j}^2 d(i,i')}{2\sum_{i=1}^{n} u_{ij}^2}.$$

Here, $d(i,i')$ is the (known) dissimilarity between objects $i$ and $i'$, while the memberships $u_{ij}$ are unknown, and are obtained using numerical minimisation techniques.

The fuzzyness of the resulting clustering can be represented by *Dunn's partition coefficient*:

$$F_K = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{K} u_{ij}^2,$$

which always lies in the interval $[\frac{1}{K}, 1]$. If the clustering is entirely fuzzy, this coefficient equals $F_K = \frac{1}{K}$, and if the clustering is crisp, the value of the coefficient is $F_K = 1$. The normalised version of this coefficient is

$$F'_K = \frac{F_K - \frac{1}{K}}{1 - \frac{1}{K}} = \frac{K\,F_K - 1}{K - 1},$$

which lies in the interval $[0, 1]$.

### Example: EU data

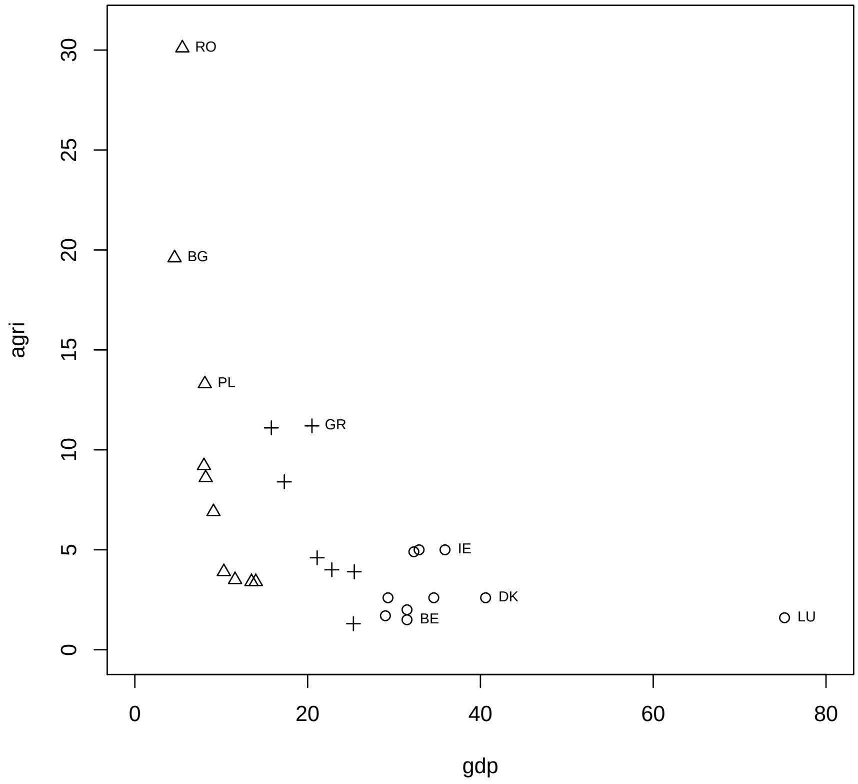
We revisit the EU data example and perform a fuzzy analysis to detect three clusters.

```
clust.fanny3 <- fanny(agri09, 3)
clust.fanny3

Fuzzy Clustering object of class 'fanny' :
m.ship.expon.        2
objective     68.22345
tolerance        1e-15
iterations          56
converged            1
maxit              500
n                   27
Membership coefficients (in %, rounded):
   [,1] [,2] [,3]
BE   81    6   13
BG   17   53   31
CZ   12   50   38
DK   62   14   24
DE   67    9   23
EE    9   65   25
IE   74    9   17
GR   17   26   57
ES   21   15   64
FR   71    8   21
IT   35   14   50
CY   16   16   68
```

```
LV   6  80  14
LT   6  79  15
LU  41  26  33
HU   6  78  16
MT  12  47  41
NL  81   7  13
AT  81   6  13
PL  10  66  24
PT  13  42  45
RO  22  43  34
SI  12  31  57
SK  10  60  30
FI  81   6  13
SE  83   5  11
UK  39  16  46
Fuzzyness coefficients:
dunn_coeff normalized
 0.5142999  0.2714498
Closest hard clustering:
BE BG CZ DK DE EE IE GR ES FR IT CY LV LT LU HU MT NL AT PL PT RO
 1  2  2  1  1  2  1  3  3  1  3  3  2  2  1  2  2  1  1  2  3  2
SI SK FI SE UK
 3  2  1  1  3

Available components:
 [1] "membership"  "coeff"       "memb.exp"    "clustering"
 [5] "k.crisp"     "objective"   "convergence" "diss"
 [9] "call"        "silinfo"     "data"
```

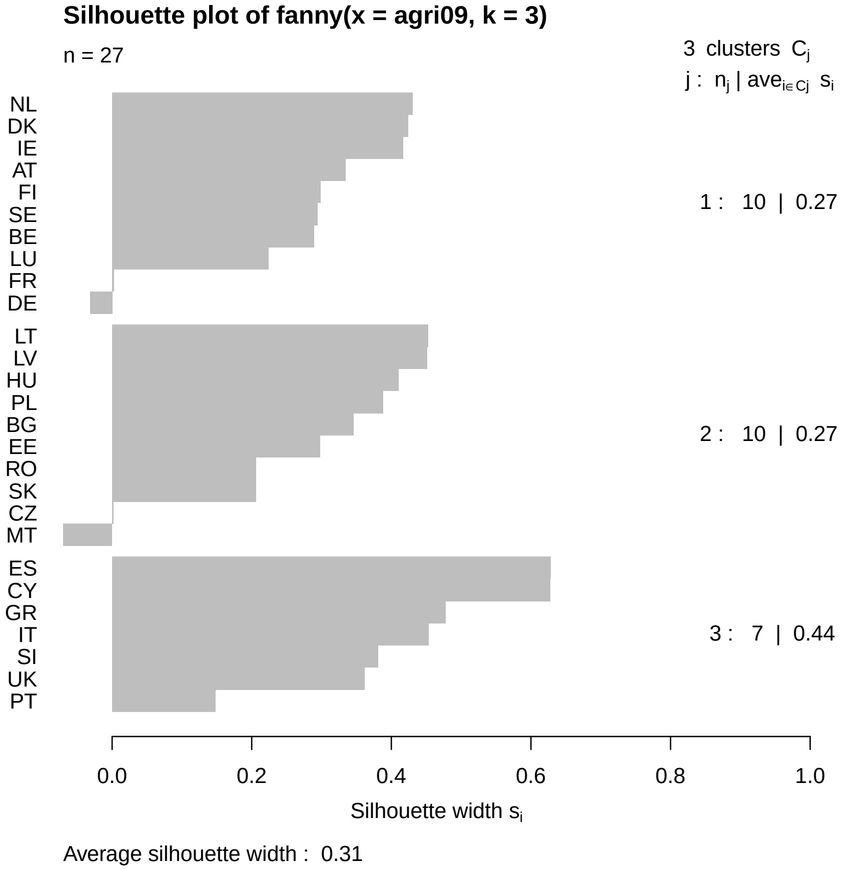
For each object, we see the membership percentage. For example, it is rather clear that Belgium belongs to cluster 1, with a membership of 81%, while there is considerable doubt whether Portugal should be in cluster 2 or cluster 3, with memberships of 42% and 45% respectively. From the normalised Dunn's partition coefficient of 0.271 we infer that the obtained clustering is rather fuzzy. We also obtain a hard partitioning, where we assign each object to the cluster to which it has the highest membership. If we do this, we see that Luxemburg is no longer isolated in its own cluster, but is joined by countries such as Belgium, Germany, the Netherlands, etc. and that the ‘poorer’ nations are also split up in a different manner. This can be seen more easily in the following scatterplot.

```
plot(agri09, pch = clust.fanny3$clustering, xlim = c(0, 80), ylim = c(0, 31))
```



To compare this clustering to the one obtained with the partitioning around medoids algorithm, we can compare the silhouette plots. The silhouette plot for the fuzzy analysis clustering is obtained by

```
plot(silhouette(clust.fanny3))
```



It is immediately obvious that the fuzzy analysis clustering is a lot worse, with average silhouette width of 0.31 compared to 0.54 obtained for 3-medoid PAM. We also see that there are two objects, Germany (DE) and Malta (MT), which have a negative silhouette coefficient. This indicates that these are actually less dissimilar to the objects in a different cluster than to the objects in the cluster to which they are assigned.

## 2.7 Hierarchical clustering algorithms

The clustering methods described in the previous sections are all examples of partitioning clustering methods, in which we supply in advance the number of clusters $K$ that we wish to detect. Sometimes, we have no idea how many clusters we are looking for, hence a partitioning algorithm is not the best approach to allocate the objects to clusters. In this case, we will use a method that creates an entire hierarchy of clusterings, ranging from one single cluster, the entire dataset, to $n$ distinct clusters, where each observation is its own cluster.

### 2.7.1 Agglomerate Nesting

In the agglomerate nesting algorithm, we start at the bottom of the cluster hierarchy, with each object in its own cluster. In each step, we combine the two clusters with the smallest between-cluster dissimilarity, until only one large cluster remains.

The dissimilarity between two clusters can be defined in various ways. The three most common ways are

1. *Group average*: the dissimilarity between clusters $C_j$ and $C_{j'}$ is the average dissimilarity between the objects from cluster $C_j$ and the objects from cluster $C_{j'}$, i.e.
$$d(C_j, C_{j'}) = \frac{1}{|C_j|\;|C_{j'}|} \sum_{i \in C_j, i' \in C_{j'}} d(i, i').$$

2. *Nearest neighbour* or *single linkage*: the dissimilarity between clusters $C_j$ and $C_{j'}$ is the smallest dissimilarity between any object from cluster $C_j$ and any object from cluster $C_{j'}$, i.e.
$$d(C_j, C_{j'}) = \min_{i \in C_j, i' \in C_{j'}} d(i, i').$$

3. *Furthest neighbour* or *complete linkage*: the dissimilarity between clusters $C_j$ and $C_{j'}$ is the largest dissimilarity between any object from cluster $C_j$ and any object from cluster $C_{j'}$, i.e.
$$d(C_j, C_{j'}) = \max_{i \in C_j, i' \in C_{j'}} d(i, i').$$

### Simple example

Assume that we have a dataset consisting of five objects $a$, $b$, $c$, $d$, and $e$, and that the data is represented as a matrix of dissimilarities:

$$\begin{array}{c|ccccc}
 & a & b & c & d & e \\
\hline
a & 0 & & & & \\
b & 2 & 0 & & & \\
c & 6 & 3 & 0 & & \\
d & 8 & 7 & 5 & 0 & \\
e & 9 & 6 & 5 & 4 & 0
\end{array}$$

A step-by-step runthrough of the algorithm can be found in Table 2.1. We observe that in the first step, clusters $\{a\}$ and $\{b\}$ are merged into a single cluster, yielding clusters $\{a, b\}$, $\{c\}$, $\{d\}$, $\{e\}$, and that this is the same for each method.

In the second step, the methods diverge however. If we use the group average or the complete linkage methods, clusters $\{d\}$ and $\{e\}$ are the least dissimilar, and will be merged, resulting in the clusters $\{a, b\}$, $\{c\}$, $\{d, e\}$. If we use the single linkage method, we decide that clusters $\{a, b\}$ and $\{c\}$ are the least dissimilar and join these two, resulting in the clusters $\{a, b, c\}$, $\{d\}$, and $\{e\}$.

In the third step, we see that clusters $\{a, b\}$ and $\{c\}$ will be joined if we use the group average method, resulting in the clusters $\{a, b, c\}$ and $\{d, e\}$. If we use the single linkage method, clusters $\{d\}$ and $\{e\}$ will be joined, also resulting in the clusters $\{a, b, c\}$ and $\{d, e\}$, but with the complete linkage method, we join clusters $\{c\}$ and $\{d, e\}$ and end up with clusters $\{a, b\}$ and $\{c, d, e\}$.

#### Graphical representation

We can graphically represent an agglomerate clustering in two ways. The first way to do so is by making a *dendrogram*. A dendrogram is basically a graphical tree in which the leaves represent the objects, and in which the vertical coordinate of the junction of two branches represents the dissimilarity between the two corresponding clusters.

The second graphical representation is the agglomerative *banner*. In this plot, the objects are listed vertically, and the width of the bars between the objects

Initial values:

| | $\boldsymbol{a}$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 0 | | | | |
| $\boldsymbol{b}$ | **2** | 0 | | | |
| $c$ | 6 | 3 | 0 | | |
| $d$ | 8 | 7 | 5 | 0 | |
| $e$ | 9 | 6 | 5 | 4 | 0 |

| | $\boldsymbol{a}$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 0 | | | | |
| $\boldsymbol{b}$ | **2** | 0 | | | |
| $c$ | 6 | 3 | 0 | | |
| $d$ | 8 | 7 | 5 | 0 | |
| $e$ | 9 | 6 | 5 | 4 | 0 |

| | $\boldsymbol{a}$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 0 | | | | |
| $\boldsymbol{b}$ | **2** | 0 | | | |
| $c$ | 6 | 3 | 0 | | |
| $d$ | 8 | 7 | 5 | 0 | |
| $e$ | 9 | 6 | 5 | 4 | 0 |

After step 1:

| | $\{a,b\}$ | $c$ | $\boldsymbol{d}$ | $e$ |
|---|---|---|---|---|
| $\{a,b\}$ | 0 | | | |
| $c$ | 4.5 | 0 | | |
| $d$ | 7.5 | 5 | 0 | |
| $\boldsymbol{e}$ | 7.5 | 5 | **4** | 0 |

| | $\boldsymbol{\{a,b\}}$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|
| $\{a,b\}$ | 0 | | | |
| $\boldsymbol{c}$ | **3** | 0 | | |
| $d$ | 7 | 5 | 0 | |
| $e$ | 6 | 5 | 4 | 0 |

| | $\{a,b\}$ | $c$ | $\boldsymbol{d}$ | $e$ |
|---|---|---|---|---|
| $\{a,b\}$ | 0 | | | |
| $c$ | 6 | 0 | | |
| $d$ | 8 | 5 | 0 | |
| $\boldsymbol{e}$ | 9 | 5 | **4** | 0 |

After step 2:

| | $\boldsymbol{\{a,b\}}$ | $c$ | $\{d,e\}$ |
|---|---|---|---|
| $\{a,b\}$ | 0 | | |
| $\boldsymbol{c}$ | **4.5** | 0 | |
| $\{d,e\}$ | 7.5 | 5 | 0 |

| | $\{a,b,c\}$ | $\boldsymbol{d}$ | $e$ |
|---|---|---|---|
| $\{a,b,c\}$ | 0 | | |
| $d$ | 5 | 0 | |
| $\boldsymbol{e}$ | 5 | **4** | 0 |

| | $\{a,b\}$ | $\boldsymbol{c}$ | $\{d,e\}$ |
|---|---|---|---|
| $\{a,b\}$ | 0 | | |
| $c$ | 6 | 0 | |
| $\boldsymbol{\{d,e\}}$ | 9 | **5** | 0 |

After step 3:

| | $\{a,b,c\}$ | $\{d,e\}$ |
|---|---|---|
| $\{a,b,c\}$ | 0 | |
| $\{d,e\}$ | 6.67 | 0 |

| | $\{a,b,c\}$ | $\{d,e\}$ |
|---|---|---|
| $\{a,b,c\}$ | 0 | |
| $\{d,e\}$ | 5 | 0 |

| | $\{a,b\}$ | $\{c,d,e\}$ |
|---|---|---|
| $\{a,b\}$ | 0 | |
| $\{c,d,e\}$ | 9 | 0 |

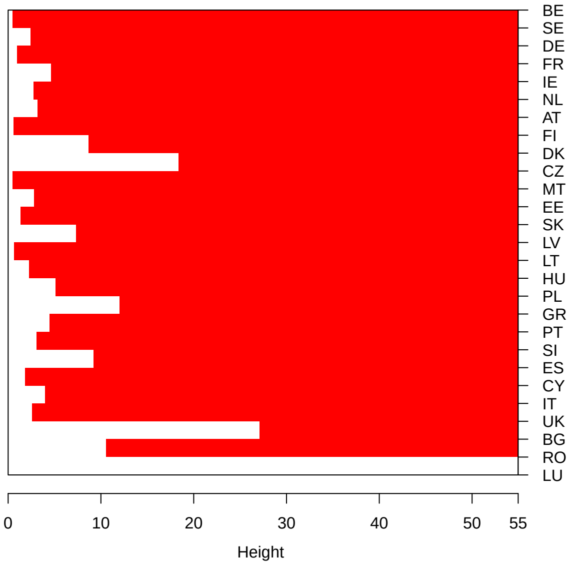
Table 2.1: The difference between the three between-cluster dissimilarity measures. The left column shows the group average method, the middle column shows the single linkage method, and the right column shows the complete linkage method.

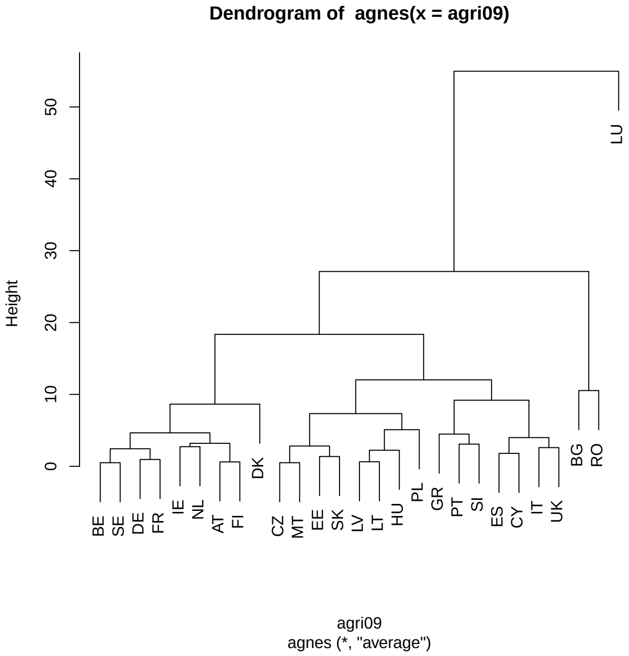
denote at which between-cluster dissimilarity the two corresponding clusters merge. For this plot, we can compute the agglomerative coefficient, which is the average width (normalised between 0 and 1) of the bars in the banner. Note that this coefficient tends to increase as the number of objects grows.

### Example: EU data

We revisit the EU data example and perform agglomerate nesting where we use the group average method to compute between-cluster dissimilarities.

```
clust.agnes.ga <- agnes(agri09)
bannerplot(clust.agnes.ga)
pltree(clust.agnes.ga)
```





From the dendrogram, we see that the last two clusters to be merged are the cluster containing only Luxemburg, and the cluster containing all the other countries, at a dissimilarity of about 55. This can also be seen from the bannerplot. If we draw an imaginary vertical line on the plot at height e.g. 40, we see that all the countries, except Luxemburg, are connected by the banner, once again indicating that the last two clusters were Luxemburg in one cluster, the other countries in the other cluster.
Setting the cutoff point to e.g. 24, we see that the big cluster containing all countries except Luxemburg was created by merging the cluster containing Bulgaria and Rumania on the one hand, and the cluster containing all other countries, except Luxemburg, on the other hand.

### 2.7.2 Divisive Analysis

In the divisive analysis algorithm, we start at the top of the cluster hierarchy, with one large cluster containing all objects. In each step, we split the largest available cluster into two smaller clusters, until we obtain $n$ clusters containing one object each.

The algorithm is described below:

1. Start with one cluster $C_1$, containing all the objects, and set $K = 1$.

2. For each cluster $C_j$, $j = 1, \ldots, K$, compute its *diameter* as

$$diam(C_j) = \max_{i,i' \in C_j} d(i, i')$$

   and take the cluster $C_j$ with the largest diameter, while setting the new cluster $C_{K+1} = \emptyset$. This new cluster is called the *splinter group*, in this iteration of the algorithm.

3. For each object $i$ in cluster $C_j$, compute its average dissimilarity to all other objects in the cluster using equation (2.5.1). Select the object $i$ with the largest average dissimilarity to the others, and move it to the splinter group, so $C_j \leftarrow C_j \setminus \{j\}$ and $C_{K+1} \leftarrow \{j\}$.

4. If there is still more than one object left in cluster $C_j$, compute the average dissimilarity from all objects $i$ left in cluster $C_j$ to the splinter group $C_{K+1}$ as in equation (2.5.2) and select the object $i$ for which

$$a(i) - d(i, C_{K+1}) = \max_{i \in C_j} (a(i) - d(i, C_{K+1})).$$

5. If $a(i) - d(i, C_{K+1}) > 0$ for the object $i$ obtained in step 4, move it to the splinter group $C_{K+1}$, so $C_j \leftarrow C_j \setminus \{j\}$ and $C_{K+1} \leftarrow C_{K+1} \cup \{j\}$, and return to step 4. Otherwise, set $K \leftarrow K + 1$ and continue to step 6.

6. Repeat steps 2 through 5 until $K = n$.

#### Simple example

Assume that we have a dataset consisting of five objects $a$, $b$, $c$, $d$, and $e$, and that the data is represented as a matrix of dissimilarities:

| | $a$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 0 | | | | |
| $b$ | 2 | 0 | | | |
| $c$ | 6 | 3 | 0 | | |
| $d$ | 8 | 7 | 5 | 0 | |
| $e$ | 9 | 6 | 5 | 4 | 0 |

We start with a single cluster $C_1 = \{a, b, c, d, e\}$, and compute, for each object, the average distance to the rest of the cluster's objects:

$$\begin{array}{lll} a(a) = \frac{2+6+8+9}{4} = 6.25 & a(b) = \frac{2+3+7+6}{4} = 4.5 & a(c) = \frac{6+3+5+5}{4} = 4.75 \\ a(d) = \frac{8+7+5+4}{4} = 6 & a(e) = \frac{9+65+4}{4} = 6. & \end{array}$$

Because object $a$ is, on average, the most dissimilar to the other objects in its cluster, we move object $a$ to the new cluster $C_2$, so now $C_1 = \{b, c, d, e\}$ and $C_2 = \{a\}$. Next, for each object left in cluster $C_1$, compute the difference between the average distance of the object to the rest of the cluster, and the average distance of the object to the cluster $C_2$:

$$\begin{array}{llll} b: & a(b) = \frac{3+7+6}{3} = 5.33 & d(b, C_2) = 2 & a(b) - d(b, C_2) = 3.33 \\ c: & a(c) = \frac{3+5+5}{3} = 4.33 & d(c, C_2) = 6 & a(c) - d(c, C_2) = -1.67 \\ d: & a(d) = \frac{7+5+4}{3} = 5.33 & d(d, C_2) = 8 & a(d) - d(d, C_2) = -2.67 \\ e: & a(e) = \frac{6+5+4}{3} = 5 & d(e, C_2) = 9 & a(e) - d(e, C_2) = -4. \end{array}$$

Since this difference is largest for object $b$ and positive, we move object $b$ from cluster $C_1$ to $C_2$, so now we have $C_1 = \{c, d, e\}$ and $C_2 = \{a, b\}$. Since we moved an object in the last step, we will compute these differences again, with the new clusters $C_1$ and $C_2$:

$$\begin{array}{llll} c: & a(c) = \frac{5+5}{2} = 5 & d(c, C_2) = \frac{6+3}{2} = 4.5 & a(c) - d(c, C_2) = 0.5 \\ d: & a(d) = \frac{5+4}{2} = 4.5 & d(d, C_2) = \frac{8+7}{2} = 7.5 & a(d) - d(d, C_2) = -3 \\ e: & a(e) = \frac{5+4}{2} = 4.5 & d(e, C_2) = \frac{9+6}{2} = 7.5 & a(e) - d(e, C_2) = -3. \end{array}$$

Because the difference is largest for object $c$ and positive, we move object $c$ from cluster $C_1$ to $C_2$, so now we have $C_1 = \{d, e\}$ and $C_2 = \{a, b, c\}$. Since we moved an object in the last step, we will compute these differences again, with the new clusters $C_1$ and $C_2$:

$$\begin{array}{llll} d: & a(d) = 4 & d(d, C_2) = \frac{8+7+5}{3} = 6.67 & a(d) - d(d, C_2) = -2.67 \\ e: & a(e) = 4 & d(e, C_2) = \frac{9+6+5}{3} = 6.67 & a(e) - d(e, C_2) = -2.67. \end{array}$$

Since none of the differences are positive, we have found the two clusters, $C_1 = \{d, e\}$ and $C_2 = \{a, b, c\}$, and we can start on determining the third cluster. Since $diam(C_1) = 4 < diam(C_2) = 6$ we will split cluster $C_2$. Compute, for each object, the average distance to the rest of the cluster's objects:

$$a(a) = \frac{2+6}{2} = 4 \quad a(b) = \frac{2+3}{2} = 2.5 \quad a(c) = \frac{6+3}{2} = 4.5$$

Because object $c$ is, on average, the most dissimilar to the other objects in its cluster, we move object $c$ to the new cluster $C_3$, so now $C_1 = \{d, e\}$, $C_2 = \{a, b\}$, and $C_3 = \{c\}$. Next, for each object left in cluster $C_2$, compute the difference between the average distance of the object to the rest of the cluster, and the average distance of the object to the cluster $C_3$:

$$\begin{array}{llll} a: & a(a) = 2 & d(a, C_3) = 6 & a(a) - d(a, C_2) = -4 \\ b: & a(b) = 2 & d(b, C_3) = 3 & a(b) - d(b, C_2) = -1. \end{array}$$

Since none of the differences are positive, we have found the three clusters in this step, $C_1 = \{d, e\}$, $C_2 = \{a, b\}$, and $C_3 = \{c\}$, and we can start on determining the fourth cluster. Now, $C_1$ has the largest diameter, which we split into clusters $C_1 = \{d\}$ and $C_4 = \{e\}$. Because the new cluster $C_1$ contains a single object, we cannot move any more objects to the cluster $C_4$, so we have found the four clusters: $C_1 = \{d\}$, $C_2 = \{a, b\}$, $C_3 = \{c\}$, and $C_4 = \{e\}$.
In the final step of the algorithm, we split the only cluster that contains two objects and end up with 5 clusters containing a single object each.

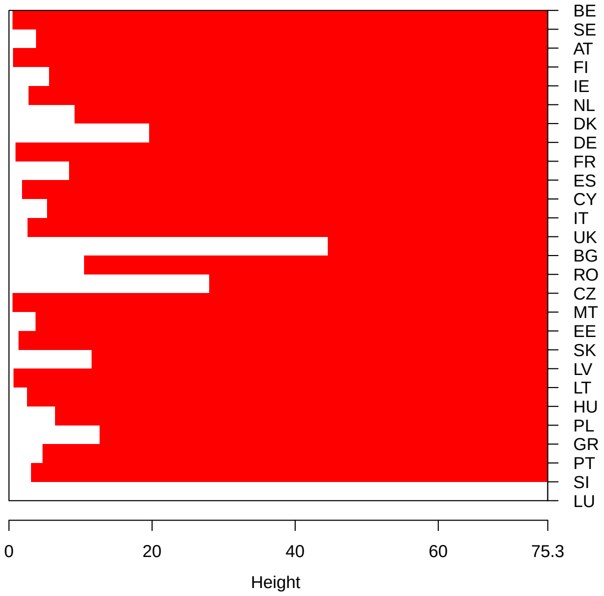
#### Graphical representation

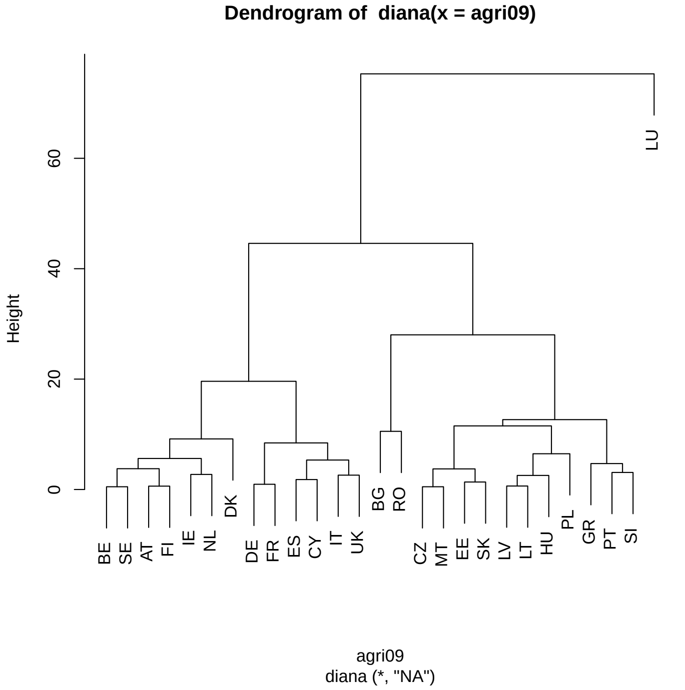
Similar to the agglomerate nesting, we can make a dendrogram of the divisive analysis clustering, where the vertical height of the junctions shows the diameter of that cluster before splitting it. We can also make a bannerplot in the same way, where the left edge of the bar goes to the diameter of that cluster before it is split by the algorithm. Finally, the divisive coefficient can also be computed, showing the average width of the bars, normalised between 0 and 1.

### Example: EU data

We revisit the EU data example and perform divisive analysis.

```
clust.diana <- diana(agri09)
bannerplot(clust.diana)
pltree(clust.diana)
```





From the dendrogram, we see that the biggest cluster has a diameter of about 75, and was split into the cluster containing only Luxemburg, and the cluster containing all the other countries. This can also be seen from the bannerplot. If we draw an imaginary vertical line on the plot at height e.g. 60, we see that all the countries, except Luxemburg, are connected by the banner, once again indicating that the first two clusters were Luxemburg in one cluster, the other countries in the other cluster.

Setting the cutoff point lower, we observe that the big cluster of 26 nations had a diameter of about 45, and was split up into a cluster of 'richer' countries, such as Belgium, Denmark, and the United Kingdom, and a cluster of poorer countries, such as Bulgaria, Romania, Hungary, and Slovenia. Observe also that the hierarchy of clusters obtained with the divisive analysis algorithm is quite different from the hierarchy obtained with the agglomerate nesting algorithm.

# Chapter 3

# Multivariate statistisch model en schatters

Als een multivariate dataset homogeen is, dan kunnen we statistische modellen gebruiken om de gegevens te analyseren en meer inzicht te bekomen in de onderliggende populatie.

## 3.1 Resultaten uit de matrixrekening

**Def.** Een vierkante matrix $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ is *orthogonaal*
$\Leftrightarrow \boldsymbol{U}\boldsymbol{U}^{\tau} = \boldsymbol{I}_p \Leftrightarrow \boldsymbol{U}^{\tau} = \boldsymbol{U}^{-1} \Leftrightarrow \boldsymbol{U}^{\tau}\boldsymbol{U} = \boldsymbol{I}_p$

Als we de rijen van $\boldsymbol{U}$ noteren als $\boldsymbol{u}_1^{\tau}, \boldsymbol{u}_2^{\tau}, \ldots, \boldsymbol{u}_p^{\tau}$ dan is dit equivalent met de eigenschap dat $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ orthonormaal zijn in $\mathbb{R}^{p \times 1}$. Een analoge redenering geldt voor de kolommen van de matrix $\boldsymbol{U}$.

**Def.** Als $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ en $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^{p \times 1}$ zodat

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \qquad \text{voor een } \lambda \in \mathbb{R}$$

dan noemt men $\boldsymbol{x}$ een *eigenvector* van $\boldsymbol{A}$ met *eigenwaarde* $\lambda$.

**Opmerkingen.**

- Meestal worden de eigenvectoren genormaliseerd ($\|\boldsymbol{x}\| = 1$), men noteert deze dan als $\boldsymbol{e} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$.
- Het bepalen van de eigenwaarden/eigenvectoren kan gebeuren via de karakteristieke vergelijking
$$(\boldsymbol{A} - \lambda \boldsymbol{I}_p)\boldsymbol{x} = 0 \Rightarrow det(\boldsymbol{A} - \lambda \boldsymbol{I}_p) = 0.$$
- Als $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ een symmetrische matrix is, dan bestaan er juist $p$ paren $(\lambda_i, \boldsymbol{e}_i)$, waarbij de eigenwaarden $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ positief, negatief of nul kunnen zijn.

**Eig. (Spectraalontbinding)** Elke symmetrische matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ kan ontbonden worden als
$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\tau} = \sum_{j=1}^{p} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j^{\tau}$$
waarbij $\boldsymbol{P} := [\boldsymbol{e}_1|\boldsymbol{e}_2|\ldots|\boldsymbol{e}_p]$ een orthogonale matrix is en
$\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.

**Def.** Een symmetrische matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ is
$$\begin{aligned} \text{positief semidefiniet (PSD)} \quad &\Leftrightarrow \quad \forall\, \boldsymbol{x} \in \mathbb{R}^{p \times 1} : \boldsymbol{x}^{\tau}\boldsymbol{A}\boldsymbol{x} \geqslant 0 \\ \text{positief definiet (PD)} \quad &\Leftrightarrow \quad \forall\, \boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^{p \times 1} : \boldsymbol{x}^{\tau}\boldsymbol{A}\boldsymbol{x} > 0 \end{aligned}$$

**Opmerkingen.**

- Voor een matrix $\boldsymbol{A} \in \text{PSD}(p)$ zijn alle eigenwaarden $\geqslant 0$. Een matrix $\boldsymbol{A}$ in $\text{PD}(p)$ heeft strikt positieve eigenwaarden, zijn inverse heeft dezelfde eigenvectoren en de eigenwaarden worden mee geïnverteerd.
- $\boldsymbol{A} \in \text{PSD}(p) \Rightarrow |\boldsymbol{A}| = \det(\boldsymbol{A}) = \prod_{j=1}^{p} \lambda_j \geqslant 0$
$\boldsymbol{A} \in \text{PD}(p) \Leftrightarrow \boldsymbol{A} \in \text{PSD}(p)$ en $|\boldsymbol{A}| > 0$
- Voor een willekeurige matrix $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ met $\det(\boldsymbol{B}) \neq 0$ geldt dat $\boldsymbol{B}\boldsymbol{B}^{\tau}$ positief definiet is.

- Voor elke matrix $\boldsymbol{A} \in \mathrm{PD}(p)$ bestaat er een $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ met $\det(\boldsymbol{B}) \neq 0$ zodat $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\tau$. Men noemt $\boldsymbol{B}$ een *wortel* van $\boldsymbol{A}$. Deze wortel is echter niet uniek.

**Def.** Als $\boldsymbol{A} \in \mathrm{PD}(p)$ een spectraaldecompositie heeft van de vorm $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\tau$, dan is de *symmetrische wortel* van $\boldsymbol{A}$

$$\boldsymbol{A}^{\frac{1}{2}} := \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^\tau = \sum_{j=1}^{p} \sqrt{\lambda_j}\boldsymbol{e}_j\boldsymbol{e}_j^\tau$$

met $\boldsymbol{\Lambda}^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$.

Enkel eigenschappen van $\boldsymbol{A}^{\frac{1}{2}}$ zijn:

- $(\boldsymbol{A}^{\frac{1}{2}})^\tau = \boldsymbol{A}^{\frac{1}{2}}$ (symmetrisch)
- $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}$ (wortel van $\boldsymbol{A}$)
- $(\boldsymbol{A}^{\frac{1}{2}})^{-1} = \boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^\tau = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}}\boldsymbol{e}_j\boldsymbol{e}_j^\tau$
- $\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{A}^{-1}$

## 3.2 Multivariate toevalsvariabelen

**Def.**

$$\boldsymbol{X} : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^{p\times 1}, \mathcal{R}^p) : \omega \to \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_p(\omega) \end{pmatrix}$$

is een toevalsveranderlijke $\Leftrightarrow \forall\, 1 \leqslant j \leqslant p : X_j$ is een reële toevalsveranderlijke.

De verdeling $P_{\boldsymbol{X}}$ wordt gegeven door

$$\forall B \in \mathcal{R}^p : P_{\boldsymbol{X}}(B) = P(\boldsymbol{X} \in B) := P(\boldsymbol{X}^{-1}(B))$$

Voor elke component $X_j : \Omega \to \mathbb{R}$ noemt men $P_{X_j}$ de *marginale verdeling* van $X_j$ ($\forall\, 1 \leqslant j \leqslant p$).

Verder kennen we het *marginale gemiddelde* van $X_j$:

$$\mu_j := \mathrm{E}[X_j] = \begin{cases} \sum_{x_j} x_j p(x_j) & \text{in het discrete geval} \\ \int_{-\infty}^{+\infty} y f_j(y) dy & \text{in het continue geval,} \end{cases}$$

de *marginale variantie* van $X_j$:

$$\sigma_{jj} = \mathrm{Var}[X_j] = \mathrm{E}[(X_j - \mu_j)^2],$$

en de covariantie tussen twee componenten $X_j$ en $X_k$:

$$\sigma_{jk} = \mathrm{Cov}[X_j, X_k] = \mathrm{E}[(X_j - \mu_j)(X_k - \mu_k)]$$

$X_j$ en $X_k$ worden *onafhankelijk* genoemd als en slechts als

$$P[X_j \leqslant x_j \text{ en } X_k \leqslant x_k] = P[X_j \leqslant x_j]P[X_k \leqslant x_k] \qquad \forall x_j, x_k \in \mathbb{R}$$

Als $\sigma_{jj} < \infty$ en $\sigma_{kk} < \infty$, dan bestaat $\sigma_{jk}$ en dan geldt:

$$X_j \text{ en } X_k \text{ zijn onafhankelijk} \quad \Rightarrow \quad \sigma_{jk} = 0$$
$$\not\Leftarrow$$

**Def.** Het *populatiegemiddelde* wordt gedefinieerd als

$$\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

**Def.** De *covariantiematrix van de populatie* wordt gedefinieerd als

$$\mathrm{Cov}[\boldsymbol{X}] = \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\tau] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

**Def.** De *correlatiematrix van de populatie* wordt gedefinieerd als

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix} = (\boldsymbol{V}^{\frac{1}{2}})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{\frac{1}{2}})^{-1} = (\boldsymbol{V}^{-\frac{1}{2}})\boldsymbol{\Sigma}(\boldsymbol{V}^{-\frac{1}{2}})$$

waarbij $\boldsymbol{V}^{\frac{1}{2}} = diag(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$.

$\boldsymbol{\Sigma}$ is steeds symmetrisch en positief semidefiniet

$\Rightarrow$ $\boldsymbol{\rho}$ is symmetrisch en positief semidefiniet.

**Result.** Als $\boldsymbol{X}$ $p$-dimensionaal is en $\boldsymbol{c} \in \mathbb{R}^{p\times 1}$, dan is $\boldsymbol{c}^\tau \boldsymbol{X}$ een univariate toevalsveranderlijke met

- $\mathrm{E}[\boldsymbol{c}^\tau \boldsymbol{X}] = \boldsymbol{c}^\tau \boldsymbol{\mu}$
- $\mathrm{Var}[\boldsymbol{c}^\tau \boldsymbol{X}] = \boldsymbol{c}^\tau \boldsymbol{\Sigma} \boldsymbol{c}$

Als $\boldsymbol{A} \in \mathbb{R}^{q\times p}$, dan is $\boldsymbol{AX}$ een $q$-dimensionale toevalsveranderlijke met

- $\mathrm{E}[\boldsymbol{AX}] = \boldsymbol{A\mu}$
- $\mathrm{Cov}[\boldsymbol{AX}] = \boldsymbol{A\Sigma A}^\tau$

Soms kunnen we de variabelen op een natuurlijke manier splitsen in twee of meer groepen. Bijvoorbeeld, 2 groepen van grootte $p-q$ en $q$. Dan schrijven we

$$\boldsymbol{X} = \left(\begin{array}{c} X_1 \\ \vdots \\ X_q \\ \hline X_{q+1} \\ \vdots \\ X_p \end{array}\right) = \left(\begin{array}{c} \boldsymbol{X}^{(1)} \\ \hline \boldsymbol{X}^{(2)} \end{array}\right).$$

Voor de kentallen wordt dat

$$\mathrm{E}[\boldsymbol{X}] = \left(\begin{array}{c} \mu_1 \\ \vdots \\ \mu_q \\ \hline \mu_{q+1} \\ \vdots \\ \mu_p \end{array}\right) = \left(\begin{array}{c} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{array}\right)$$

en

$$\begin{aligned} \mathrm{Cov}[\boldsymbol{X}] &= \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\tau] \\ &= \left(\begin{array}{c|c} (\boldsymbol{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\boldsymbol{X}^{(1)}-\boldsymbol{\mu}^{(1)})^\tau & (\boldsymbol{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\boldsymbol{X}^{(2)}-\boldsymbol{\mu}^{(2)})^\tau \\ \hline (\boldsymbol{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\boldsymbol{X}^{(1)}-\boldsymbol{\mu}^{(1)})^\tau & (\boldsymbol{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\boldsymbol{X}^{(2)}-\boldsymbol{\mu}^{(2)})^\tau \end{array}\right) \\ &= \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right). \end{aligned}$$

$\mathbf{\Sigma}_{11} \in \mathbb{R}^{q\times q}$ en $\mathbf{\Sigma}_{22} \in \mathbb{R}^{(p-q)\times(p-q)}$ zijn symmetrische matrices, $\mathbf{\Sigma}_{12} \in \mathbb{R}^{q\times(p-q)}$ en $\mathbf{\Sigma}_{21} \in \mathbb{R}^{(p-q)\times q}$ hoeven noch symmetrisch noch vierkant te zijn. We noemen

$$\mathbf{\Sigma}_{12} = \mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}).$$

## 3.3 Multivariate schatters

### 3.3.1 Multivariate steekproeven

$$\forall\, 1 \leqslant i \leqslant n: \quad \boldsymbol{X}_i : (\Omega, \mathcal{A}, P) \to \boldsymbol{X}_i(\omega) = \begin{pmatrix} X_{i1}(\omega) \\ \vdots \\ X_{ip}(\omega) \end{pmatrix} \in \mathbb{R}^{p\times 1}$$

met $\boldsymbol{X}_i : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^{p\times 1}, \mathcal{R}^p)$ is een toevalsveranderlijke.

**Def.** $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is een *steekproef*
$\Leftrightarrow \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ zijn *iid*, onafhankelijk en gelijk verdeeld (met verdeling $P_{\boldsymbol{X}_1}$).

Als $P_{\boldsymbol{X}_1}$ een dichtheid $f(\boldsymbol{x}_1) = f(x_{11}, \ldots, x_{1p})$ heeft
$\Rightarrow (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ heeft dichtheid $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)\ldots f(\boldsymbol{x}_n)$.

### 3.3.2 De schatters $\overline{\boldsymbol{X}}$ en $\boldsymbol{S}$ zijn onvertekend

Niet-parametrisch model: $\mathcal{P} = \{$kansmaten $P$ op $(\mathbb{R}^p, \mathcal{R}^p)$ met 2de moment$\}$

**Eig.** Zij $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} P_{\boldsymbol{X}_1}$ zodat $\mathrm{E}[\boldsymbol{X}_1] = \boldsymbol{\mu}$ en $\mathrm{Cov}[\boldsymbol{X}_1] = \boldsymbol{\Sigma}$ bestaan. Dan is

1. $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ een onvertekende (zuivere) schatter van $\boldsymbol{\mu}$ en $\mathrm{Cov}[\overline{\boldsymbol{X}}] = \frac{1}{n}\boldsymbol{\Sigma}$.
2. $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\tau$ een onvertekende (zuivere) schatter van $\boldsymbol{\Sigma}$.

*Proof.* 1.

$$\begin{aligned} \mathrm{E}[\overline{\boldsymbol{X}}] &= \mathrm{E}[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[\boldsymbol{X}_i] \\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\mu} = \boldsymbol{\mu} \end{aligned}$$

$\Rightarrow$ zuivere schatter

$$\begin{aligned}
\mathrm{Cov}[\overline{\boldsymbol{X}}] &= \mathrm{E}[(\overline{\boldsymbol{X}}-\boldsymbol{\mu})(\overline{\boldsymbol{X}}-\boldsymbol{\mu})^\tau] \\
&= \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i-\boldsymbol{\mu})\right)\left(\frac{1}{n}\sum_{l=1}^n(\boldsymbol{X}_l-\boldsymbol{\mu})^\tau\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^n \underbrace{\mathrm{E}[(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_l-\boldsymbol{\mu})^\tau]}_{=0 \text{ als } i\neq l \text{ (onafhankelijkheid)}} \\
&= \frac{1}{n^2}\sum_{i=1}^n \underbrace{\mathrm{E}[(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\tau]}_{=\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{\Sigma}
\end{aligned}$$

2.

$$\begin{aligned}
\boldsymbol{W} &= \sum_{i=1}^n(\boldsymbol{X}_i-\overline{\boldsymbol{X}})(\boldsymbol{X}_i-\overline{\boldsymbol{X}})^\tau \\
&= \sum_{i=1}^n(\boldsymbol{X}_i-\overline{\boldsymbol{X}})\boldsymbol{X}_i^\tau + \left(\sum_{i=1}^n(\boldsymbol{X}_i-\overline{\boldsymbol{X}})\right)(-\overline{\boldsymbol{X}})^\tau \\
&= \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\tau - n\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^\tau \\
\Rightarrow \mathrm{E}[\boldsymbol{W}] &= \sum_{i=1}^n \mathrm{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\tau] - n\mathrm{E}[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^\tau]
\end{aligned}$$

Als $\boldsymbol{V}$ een toevalsveranderlijke is met $\mathrm{E}[\boldsymbol{V}] = \boldsymbol{\mu_V}$ en $\mathrm{Cov}[\boldsymbol{V}] = \boldsymbol{\Sigma_V}$, dan kunnen we nagaan dat $\mathrm{E}[\boldsymbol{V}\boldsymbol{V}^\tau] = \boldsymbol{\Sigma_V} + \boldsymbol{\mu_V}\boldsymbol{\mu_V}^\tau$. In het bijzonder is

$$\mathrm{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\tau] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\tau \quad \text{en} \quad \mathrm{E}[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^\tau] = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\tau.$$

En dus

$$\mathrm{E}[\boldsymbol{W}] = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}^\tau - n(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\tau) = (n-1)\boldsymbol{\Sigma}$$

waaruit volgt

$$\mathrm{E}[\boldsymbol{S}] = \frac{1}{n-1}\mathrm{E}[\boldsymbol{W}] = \boldsymbol{\Sigma}.$$

□

Merk op dat de empirische correlatiematrix $\boldsymbol{R}$ geen onvertekende schatter is van $\boldsymbol{\rho}$, maar wel een goede benadering.

### 3.3.3 Affiene equivariantie van $\overline{\boldsymbol{X}}$ en $\boldsymbol{S}$

**Eig.** $\overline{\boldsymbol{X}}$ en $\boldsymbol{S}$ zijn *affien equivariant.* Neem $\boldsymbol{A} \in \mathbb{R}^{q\times p}$, en $\boldsymbol{\delta} \in \mathbb{R}^{q\times 1}$, stel $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\delta}$, dan is

- $\overline{\boldsymbol{Y}} = \boldsymbol{A}\overline{\boldsymbol{X}} + \boldsymbol{\delta}$
- $\boldsymbol{S}_{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{X}}\boldsymbol{A}^{\tau}$

## 3.4 Mahalanobis afstand

De Euclidische afstand tussen twee punten $\boldsymbol{x}$ en $\boldsymbol{y} \in \mathbb{R}^p$ is gedefinieerd als

**Def.**

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\tau (\boldsymbol{x} - \boldsymbol{y})}$$

met bijhorende norm

**Def.**

$$\|\boldsymbol{x}\|_E = d_E(\boldsymbol{x}, 0) = \sqrt{\boldsymbol{x}^\tau \boldsymbol{x}}$$

Punten die op gelijke Euclidische afstand $c$ van de oorsprong liggen, behoren tot de sfeer $S : \sum_{i=1}^{p} x_i^2 = c^2$. Wanneer de punten $\boldsymbol{x}$ en $\boldsymbol{y}$ observaties zijn voor $p$ variabelen, volgt daaruit dat elk van de $p$ variabelen evenveel gewicht heeft bij de berekening van de afstand. Bij statistiek is het echter gebruikelijk om variabelen met een grote spreiding minder gewicht te geven dan variabelen met een kleine spreiding. Dit kan men bekomen door de variabelen te herleiden naar een eenheidsschaal:

$$\boldsymbol{x}_s = (\frac{x_1}{s_1}, \ldots, \frac{x_p}{s_p}) \qquad \text{en} \qquad \boldsymbol{y}_s = (\frac{y_1}{s_1}, \ldots, \frac{y_p}{s_p})$$

waarbij $s_i$ de standaardafwijking is van de $i$-de variabele.
De afstand tusen deze gestandaardiseerde punten wordt dan

$$d_E(\boldsymbol{x}_s, \boldsymbol{y}_s) = \sqrt{\sum_{i=1}^{p} \left(\frac{x_i - y_i}{s_i}\right)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\tau \boldsymbol{D}^{-1} (\boldsymbol{x} - \boldsymbol{y})} = d_{\boldsymbol{D}}(\boldsymbol{x}, \boldsymbol{y})$$

met $\boldsymbol{D} = diag(s_1^2, \ldots, s_p^2)$. De bijhorende norm wordt

$$\|\boldsymbol{x}\|_{\boldsymbol{D}} = \|\boldsymbol{x}_s\|_E = \sqrt{\boldsymbol{x}^\tau \boldsymbol{D}^{-1} \boldsymbol{x}}.$$

Observaties op gelijke afstand van de oorsprong liggen op de ellipsoïde met als vergelijking $E : \|\boldsymbol{x}\|_{\boldsymbol{D}}^2 = \sum_{i=1}^{p} (\frac{x_i}{s_i})^2 = c^2$. Dit is een ellipsoïde waarvan de assen evenwijdig liggen met de coördinaatassen, de grootste as ligt in de richting van de variabele met de grootste spreiding.

Tenslotte kunnen we ook nog de correlatie tussen de verschillende variabelen in rekening brengen. Hiertoe zouden de assen van de ellipsoïde het lineaire verband tussen de verschillende variabelen moeten weerspiegelen. Door de matrix $\boldsymbol{D}$ te vervangen door de volledige covariantiematrix $\boldsymbol{S}$ van de gegevens, wordt de ellipsoïde geroteerd.

**Def.** De *Mahalanobis afstand* of *statistische afstand* tussen twee punten $\boldsymbol{x}$ en $\boldsymbol{y} \in \mathbb{R}^p$ wordt gedefinieerd als

$$d_{\boldsymbol{S}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\tau \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{y})}$$

De statistische norm van $\boldsymbol{x}$ is

$$\|\boldsymbol{x}\|_{\boldsymbol{S}} = \sqrt{\boldsymbol{x}^\tau \boldsymbol{S}^{-1} \boldsymbol{x}}$$

**Opmerking.**

De matrix $\boldsymbol{S}$ is steeds positief semidefiniet vermits

$$\boldsymbol{z}^\tau \boldsymbol{S} \boldsymbol{z} = \frac{1}{n-1} \sum_{i=1}^{n} \boldsymbol{z}^\tau (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\tau \boldsymbol{z} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{z}^\tau (\boldsymbol{x}_i - \overline{\boldsymbol{x}}))^2 \geqslant 0$$

Als bovendien $|\boldsymbol{S}| > 0$, dan is $\boldsymbol{S} \in \mathrm{PD}(p) \Rightarrow \|\boldsymbol{x}\|_{\boldsymbol{S}}$ en $d(\boldsymbol{x}, \boldsymbol{y})_{\boldsymbol{S}}$ bestaan.

**Eig.**

- $\|\boldsymbol{x}\|_{\boldsymbol{S}}$ is een norm op de vectorruimte $\mathbb{R}^{p \times 1}$.
- $d(\boldsymbol{x}, \boldsymbol{y})_{\boldsymbol{S}}$ is een metriek op $\mathbb{R}^{p \times 1}$.

Observaties op gelijke afstand $c$ van de oorsprong voldoen nu aan de vergelijking $\|\boldsymbol{x}\|_{\boldsymbol{S}}^2 = \boldsymbol{x}^\tau \boldsymbol{S}^{-1} \boldsymbol{x} = c^2$. $\boldsymbol{S} \in \mathrm{PD}(p)$ heeft een spectraaldecompositie $\boldsymbol{S} = \boldsymbol{P \Lambda P}^\tau$ met $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p > 0$. Bijgevolg is $\boldsymbol{S}^{-1}$ ook positief definiet met spectraaldecompositie

$$\boldsymbol{S}^{-1} = \boldsymbol{P \Lambda}^{-1} \boldsymbol{P}^\tau = \frac{1}{\lambda_1} \boldsymbol{e}_1 \boldsymbol{e}_1^\tau + \ldots + \frac{1}{\lambda_p} \boldsymbol{e}_p \boldsymbol{e}_p^\tau$$

zodat

$$\|\boldsymbol{x}\|_{\boldsymbol{S}}^2 = \frac{1}{\lambda_1} (\boldsymbol{e}_1^\tau \boldsymbol{x})^2 + \frac{1}{\lambda_2} (\boldsymbol{e}_2^\tau \boldsymbol{x})^2 + \ldots + \frac{1}{\lambda_p} (\boldsymbol{e}_p^\tau \boldsymbol{x})^2 = c^2$$

Gebruikmakend van de transformatie

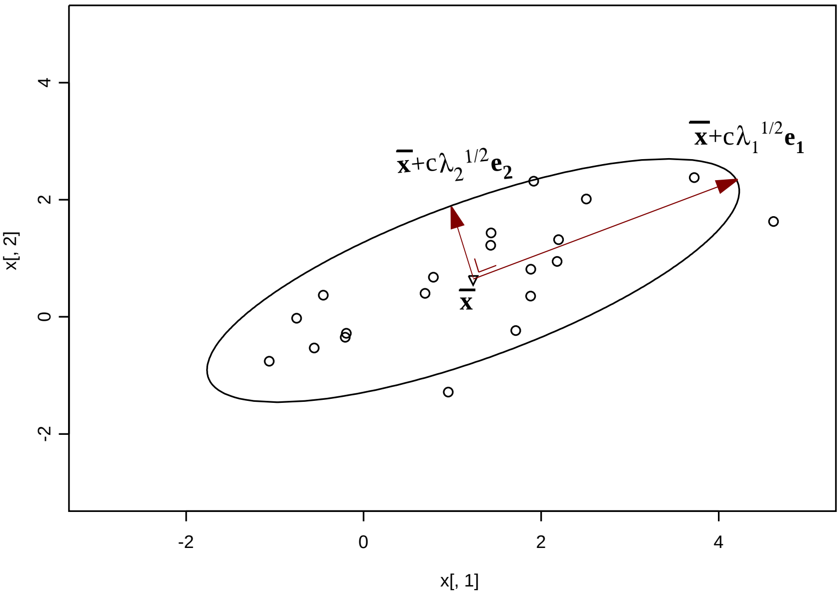
$$y_1 = \boldsymbol{e}_1^\tau \boldsymbol{x}, y_2 = \boldsymbol{e}_2^\tau \boldsymbol{x}, \ldots, y_p = \boldsymbol{e}_p^\tau \boldsymbol{x}$$

vinden we

$$\frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \ldots + \frac{1}{\lambda_p}y_p^2 = c^2$$

wat de standaard vergelijking is van een ellipsoide met assen gelegen in de richting van de eigenvectoren $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$, en waarbij de lengte van de halve as in elke richting $\boldsymbol{e}_h$ gegeven wordt door $c\sqrt{\lambda_h}$.

Vaak is men geïnteresseerd in de afstand van een observatie tot het centrum van de gegevens, wat niet steeds overeenkomt met de oorsprong. Observaties met gelijke afstand $d_{\boldsymbol{S}}(\boldsymbol{x}, \overline{\boldsymbol{x}})$ bevinden zich op een ellipsoïde met centrum $\overline{\boldsymbol{x}}$.



Hier zijn $\lambda_1 \geqslant \lambda_2$ de eigenwaarden van $\boldsymbol{S}$ met bijhorende eigenvectoren $\boldsymbol{e}_1$ en $\boldsymbol{e}_2$.

## 3.5 De veralgemeende variantie en de totale variantie

De covariantiematrix $\boldsymbol{S}$ wordt soms samengevat in één enkele waarde. hiervoor gebruikt men de *veralgemeneende variantie* of *de totale variantie.*

**Def.** (S.Wilks 1932) De *veralgemeende variantie* $:= |\boldsymbol{S}|$.

Voor $p = 1$ vinden we $|\boldsymbol{S}| = s_{11} = s^2$, vandaar de benaming 'veralgemeende' variantie.

Als $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ de eigenwaarden zijn van $\boldsymbol{S}$, dan weten we dat

$$|\boldsymbol{S}| = \lambda_1 \lambda_2 \ldots \lambda_p$$

en

$$|\boldsymbol{S}| \neq 0 \Leftrightarrow \lambda_p > 0 \Leftrightarrow \boldsymbol{S} \in \mathrm{PD}(p) \Leftrightarrow \boldsymbol{S}^{-1} \in \mathrm{PD}(p).$$

Uiteraard kan het begrip veralgemeende variantie ook voor een populatieverdeling gebruikt worden, dus $|\boldsymbol{\Sigma}|$.

**Meetkundige interpretatie**

Als $|\boldsymbol{S}| \neq 0$, bestaat de veralgemeende afstand

$$\|\boldsymbol{y}\|_{\boldsymbol{S}} = \sqrt{\boldsymbol{y}^{\tau} \boldsymbol{S}^{-1} \boldsymbol{y}} \quad \text{en} \quad d_{\boldsymbol{S}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{S}}.$$

Alle punten $\boldsymbol{y}$ op constante afstand $c$ van $\overline{\boldsymbol{x}}$ voldoen aan

$$\|\boldsymbol{y} - \overline{\boldsymbol{x}}\|_{\boldsymbol{S}}^2 = (\boldsymbol{y} - \overline{\boldsymbol{x}})^{\tau} \boldsymbol{S}^{-1} (\boldsymbol{y} - \overline{\boldsymbol{x}}) = c^2.$$

Dit is een ellipsoïde $E$ in $\mathbb{R}^{p \times 1}$ met centrum $\overline{\boldsymbol{x}}$. Het volume van deze ellipsoïde wordt gegeven door

$$\mathrm{Vol}(E) = k_p |\boldsymbol{S}|^{\frac{1}{2}} c^p \qquad \text{met } k_p = \mathrm{Vol}(S(0,1)) = \frac{2\pi^{p/2}}{p\Gamma(\frac{p}{2})}$$

dus $|\boldsymbol{S}| \propto \mathrm{Vol}^2(E)$.

**Opmerking.** De veralgemeende variantie is slechts een enkel getal, natuurlijk moeten we de volledige matrix $\boldsymbol{S}$ zelf kennen om de vorm van $E$ en dus de associatie tussen de $p$ variabelen te kennen.

**Result.**

$$\begin{aligned}|\boldsymbol{S}| = 0 \;&\Leftrightarrow\; \text{rang}(\boldsymbol{S}) < p \Leftrightarrow \lambda_p = 0\\ &\Leftrightarrow\; \text{alle } \boldsymbol{x}_i - \overline{\boldsymbol{x}} \text{ liggen op een vectorieel hypervlak in } \mathbb{R}^{p\times 1}\\ &\quad\;\text{dat door } \mathbf{0} \text{ gaat } (= \boldsymbol{e}_p^{\perp}).\\ &\Leftrightarrow\; \text{alle } \boldsymbol{x}_i \text{ liggen op een affien hypervlak in } \mathbb{R}^{p\times 1}\\ &\quad\;\text{dat door } \overline{\boldsymbol{x}} \text{ gaat } (= \overline{\boldsymbol{x}} + \boldsymbol{e}_p^{\perp}).\\ &\Leftrightarrow\; \text{alle } \boldsymbol{x}_i \text{ liggen op een affien hypervlak } H \subset \mathbb{R}^{p\times 1}\\ &\quad(\Rightarrow \overline{\boldsymbol{x}} \in H).\end{aligned}$$

In dat geval kan men een of meerdere variabelen laten vallen en dus de dimensie $p$ reduceren.

Indien we willen werken met $|\boldsymbol{S}|$ is het dus ook niet interessant om minder waarnemingen dan dimensies te hebben:

**Result.** Als $n \leqslant p \Rightarrow |\boldsymbol{S}| = 0$

*Proof.*

$$\begin{aligned}&\textstyle\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \mathbf{0}\\ \Rightarrow\;& \text{rang}(\underbrace{\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \ldots, \boldsymbol{x}_n - \overline{\boldsymbol{x}}}_{=\boldsymbol{A}(p\times n)}) \leqslant n-1\\ \Rightarrow\;& \text{rang}(\boldsymbol{S}) = \text{rang}(\boldsymbol{A}\boldsymbol{A}^{\tau}) \leqslant n-1 < p\\ \Rightarrow\;& |\boldsymbol{S}| = 0\end{aligned}$$

□

**Verband met $|\boldsymbol{R}|$**

We kunnen de variabele $j$ $(1 \leqslant j \leqslant p)$ standaardiseren door de observaties $x_{1j}, \ldots, x_{nj}$ te vervangen door

$$z_{1j} = \frac{x_{1j} - \overline{x}_{.j}}{\sqrt{s_{jj}}}, \ldots, z_{nj} = \frac{x_{nj} - \overline{x}_{.j}}{\sqrt{s_{jj}}}.$$

**Result.**

$$\boldsymbol{S}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)=\boldsymbol{R}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$$
$$|\boldsymbol{R}|=\frac{1}{s_{11}s_{22}\ldots s_{pp}}|\boldsymbol{S}|$$

Als we de eenheden van de variabelen wijzigen ($x_{ij}\rightarrow y_{ij}=a_jx_{ij}+b_j$) vinden we dezelfde gestandaardiseerde waarden $z_{ij}$ terug, en dus

**Result.**

$$|\boldsymbol{R}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)|=|\boldsymbol{R}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)|$$

Dus is $|\boldsymbol{R}|$ invariant voor een wijziging van de meeteenheden.

Naast de veralgemeende variantie definieert men ook de totale variantie.

**Def.** De *totale variantie* is

$$tr(\boldsymbol{S})=s_{11}+\ldots+s_{pp}\qquad\text{populatie: } tr(\boldsymbol{\Sigma})$$

Vanzelfsprekend is steeds $tr(\boldsymbol{R})=p$.
Net zoals de veralgemeende variantie kan de totale variantie ook uitgedrukt worden in functie van de eigenwaarden

**Result.** Voor de totale variantie geldt

$$tr(\boldsymbol{S})=s_{11}+s_{22}+\ldots+s_{pp}=\lambda_1+\lambda_2+\ldots+\lambda_p$$

*Proof.*
Door de spectraaldecompositie kan $\boldsymbol{S}$ geschreven worden als $\boldsymbol{S}=\sum_{j=1}^{p}\lambda_j\boldsymbol{e}_j\boldsymbol{e}_j^\tau=\boldsymbol{P\Lambda P}^\tau$. $\boldsymbol{\Lambda}$ is de diagonaal matrix met de eigenwaarden en $\boldsymbol{P}=[\boldsymbol{e}_1,\ldots,\boldsymbol{e}_p]$, dus $\boldsymbol{PP}^\tau=\boldsymbol{I}=\boldsymbol{P}^\tau\boldsymbol{P}$. We bekomen dan

$$tr(\boldsymbol{S})=tr(\boldsymbol{P\Lambda P}^\tau)=tr(\boldsymbol{\Lambda PP}^\tau)=tr(\boldsymbol{\Lambda})=\lambda_1+\lambda_2+\ldots+\lambda_p.$$

□

# Chapter 4

# De multivariate normale verdeling

## 4.1 Definitie

De multivariate normale verdeling is een verdeling op $(\mathbb{R}^{p\times 1}, \mathcal{R}^p)$ dus '$p$-variaat'. Notaties:

$$\mathbf{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p\times 1} = \mathbb{R}^{p\times 1} \quad \text{en} \quad \boldsymbol{I} = \boldsymbol{I}_p = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \in \mathbb{R}^{p\times p}$$

**Def.** De $p$-variate **standaard** normale verdeling is

$$N_p(\mathbf{0}, \boldsymbol{I}_p) = \text{verdeling van } \boldsymbol{Z} = (Z_1, \dots, Z_p)^\tau \text{ als alle}$$
$$Z_i \sim N(0,1) \text{ onafhankelijk zijn.}$$

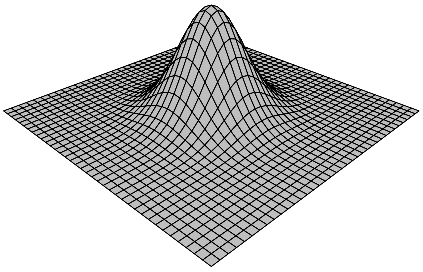
(De univariate $N(0,1)$ kunnen we dus als $N_1(0,1)$ noteren.)

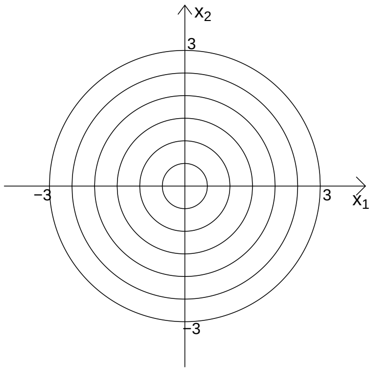
Dichtheidsfunktie:

$$\begin{aligned} f_{\mathbf{0},\boldsymbol{I}}(\boldsymbol{x}) = \phi(x_1)\dots\phi(x_p) &= \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(x_1^2 + \cdots + x_p^2)} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\boldsymbol{x}^\tau \boldsymbol{x}} = \boxed{\frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\|\boldsymbol{x}\|^2}} \end{aligned}$$

Men noemt deze verdeling **sferisch symmetrisch** omdat $f_{\boldsymbol{0},\boldsymbol{I}}(\boldsymbol{x})$ enkel van $\|\boldsymbol{x}\|$ afhangt. De contours van $f_{\boldsymbol{0},\boldsymbol{I}}$ worden gegeven door

$$f_{\boldsymbol{0},\boldsymbol{I}}(\boldsymbol{x}) = c \Leftrightarrow \|\boldsymbol{x}\|^2 = c \Leftrightarrow \boldsymbol{x} \in S(\boldsymbol{0}, c).$$





- $\text{Mod}[\boldsymbol{Z}] = \text{argmax}_{\boldsymbol{x}}\, f_{\boldsymbol{0},\boldsymbol{I}}(\boldsymbol{x}) = 0$

- $\text{E}[\boldsymbol{Z}] = \boldsymbol{0}$

- $\text{Cov}[\boldsymbol{Z}] = \boldsymbol{I}_p$

Om de **algemene** normale verdeling in te voeren, beschouwen we een $p$-variate stochastiek

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$$

waarbij $\boldsymbol{\mu} \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ met $det(\boldsymbol{A}) \neq 0$. Noteer $\boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{A}^{\tau}$, dan is

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{|det(\boldsymbol{A})|} f_{\boldsymbol{Z}}(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \\ &= \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\tau}(\boldsymbol{A}^{-1})^{\tau}\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \end{aligned}$$

Deze dichtheidsfunktie hangt dus enkel af van $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$, en niet meer van $\boldsymbol{A}$ zelf.

**Def.** Voor elke $\boldsymbol{\mu} \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p} \in PD(p)$ definieert men de verdeling $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ door haar dichtheidsfunktie

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}}$$

(voor alle $\boldsymbol{x} \in \mathbb{R}^{p\times 1}$)

Men noemt deze verdeling **elliptisch symmetrisch** omdat de dichtheid constant is op ellipsoïdes :

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x) = c > 0 \Leftrightarrow (\boldsymbol{x}-\boldsymbol{\mu})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$$

met centrum $\boldsymbol{\mu}$, waarvan $\boldsymbol{\Sigma}$ de vorm en $c$ de grootte bepalen.

Bepaal de eigenwaarden $\lambda_1 \geqslant \cdots \geqslant \lambda_p > 0$ van $\boldsymbol{\Sigma}$, met bijhorende orthonormale eigenvectoren $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$. Dan vormen

$$\boldsymbol{\mu} \pm \boldsymbol{e}_1\sqrt{c\lambda_1}, \ldots, \boldsymbol{\mu} \pm \boldsymbol{e}_p\sqrt{c\lambda_p}$$

de **hoofdassen** van de ellipsoïde. We verifiëren dat

$$\begin{aligned}&(\boldsymbol{\mu} \pm \boldsymbol{e}_j\sqrt{c\lambda_j} - \boldsymbol{\mu})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} \pm \boldsymbol{e}_j\sqrt{c\lambda_j} - \boldsymbol{\mu})\\ &= c\lambda_j\boldsymbol{e}_j^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j = c\lambda_j\boldsymbol{e}_j^{\tau}\lambda_j^{-1}\boldsymbol{e}_j = c\end{aligned}$$

## Eigenschappen van $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Als $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dan geldt voor een willekeurige wortel $\boldsymbol{A}$ van $\boldsymbol{\Sigma}$ dat

$$\boldsymbol{Z} := \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$$

- $\mathrm{Mod}[\boldsymbol{X}] = \mathrm{Mod}(f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \boldsymbol{\mu}$

- $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$

- $\mathrm{Cov}[\boldsymbol{X}] = \boldsymbol{A}\mathrm{Cov}[\boldsymbol{Z}]\boldsymbol{A}^{\tau} = \boldsymbol{\Sigma}$

**Eig.** Als $\boldsymbol{X}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ en $\boldsymbol{X}_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ onafhankelijk:

$$\boldsymbol{X}_1 + \boldsymbol{X}_2 \sim N_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$$

*Proof.* Zie cursus Wiskundige Statistiek. □

## 4.2 Karakterisatie door lineaire combinaties

We nemen lineaire combinaties $\boldsymbol{x} \to \boldsymbol{\beta}^\tau \boldsymbol{x}$, dit komt neer op projecties op rechten.

**Stelling.** Neem $\boldsymbol{X} : (\Omega, \mathcal{A}, P) \to \mathbb{R}^{p\times 1}, \boldsymbol{\mu} \in \mathbb{R}^{p\times 1}, \boldsymbol{\Sigma} \in PD(p)$, dan geldt:

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;\Leftrightarrow\; \forall\, \mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^{p\times 1} : \quad \boldsymbol{\beta}^\tau \boldsymbol{X} \sim N_1(\boldsymbol{\beta}^\tau \boldsymbol{\mu}, \boldsymbol{\beta}^\tau \boldsymbol{\Sigma} \boldsymbol{\beta})$$

*Proof.* Zie cursus Wiskundige Statistiek. □

**Gevolg.** Als $\boldsymbol{X} : (\Omega, \mathcal{A}, P) \to \mathbb{R}^{p\times 1}$ en $\forall\, \mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is de variabele $\boldsymbol{\beta}^\tau \boldsymbol{X}$ normaal verdeeld, dan is $\boldsymbol{X}$ multivariaat normaal verdeeld.

*Proof.* Neem $\boldsymbol{\beta}_{(j)}^\tau = [0 \dots \underbrace{1}_{\text{plaats } j} \dots 0]$, dan

$$0 < \sigma_{jj} = \mathrm{Var}[X_j] = \mathrm{Var}[\boldsymbol{\beta}_{(j)}^\tau \boldsymbol{X}] < \infty \qquad (\forall\, 1 \leqslant j \leqslant p)$$

en dus $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{X}]$ en $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{X}]$ bestaan. Voor elke $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ vinden we dan

$$\begin{cases} \mathrm{E}[\boldsymbol{\beta}^\tau \boldsymbol{X}] = \boldsymbol{\beta}^\tau \mathrm{E}[\boldsymbol{X}] = \boldsymbol{\beta}^\tau \boldsymbol{\mu} \\ \mathrm{Var}[\boldsymbol{\beta}^\tau \boldsymbol{X}] = \mathrm{E}[(\boldsymbol{\beta}^\tau \boldsymbol{X} - \boldsymbol{\beta}^\tau \boldsymbol{\mu})(\boldsymbol{\beta}^\tau \boldsymbol{X} - \boldsymbol{\beta}^\tau \boldsymbol{\mu})^\tau] = \boldsymbol{\beta}^\tau \boldsymbol{\Sigma} \boldsymbol{\beta} \end{cases}$$

en dan kunnen we de vorige stelling toepassen. □

Voor meerdere lineaire combinaties vinden we

**Stelling.** Als $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), q \leqslant p, \boldsymbol{B} \in \mathbb{R}^{q\times p}$ met $rang(\boldsymbol{B}) = q$ en $\boldsymbol{\delta} \in \mathbb{R}^{q\times 1}$:

$$\boldsymbol{BX} + \boldsymbol{\delta} = \begin{pmatrix} b_{11}X_1 + \ldots + b_{1p}X_p + d_1 \\ \vdots \\ b_{q1}X_1 + \ldots + b_{qp}X_p + d_q \end{pmatrix} \sim N_q(\boldsymbol{B\mu} + \boldsymbol{\delta}, \boldsymbol{B\Sigma B}^\tau)$$

*Proof.* Voor elke vector $\mathbf{0} \neq \boldsymbol{a} \in \mathbb{R}^{q\times 1}$ weten we (vorige stelling)

$$\begin{aligned} \boldsymbol{a}^\tau(\boldsymbol{BX}) = (\boldsymbol{B}^\tau\boldsymbol{a})^\tau\boldsymbol{X} \quad &\sim \quad N_1((\boldsymbol{B}^\tau\boldsymbol{a})^\tau\boldsymbol{\mu}, \underbrace{(\boldsymbol{B}^\tau\boldsymbol{a})^\tau\boldsymbol{\Sigma}(\boldsymbol{B}^\tau\boldsymbol{a})}_{>0,\text{ want } \boldsymbol{B}^\tau\boldsymbol{a}\neq 0}) \\ & \quad\quad \| \\ & \quad\quad N_1(\boldsymbol{a}^\tau(\boldsymbol{B\mu}), \boldsymbol{a}^\tau(\boldsymbol{B\Sigma B}^\tau)\boldsymbol{a}) \end{aligned}$$

Door de vorige stelling in omgekeerde richting te gebruiken, geeft dit:

$$\boldsymbol{BX} \sim N_q(\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}^\tau) \qquad \text{met } \boldsymbol{B\Sigma B}^\tau \in PD(q)$$

en dus $\boldsymbol{BX} + \boldsymbol{\delta} \sim N_q(\boldsymbol{B\mu} + \boldsymbol{\delta}, \boldsymbol{B\Sigma B}^\tau)$. □

__Voorbeeld.__

Als $\boldsymbol{X}$ verdeeld is als $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, vindt dan de verdeling van

$$\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \boldsymbol{BX}$$

Wegens voorgaande eigenschappen is de verdeling van $\boldsymbol{BX}$ opnieuw multivariaat normaal met gemiddelde

$$\boldsymbol{B\mu} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{pmatrix}$$

en covariantiematrix

$$\begin{aligned}
\boldsymbol{B\Sigma B}^\tau &= \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_{22} - \sigma_{23} & \sigma_{23} - \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} \\ \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{pmatrix}
\end{aligned}$$

## 4.3 Marginale verdelingen

**Stelling.** Stel dat $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan

1. $\forall\, 1 \leqslant j \leqslant p : \quad X_j \sim N_1(\mu_j, \sigma_{jj})$
2. Als we $\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ als volgt opdelen

$$\boldsymbol{X} = \left(\begin{array}{c} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{array}\right) \begin{array}{l} \}q \\ \}p-q \end{array} \qquad \boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right) \qquad \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$$

$\Rightarrow \boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

*Proof.* 1. Neem de lineaire combinatie $\boldsymbol{\beta}^\tau_{(j)}\boldsymbol{X}$ met

$$\boldsymbol{\beta}^\tau_{(j)} = [0, \ldots, \underbrace{1}_{\text{plaats } j}, \ldots, 0].$$

$$\Rightarrow \boldsymbol{\beta}_{(j)}^{\tau}\boldsymbol{X} = X_j, \boldsymbol{\beta}_{(j)}^{\tau}\boldsymbol{\mu} = \mu_j, \boldsymbol{\beta}_{(j)}\boldsymbol{\Sigma}\boldsymbol{\beta}_{(j)}^{\tau} = \sigma_{jj}.$$

2. Neem de matrix
$$\boldsymbol{B} = \left( \boldsymbol{I}_q \;\middle|\; \underbrace{\boldsymbol{0}}_{q\times(p-q)} \right)$$
$$\Rightarrow \boldsymbol{BX} = \boldsymbol{X}_1,\ \boldsymbol{B\mu} = \boldsymbol{\mu}_1,\ \boldsymbol{B\Sigma B}^{\tau} = \boldsymbol{\Sigma}_{11} \text{ dus } \boldsymbol{\Sigma}_{11} \in PD(q).$$

□

**Opmerking.**

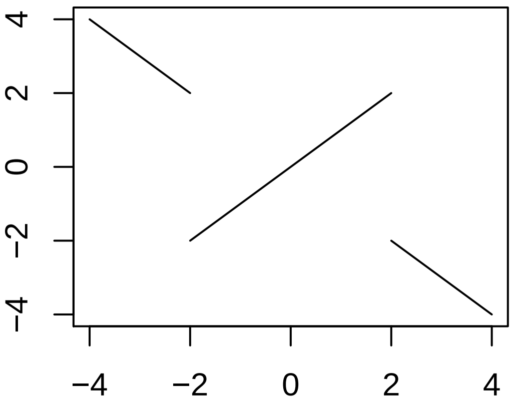
Het is niet voldoende dat alle marginalen normaal verdeeld zijn opdat ook $\boldsymbol{X}$ normaal zou zijn. Neem bijvoorbeeld

$$\begin{aligned} X_1 &\sim N_1(0,1) \\ X_2 &= \begin{cases} X_1 & \text{als } |X_1| \leqslant 2 \\ -X_1 & \text{als } |X_1| > 2 \end{cases} \end{aligned}$$

In dat geval is ook

$$X_2 \sim N_1(0,1)$$

maar $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ volgt geen bivariate normaalverdeling. Immers, het bereik van $\boldsymbol{X}$ is:



In dit voorbeeld bezit bijvoorbeeld de projectie $\boldsymbol{\beta}^{\tau}\boldsymbol{X}$ met $\boldsymbol{\beta} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ geen univariate normale verdeling.

**Voorbeeld.**

Stel $\boldsymbol{X}$ is verdeeld als $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Wat is dan de verdeling van $\begin{pmatrix} X_2 \\ X_4 \end{pmatrix}$?

We kunnen $\boldsymbol{X}$ partitioneren als

$$\tilde{\boldsymbol{X}} = \left(\begin{array}{c} X_2 \\ X_4 \\ \hline X_1 \\ X_3 \\ X_5 \end{array}\right), \qquad \tilde{\boldsymbol{\mu}} = \left(\begin{array}{c} \mu_2 \\ \mu_4 \\ \hline \mu_1 \\ \mu_3 \\ \mu_5 \end{array}\right), \qquad \tilde{\boldsymbol{\Sigma}} = \left(\begin{array}{cc|ccc} \sigma_{22} & \sigma_{24} & \sigma_{12} & \sigma_{23} & \sigma_{25} \\ \sigma_{24} & \sigma_{44} & \sigma_{14} & \sigma_{34} & \sigma_{45} \\ \hline \sigma_{12} & \sigma_{14} & \sigma_{11} & \sigma_{13} & \sigma_{15} \\ \sigma_{23} & \sigma_{34} & \sigma_{13} & \sigma_{33} & \sigma_{35} \\ \sigma_{25} & \sigma_{45} & \sigma_{15} & \sigma_{35} & \sigma_{55} \end{array}\right)$$

Dus weten we

$$\begin{pmatrix} X_2 \\ X_4 \end{pmatrix} \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = N_2\left(\begin{pmatrix} \mu_2 \\ \mu_4 \end{pmatrix}, \begin{pmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{pmatrix}\right)$$

De normale verdeling voor een willekeurige subset van variabelen kan dus bekomen worden door de juiste selectie van gemiddeldes en covarianties uit de oorspronkelijke $\boldsymbol{\Sigma}$ te nemen.

## 4.4 Covariantie tussen multivariate componenten

In het algemeen geldt voor $X_1, X_2 \in L^2(\Omega, \mathcal{A}, P)$

$$X_1, X_2 \text{ onafhankelijk} \quad \Rightarrow \quad \sigma_{12} = \operatorname{Cov}(X_1, X_2) = 0 \quad \text{niet gecorreleerd}$$
$$\not\Leftarrow$$

In het specifieke geval van een bivariaat normale verdeling vinden we

**Result.** Als $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan geldt

$$X_1, X_2 \text{ zijn onafhankelijk } \Leftrightarrow \operatorname{Cov}(X_1, X_2) = 0$$

*Proof.* 1. $\boxed{\Rightarrow}$

gekend

2. $\boxed{\Leftarrow}$

$$\sigma_{12} = 0 \Rightarrow \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix} \qquad \text{en } \rho_{12} = 0$$

$$\begin{aligned} \Rightarrow f(x_1, x_2) &= \frac{1}{\sqrt{2\pi\sigma_{11}}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2} \frac{1}{\sqrt{2\pi\sigma_{22}}} e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2} \\ &= f_{N_1(\mu_1,\sigma_{11})}(x_1) f_{N_1(\mu_2,\sigma_{22})}(x_2) \end{aligned}$$

$\Rightarrow X_1$ en $X_2$ zijn onafhankelijk.

□

**Interpretatie.**

Als de assen van de contourellipsen $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{z}) = c$ parallel lopen met de coördinaatassen (dus horizontaal of verticaal), dan zijn de variabelen onafhankelijk. Als de assen niet parallel zijn met de coördinaatassen dan hebben we afhankelijkheid.

We kunnen dit veralgemenen naar het multivariate geval:

**Result.**

1. Als $\boldsymbol{X}_1 : (\Omega, \mathcal{A}, P) \to \mathbb{R}^{p_1 \times 1}$ en $\boldsymbol{X}_2 : (\Omega, \mathcal{A}, P) \to \mathbb{R}^{p_2 \times 1}$ onafhankelijk zijn en $\boldsymbol{\Sigma}_{\boldsymbol{X}_1}$ en $\boldsymbol{\Sigma}_{\boldsymbol{X}_2}$ bestaan
$\Rightarrow \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbf{0} \in \mathbb{R}^{p_1 \times p_2}$.

2. Als $\left(\begin{array}{c} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{array}\right) \sim N_{p_1+p_2}\left(\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array}\right), \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{array}\right)\right)$ dan

$$\boldsymbol{X}_1 \text{ en } \boldsymbol{X}_2 \text{ zijn onafhankelijk } \Leftrightarrow \boldsymbol{\Sigma}_{12} = \mathbf{0} \in \mathbb{R}^{p_1 \times p_2}$$

3. Als $\boldsymbol{X}_1 \sim N_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ en $\boldsymbol{X}_2 \sim N_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ onafhankelijk zijn

$$\Rightarrow \left(\begin{array}{c} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{array}\right) \sim N_{p_1+p_2}\left(\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array}\right), \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{array}\right)\right)$$

*Proof.* Zoals in het bivariate geval ($p_1 = p_2 = 1$). □

**Voorbeeld.**

Stel $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ met

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Vermits $X_1$ en $X_2$ covariantie $\sigma_{12} = 1$ hebben, zijn ze niet onafhankelijk. Als we $\boldsymbol{X}$ en $\boldsymbol{\Sigma}$ partitioneren als

$$\boldsymbol{X} = \left(\begin{array}{c} X_1 \\ X_2 \\ \hline X_3 \end{array}\right), \qquad \boldsymbol{\Sigma} = \left(\begin{array}{cc|c} 4 & 1 & 0 \\ 1 & 3 & 0 \\ \hline 0 & 0 & 2 \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$$

dan zien we dat $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ en $X_3$ covariantiematrix $\boldsymbol{\Sigma}_{12} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ hebben. Bijgevolg zijn $(X_1, X_2)$ en $X_3$ onafhankelijk. Daaruit volgt dat $X_3$ ook onafhankelijk is van $X_1$ en van $X_2$ afzonderlijk.

## 4.5 Conditionele verdelingen

**Def.** Als $X$ en $Y$ stochastische veranderlijken zijn met gemeenschappelijke dichtheid $f_{X,Y}(x,y)$ en marginale dichtheden $f_X$ en $f_Y$, dan definiëren we de conditionele dichtheid als

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{als } f_X(x) > 0 \\ 0 & \text{als } f_X(x) = 0 \end{cases}$$

De voorwaardelijke verwachtingswaarde van $Y$ gegeven $X = x$ wordt gedefinieerd als

$$\mathrm{E}[Y|X = x] = \int y f_{Y|X}(y|x) dy$$

**Eig.** Als $\boldsymbol{X} = \left(\begin{array}{c} \boldsymbol{X}_1 \\ \hline \boldsymbol{X}_2 \end{array}\right) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ met $\boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right)$, $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$ dan is $P_{\boldsymbol{X}_2|\boldsymbol{X}_1=\boldsymbol{x}_1}$ normaal, met

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1] &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \\ \mathrm{Cov}[\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1] &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{aligned}$$

(merk op dat $\boldsymbol{\Sigma} \in PD(p) \Rightarrow \boldsymbol{\Sigma}_{11} \in PD \Rightarrow \boldsymbol{\Sigma}_{11}^{-1}$ bestaat)

*Proof.* Stel

$$\boldsymbol{B}(p \times p) = \left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{0} \\ \hline -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I} \end{array}\right) \begin{array}{l} \}p-q \\ \}q \end{array}$$

$$\begin{aligned} \Rightarrow \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\mu}) &= \left(\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{0} \\ \hline -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I} \end{array}\right) \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) \end{pmatrix} \end{aligned}$$

moet een normale verdeling volgen met

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})] &= \boldsymbol{B}\mathrm{E}[\boldsymbol{X}-\boldsymbol{\mu}] = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \\
\mathrm{Cov}[\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})] &= \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^{\tau} \\
&= \left(\begin{array}{c|c} \boldsymbol{I} & \mathbf{0} \\ \hline -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I} \end{array}\right) \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right) \boldsymbol{B}^{\tau} \\
&= \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}+\boldsymbol{\Sigma}_{21} & -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}+\boldsymbol{\Sigma}_{22} \end{array}\right) \boldsymbol{B}^{\tau} \\
&= \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{array}\right) \left(\begin{array}{c|c} \boldsymbol{I} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \hline \mathbf{0} & \boldsymbol{I} \end{array}\right) \\
&= \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{11}(-\boldsymbol{\Sigma}_{11}^{-1})\boldsymbol{\Sigma}_{12}+\boldsymbol{\Sigma}_{12} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{array}\right) \\
&= \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{array}\right)
\end{aligned}
$$

De twee componenten van $\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})$ hebben covariantie $=\mathbf{0}$, dus zijn $\boldsymbol{X}_2-\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_1)$ en $\boldsymbol{X}_1-\boldsymbol{\mu}_1$ onafhankelijk. Bijgevolg geldt

$$
\begin{aligned}
& \boldsymbol{X}_2-\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_1) \sim N_q(\mathbf{0},\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \\
\Rightarrow\quad & \boldsymbol{X}_2-\underbrace{\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)}_{\text{constante}}|\boldsymbol{X}_1=\boldsymbol{x}_1 \sim N_q(\mathbf{0},\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \\
\Rightarrow\quad & \boldsymbol{X}_2|\boldsymbol{X}_1=\boldsymbol{x}_1 \sim N_q(\underbrace{\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)}_{\text{lineair in } \boldsymbol{x}_1},\underbrace{\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}_{\text{hangt niet af van } \boldsymbol{x}_1})
\end{aligned}
$$

□

## 4.6 Verdeling van de veralgemeende afstand

**Eig.** Als $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan

1. $d^2_{\boldsymbol{\Sigma}}(\boldsymbol{X}, \boldsymbol{\mu}) = (\boldsymbol{X} - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p$
2. $P[\boldsymbol{X} \in E_{1-\alpha}] = 1 - \alpha$ met $E_{1-\alpha}$ de ellipsoïde

$$E_{1-\alpha} = \{\boldsymbol{y} \in \mathbb{R}^{p\times 1}; (\boldsymbol{y} - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \leqslant \chi^2_{p,\alpha}\}$$

*Proof.* 1. $\exists\, \boldsymbol{A} \in \mathbb{R}^{p\times p}$ met $|\boldsymbol{A}| \neq 0$ zodat $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\tau$ dus

$$\begin{aligned}
\boldsymbol{Z} &:= \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p) \Rightarrow \boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} \\
&\Rightarrow (\boldsymbol{X} - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = (\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} - \boldsymbol{\mu}) \\
&= \boldsymbol{Z}^\tau \boldsymbol{A}^\tau (\boldsymbol{A}^\tau)^{-1} \boldsymbol{A}^{-1} \boldsymbol{A}\boldsymbol{Z} = \boldsymbol{Z}^\tau \boldsymbol{Z} \\
&= \sum_{j=1}^{p} Z_j^2 \sim \chi^2_p \quad \text{omdat } Z_1, \ldots, Z_p \ iid \sim N_1(0, 1)
\end{aligned}$$

2. $P[\boldsymbol{X} \in E_{1-\alpha}] = P[\underbrace{d^2_{\boldsymbol{\Sigma}}(\boldsymbol{X}, \boldsymbol{\mu})}_{\sim \chi^2_p} \leqslant \chi^2_{p,\alpha}] = 1 - \alpha$

□

Men noemt $E_{1-\alpha}$ de ellipsoïde met waarschijnlijkheid (tolerantie) $1 - \alpha$. De grens van $E_{1-\alpha}$ is een contour van $f_{N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$:

$$\partial E_{1-\alpha} = \{\boldsymbol{y} \in \mathbb{R}^{p\times 1}; d^2_{\boldsymbol{\Sigma}}(\boldsymbol{y}, \boldsymbol{\mu}) = \underbrace{\chi^2_{p,\alpha}}_{\text{constante}}\}$$

# Chapter 5

# Schatters in het parametrische model $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$

## 5.1 Schatters voor $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$

Stel $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ met $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$ onbekend.
De gezamenlijke dichtheid van de steekproef is dan

$$\begin{aligned} f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) &= \prod_{i=1}^{n} f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}_i) \\ &= \prod_{i=1}^{n} \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})} \right\} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})} \end{aligned}$$

Voor een vaste steekproef $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ kunnen we deze dichtheid beschouwen als een functie van de onbekende parameters $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$. We noemen dit dan de **likelihood functie** en noteren ze door $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Een algemene schattingsmethode voor onbekende parameters is de maximum likelihood (ML) methode die de waarden van de onbekende parameters bepaald zodat de likelihood functie maximaal is. Deze methode bepaald dus de parameterwaarden waarvoor de bekomen steekproef het meest aannemelijk is (de hoogst mogelijke dichtheid heeft). De ML schatters zijn dus gedefinieerd als:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) := \underset{(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n}_{\text{gegeven}})$$

In de appendix wordt het volgende aangetoond:

**Stelling.** Indien $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan is de ML schatter $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ gelijk aan

$$\begin{cases} \hat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} \\ \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\tau} = \frac{n-1}{n}\boldsymbol{S} \end{cases}$$

**Opmerkingen:**

- $E\left[\hat{\boldsymbol{\mu}}\right] = E\left[\overline{\boldsymbol{X}}\right] = \boldsymbol{\mu}$ zuivere schatter
- $E\left[\hat{\boldsymbol{\Sigma}}\right] = E\left[\frac{n-1}{n}\boldsymbol{S}\right] = \frac{n-1}{n}E\left[\boldsymbol{S}\right] = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$ onzuivere schatter
(de vertekening is klein want $\frac{n-1}{n} \approx 1$)

## 5.2 Multivariate centrale limietstelling

De multivariate centrale limietstelling is een cruciaal resultaat in de statistiek dat toelaat om statistische inferentie methoden te ontwikkelen.

**Stelling. Multivariate centrale limietstelling**
Stel $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ iid waarbij $\boldsymbol{\mu} := \mathrm{E}[\boldsymbol{X}_1] \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{\Sigma} := \mathrm{Cov}[\boldsymbol{X}_1] \in \mathbb{R}^{p\times p}$ bestaan, met $|\boldsymbol{\Sigma}| > 0$. Dan geldt:

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Dus is de schatter $\overline{\boldsymbol{X}}_n$ asymptotisch normaal, zelfs wanneer de observaties niet afkomstig zijn uit een normale verdeling.

Gevolg: Verdeling van $n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\tau \boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})$.
Indien $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan weten we dat $\overline{\boldsymbol{X}}_n \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$, dus de veralgemeende afstand voldoet aan:

$$d^2_{\frac{1}{n}\boldsymbol{\Sigma}}(\overline{\boldsymbol{X}}_n, \boldsymbol{\mu}) = \underbrace{(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\tau (\frac{1}{n}\boldsymbol{\Sigma})^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})}_{n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})} \sim \chi^2_p$$

Indien de toevalsvectoren $\boldsymbol{X}_i$ niet meer normaal verdeeld zijn, maar $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ wel iid zijn en $\boldsymbol{\mu} := \mathrm{E}[\boldsymbol{X}_1] \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{\Sigma} := \mathrm{Cov}[\boldsymbol{X}_1]$ bestaan, dan weten we dat voor $n \gg p$ geldt:

1. $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \approx N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ (CLS voor $\overline{\boldsymbol{X}}_n$)
   $\Rightarrow \overline{\boldsymbol{X}}_n \approx N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$
2. $\boldsymbol{S}_n \approx \boldsymbol{\Sigma}$ (consistentie van $\boldsymbol{S}_n$, zie cursus Wiskundige Statistiek)
   $\Rightarrow \boldsymbol{S}_n^{-1} \approx \boldsymbol{\Sigma}^{-1}$

Dus kunnen we volgende benadering gebruiken

**Result.**

$$n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^\tau \boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \approx \chi^2_p$$

In de praktijk is deze benadering goed wanneer $\frac{n}{p} \geqslant 5$.

# Chapter 6

# Testen van normaliteit

## 6.1 Inleiding

Veel technieken voor multivariate analyse zijn gebaseerd op de veronderstelling dat de steekproef getrokken is uit de normaalverdeling $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (met een $\boldsymbol{\mu} \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{\Sigma} \in PD(p)$ willekeurig). Door enkel gebruik te maken van de gegevens $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ willen we nagaan of de hypothese van normaliteit voldaan is.

Er bestaan ontzettend veel testen voor normaliteit, maar wij zullen ons hier beperken tot drie criteria:

1. **Univariate marginalen**
   Lijkt de marginale verdeling van elke component $j$ normaal verdeeld? Kan men dus $\forall\, 1 \leqslant j \leqslant p$ $x_{1j}, \dots, x_{nj}$ benaderen door een univariate normaalverdeling?

2. **Bivariate marginalen**
   Lijken de grafieken van de koppels van variabelen normaal verdeeld? Bezitten de koppels $\{(x_{1j}, x_{1k}), \dots, (x_{nj}, x_{nk})\}$ $\forall\, j \neq k$ elliptische omtreklijnen?

3. **Radiale marginalen**
   Zijn de kwadratische veralgemeende afstanden $MD^2(\boldsymbol{x}_1), \dots, MD^2(\boldsymbol{x}_n)$ $\chi^2$-verdeeld? Men noemt $MD(\boldsymbol{x}_i) = d_{\boldsymbol{S}}(\boldsymbol{x}_i, \overline{\boldsymbol{x}}) = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\tau \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})}$ de *Mahalanobis afstand.*

## 6.2 Univariate marginalen

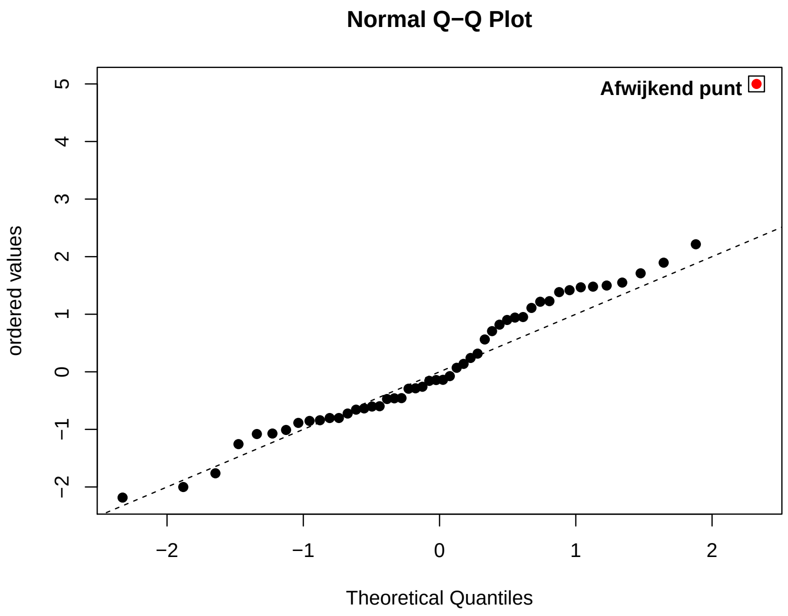
In het algemeen wil men (eerst visueel) nagaan of een univariate steekproef $x_1, \ldots, x_n \overset{?}{\approx} F$. Hiervoor is het zeer handig om gebruik te maken van de "kwantiel-kwantiel"grafiek ("QQ-plot") van Wilk en Gnanadesikan (1968).

- Op de horizontale as: het kwantiel $q_i = F^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right)$
- Op de vertikale as: de ordestatistiek $x_{i:n}\left[\equiv \hat{F}_n^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right)\right]$

**Result.** De QQ-plot is dus de grafiek van

$$\left\{\left(F^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right), x_{i:n}\right); 1 \leqslant i \leqslant n\right\}$$

<u>**Voorbeeld:**</u> Nagaan of $x_1, \ldots, x_{50} \sim N_1(0,1)$



Alle punten, behalve één, liggen dicht bij de bissectrice
⇒ goede fit, behalve voor 1 observatie.

Om de normaliteit $x_1, \ldots, x_n \overset{?}{\approx} N_1(\mu, \sigma^2)$ na te gaan met de QQ-plot, moeten we $\mu$ en $\sigma$ niet kennen. We kunnen gewoon $q_i = \Phi^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right)$ op de horizontale as plaatsen.

Reden:

$X_i \sim N_1(\mu, \sigma^2)$
$\Leftrightarrow Y_i := \frac{X_i - \mu}{\sigma} \sim N_1(0,1)$
$\Rightarrow$ QQ-plot van $\left(\Phi^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right), y_{i:n}\right)$ is lineair
$\Rightarrow$ QQ-plot van $\left(\Phi^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right), \underbrace{x_{i:n}}_{=\sigma y_{i:n}+\mu}\right)$ is lineair

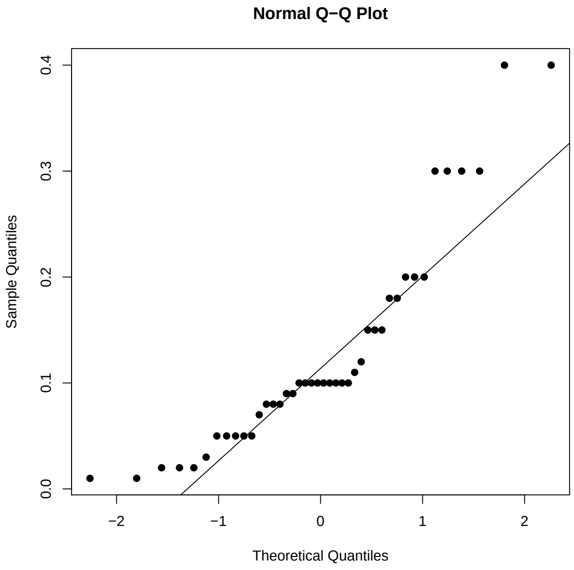
**Voorbeeld:**

Kwaliteitscontrole van magnetrons.

$X_i$ = uitgezonden straling met gesloten deur, $n = 42$.

| Oven | straling | Oven | straling | Oven | straling |
|---|---|---|---|---|---|
| 1 | 0.15 | 15 | 0.10 | 29 | 0.08 |
| 2 | 0.09 | 16 | 0.10 | 30 | 0.18 |
| 3 | 0.18 | 17 | 0.02 | 31 | 0.10 |
| 4 | 0.10 | 18 | 0.10 | 32 | 0.20 |
| 5 | 0.05 | 19 | 0.01 | 33 | 0.11 |
| 6 | 0.12 | 20 | 0.40 | 34 | 0.30 |
| 7 | 0.08 | 21 | 0.10 | 35 | 0.02 |
| 8 | 0.05 | 22 | 0.05 | 36 | 0.20 |
| 9 | 0.08 | 23 | 0.03 | 37 | 0.20 |
| 10 | 0.10 | 24 | 0.05 | 38 | 0.30 |
| 11 | 0.07 | 25 | 0.15 | 39 | 0.30 |
| 12 | 0.02 | 26 | 0.10 | 40 | 0.40 |
| 13 | 0.01 | 27 | 0.15 | 41 | 0.30 |
| 14 | 0.10 | 28 | 0.09 | 42 | 0.05 |

Aan de hand van de QQ-plot merken we dat de observaties afwijken van de normale verdeling (scheef naar rechts).



De *Shapiro-Wilk test* komt ongeveer overeen met het berekenen van de correlatie tussen de $q_i$ en de $x_{i:n}$ binnen de QQ-plot. De nulhypothese stelt dat de variabele normaal verdeeld is. Software levert de P-waarde, op basis waarvan men de nulhypothese al dan niet kan verwerpen.

```
shapiro.test(radclosed)

Shapiro-Wilk normality test


data:  radclosed
W = 0.85793, p-value = 9.902e-05
```

In dit voorbeeld levert de Shapiro-Wilk test dus een P-waarde gelijk aan 0.000099, en bijgevolg verwerpen we de aanname van normaliteit.

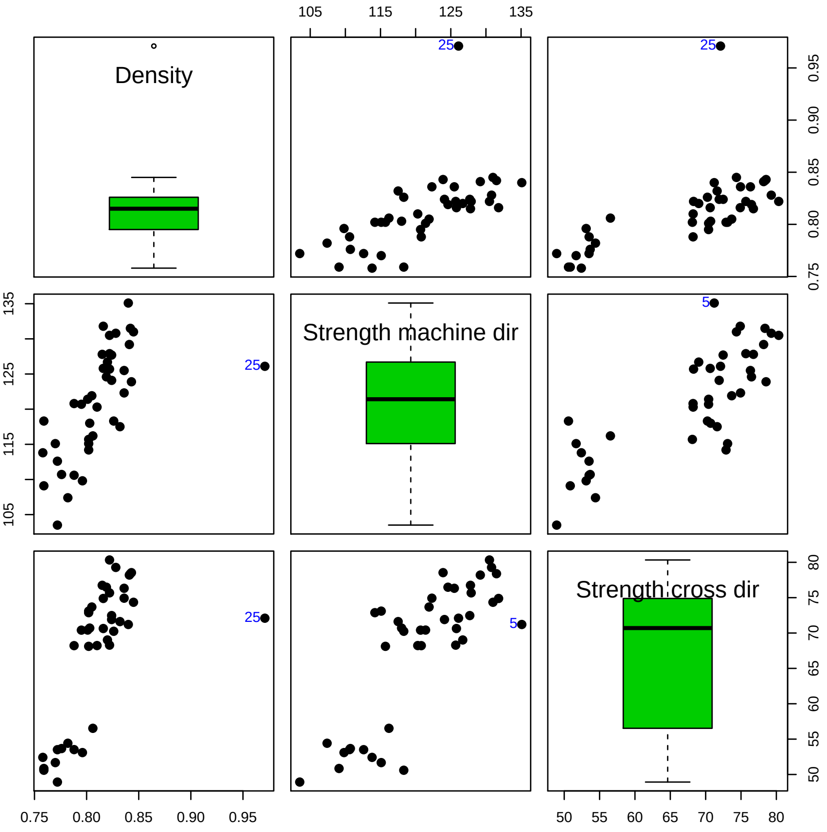
## 6.3 Paarsgewijze plots

Paarsgewijze plots worden gebruikt om te kijken of

- contouren elliptisch zijn
- er afwijkende punten zijn
- er verschillende groepen (clusters) zijn

**Voorbeeld:** $(n = 41,\ p = 3)$

```
scatterplotMatrix(papdata,diagonal="boxplot",
            var.labels=c("Density","Strength machine dir","Strength cross dir"),
            smoother=F,reg.line=F,id.n=1,id.col=4,pch=19,cex=1.25)
```



Hieruit volgt dat de gegevens niet uit een normale verdeling komen.

## 6.4 Verdeling van de Mahalanobis afstanden

Indien $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ en $n \gg p$ (minstens $\frac{n}{p} \geqslant 5$), dan is het kwadraat van de Mahalanobis afstanden

$$\begin{aligned} MD^2(\boldsymbol{X}_i) &:= d_S^2(\boldsymbol{X}_i, \overline{\boldsymbol{X}}) = (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\tau \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}) \\ &\overset{\text{benaderend}}{\approx} (\boldsymbol{X}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}) \overset{\text{Stelling}}{\sim} \chi_p^2 \end{aligned}$$

Men kan de empirische verdeling van $MD^2(\boldsymbol{x}_i)$ visueel voorstellen door de QQ-plot te gebruiken.

$$\boxed{\Big(\underbrace{F_{\chi_p^2}^{-1}\left(\frac{i-\frac{1}{3}}{n+\frac{1}{3}}\right)}_{=\chi^2_{p,1-(i-\frac{1}{3})/(n+\frac{1}{3})}=:q_i}, MD^2_{i:n}\Big)} \qquad (\chi^2\text{-plot})$$

In de praktijk kan men elk punt boven de horizontale rechte $\chi^2_{p,0.025}$ als ongewoon beschouwen.

**Opmerking:**

Men kan ook volgende QQ-plot tekenen

$$\left(\sqrt{\chi^2_{p,1-(i-\frac{1}{3})/(n+\frac{1}{3})}}, MD(x_i)\right) \text{ (afstanden zonder kwadraten)}$$

met cutoff $\sqrt{\chi^2_{p,0.025}}$.

**Voorbeeld:** ($n = 30$ houtplanken, $p = 4$)

Onbuigzaamheid bij schokken ($x_1$), trillingen ($x_2$) en twee statische krachten ($x_3$ en $x_4$). De gestandaardiseerde metingen zijn

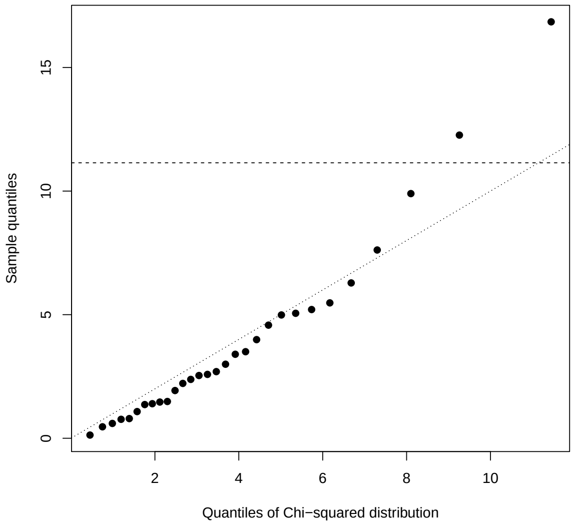
$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{s_{jj}}} \quad (i = 1, \ldots, 30 \text{ en } j = 1, \ldots, 4)$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $MD(x_i)$ |
|---|---|---|---|---|---|---|---|---|
| 1889 | 1651 | 1561 | 1778 | -0.05 | -0.31 | 0.17 | 0.16 | 0.60 |
| 2403 | 2048 | 2087 | 2197 | 1.53 | 0.94 | 1.91 | 1.46 | 5.48 |
| 2119 | 1700 | 1815 | 2222 | 0.66 | -0.16 | 1.01 | 1.54 | 7.62 |
| 1645 | 1627 | 1110 | 1533 | -0.80 | -0.38 | -1.32 | -0.59 | 5.21 |
| 1976 | 1916 | 1614 | 1883 | 0.22 | 0.52 | 0.35 | 0.49 | 1.40 |
| 1712 | 1712 | 1439 | 1546 | -0.60 | -0.12 | -0.23 | -0.55 | 2.22 |
| 1943 | 1685 | 1271 | 1671 | 0.11 | -0.20 | -0.79 | -0.17 | 4.99 |
| 2104 | 1820 | 1717 | 1874 | 0.61 | 0.22 | 0.69 | 0.46 | 1.49 |
| 2983 | 2794 | 2412 | 2581 | 3.31 | 3.28 | 2.98 | 2.65 | 12.26 |
| 1745 | 1600 | 1384 | 1508 | -0.50 | -0.47 | -0.41 | -0.67 | 0.77 |
| 1710 | 1591 | 1518 | 1667 | -0.60 | -0.50 | 0.03 | -0.18 | 1.93 |
| 2046 | 1907 | 1627 | 1898 | 0.43 | 0.49 | 0.39 | 0.54 | 0.46 |
| 1840 | 1841 | 1595 | 1741 | -0.20 | 0.29 | 0.28 | 0.05 | 2.70 |
| 1867 | 1685 | 1493 | 1678 | -0.12 | -0.20 | -0.05 | -0.15 | 0.13 |
| 1859 | 1649 | 1389 | 1714 | -0.14 | -0.32 | -0.40 | -0.03 | 1.08 |
| 1954 | 2149 | 1180 | 1281 | 0.15 | 1.25 | -1.09 | -1.38 | 16.85 |
| 1325 | 1170 | 1002 | 1176 | -1.79 | -1.82 | -1.67 | -1.70 | 3.50 |
| 1419 | 1371 | 1252 | 1308 | -1.50 | -1.19 | -0.85 | -1.29 | 3.99 |
| 1828 | 1634 | 1602 | 1755 | -0.24 | -0.36 | 0.31 | 0.09 | 1.36 |
| 1725 | 1594 | 1313 | 1646 | -0.56 | -0.49 | -0.65 | -0.24 | 1.46 |
| 2276 | 2189 | 1547 | 2111 | 1.14 | 1.38 | 0.12 | 1.20 | 9.90 |
| 1899 | 1614 | 1422 | 1477 | -0.02 | -0.43 | -0.29 | -0.77 | 5.06 |
| 1633 | 1513 | 1290 | 1516 | -0.84 | -0.74 | -0.72 | -0.65 | 0.80 |
| 2061 | 1867 | 1646 | 2037 | 0.48 | 0.37 | 0.45 | 0.97 | 2.54 |
| 1856 | 1493 | 1356 | 1533 | -0.15 | -0.81 | -0.51 | -0.59 | 4.58 |
| 1727 | 1412 | 1238 | 1469 | -0.55 | -1.06 | -0.89 | -0.79 | 3.40 |
| 2168 | 1896 | 1701 | 1834 | 0.81 | 0.46 | 0.63 | 0.34 | 2.38 |
| 1655 | 1675 | 1414 | 1597 | -0.77 | -0.23 | -0.31 | -0.40 | 3.00 |
| 2326 | 2301 | 2065 | 2234 | 1.29 | 1.73 | 1.83 | 1.58 | 6.28 |
| 1490 | 1382 | 1214 | 1284 | -1.28 | -1.15 | -0.97 | -1.37 | 2.58 |

**Opmerking:**

$$MD^2(\boldsymbol{z}_i) = (\boldsymbol{z}_i - \overline{\boldsymbol{z}})^\tau S_{\boldsymbol{z}}^{-1}(\boldsymbol{z}_i - \overline{\boldsymbol{z}}) \overset{!}{=} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\tau S_{\boldsymbol{x}}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$$

Dus het standaardiseren verandert niets aan $MD^2$.



De horizontale lijn is ter hoogte van $\chi^2_{4,0.025} = 11.14$. Hieruit blijkt dat er 2 afwijkende punten zijn (9 en 16).

# Chapter 7

# Datatransformaties

## 7.1 Overzicht

Indien de verdeling van de $\boldsymbol{x}_i \in \mathbb{R}^{p \times 1}$ niet normaal is, kan men deze normaal proberen te maken door middel van een transformatie.

Veelal beperkt men zich tot de verdelingen van de componenten (dus de univariate marginalen) in de hoop dat het uiteindelijke resultaat bij benadering multivariaat normaal verdeeld is. Dit moet wel worden nagegaan in elke situatie.

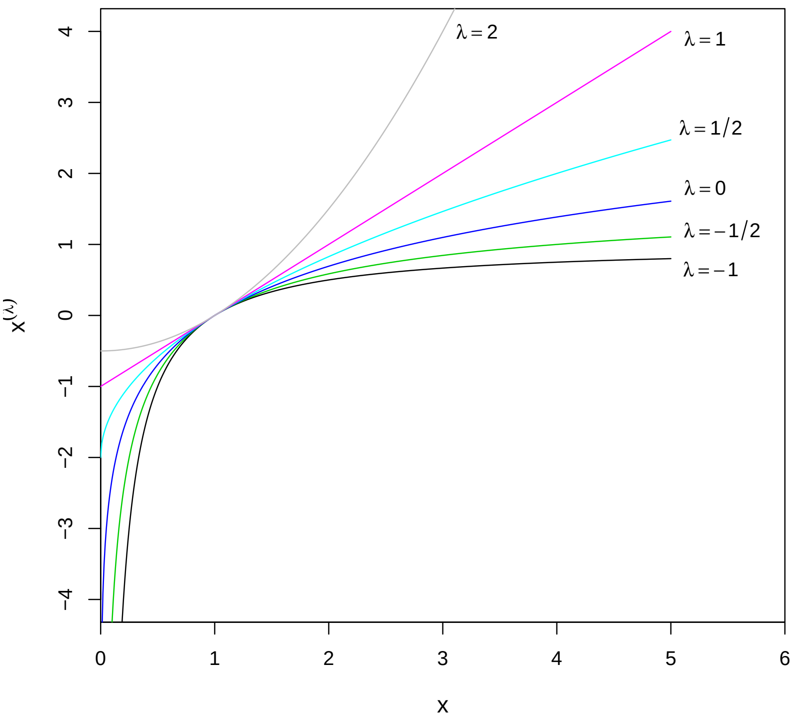
Men gebruikt vaak:

| Oorspronkelijke variabele | Getransformeerde variabele |
|---|---|
| $y$ = absolute frequentie<br>$\in \{0, 1, 2, 3, \ldots\}$ | $\sqrt{y}$ |
| $\hat{p}$ = percentage (relatieve frequentie)<br>$\in ]0, 1[$ | $probit(\hat{p}) := \Phi^{-1}(\hat{p})$<br>of $logit(\hat{p}) := \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ |
| $r$ = correlatie<br>$\in ]-1, 1[$ | Fisher's $z := \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ |
| $x$ = positieve continue variabele<br>$\in \mathbb{R}_0^+$ | $x^{(\lambda)}$ = transformatie van<br>Box en Cox (1964) |

## 7.2 Transformatie van Box en Cox

**Def.** Voor een toevalsveranderlijke $X > 0$ en $\lambda \in \mathbb{R}$:

$$X^{(\lambda)} := \begin{cases} \frac{X^\lambda - 1}{\lambda} & \text{als } \lambda \neq 0 \\ \ln(X) & \text{als } \lambda = 0 \end{cases}$$



Deze transformatie $(\mathbb{R}_0^+, \mathbb{R}) \to \mathbb{R} : (x, \lambda) \to x^{(\lambda)}$ is monotoon en afleidbaar wat betreft $x$ en monotoon en continu wat betreft $\lambda$.

Neem bijvoorbeeld een willekeurige $x > 0$, dan

$$\begin{aligned} \lim_{\lambda \to 0} x^{(\lambda)} &= \lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{e^{\lambda \ln(x)} - 1}{\lambda} \\ &\overset{\text{l'Hôpital}}{=} \lim_{\lambda \to 0} \frac{e^{\lambda \ln(x)} \ln(x) - 0}{1} = \ln(x) = x^{(0)} \end{aligned}$$

**Opmerking.** $X^{(\lambda)}$ kan enkel *exact* normaal verdeeld zijn $\Leftrightarrow \lambda = 0 \Leftrightarrow X$ is lognormaal.

*Proof.* Het beeld van $\mathbb{R}_0^+ \to \mathbb{R} : x \to x^{(\lambda)}$ is $\mathbb{R} \Leftrightarrow \lambda = 0$ want

$$\begin{cases} \sup_{x>0} x^{(\lambda)} = \frac{-1}{\lambda} \text{ is eindig} & \text{als } \lambda < 0 \\ \inf_{x>0} x^{(\lambda)} = \frac{-1}{\lambda} \text{ is eindig} & \text{als } \lambda > 0 \end{cases}$$

□

Men hoopt dat er een $\lambda \in \mathbb{R}$ bestaat zodat $X^{(\lambda)} \approx N_1(\mu, \sigma^2)$ voor een zekere $\mu \in \mathbb{R}$ en $\sigma > 0$. Als dit exact waar was, zou de dichtheid van $X$ gelijk zijn aan:

$$\begin{aligned} f_{\lambda,\mu,\sigma}(x) &= \frac{d}{dx}\Phi\left(\frac{x^{(\lambda)} - \mu}{\sigma}\right) \\ &= \phi\left(\frac{x^{(\lambda)} - \mu}{\sigma}\right)\frac{1}{\sigma}\frac{d}{dx}\left(\frac{x^{\lambda} - 1}{\lambda}\right) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{(x^{(\lambda)}-\mu)^2}{2\sigma^2}}\frac{1}{\sigma}\frac{1}{\lambda}\lambda x^{\lambda-1} \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x^{(\lambda)}-\mu)^2}{2\sigma^2}}e^{(\lambda-1)\ln(x)} \end{aligned}$$

Voor een steekproef $X_1, \ldots, X_n$ vinden we dan

$$f_{\lambda,\mu,\sigma}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{\lambda,\mu,\sigma}(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i^{(\lambda)}-\mu)^2}{2\sigma^2}}e^{(\lambda-1)\ln(x_i)}$$

Box en Cox (1964) hebben voorgesteld om de parameters $\lambda, \mu$ en $\sigma$ te schatten aan de hand van de maximum likelihood methode. Dit betekent dat we voor de parameters de waarden bepalen waarvoor de verzamelde steekproef het meest aannemelijk wordt:

$$\begin{aligned}\max_{\lambda,\mu,\sigma} \prod_{i=1}^{n} f_{\lambda,\mu,\sigma}(x_i) &\Leftrightarrow \max_{\lambda,\mu,\sigma} \sum_{i=1}^{n} \ln\left(f_{\lambda,\mu,\sigma}(x_i)\right) \\ &\Leftrightarrow \max_{\lambda} \max_{\mu,\sigma} \sum_{i=1}^{n} \left\{ -\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{(x_i^{(\lambda)} - \mu)^2}{2\sigma^2} + (\lambda - 1)\ln(x_i) \right\}\end{aligned}$$

Voor elke $\lambda$ is deze som maximaal voor (analoog aan het vinden van de MLE voor $x_i^{(\lambda)} \sim N_1(\mu, \sigma^2)$):

- $\hat{\mu}_\lambda = \overline{x^{(\lambda)}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)}$
- $\hat{\sigma}_\lambda = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2}$

Nu moeten we enkel nog de $\hat{\lambda}$ vinden:

$$\begin{aligned}\hat{\lambda} &= \operatorname*{argmax}_{\lambda} \prod_{i=1}^{n} f_{\lambda,\hat{\mu}_\lambda,\hat{\sigma}_\lambda}(x_i) \\ &= \operatorname*{argmax}_{\lambda} \sum_{i=1}^{n} \left\{ -\ln(\hat{\sigma}_\lambda) - \frac{1}{2}\ln(2\pi) - \frac{(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2}{2\hat{\sigma}_\lambda^2} + (\lambda - 1)\ln(x_i) \right\} \\ &= \operatorname*{argmax}_{\lambda} \left\{ -\frac{n}{2}\ln(\hat{\sigma}_\lambda^2) - \frac{1}{2\hat{\sigma}_\lambda^2}\sum_{i=1}^{n}(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2 + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i) \right\} \\ &= \operatorname*{argmax}_{\lambda} \left\{ -\frac{n}{2}\ln(\hat{\sigma}_\lambda^2) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i) \right\}\end{aligned}$$

**Result.**

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} \left\{ -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2\right] + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i) \right\}$$

Bijgevolg kan men de grafiek maken van

$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2\right] + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i)$$

in functie van $\lambda$, en kijken in welke waarde ($\hat{\lambda}$) deze zijn maximum bereikt.

**Opmerkingen:**

1. Kleine aanpassingen van $\hat{\lambda}$ leiden meestal niet tot een zichtbaar verschil in de getransformeerde variabele. Daarom kan je best een waarde voor $\lambda$ kiezen in de buurt van het optimum $\hat{\lambda}$ die leidt tot een eenvoudige transformatie, bv. $\lambda = -1, -1/2, 1/3, 0, 1/3, 1/2, \ldots$.

2. Heel deze berekening is gebaseerd op de **veronderstelling** dat er een $\lambda$ **bestaat**, waarvoor $X^{(\lambda)}$ benaderend normaal verdeeld is. Als dit niet het geval is, kunnen de observaties $x_1^{(\hat{\lambda})}, \ldots, x_n^{(\hat{\lambda})}$ sterk afwijken van de normaalverdeling! Na het berekenen van $\hat{\lambda}$, moet men dus **altijd controleren** of $x_1^{(\hat{\lambda})}, \ldots, x_n^{(\hat{\lambda})}$ normaal verdeeld zijn. Hiervoor kunnen bijvoorbeeld de QQ-plot en de Shapiro-Wilk test worden gebruikt.

**Voorbeeld:**

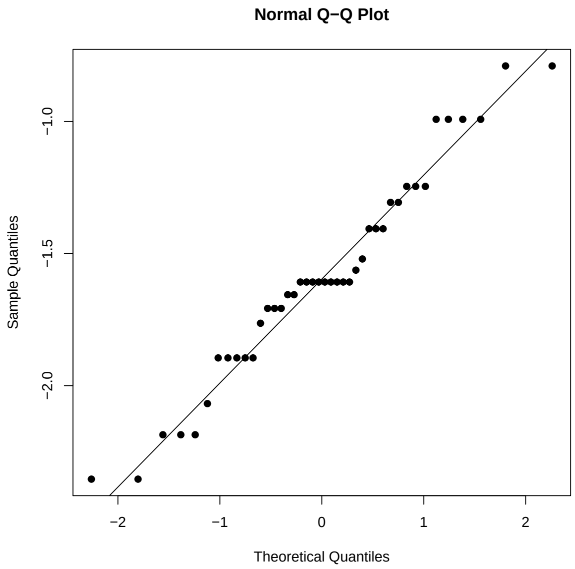
Voor de gegevens van de magnetrons ($X_i$ = uitgezonden straling met gesloten deur, $n = 42$) heeft de QQ-plot ons al aangetoond dat deze niet normaal verdeeld zijn.

Laten we nu $l(\lambda)$ berekenen voor verschillende waarden van $\lambda$:

| $\lambda$ | $l(\lambda)$ | $\lambda$ | $l(\lambda)$ | $\lambda$ | $l(\lambda)$ |
|---|---|---|---|---|---|
| -1.00 | 70.52 | 0.00 | 104.83 | 1.00 | 97.10 |
| -0.90 | 75.65 | 0.10 | 105.84 | 1.10 | 94.64 |
| -0.80 | 80.46 | 0.20 | 106.39 | 1.20 | 91.96 |
| -0.70 | 84.94 | **0.30** | **106.51** | 1.30 | 89.10 |
| -0.60 | 89.06 | 0.40 | 106.20 | 1.40 | 86.07 |
| -0.50 | 92.79 | 0.50 | 105.50 | 1.50 | 82.88 |
| -0.40 | 96.10 | 0.60 | 104.43 | | |
| -0.30 | 98.97 | 0.70 | 103.03 | | |
| -0.20 | 101.39 | 0.80 | 101.33 | | |
| -0.10 | 103.35 | 0.90 | 99.34 | | |

Hieruit volgt dat $\hat{\lambda} \approx 0.30$ een optimale keuze is. We kunnen bv. $\hat{\lambda} = 1/3$ nemen.

De QQ-plot van $x_{i:n}^{(1/3)}$ ziet eruit als volgt:



```
library(car)
radtransform=bcPower(radclosed,1/3)
shapiro.test(radtransform)

Shapiro-Wilk normality test


data:  radtransform
W = 0.96457, p-value = 0.2147
```

De Shapiro-Wilk test levert een P-waarde $= 0.21$, dus normaliteit wordt niet meer verworpen.

## 7.3 Multivariate gegevens

Voor *multivariate* gegevens

$$\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix}, \ldots, \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{np} \end{pmatrix}$$

moet men dus alle marginalen transformeren (met $\lambda_1, \ldots, \lambda_p$) zodat

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \to \begin{pmatrix} x_{i1}^{(\lambda_1)} \\ x_{i2}^{(\lambda_2)} \\ \vdots \\ x_{ip}^{(\lambda_p)} \end{pmatrix}$$

Men kan $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ selecteren op twee manieren:

1. Voor elke variabele $1 \leqslant k \leqslant p$ berekent men

$$\begin{aligned} \hat{\lambda}_k &:= \underset{\lambda_k}{\operatorname{argmax}}\, l_k(\lambda_k) \\ &= \underset{\lambda_k}{\operatorname{argmax}} \left\{ -\frac{n}{2} \ln \left[ \frac{1}{n} \sum_{i=1}^{n} (x_{ik}^{(\lambda_k)} - \overline{x_k^{(\lambda_k)}})^2 \right] + (\lambda_k - 1) \sum_{i=1}^{n} \ln(x_{ik}) \right\} \end{aligned}$$

2. Voor alle variabelen tegelijk berekent men (numeriek)

$$\begin{aligned} (\hat{\lambda}_1, \ldots, \hat{\lambda}_p) &:= \underset{\lambda_1,\ldots,\lambda_p}{\operatorname{argmax}}\ l(\lambda_1, \ldots, \lambda_p) \\ &= \underset{\lambda_1,\ldots,\lambda_p}{\operatorname{argmax}} \left\{ -\frac{n}{2} \ln |\boldsymbol{S}_{\lambda_1,\ldots,\lambda_p}| + \sum_{k=1}^{p} (\lambda_k - 1) \sum_{i=1}^{n} \ln(x_{ik}) \right\} \end{aligned}$$

   met $\boldsymbol{S}_{\lambda_1,\ldots,\lambda_p}$ de empirische covariantiematrix van de getransformeerde gegevens $(x_{i1}^{(\lambda_1)}, x_{i2}^{(\lambda_2)}, \ldots, x_{ip}^{(\lambda_p)})^\tau,\ i = 1, \ldots, p$.

Deze twee methodes geven vaak gelijkaardige resultaten, maar de eerste methode is eenvoudiger om te berekenen.

**Voorbeeld:**

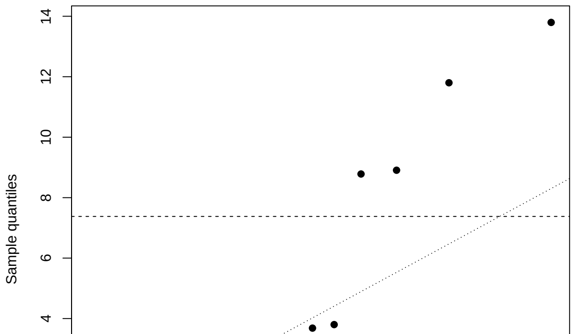
Voor de 42 magnetrons beschouwen we nu

- $X_1$ = uitgezonden straling met gesloten deur (zoals vroeger)
- $X_2$ = uitgezonden straling met open deur (nieuw)

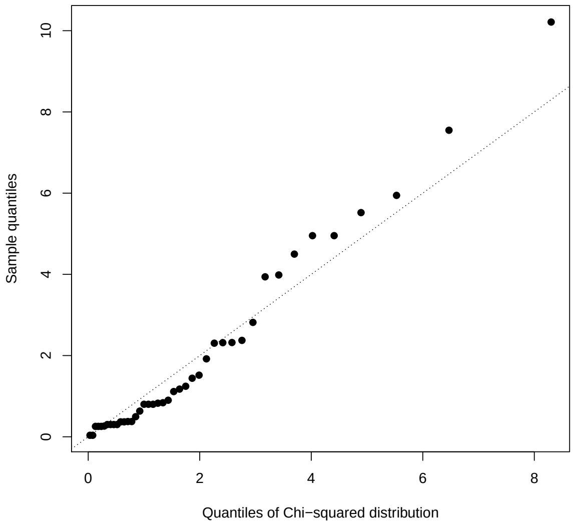
**Tabel:** $X_2$ = uitgezonden straling met open deur

| Oven | straling | Oven | straling | Oven | straling |
|---|---|---|---|---|---|
| 1 | 0.30 | 15 | 0.12 | 29 | 0.09 |
| 2 | 0.09 | 16 | 0.20 | 30 | 0.28 |
| 3 | 0.30 | 17 | 0.04 | 31 | 0.10 |
| 4 | 0.10 | 18 | 0.10 | 32 | 0.10 |
| 5 | 0.10 | 19 | 0.01 | 33 | 0.10 |
| 6 | 0.12 | 20 | 0.60 | 34 | 0.30 |
| 7 | 0.09 | 21 | 0.12 | 35 | 0.12 |
| 8 | 0.10 | 22 | 0.10 | 36 | 0.25 |
| 9 | 0.09 | 23 | 0.05 | 37 | 0.20 |
| 10 | 0.10 | 24 | 0.05 | 38 | 0.40 |
| 11 | 0.07 | 25 | 0.15 | 39 | 0.33 |
| 12 | 0.05 | 26 | 0.30 | 40 | 0.32 |
| 13 | 0.01 | 27 | 0.15 | 41 | 0.12 |
| 14 | 0.45 | 28 | 0.09 | 42 | 0.12 |

We weten reeds dat $X_1$ niet normaal verdeeld is. Ook $X_2$ is niet normaal verdeeld, wat we analoog kunnen opmerken uit de normale kwantielplot voor $X_2$ (niet getoond). De kwantielplot van de kwadratische Mahalanobis afstanden levert ook geen rechtlijnig verband:



Wanneer we de marginale verdelingen elk afzonderlijk transformeren volgens de univariate Box-Cox transformatie leverde dit $\hat{\lambda}_1 = 1/3$ en analoog vinden we dat ook $\hat{\lambda}_2 = 1/3$ genomen kan worden. De kwantielplot van de kwadratische Mahalanobis afstanden levert een min of meer rechtlijnig verband:

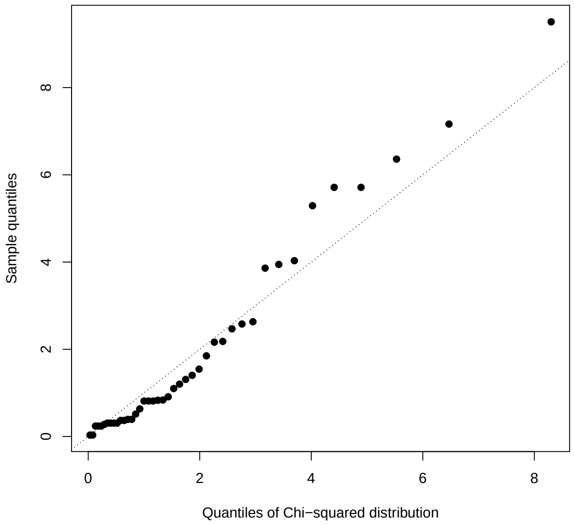


Voor de tweede methode moeten we $(\hat{\lambda}_1, \hat{\lambda}_2)$ vinden zodat $l(\lambda_1, \lambda_2)$ maximaal is.

**Tabel:** Waarden voor $l(\lambda_1, \lambda_2)$ voor een aantal gekozen $\lambda_1$ (horizontaal) en $\lambda_2$ (vertikaal)

| | 0.00 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 |
|---|---|---|---|---|---|---|
| 0.00 | 224.76 | 224.96 | 224.51 | 223.49 | 221.99 | 220.07 |
| 0.10 | 225.20 | 225.78 | 225.68 | 224.96 | 223.70 | 221.97 |
| 0.20 | 224.67 | 225.60 | **225.83** | 225.42 | 224.42 | 222.91 |
| 0.30 | 223.36 | 224.57 | 225.10 | 224.98 | 224.25 | 222.97 |
| 0.40 | 221.46 | 222.94 | 223.68 | 223.80 | 223.32 | 222.27 |
| 0.50 | 219.15 | 220.75 | 221.72 | 222.05 | 221.77 | 220.92 |

De tweede methode levert $(\hat{\lambda}_1, \hat{\lambda}_2) = (0.20, 0.20)$ als optimale keuze. We kunnen bv. $(\hat{\lambda}_1, \hat{\lambda}_2) = (1/4, 1/4)$ als eenvoudige keuze dicht bij dit optimum. Dit verschilt licht van de vorige methode. De kwantielplot van de kwadratische Mahalanobis afstanden is niet erg verschillend van deze gebaseerd op $(\hat{\lambda}_1, \hat{\lambda}_2) = (1/3, 1/3)$.



Merk op dat we bij de tweede methode de hele matrix $\boldsymbol{S}_{\lambda_1,\lambda_2}$ gebruiken, dus ook de empirische covariantie tussen $x_{i1}^{(\lambda_1)}$ en $x_{i2}^{(\lambda_2)}$, die niet gebruikt wordt in de eerste methode.

# Chapter 8

# Inferentie voor de parameter $\boldsymbol{\mu}$

## 8.1 De $T^2$-test van Hotelling

Stel $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ met $\boldsymbol{\mu} \in \mathbb{R}^{p\times 1}$ en $\boldsymbol{\Sigma} \in PD(p)$ onbekend.

Men wil testen:

$$\begin{cases} H_0: & \boldsymbol{\mu} = \boldsymbol{\mu}_0 \\ H_1: & \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \end{cases}$$

voor een gegeven $\boldsymbol{\mu}_0 \in \mathbb{R}^{p\times 1}$.

In het univariate geval ($p = 1$) berekenen we hiervoor

- $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- $S_{XX} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

We weten dan dat

**Result.** (Student $\approx$ 1905)

$$T := \frac{\overline{X} - \mu_0}{\sqrt{\frac{S_{XX}}{n}}} \sim t_{n-1} \qquad | H_0$$

waaruit volgt dat

**Result.** (Student ≈ 1905)

$$T^2 := \left(\overline{X} - \mu_0\right) \left(\frac{1}{n} S_{XX}\right)^{-1} \left(\overline{X} - \mu_0\right) \sim F_{1,n-1} \qquad | H_0$$

want $T \sim t_{n-1} \Rightarrow T^2 \sim F_{1,n-1}$.
Hieruit volgt de (tweezijdige) test van Student:

$$\boxed{H_0 \text{ verwerpen op niveau } \alpha \Leftrightarrow T^2 > \left(t_{n-1,\frac{\alpha}{2}}\right)^2 = F_{1,n-1,\alpha}}$$

met $t_{n-1,\alpha}$ het $(1-\alpha)$-kwantiel van de $t_{n-1}$-verdeling (en analoog voor $F_{1,n-1}$).
In het algemene geval ($p \geqslant 1$) is het een logische veralgemening om de statistiek $T^2$ van Hotelling te berekenen:

$$\boxed{T^2 := \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)^\tau \left(\frac{1}{n}\boldsymbol{S}\right)^{-1} \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)}$$

Nu geldt volgend resultaat.

**Stelling. (Harold Hotelling 1931)**
Indien $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ en $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, dan geldt

$$\frac{n-p}{(n-1)p} T^2 \sim F_{p,n-p}$$

(Zonder bewijs.)
Hieruit volgt de test van Hotelling:

$$\boxed{H_0 \text{ verwerpen op niveau } \alpha \Leftrightarrow T^2 > \underbrace{\frac{(n-1)p}{n-p} F_{p,n-p,\alpha}}_{=F_{1,n-1,\alpha} \text{ als } p=1}}$$

**Voorbeeld:**

Transpiratie van 20 gezonde vrouwen is onderzocht. Drie componenten, $X_1 =$ zweetgemiddelde, $X_2 =$ sodiumgehalte en $X_3 =$ kaliumgehalte werden gemeten en de resultaten staan in onderstaande tabel.

| | $X_1$ | $X_2$ | $X_3$ | | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.7 | 48.5 | 9.3 | 11 | 3.9 | 36.9 | 12.7 |
| 2 | 5.7 | 65.1 | 8.0 | 12 | 4.5 | 58.8 | 12.3 |
| 3 | 3.8 | 47.2 | 10.9 | 13 | 3.5 | 27.8 | 9.8 |
| 4 | 3.2 | 53.2 | 12.0 | 14 | 4.5 | 40.2 | 8.4 |
| 5 | 3.1 | 55.5 | 9.7 | 15 | 1.5 | 13.5 | 10.1 |
| 6 | 4.6 | 36.1 | 7.9 | 16 | 8.5 | 56.4 | 7.1 |
| 7 | 2.4 | 24.8 | 14.0 | 17 | 4.5 | 71.6 | 8.2 |
| 8 | 7.2 | 33.1 | 7.6 | 18 | 6.5 | 52.8 | 10.9 |
| 9 | 6.7 | 47.4 | 8.5 | 19 | 4.1 | 44.1 | 11.2 |
| 10 | 5.4 | 54.1 | 11.3 | 20 | 5.5 | 40.9 | 9.4 |

Test de hypothese $H_0 : \boldsymbol{\mu} = (4 \quad 50 \quad 10)^\tau$ tov $H_1 : \boldsymbol{\mu} \neq (4 \quad 50 \quad 10)^\tau$ op significantieniveau $\alpha = 0.10$.

We krijgen volgende berekeningen

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 4.640 \\ 45.400 \\ 9.965 \end{pmatrix} \qquad \boldsymbol{S} = \begin{pmatrix} 2.879 & 10.002 & -1.810 \\ 10.002 & 199.798 & -5.627 \\ -1.810 & -5.627 & 3.628 \end{pmatrix}$$

$$\boldsymbol{S}^{-1} = \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}$$

Als we de geobserveerde $T^2 = 9.74$ vergelijken met de kritieke waarde

$$\frac{(n-1)p}{(n-p)} F_{p,n-p,0.1} = \frac{19(3)}{17} F_{3,17,0.1} = 3.353(2.44) = 8.18,$$

zien we dat $T^2$ groter is en dus verwerpen we $H_0$ op het 10% niveau. We kunnen ook de P-waarde berekenen als $P(F > \frac{n-p}{(n-1)p} T^2)$ met $F \sim F_{p,n-p}$. Hier wordt dit

$$\text{P-waarde} = P(F > 2.82) = 0.07$$

wat dezelfde conclusie oplevert.

**Eig.** De $T^2$-test is affien invariant.

*Proof.* Beschouw een niet-singuliere affiene afbeelding

$$\boldsymbol{X}_{p\times 1} \longrightarrow \boldsymbol{Y}_{p\times 1} = \boldsymbol{A}_{p\times p}\boldsymbol{X}_{p\times 1} + \boldsymbol{\beta}_{p\times 1} \qquad det(\boldsymbol{A}) \neq 0$$

Dus $\overline{\boldsymbol{Y}} = \boldsymbol{A}\overline{\boldsymbol{X}} + \boldsymbol{\beta}, \boldsymbol{S}_{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^{\tau}, \boldsymbol{\mu}_{\boldsymbol{Y}} = \mathrm{E}\left[\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\beta}\right] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\beta}$ en $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\tau}$. We weten ook dat $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{Y}})$.
Dus

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 \Leftrightarrow \boldsymbol{\mu}_{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{\mu}_0 + \boldsymbol{\beta} =: (\boldsymbol{\mu}_{\boldsymbol{Y}})_0$$

Om te testen of $H_0 : \boldsymbol{\mu}_{\boldsymbol{Y}} = (\boldsymbol{\mu}_{\boldsymbol{Y}})_0$ met Hotelling berekenen we

$$\begin{aligned}
T^2_{\boldsymbol{Y}} &= \left(\overline{\boldsymbol{Y}} - (\boldsymbol{\mu}_{\boldsymbol{Y}})_0\right)^{\tau} \left(\frac{1}{n}\boldsymbol{S}_{\boldsymbol{Y}}\right)^{-1} \left(\overline{\boldsymbol{Y}} - (\boldsymbol{\mu}_{\boldsymbol{Y}})_0\right) \\
&= \left(\boldsymbol{A}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)\right)^{\tau} \left(\frac{1}{n}\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^{\tau}\right)^{-1} \boldsymbol{A}\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right) \\
&= \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)^{\tau} n\boldsymbol{A}^{\tau}(\boldsymbol{A}^{\tau})^{-1}\boldsymbol{S}^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right) \\
&= \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)^{\tau} \left(\frac{1}{n}\boldsymbol{S}\right)^{-1} \left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\right) \\
&\overset{!}{=} T^2
\end{aligned}$$

Ook de kritieke waarde

$$c_{1-\alpha} = \frac{(n-1)p}{(n-p)} F_{p,n-p,\alpha}$$

blijft dezelfde, dus het besluit verkregen met de test van Hotelling blijft invariant. □

## 8.2 Betrouwbaarheidsgebied voor $\boldsymbol{\mu}$

In het univariate geval $X_1, \ldots, X_n \overset{iid}{\sim} N_1(\mu, \sigma_{xx})$ heeft men

$$P\left[\left|\frac{\overline{X}-\mu}{\sqrt{\frac{S_{XX}}{n}}}\right| \leqslant t_{n-1,\frac{\alpha}{2}}\right] = 1-\alpha$$

waaruit volgt dat het $1-\alpha$ betrouwbaarheidsgebied $R$ voor $\mu$ volgend interval is:

$$\left[\overline{X} - t_{n-1,\frac{\alpha}{2}}\sqrt{\frac{S_{XX}}{n}}; \overline{X} + t_{n-1,\frac{\alpha}{2}}\sqrt{\frac{S_{XX}}{n}}\right].$$

In het multivariate geval $\boldsymbol{X}_i \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ hebben we gezien dat

$$P\left[(\overline{\boldsymbol{X}}-\boldsymbol{\mu})^\tau \left(\frac{1}{n}\boldsymbol{S}\right)^{-1} (\overline{\boldsymbol{X}}-\boldsymbol{\mu}) \leqslant \frac{(n-1)p}{n-p} F_{p,n-p,\alpha}\right] = 1-\alpha.$$

**Result.** Dus het $1-\alpha$ betrouwbaarheidsgebied voor $\boldsymbol{\mu}$ is de *ellipsoïde*

$$\begin{aligned} R &= \left\{\boldsymbol{\mu};\ (\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau \left(\frac{1}{n}\boldsymbol{S}\right)^{-1} (\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \leqslant \underbrace{\frac{(n-1)p}{n-p} F_{p,n-p,\alpha}}_{=:c^2}\right\} \\ &= \left\{\boldsymbol{\mu};\ d^2_{\frac{1}{n}\boldsymbol{S}}(\overline{\boldsymbol{x}}, \boldsymbol{\mu}) \leqslant \frac{(n-1)p}{n-p} F_{p,n-p,\alpha}\right\} \\ &= \{\boldsymbol{\mu}_0;\ \text{de test van Hotelling verwerpt } \boldsymbol{\mu}_0 \text{ niet}\} \end{aligned}$$

Indien $(\lambda_j, \boldsymbol{e}_j)$ respectievelijk de eigenwaarden en eigenvectoren zijn van $\boldsymbol{S}$, dan zijn de hoofdassen van $R$

$$\overline{\boldsymbol{x}} \pm \boldsymbol{e}_j c \frac{\sqrt{\lambda_j}}{\sqrt{n}} = \overline{\boldsymbol{x}} \pm \boldsymbol{e}_j \sqrt{\lambda_j} \sqrt{\frac{(n-1)p}{n(n-p)} F_{p,n-p,\alpha}}$$

We merken op

$$\underbrace{\text{tolerantie-ellips}}_{\text{voor een nieuwe observatie(vroeger)}} \quad \neq \quad \underbrace{\text{betrouwbaarheidsellips}}_{\text{voor } \boldsymbol{\mu} \text{ (hier)}}$$

**Voorbeeld:**

Data i.v.m. de straling van magnetrons (zie hoofdstuk 6).

Neem

$$x_1 = \sqrt[4]{\text{gemeten straling met gesloten deur}}$$
$$x_2 = \sqrt[4]{\text{gemeten straling met open deur.}}$$

Voor de $n = 42$ paren van getransformeerde observaties, vinden we

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 0.564 \\ 0.603 \end{pmatrix} \quad \boldsymbol{S} = \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix} \quad \boldsymbol{S}^{-1} = \begin{pmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{pmatrix}$$

De eigenwaarden en eigenvectoren van $\boldsymbol{S}$ zijn

$$\lambda_1 = 0.026 \qquad \boldsymbol{e}_1^\tau = (0.704 \quad 0.710)$$
$$\lambda_2 = 0.002 \qquad \boldsymbol{e}_2^\tau = (-0.710 \quad 0.704)$$

De 95% betrouwbaarheidsellips voor $\boldsymbol{\mu}$ bestaat uit alle waarden $(\mu_1, \mu_2)$ die voldoen aan

$$42 \begin{pmatrix} 0.564 - \mu_1 \\ 0.603 - \mu_2 \end{pmatrix}^\tau \begin{pmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{pmatrix} \begin{pmatrix} 0.564 - \mu_1 \\ 0.603 - \mu_2 \end{pmatrix} \leqslant \frac{2(41)}{40} F_{2,40,0.05}$$

Of, vermits $F_{2,40,0.05} = 3.23$,

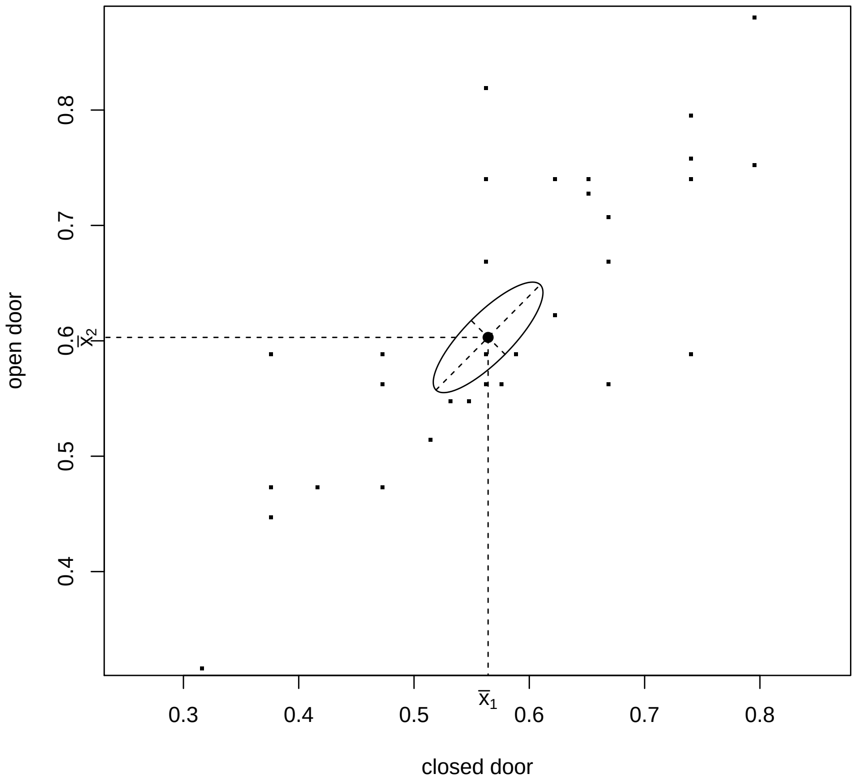
$$42(203.018)(0.564 - \mu_1)^2 + 42(200.228)(0.603 - \mu_2)^2 \\ - 84(163.391)(0.564 - \mu_1)(0.603 - \mu_2) \leqslant 6.62$$

Om te kijken of $\boldsymbol{\mu}_0^\tau = (0.562 \quad 0.589)$ in het betrouwbaarheidsgebied ligt, berekenen we

$$42(203.018)(0.564 - 0.562)^2 + 42(200.228)(0.603 - 0.589)^2 \\ - 84(163.391)(0.564 - 0.562)(0.603 - 0.589) = 1.30 \leqslant 6.62$$

en we concluderen dat $\boldsymbol{\mu}_0$ erin ligt.

Analoog wordt $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ niet verworpen ten opzichte van $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ op niveau $\alpha = 0.05$.



Het centrum van de gezamenlijke 95% betrouwbaarheidsellips is $\overline{\boldsymbol{x}} = (0.564 \quad 0.603)^\tau$ en de halflengtes van de grote en de kleine as wordt respectievelijk gegeven door

$$\sqrt{\lambda_1}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p,\alpha}} = \sqrt{0.026}\sqrt{\frac{2(41)}{42(40)}(3.23)} = 0.064 \quad \text{en}$$

$$\sqrt{\lambda_2}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p,\alpha}} = \sqrt{0.002}\sqrt{\frac{2(41)}{42(40)}(3.23)} = 0.018$$

De assen volgen $\boldsymbol{e}_1 = (0.704 \quad 0.710)^\tau$ en $\boldsymbol{e}_2 = (-0.710 \quad 0.704)^\tau$ wanneer deze vectoren getekend worden vanuit $\overline{\boldsymbol{x}}$.

## 8.3 Simultane betrouwbaarheidsintervallen

Een multivariaat betrouwbaarheidsgebied geeft informatie over de mogelijke waarden voor de volledige vector $\boldsymbol{\mu}$. Meestal zijn we echter ook geïnteresseerd in betrouwbaarheidsintervallen voor de afzonderlijke componenten van $\boldsymbol{\mu}$.

Als $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dan geldt voor elke $\boldsymbol{u} \in \mathbb{R}^p$ : $\boldsymbol{u}^\tau \boldsymbol{X} \sim N(\boldsymbol{u}^\tau \boldsymbol{\mu}, \boldsymbol{u}^\tau \boldsymbol{\Sigma} \boldsymbol{u})$. We weten voor deze univariate componenten dat

$$\frac{\sqrt{n}(\boldsymbol{u}^\tau \overline{\boldsymbol{X}} - \boldsymbol{u}^\tau \boldsymbol{\mu})}{\sqrt{\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}} \sim t_{n-1}$$

zodat een $100(1-\alpha)\%$ betrouwbaarheidsinterval gegeven wordt door

$$\boxed{\left[\boldsymbol{u}^\tau \overline{\boldsymbol{X}} - t_{n-1,\alpha/2}\sqrt{\frac{\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}{n}}, \boldsymbol{u}^\tau \overline{\boldsymbol{X}} + t_{n-1,\alpha/2}\sqrt{\frac{\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}{n}}\right]}$$

Voor elke $\boldsymbol{u} \in \mathbb{R}^p$ heeft dit BTI een betrouwbaarheid $1-\alpha$. De kans dat $\boldsymbol{u}^\tau \boldsymbol{\mu}$ zich gelijktijdig $\forall \boldsymbol{u}$ in het bijhorende interval bevindt, is echter minder dan $1-\alpha$. De betrouwbaarheid die bekomen wordt door deze combinatie van klassieke univariate betrouwbaarheidsintervallen te nemen, is dus *kleiner dan* $1-\alpha$.

Projecties van het multivariate betrouwbaarheidsgebied voor $\boldsymbol{\mu}$ op alle richtingen $\boldsymbol{u}$ levert betrouwbaarheidsintervallen met een gezamenlijke betrouwbaarheid $1-\alpha$:

**Eig.** Als $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dan worden $100(1-\alpha)\%$ betrouwbaarheidsintervallen voor $\boldsymbol{\mu}$ gegeven door

$$\left[\boldsymbol{u}^\tau \overline{\boldsymbol{X}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p,\alpha}\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}, \boldsymbol{u}^\tau \overline{\boldsymbol{X}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p,\alpha}\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}\right]$$

voor elke $\boldsymbol{u} \in \mathbb{R}^p$.

Wanneer we echter slechts enkele projecties beschouwen (de univariate componenten van $\boldsymbol{X}$ bijvoorbeeld) dan geeft deze methode ons BTI's met een gezamenlijke betrouwbaarheid die minstens gelijk is aan $1-\alpha$, maar meestal is deze *groter dan* $1-\alpha$.

Als we geïnteresseerd zijn in de betrouwbaarheidsintervallen voor een klein aantal richtingen, moeten we bovenstaande intervallen dus inkorten zodat de gezamenlijke betrouwbaarheid opnieuw gelijk wordt aan $1 - \alpha$. Een benaderende oplossing van dit probleem wordt gegeven door de *Bonferroni methode.*

Stel dat we univariate betrouwbaarheidsintervallen $B_i$ berekenen in $m$ richtingen $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$. Voor elke $i$ hebben we een gekend betrouwbaarheidsniveau $1 - \alpha_i$. Dan geldt

$$\begin{aligned}
P(\boldsymbol{u}_i^\tau \boldsymbol{\mu} \in B_i, \forall i) &= 1 - P(\exists i : \boldsymbol{u}_i^\tau \boldsymbol{\mu} \notin B_i) \\
&\geqslant 1 - \sum_{i=1}^{m} P(\boldsymbol{u}_i^\tau \boldsymbol{\mu} \notin B_i) \\
&= 1 - \sum_{i=1}^{m} (1 - P(\boldsymbol{u}_i^\tau \boldsymbol{\mu} \in B_i)) \\
&= 1 - \sum_{i=1}^{m} \alpha_i
\end{aligned}$$

Wanneer we $\alpha_i = \frac{\alpha}{m}$ nemen $\forall i$, dan bekomen we dus betrouwbaarheidsintervallen met een gezamenlijke betrouwbaarheid die minstens gelijk is aan $1 - \alpha$ en die afzonderlijk een betrouwbaarheid hebben van $1 - \frac{\alpha}{m}$.

Als we dit toepassen op de univariate BTI's, krijgen we betrouwbaarheidsintervallen

$$\boxed{\left[\boldsymbol{u}^\tau \overline{\boldsymbol{X}} - t_{n-1,\alpha/(2m)} \sqrt{\frac{\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}{n}}, \boldsymbol{u}^\tau \overline{\boldsymbol{X}} + t_{n-1,\alpha/(2m)} \sqrt{\frac{\boldsymbol{u}^\tau \boldsymbol{S} \boldsymbol{u}}{n}}\right]}$$

met een gezamenlijke betrouwbaarheid groter of gelijk aan $1 - \alpha$.

Wanneer we deze Bonferroni intervallen vergelijken met de simultane betrouwbaarheidsintervallen dan merken we op dat

$$\frac{\text{lengte Bonferroni interval}}{\text{lengte simultaan interval}} = \frac{t_{n-1,\alpha/(2m)}}{\sqrt{\frac{p(n-1)}{n-p} F_{p,n-p,\alpha}}}.$$

Als $m$ klein is, kan men aantonen dat deze fractie kleiner is dan 1. De Bonferroni intervallen zijn bijgevolg inderdaad korter dan de gezamenlijke betrouwbaarheidsintervallen en hebben toch een gegarandeerde betrouwbaarheid van $1 - \alpha$ of meer.

**Voorbeeld:**

Voor de data i.v.m. de straling van magnetrons vinden we voor simulatane 95% betrouwbaarheidsintervallen voor de twee componenten dat

$$\frac{\text{lengte Bonferroni interval}}{\text{lengte simultaan interval}} = \frac{t_{41,0.05/4}}{\sqrt{\frac{2(41)}{40}F_{2,40,0.05}}} =$$

# Chapter 9

# Principal component analysis

## 9.1 Introduction

Principal component analysis (PCA) is concerned with explaining the variance-covariance structure of the data through a few *linear combinations* of the original variables. Its general objectives are:

- data reduction
- interpretation.

$\boxed{\text{Data reduction.}}$

Although the original data set contains $p$ variables, often much of the variability can be accounted for by a smaller number ($m$) of principal components. When there is (almost) as much information in the $m$ components as there is in the original $p$ variables, the original data set consisting of $n$ observations on $p$ variables can be reduced to one consisting of $n$ observations on $m$ principal components.

$\boxed{\text{Interpretation.}}$

A PCA can show relationships that were not previously suspected, and it allows interpretations that would not ordinarily result.

## 9.2 Construction of population principal components

Let the random vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]^\tau$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$.
Consider the linear combinations

$$\begin{aligned}
Y_1 &= \boldsymbol{l}_1^\tau \boldsymbol{X} &&= l_{11}X_1 + l_{21}X_2 + \cdots + l_{p1}X_p \\
Y_2 &= \boldsymbol{l}_2^\tau \boldsymbol{X} &&= l_{12}X_1 + l_{22}X_2 + \cdots + l_{p2}X_p \\
&\vdots && \vdots \\
Y_h &= \boldsymbol{l}_h^\tau \boldsymbol{X} &&= l_{1h}X_1 + l_{2h}X_2 + \cdots + l_{ph}X_p \qquad (1) \\
&\vdots && \vdots \\
Y_p &= \boldsymbol{l}_p^\tau \boldsymbol{X} &&= l_{1p}X_1 + l_{2p}X_2 + \cdots + l_{pp}X_p
\end{aligned}$$

The variances and covariances of the linear combinations are:

$$\begin{aligned}
\text{Var}[Y_h] &= \boldsymbol{l}_h^\tau \boldsymbol{\Sigma} \boldsymbol{l}_h && h = 1, 2, \ldots, p && (2) \\
\text{Cov}[Y_k, Y_h] &= \boldsymbol{l}_h^\tau \boldsymbol{\Sigma} \boldsymbol{l}_k && h, k = 1, 2, \ldots, p && (3)
\end{aligned}$$

Basic idea: The principal components are *uncorrelated* linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances in (2) are as large as possible.

The first principal component is the linear combination with maximum variance. It maximizes $\text{Var}[Y_1] = \boldsymbol{l}_1^\tau \boldsymbol{\Sigma} \boldsymbol{l}_1$ under the constraint $\boldsymbol{l}_1^\tau \boldsymbol{l}_1 = 1$.
We define

$Y_1 =$ first PC $=$ linear combination $\boldsymbol{l}_1^\tau \boldsymbol{X}$ that maximizes $\text{Var}[\boldsymbol{l}_1^\tau \boldsymbol{X}]$
subject to $\boldsymbol{l}_1^\tau \boldsymbol{l}_1 = 1$.

$Y_2 =$ second PC $=$ linear combination $\boldsymbol{l}_2^\tau \boldsymbol{X}$ that maximizes $\text{Var}[\boldsymbol{l}_2^\tau \boldsymbol{X}]$
subject to $\boldsymbol{l}_2^\tau \boldsymbol{l}_2 = 1$ and $\text{Cov}[\boldsymbol{l}_1^\tau \boldsymbol{X}, \boldsymbol{l}_2^\tau \boldsymbol{X}] = 0$.

$\vdots$

$Y_h =$ $h$th PC $=$ linear combination $\boldsymbol{l}_h^\tau \boldsymbol{X}$ that maximizes $\text{Var}[\boldsymbol{l}_h^\tau \boldsymbol{X}]$
subject to $\boldsymbol{l}_h^\tau \boldsymbol{l}_h = 1$ and $\text{Cov}[\boldsymbol{l}_h^\tau \boldsymbol{X}, \boldsymbol{l}_k^\tau \boldsymbol{X}] = 0$ for $k < h$.

$\vdots$

$Y_p =$ $p$th PC $=$ linear combination $\boldsymbol{l}_p^\tau \boldsymbol{X}$ that maximizes $\text{Var}[\boldsymbol{l}_p^\tau \boldsymbol{X}]$
subject to $\boldsymbol{l}_p^\tau \boldsymbol{l}_p = 1$ and $[\text{Cov}[\boldsymbol{l}_p^\tau \boldsymbol{X}, \boldsymbol{l}_k^\tau \boldsymbol{X}] = 0$ for $k < p$.

**Result 1.** Let $\boldsymbol{\Sigma}$ be a positive definite matrix with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ and associated normalized eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$. Then

$$\begin{aligned} &\max_{\boldsymbol{x}\neq 0} \tfrac{\boldsymbol{x}^\tau \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\tau \boldsymbol{x}} && = \lambda_1 && (\text{attained when } \boldsymbol{x} = \boldsymbol{e}_1) \\ &\max_{\boldsymbol{x}\perp \boldsymbol{e}_1,\ldots,\boldsymbol{e}_k} \tfrac{\boldsymbol{x}^\tau \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\tau \boldsymbol{x}} && = \lambda_{k+1} && (\text{attained when } \boldsymbol{x} = \boldsymbol{e}_{k+1}, k = 1, 2, \ldots, p-1) \end{aligned}$$

*Proof.*

Let $\boldsymbol{P}$ be the orthogonal matrix whose columns are the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ and $\boldsymbol{\Lambda}$ be the diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$ along the main diagonal, so $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\tau$. Let $\boldsymbol{y} = \boldsymbol{P}^\tau \boldsymbol{x}$.

Consequently, $\boldsymbol{x} \neq 0$ implies $\boldsymbol{y} \neq 0$. Thus,

$$\begin{aligned} \frac{\boldsymbol{x}^\tau \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\tau \boldsymbol{x}} &= \frac{\boldsymbol{x}^\tau \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\tau \boldsymbol{P}\boldsymbol{P}^\tau \boldsymbol{x}} = \frac{\boldsymbol{x}^\tau \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\tau \boldsymbol{x}}{\boldsymbol{y}^\tau \boldsymbol{y}} = \frac{\boldsymbol{y}^\tau \boldsymbol{\Lambda} \boldsymbol{y}}{\boldsymbol{y}^\tau \boldsymbol{y}} \\ &= \frac{\sum\limits_{j=1}^{p} \lambda_j y_j^2}{\sum\limits_{j=1}^{p} y_j^2} \leqslant \lambda_1 \frac{\sum\limits_{j=1}^{p} y_j^2}{\sum\limits_{j=1}^{p} y_j^2} = \lambda_1 \end{aligned}$$

Setting $\boldsymbol{x} = \boldsymbol{e}_1$ gives

$$\boldsymbol{y} = \boldsymbol{P}^\tau \boldsymbol{e}_1 = [1, 0, \ldots, 0]^\tau$$

For this choice of $\boldsymbol{x}$, we get $\frac{\boldsymbol{y}^\tau \boldsymbol{\Lambda} \boldsymbol{y}}{\boldsymbol{y}^\tau \boldsymbol{y}} = \lambda_1$, or

$$\frac{\boldsymbol{e}_1^\tau \boldsymbol{\Sigma} \boldsymbol{e}_1}{\boldsymbol{e}_1^\tau \boldsymbol{e}_1} = \lambda_1$$

A similar argument produces the second part of the result.

Now, $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y} = y_1 \boldsymbol{e}_1 + y_2 \boldsymbol{e}_2 + \ldots + y_p \boldsymbol{e}_p$, so $\boldsymbol{x} \perp \boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ implies

$$0 = \boldsymbol{e}_j^\tau \boldsymbol{x} = y_1 \boldsymbol{e}_j^\tau \boldsymbol{e}_1 + y_2 \boldsymbol{e}_j^\tau \boldsymbol{e}_2 + \ldots + y_p \boldsymbol{e}_j^\tau \boldsymbol{e}_p = y_j, \qquad j \leqslant k$$

Therefore, for $\boldsymbol{x}$ perpendicular to the first $k$ eigenvectors $\boldsymbol{e}_j$, the left-hand side of the inequality becomes

$$\frac{\boldsymbol{x}^\tau \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^\tau \boldsymbol{x}} = \frac{\sum\limits_{j=k+1}^{p} \lambda_j y_j^2}{\sum\limits_{j=k+1}^{p} y_j^2}$$

Taking $y_{k+1} = 1, y_{k+2} = \ldots = y_p = 0$ gives the asserted maximum. □

**Result 2.** Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]^\tau$. Let $\boldsymbol{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ and $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p\}$ is orthonormal. ( If some $\lambda_h$ are equal, the choices of the corresponding coefficient vectors $\boldsymbol{e}_h$ and $Y_h$ are not unique.) Denote the coordinates of $\boldsymbol{e}_h$ as $\boldsymbol{e}_h = [e_{1h}, e_{2h}, \ldots, e_{ph}]^\tau$. The $h$th *principal component* is then given by

$$Y_h \quad = \boldsymbol{e}_h^\tau \boldsymbol{X} \quad = e_{1h}X_1 + e_{2h}X_2 + \ldots + e_{ph}X_p \qquad h = 1, 2, \ldots, p \tag{4}$$

i.e. put $\boldsymbol{l}_1 = \boldsymbol{e}_1, \ldots, \boldsymbol{l}_p = \boldsymbol{e}_p$. With these choices,

$$\begin{aligned} \mathrm{Var}[Y_h] &= \boldsymbol{e}_h^\tau \boldsymbol{\Sigma} \boldsymbol{e}_h &= \lambda_h \qquad & h = 1, 2, \ldots, p \qquad (5) \\ \mathrm{Cov}[Y_k, Y_h] &= \boldsymbol{e}_h^\tau \boldsymbol{\Sigma} \boldsymbol{e}_k &= 0 \qquad & h \neq k \end{aligned}$$

*Proof.* From Result 1 it follows that

$$\begin{aligned} \max_{l \neq 0} \frac{\boldsymbol{l}^\tau \boldsymbol{\Sigma} \boldsymbol{l}}{\boldsymbol{l}^\tau \boldsymbol{l}} &= \lambda_1 = \frac{\boldsymbol{e}_1^\tau \boldsymbol{\Sigma} \boldsymbol{e}_1}{\boldsymbol{e}_1^\tau \boldsymbol{e}_1} = \boldsymbol{e}_1^\tau \boldsymbol{\Sigma} \boldsymbol{e}_1 = \max_{\boldsymbol{l}^\tau \boldsymbol{l} = 1} \boldsymbol{l}^\tau \boldsymbol{\Sigma} \boldsymbol{l} = \mathrm{Var}[Y_1] \\ \max_{l \perp \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k} \frac{\boldsymbol{l}^\tau \boldsymbol{\Sigma} \boldsymbol{l}}{\boldsymbol{l}^\tau \boldsymbol{l}} &= \lambda_{k+1} = \frac{\boldsymbol{e}_{k+1}^\tau \boldsymbol{\Sigma} \boldsymbol{e}_{k+1}}{\boldsymbol{e}_{k+1}^\tau \boldsymbol{e}_{k+1}} = \boldsymbol{e}_{k+1}^\tau \boldsymbol{\Sigma} \boldsymbol{e}_{k+1} \\ &= \max_{\substack{l \perp \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k \\ \boldsymbol{l}^\tau l = 1}} \boldsymbol{l}^\tau \boldsymbol{\Sigma} \boldsymbol{l} = \mathrm{Var}[Y_{k+1}] \end{aligned}$$

For any two eigenvectors $\boldsymbol{e}_h$ and $\boldsymbol{e}_k$ with $h \neq k$ we have $\boldsymbol{e}_h^\tau \boldsymbol{e}_k = 0$.
So we conclude

$$\mathrm{Cov}[Y_h, Y_k] = \boldsymbol{e}_h^\tau \boldsymbol{\Sigma} \boldsymbol{e}_k = \boldsymbol{e}_h^\tau \lambda_k \boldsymbol{e}_k = 0$$

for any $h \neq k$. □

Result 2 has some important corollaries. Recall that the total variance of a distribution is defined as the trace of the covariance matrix, $tr(\boldsymbol{\Sigma})$. We thus obtain that

$$\begin{aligned} tr(\boldsymbol{\Sigma}) &= \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} &= \sum_{j=1}^{p} \mathrm{Var}[X_j] \\ &= \lambda_1 + \lambda_2 + \ldots + \lambda_p &= \sum_{h=1}^{p} \mathrm{Var}[Y_h] \end{aligned}$$

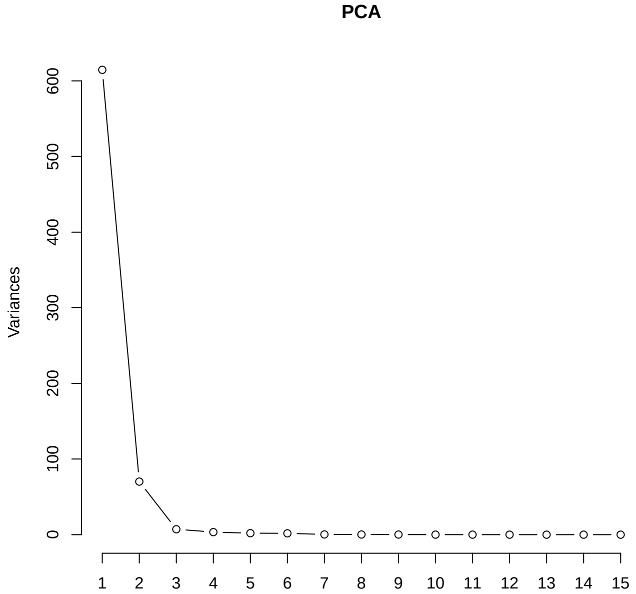
The proportion of the total variance due to the $h$th principal component is therefore equal to:

$$\frac{\lambda_h}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad h = 1, 2, \ldots, p$$

If most of the total population variance, can be attributed to the first one, two or three components, then these components can "replace" the original $p$ variables without much loss of information. There are several criteria to select this number of principal components, e.g.

1. 80% or 90% of the total variance.
2. Plot $\lambda_j$ versus index $j$. Such a plot is called a *screeplot*. One often chooses the number of components based on the so-called elbow in the plot.

Example:



Consider the coefficient vector $\boldsymbol{e}_h = [e_{1h}, \ldots, e_{jh}, \ldots, e_{ph}]^\tau$. The magnitude of $e_{jh}$ measures the importance of the $j$th variable to the $h$th principal component, irrespective of the other variables. In particular, $e_{jh}$ is proportional to the correlation coefficient between $X_j$ and $Y_h$.

**Result 3.** If $Y_1 = \boldsymbol{e}_1^\tau \boldsymbol{X}, Y_2 = \boldsymbol{e}_2^\tau \boldsymbol{X}, \ldots, Y_p = \boldsymbol{e}_p^\tau \boldsymbol{X}$ are the principal components obtained from the covariance matrix $\boldsymbol{\Sigma}$, then

$$\rho_{(X_j,Y_h)} \quad = \frac{e_{jh}\sqrt{\lambda_h}}{\sqrt{\sigma_{jj}}} \qquad j, h = 1, 2, \ldots, p \qquad (6)$$

are the correlation coefficients between the variables $X_j$ and the components $Y_h$.

*Proof.*
Set $\boldsymbol{l}_j = [0, \ldots, 0, 1, 0, \ldots, 0]^\tau$ then

$$X_j = \boldsymbol{l}_j^\tau \boldsymbol{X}$$

$$\text{Cov}[X_j, Y_h] = \text{Cov}[\boldsymbol{l}_j^\tau \boldsymbol{X}, \boldsymbol{e}_h^\tau \boldsymbol{X}] = \boldsymbol{l}_j^\tau \boldsymbol{\Sigma} \boldsymbol{e}_h$$

Since

$$\boldsymbol{\Sigma} \boldsymbol{e}_h = \lambda_h \boldsymbol{e}_h$$

it follows that

$$\text{Cov}[X_j, Y_h] = \boldsymbol{l}_j^\tau \lambda_h \boldsymbol{e}_h = \lambda_h e_{jh}.$$

Then since

$$\text{Var}[Y_h] = \lambda_h \qquad \text{and} \qquad \text{Var}[X_j] = \sigma_{jj}$$

it follows that

$$\rho_{(X_j,Y_h)} = \frac{\text{Cov}[X_j, Y_h]}{\sqrt{\text{Var}[Y_h]}\sqrt{\text{Var}[X_j]}} = \frac{e_{jh}\sqrt{\lambda_h}}{\sqrt{\sigma_{jj}}} \qquad j, k = 1, 2, \ldots, p$$

□

**Example 1.** Suppose the random variables $X_1, X_2$, and $X_3$ have the covariance matrix

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

It may be verified that the eigenvalue-eigenvector pairs are

$$\begin{aligned} \lambda_1 &= 5.83, & \boldsymbol{e}_1^\tau &= [0.383, -0.924, 0] \\ \lambda_2 &= 2.00, & \boldsymbol{e}_2^\tau &= [0, 0, 1] \\ \lambda_3 &= 0.17, & \boldsymbol{e}_3^\tau &= [0.924, 0.383, 0] \end{aligned}$$

The principal components become

$$\begin{aligned} Y_1 &= \boldsymbol{e}_1^\tau \boldsymbol{X} &=& \; 0.383X_1 - 0.924X_2 \\ Y_2 &= \boldsymbol{e}_2^\tau \boldsymbol{X} &=& \; X_3 \\ Y_3 &= \boldsymbol{e}_3^\tau \boldsymbol{X} &=& \; 0.924X_1 + 0.383X_2 \end{aligned}$$

The variance of the first principal component and the covariance between the first and the second component are

$$\begin{aligned} \mathrm{Var}[Y_1] &= \mathrm{Var}[0.383X_1 - 0.924X_2] \\ &= 5.83 \\ &= \lambda_1 \\ \mathrm{Cov}[Y_1, Y_2] &= \mathrm{Cov}[0.383X_1 - 0.924X_2, X_3] \\ &= 0 \end{aligned}$$

It is clear that

$$\begin{aligned} \sigma_{11} + \sigma_{22} + \sigma_{33} &= 1 + 5 + 2 \\ &= \lambda_1 + \lambda_2 + \lambda_3 \\ &= 5.83 + 2.00 + 0.17 \end{aligned}$$

The first two components account for a proportion $(5.83 + 2)/8 = 0.98$ of the population variance. In this case the components $Y_1$ and $Y_2$ could replace the three original variables with little loss of information.

Finally, using Result 3, we have that

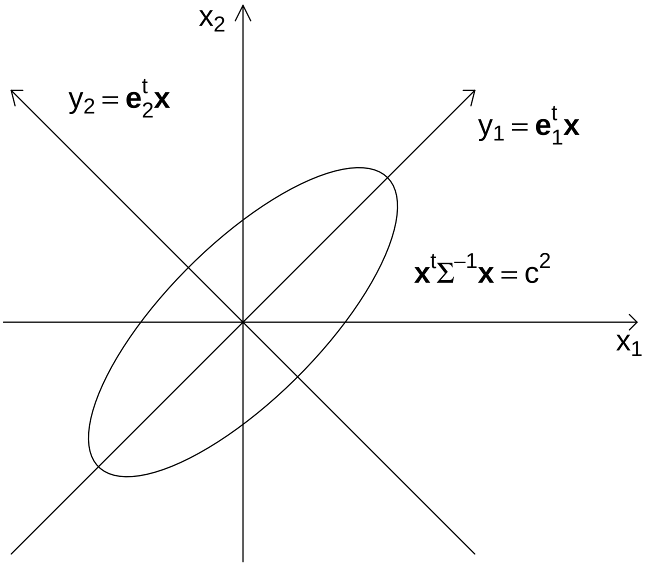
$$\begin{aligned} \rho_{X_1,Y_1} &= 0.925 \\ \rho_{X_2,Y_1} &= -0.998 \\ \rho_{X_1,Y_2} &= 0 \\ \rho_{X_2,Y_2} &= 0 \\ \rho_{X_3,Y_2} &= 1 \end{aligned}$$

We conclude that $X_1$ and $X_2$ are about equally important to the first principal component.

## 9.3 Geometrical interpretation

The principal components are determined by the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ of $\boldsymbol{\Sigma}$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by orthogonally transforming the original system, with $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ as the new coordinate axes. The new axes represent the directions with maximum variability. We have seen that these axes correspond to the axes of the ellipsoid formed by the points at equal Mahalanobis distance of the origin:

**Result 4.** Consider the $p$ dimensional ellipsoid $\boldsymbol{X}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = c^2$.
The principal components define the axes of the ellipsoid.



$x_2$
$y_2 = \mathbf{e}_2^t\mathbf{x}$
$y_1 = \mathbf{e}_1^t\mathbf{x}$
$\mathbf{x}^t\Sigma^{-1}\mathbf{x} = c^2$
$x_1$

## 9.4 Principal components obtained from standardized variables

Consider the standardized variables

$$\boldsymbol{Z} = (\boldsymbol{V}^{\frac{1}{2}})^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

with the diagonal standard deviation matrix

$$\boldsymbol{V}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

We know that $E[\boldsymbol{Z}] = 0$ and $\text{Cov}[\boldsymbol{Z}] = (\boldsymbol{V}^{\frac{1}{2}})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{\frac{1}{2}})^{-1} = \text{Corr}[\boldsymbol{X}] = \boldsymbol{\rho}$. The principal components of $\boldsymbol{Z}$ can be obtained from the eigenvectors of the correlation matrix $\boldsymbol{\rho}$ of $\boldsymbol{X}$.

**Result 5.** The $h$th principal component of the standardized variables $\boldsymbol{Z} = [Z_1, Z_2, \dots, Z_p]^\tau$, with $\text{Cov}[\boldsymbol{Z}] = \boldsymbol{\rho}$, and $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \dots, (\lambda_p, \boldsymbol{e}_p)$ the eigenvalue-eigenvector pairs for $\boldsymbol{\rho}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_p \geqslant 0$,
is given by

$$Y_h = \boldsymbol{e}_h^\tau \boldsymbol{Z} = \boldsymbol{e}_h^\tau (\boldsymbol{V}^{\frac{1}{2}})^{-1}(\boldsymbol{X} - \boldsymbol{\mu}), \qquad h = 1, 2, \dots, p$$

Moreover,

$$\sum_{h=1}^{p} \text{Var}[Y_h] = \sum_{j=1}^{p} \text{Var}[Z_j] = p$$

and

$$\rho_{(Z_j, Y_h)} = e_{jh}\sqrt{\lambda_h}, \qquad j, h = 1, 2, \dots, p$$

*Proof.*
This follows immediately from Results 2, 3.5, and 3. □

The proportion of total variance explained by the $h$th principal component of $\boldsymbol{Z}$ is therefore equal to

$$\frac{\lambda_h}{p} \qquad h = 1, 2, \dots, p.$$

**Example 2.** Consider the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

and the resulting correlation matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

The eigenvalue-eigenvector pairs from $\boldsymbol{\Sigma}$ are

$$\begin{aligned} \lambda_1 &= 100.16, & \boldsymbol{e}_1 = &\quad [0.040, 0.999]^{\tau} \\ \lambda_2 &= 0.84, & \boldsymbol{e}_2 = &\quad [0.999, -0.040]^{\tau} \end{aligned}$$

The eigenvalue-eigenvector pairs from $\boldsymbol{\rho}$ are

$$\begin{aligned} \lambda_1 &= 1.4, & \boldsymbol{e}_1 = &\quad [0.707, 0.707]^{\tau} \\ \lambda_2 &= 0.6, & \boldsymbol{e}_2 = &\quad [0.707, -0.707]^{\tau} \end{aligned}$$

The principal components are

$$\boldsymbol{\Sigma} : \left\{ \begin{aligned} Y_1 &= 0.040X_1 + 0.999X_2 \\ Y_2 &= 0.999X_1 - 0.040X_2 \end{aligned} \right\}$$

and

$$\boldsymbol{\rho} : \left\{ \begin{aligned} Y_1 &= 0.707Z_1 + 0.707Z_2 &= 0.707\left(\tfrac{X_1-\mu_1}{\sqrt{\sigma_{11}}}\right) + 0.0707\left(\tfrac{X_2-\mu_2}{\sqrt{\sigma_{22}}}\right) \\ Y_2 &= 0.707Z_1 - 0.707Z_2 &= 0.707\left(\tfrac{X_1-\mu_1}{\sqrt{\sigma_{11}}}\right) - 0.0707\left(\tfrac{X_2-\mu_2}{\sqrt{\sigma_{22}}}\right) \end{aligned} \right\}$$

We see that $X_2$ completely dominates the first principal component of $\boldsymbol{\Sigma}$. This first principal component explains a proportion $\lambda_1/(\lambda_1 + \lambda_2) = 100.16/101 = 0.992$ of the total population variance.

In contrast, the variables $Z_1$ and $Z_2$ contribute equally to the principal components of $\boldsymbol{\rho}$.

$$\begin{aligned} \rho_{Z_1,Y_1} &= e_{11}\sqrt{\lambda_1} = & 0.707\sqrt{1.4} &= 0.837 \\ \rho_{Z_2,Y_1} &= e_{21}\sqrt{\lambda_1} = & 0.707\sqrt{1.4} &= 0.837 \end{aligned}$$

In this case, the first principal component explains a proportion $\frac{\lambda_1}{p} = 0.7$ of the total standardized population variance.

We conclude that the relative importance of the variables is affected by the standardization.

**Conclusion.** The principal components of $\boldsymbol{\Sigma}$ differ from those of $\boldsymbol{\rho}$.

The variables are often standardized when they have different units or widely different scales.

## 9.5 Sample principal components

Let $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}]^\tau$. Assume the data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ represent $n$ independent observations from some elliptic $p$-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. These data yield the sample mean vector $\bar{\boldsymbol{x}}$, the sample covariance matrix $\boldsymbol{S}$, and the sample correlation matrix $\boldsymbol{R}$.
We know that the $n$ values of any linear combination

$$\boldsymbol{l}_1^\tau \boldsymbol{x}_i = l_{11}x_{i1} + l_{21}x_{i2} + \ldots + l_{p1}x_{ip} \qquad i = 1, 2, \ldots, n$$

have sample mean $\boldsymbol{l}_1^\tau \bar{\boldsymbol{x}}$ and sample variance $\boldsymbol{l}_1^\tau \boldsymbol{S} \boldsymbol{l}_1$. The pairs of values $(\boldsymbol{l}_1^\tau \boldsymbol{x}_i, \boldsymbol{l}_2^\tau \boldsymbol{x}_i)$ have sample covariance $\boldsymbol{l}_1^\tau \boldsymbol{S} \boldsymbol{l}_2$.
We obtain the following results concerning sample principal components.

**Result 6.** If $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$, the $h$th sample principal component is given by

$$\hat{y}_h = \hat{\boldsymbol{e}}_h^\tau \boldsymbol{x} = \hat{e}_{1h}x_1 + \hat{e}_{2h}x_2 + \ldots + \hat{e}_{ph}x_p \qquad h = 1, 2, \ldots, p$$

where $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_p \geqslant 0$ and $\boldsymbol{x}$ is any observation on the variables $X_1, X_2, \ldots, X_p$. Also

$$\begin{aligned} \underset{i=1}{\overset{n}{\mathrm{var}}}(\hat{y}_{ih}) &= \hat{\lambda}_h \qquad && \text{for } h = 1, 2, \ldots, p \\ \underset{i=1}{\overset{n}{\mathrm{cov}}}(\hat{y}_{ih}, \hat{y}_{ik}) &= 0 \qquad && \text{for } h \neq k \end{aligned}$$

**Result 7.**

$$\text{Total sample variance } = \mathrm{tr}(\boldsymbol{S}) = \sum_{j=1}^{p} s_{jj} = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p = \sum_{h=1}^{p} \hat{\lambda}_h$$

and

$$r_{x_j, \hat{y}_h} = \frac{\hat{e}_{jh}\sqrt{\hat{\lambda}_h}}{\sqrt{s_{jj}}} \qquad j, h = 1, 2, \ldots, p$$

The sample principal components of the standardized observations are given by Result 6, with the matrix $\boldsymbol{R}$ instead of $\boldsymbol{S}$. Again the principal components of $\boldsymbol{S}$ and $\boldsymbol{R}$ are not the same.

**Comment.** The observations $\boldsymbol{x}_i$ are almost always "centered" by subtracting $\bar{\boldsymbol{x}}$. This has no effect on the sample covariance matrix $\boldsymbol{S}$ and gives the $h$th principal component

$$\hat{y}_h = \hat{\boldsymbol{e}}_h^{\tau}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \qquad \text{for } h = 1, 2, \ldots, p$$

for any observation vector $\boldsymbol{x}$. If we consider the values of the $h$th component

$$\hat{y}_{ih} = \hat{\boldsymbol{e}}_h^{\tau}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \qquad \text{for } h = 1, 2, \ldots, p$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{ih}) = \bar{\hat{y}}_h = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{e}}_h^{\tau}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \frac{1}{n}\hat{\boldsymbol{e}}_h^{\tau}(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) = 0$$

The sample mean of each principal component is thus zero.

**Example 3.** We have a study of size and shape relationships for turtles which measures carapace length, width and height. In studies of size-and-shape relationships, one often uses a logarithmic transformation. The natural logarithms of the measurements of 24 male turtles have sample mean vector $\bar{\boldsymbol{x}}^{\tau} = [4.725, 4.478, 3.703]$ and covariance matrix

$$\boldsymbol{S} = 10^{-3}\begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

A principal component analysis yields the following result.

```
males.pca=prcomp(males)
males.pca

Standard deviations (1, .., p=3):
[1] 0.15265434 0.02446027 0.01896934


Rotation (n x k) = (3 x 3):
         PC1        PC2        PC3
V1 0.6831023 -0.1594791  0.7126974
V2 0.5102195 -0.5940118 -0.6219534
V3 0.5225392  0.7884900 -0.3244015

summary(males.pca)

Importance of components:
                          PC1     PC2     PC3
Standard deviation     0.1527 0.02446 0.01897
Proportion of Variance 0.9605 0.02466 0.01483
Cumulative Proportion  0.9605 0.98517 1.00000
```

The first principal component has an interesting subject-matter interpretation. Since

$$\begin{aligned}\hat{y}_1 &= 0.683 \ln(\text{length}) + 0.510 \ln(\text{width}) + 0.523 \ln(\text{height}) \\ &= \ln[(\text{length})^{0.683}(\text{width})^{0.510}(\text{height})^{0.523}]\end{aligned}$$

the first PC may be viewed as a multiple of ln(volume).

**Example 4.** The weekly rates of return for five stocks listed on the New York Stock Exchange were determined for the period January 1975 through December 1976. Let $x_1, x_2, \ldots, x_5$ denote observed weekly rates of return for the five stocks. Then

$$\bar{\boldsymbol{x}} = [0.0054, 0.0048, 0.0057, 0.0063, 0.0037]^{\tau}$$

and

$$\boldsymbol{R} = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{bmatrix}$$

The eigenvalues and corresponding normalized eigenvectors of $\boldsymbol{R}$ are:

$$\begin{aligned} \hat{\lambda}_1 &= 2.857 & \hat{\boldsymbol{e}}_1 &= [0.464, 0.457, 0.470, 0.421, 0.421]^\tau \\ \hat{\lambda}_2 &= 0.809 & \hat{\boldsymbol{e}}_2 &= [0.240, 0.509, 0.260, -0.526, -0.582]^\tau \\ \hat{\lambda}_3 &= 0.540 & \hat{\boldsymbol{e}}_3 &= [-0.612, 0.178, 0.335, 0.541, -0.435]^\tau \\ \hat{\lambda}_4 &= 0.452 & \hat{\boldsymbol{e}}_4 &= [0.387, 0.206, -0.662, 0.472, -0.382]^\tau \\ \hat{\lambda}_5 &= 0.343 & \hat{\boldsymbol{e}}_5 &= [-0.451, 0.676, -0.400, -0.176, 0.385]^\tau \end{aligned}$$

Using the standardized variables, we obtain the first two sample principal components

$$\begin{aligned} \hat{y}_1 &= \hat{\boldsymbol{e}}_1^\tau \boldsymbol{z} &&= 0.464z_1 + 0.457z_2 + 0.470z_3 + 0.421z_4 + 0.421z_5 \\ \hat{y}_2 &= \hat{\boldsymbol{e}}_2^\tau \boldsymbol{z} &&= 0.240z_1 + 0.509z_2 + 0.260z_3 - 0.526z_4 - 0.582z_5 \end{aligned}$$

These components account for

$$(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}) \times 100\% = 73\%$$

of the total sample variance. The first component is an equally weighted sum, or "index", of the five stocks. This component might be called a market component. The second component represents a contrast between the first three stocks (which were chemical stocks) and the last two stocks (oil stocks). It might be called an industry component.

**Example 5.** The body weight (in grams) for $n = 150$ female mice were obtained immediately after birth of their first four litters. The sample mean vector and sample correlation matrix were

$$\bar{\boldsymbol{x}} = [39.88, 45.08, 48.11, 49.95]^\tau$$

$$\boldsymbol{R} = \begin{bmatrix} 1.000 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.000 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.000 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.000 \end{bmatrix}$$

The eigenvalues of this matrix are

$$\hat{\lambda}_1 = 3.058, \quad \hat{\lambda}_2 = 0.382, \quad \hat{\lambda}_3 = 0.342, \text{and} \quad \hat{\lambda}_4 = 0.217$$

and $\boldsymbol{e}_1 = [0.493, 0.522, 0.487, 0.497]^\tau$.

The first principal component accounts for $100(\hat{\lambda}_1/p)\% = 76\%$ of the total variance. The average post-birth weights increase over time. The variation in weights is fairly well explained by the first principal component with nearly equal coefficients.

**Comment.** An unusually small value for the last eigenvalue can indicate an unnoticed linear dependency in the data set.

Consider a situation where $x_1$, $x_2$, and $x_3$ are subtest scores and the total score $x_4$ is the sum $x_1 + x_2 + x_3$. Although the linear combination

$$[1, 1, 1, -1]\boldsymbol{x} = x_1 + x_2 + x_3 - x_4$$

is always zero, rounding error in the computation of eigenvalues may lead to a small nonzero value.

Thus eigenvalues very close to zero are important!

## 9.6 Graphing the principal components

Plots of the principal components can

- reveal suspect observations
- provide checks on the assumption of normality

The last principal components can help to detect suspect observations. Each observation $\boldsymbol{x}_i$ can be expressed as a linear combination

$$\begin{aligned} \boldsymbol{x}_i &= (\boldsymbol{x}_i^\tau \hat{\boldsymbol{e}}_1)\hat{\boldsymbol{e}}_1 + (\boldsymbol{x}_i^\tau \hat{\boldsymbol{e}}_2)\hat{\boldsymbol{e}}_2 + \ldots + (\boldsymbol{x}_i^\tau \hat{\boldsymbol{e}}_p)\hat{\boldsymbol{e}}_p \\ &= \hat{y}_{i1}\hat{\boldsymbol{e}}_1 + \hat{y}_{i2}\hat{\boldsymbol{e}}_2 + \ldots + \hat{y}_{ip}\hat{\boldsymbol{e}}_p \end{aligned}$$

of the complete set of eigenvectors $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_p$ of $\boldsymbol{S}$. The magnitudes of the last principal components determine how well the first few components fit the observations, because

$$\hat{y}_{i1}\hat{\boldsymbol{e}}_1 + \hat{y}_{i2}\hat{\boldsymbol{e}}_2 + \ldots + \hat{y}_{i,q-1}\hat{\boldsymbol{e}}_{q-1}$$

differs from $\boldsymbol{x}_i$ by

$$\hat{y}_{iq}\hat{\boldsymbol{e}}_q + \ldots + \hat{y}_{ip}\hat{\boldsymbol{e}}_p$$

whose squared length is

$$\hat{y}_{iq}^2 + \ldots + \hat{y}_{ip}^2.$$

Suspect observations will often be such that at least one of the coordinates $\hat{y}_{iq}, \ldots, \hat{y}_{ip}$ contributing to this squared length will be large.

Method

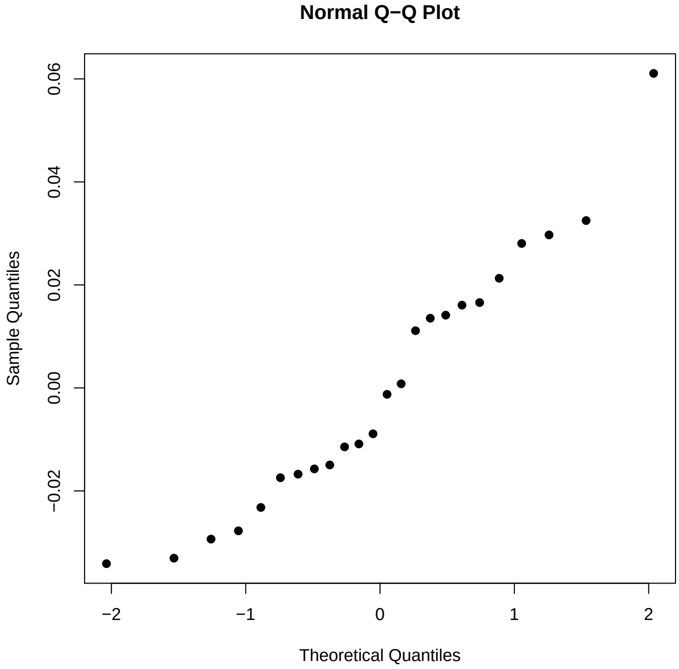
1. To help check the normal assumption, construct scatter diagrams for pairs of the first few principal components. Also make Q-Q plots of the sample values generated by each principal component.

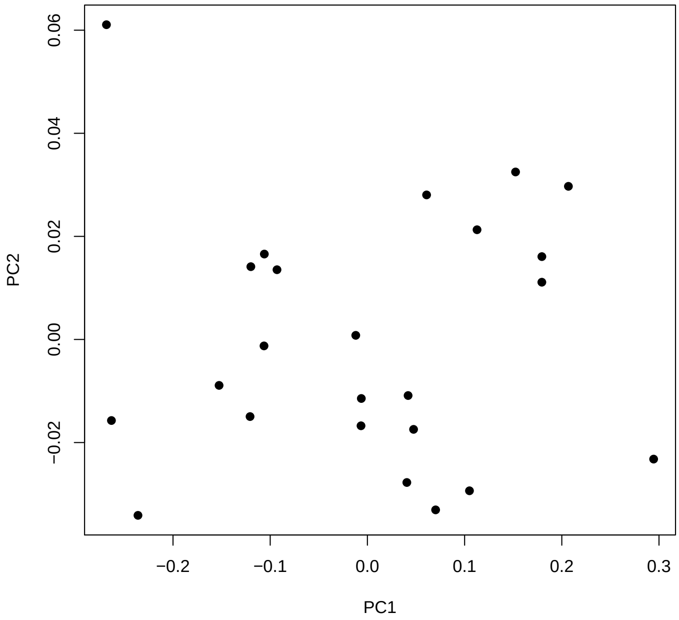
2. Construct scatter diagrams and Q-Q plots of the last few principal components to identify suspect observations.

**Example 6.** If we consider the male turtle data discussed in Example 3, the three sample principal components are

$$\begin{aligned}
\hat{y}_1 &= 0.683(x_1 - 4.725) + 0.510(x_2 - 4.478) + 0.523(x_3 - 3.703) \\
\hat{y}_2 &= -0.159(x_1 - 4.725) - 0.594(x_2 - 4.478) + 0.788(x_3 - 3.703) \\
\hat{y}_3 &= -0.713(x_1 - 4.725) + 0.622(x_2 - 4.478) + 0.324(x_3 - 3.703)
\end{aligned}$$

where $x_1$ =ln(length), $x_2$ =ln(width), and $x_3$ =ln(height). The plots below show the Q-Q plot of $\hat{y}_2$ and the scatterplot of $(\hat{y}_1, \hat{y}_2)$. The observation of the first turtle lies in the upper right corner of the Q-Q plot and in the upper left corner of the scatterplot. This point is suspect, and therefore it should be checked for recording errors. Apart from the first turtle, the scatterplot appears to be reasonably elliptical.




PC2
0.06
0.04
0.02
0.00
−0.02
−0.2
−0.1
0.0
0.1
0.2
0.3
PC1

## 9.7 Approximation using principal components

Let us consider approximations of the form $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]^\tau$ to the centered data matrix

$$\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]^\tau = [\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_n - \bar{\boldsymbol{x}}]^\tau$$

The error of approximation is the "error sum of squares":

$$\text{ESS}(\boldsymbol{A}) = \sum_{i=1}^{n} \|\boldsymbol{g}_i - \boldsymbol{a}_i\|^2 = \sum_{i=1}^{n} (\boldsymbol{g}_i - \boldsymbol{a}_i)^\tau (\boldsymbol{g}_i - \boldsymbol{a}_i) = \|\boldsymbol{G} - \boldsymbol{A}\|_F^2$$

where $\| \ldots \|_F$ is the Frobenius norm of a matrix.

**Result 8.** Among all $n \times p$ matrices $\boldsymbol{A}$ with $rank(\boldsymbol{A}) \leqslant m < min(p, n)$, ESS($\boldsymbol{A}$) is minimized by the choice

$$\hat{\boldsymbol{A}}_m = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]^\tau \hat{\boldsymbol{E}} \hat{\boldsymbol{E}}^\tau = [\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_m] \hat{\boldsymbol{E}}^\tau$$

with $\hat{\boldsymbol{E}} = [\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_m]$ where $\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_m$ are the first $m$ eigenvectors of $\boldsymbol{S}$, hence $rank(\hat{\boldsymbol{E}}) = m$. The $i$th row of $\hat{\boldsymbol{A}}_m$ is

$$\hat{\boldsymbol{a}}_i^\tau = \hat{y}_{i1} \hat{\boldsymbol{e}}_1^\tau + \hat{y}_{i2} \hat{\boldsymbol{e}}_2^\tau + \ldots + \hat{y}_{im} \hat{\boldsymbol{e}}_m^\tau$$

where

$$[\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{im}]^\tau = [\boldsymbol{g}_i^\tau \hat{\boldsymbol{e}}_1, \boldsymbol{g}_i^\tau \hat{\boldsymbol{e}}_2, \ldots, \boldsymbol{g}_i^\tau \hat{\boldsymbol{e}}_m]^\tau$$

are the values of the first $m$ sample principal components for the $i$th observation. Moreover,

$$ESS(\hat{\boldsymbol{A}}_m) = (n-1)(\hat{\lambda}_{m+1} + \ldots + \hat{\lambda}_p)$$

where $\hat{\lambda}_{m+1} \geqslant \ldots \geqslant \hat{\lambda}_p$ are the smallest $p - m$ eigenvalues of $\boldsymbol{S}$.

This means that the space generated by $\{\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_m\}$ is also the result of a least squares optimization.

## 9.8 The connection between PCA and orthogonal regression

Let us consider a set of $m$ orthonormal column vectors $\boldsymbol{l}_1, \dots, \boldsymbol{l}_m$ in $I\!R^{p\times1}$. The $m$-dimensional subspace through the origin determined by $\boldsymbol{l}_1, \boldsymbol{l}_2, \dots, \boldsymbol{l}_m$ is then

$$\text{vec}\{\boldsymbol{l}_1, \dots, \boldsymbol{l}_m\} = \{\boldsymbol{L\beta}; \boldsymbol{\beta} \in I\!R^{m\times1}\}$$

Translating this subspace to pass through some point $\boldsymbol{c} \in I\!R^{p\times1}$ yields the affine subspace

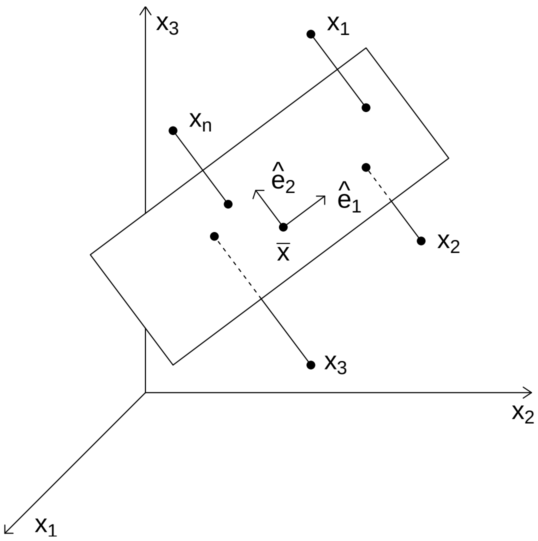
$$\boldsymbol{H} = \boldsymbol{c} + \text{vec}\{\boldsymbol{l}_1, \dots, \boldsymbol{l}_m\} = \{\boldsymbol{c} + \boldsymbol{L\beta}; \boldsymbol{\beta} \in I\!R^{m\times1}\}$$

We are going to select the $m$-dimensional affine subspace $\boldsymbol{H}$ that minimizes the sum of squared distances between the observations $\boldsymbol{x}_i$ and $\boldsymbol{H}$. This is **orthogonal regression.** The word 'orthogonal' comes from the fact that the distance $d(\boldsymbol{x}_i, \boldsymbol{H})$ is measured in the direction orthogonal to $\boldsymbol{H}$ (whereas in classical regression we use the vertical distance, i.e. the absolute residual $|\boldsymbol{r}_i|$).

Suppose that we approximate $\boldsymbol{x}_i$ by $\boldsymbol{c} + \boldsymbol{L}\boldsymbol{b}_i$ with $\sum_{i=1}^n \boldsymbol{b}_i = \boldsymbol{0}$. (If $\sum_{i=1}^n \boldsymbol{b}_i = n\bar{\boldsymbol{\beta}} \neq \boldsymbol{0}$, use $\boldsymbol{c} + \boldsymbol{L}\boldsymbol{b}_i = (\boldsymbol{c} + \boldsymbol{L}\bar{\boldsymbol{\beta}}) + \boldsymbol{L}(\boldsymbol{b}_i - \bar{\boldsymbol{\beta}}) = c^* + \boldsymbol{L}\boldsymbol{b}_i^*$ instead.) Then

$$\begin{aligned}
&\textstyle\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{c} - \boldsymbol{L}\boldsymbol{b}_i)^\tau (\boldsymbol{x}_i - \boldsymbol{c} - \boldsymbol{L}\boldsymbol{b}_i)\\
&= \textstyle\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}} - \boldsymbol{L}\boldsymbol{b}_i + \bar{\boldsymbol{x}} - \boldsymbol{c})^\tau (\boldsymbol{x}_i - \bar{\boldsymbol{x}} - \boldsymbol{L}\boldsymbol{b}_i + \bar{\boldsymbol{x}} - \boldsymbol{c})\\
&= \textstyle\sum_{i=1}^n (\boldsymbol{g}_i - \boldsymbol{L}\boldsymbol{b}_i)^\tau (\boldsymbol{g}_i - \boldsymbol{L}\boldsymbol{b}_i) + n(\bar{\boldsymbol{x}} - \boldsymbol{c})^\tau (\bar{\boldsymbol{x}} - \boldsymbol{c})\\
&\overset{(*)}{\geqslant} \textstyle\sum_{i=1}^n (\boldsymbol{g}_i - \hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^\tau \boldsymbol{g}_i)^\tau (\boldsymbol{g}_i - \hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^\tau \boldsymbol{g}_i) + n(\bar{\boldsymbol{x}} - \boldsymbol{c})^\tau (\bar{\boldsymbol{x}} - \boldsymbol{c})\\
&\geqslant \textstyle\sum_{i=1}^n (\boldsymbol{g}_i - \hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^\tau \boldsymbol{g}_i)^\tau (\boldsymbol{g}_i - \hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^\tau \boldsymbol{g}_i)
\end{aligned}$$

where (*) holds by Result 8, since $\text{rank}[\boldsymbol{L\beta}_1, \dots, \boldsymbol{L\beta}_n] \leqslant m$. If we take $\boldsymbol{c} = \bar{\boldsymbol{x}}$ the lower bound is reached, so the best affine subspace passes through the sample mean. The subspace is thus determined by the first $m$ eigenvectors of $\boldsymbol{S}$, namely $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \dots, \hat{\boldsymbol{e}}_m$. Moreover, the coefficient of $\hat{\boldsymbol{e}}_k$ is $\hat{\boldsymbol{e}}_k^\tau (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \hat{y}_{ik}$, which is the $k$th sample principal component of the $i$th observation $\boldsymbol{x}_i$.



The name 'orthogonal regression' is often used when fitting a hyperplane, i.e. $m = p - 1$. In this case we can write the approximating hyperplane $\boldsymbol{H}$ as

$$\boldsymbol{H} = \bar{\boldsymbol{x}} + \hat{\boldsymbol{e}}_p^{\perp}$$

where $\hat{\boldsymbol{e}}_p^{\perp}$ is the orthogonal complement of the last eigenvector $\hat{\boldsymbol{e}}_p$. In this case, the sum of squared distances to $\boldsymbol{H}$ becomes

$$\sum_{i=1}^{n} d^2(\boldsymbol{x}_i, \boldsymbol{H}) = \mathrm{ESS}(\hat{\boldsymbol{A}}_{p-1}) = (n-1)\hat{\lambda}_p = (n-1)\operatorname*{var}_{i=1}^{n}(y_{ip}).$$

**Remark.** In two dimensions ($p = 2$), the orthogonal regression line therefore corresponds to the first principal component.

# Chapter 10

# Classification methods

## 10.1 Expected Cost of Misclassification

Assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is a $p$-variate random variable of explanatory variables, and $Y$ is the response variable taking only values 0 or 1. For an observation $\boldsymbol{x}$, this is the same as saying that the observation is drawn from a population $\pi_1$ with overall (*prior*) probability $P(Y = 1) = p_1$, or from a population $\pi_2$ with overall (*prior*) probability $P(Y = 0) = p_2 = 1 - p_1$. Furthermore, we assume, for sake of convenience, that the explanatory variables are continuous. These two populations can be described by probability density functions $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$. Some examples are found in the table below.

| Populations $\pi_1$ and $\pi_2$ | Measured variables $\boldsymbol{X}$ |
|---|---|
| Solvent and distressed banks | Total assets, cost of stocks and bonds, market value of stocks and bonds, loss expenses, surplus |
| Nonulcer dyspeptics (those with upset stomach problems) and controls ("normal") | Measures of anxiety, dependence, guilt, perfectionism |
| Steel beams that will break under a certain amount of stress, and those that won't | Composition of the steel, dimensions of the beam's profile |

Our goal here is to find regions $R_1$ and $R_2 = \Omega \setminus R_1$, where $\Omega \subseteq \mathbb{R}^p$ is the sample space, such that we assign $\boldsymbol{x}$ to $\pi_1$ if $\boldsymbol{x} \in R_1$, and to $\pi_2$ if $\boldsymbol{x} \in R_2$. Typically, we have $\Omega = \mathbb{R}^p$.

Since we do not know in advance which population an observation $\boldsymbol{x}$ belongs to, it is possible to misclassify this observation to $\pi_2$ while it in fact is from $\pi_1$ and vice versa. This issue arises when the two populations overlap. Hence, a good classification procedure should result in few misclassified observations, and thus should have a low probability of misclassification. This is not the only feature that a good classification rule needs to possess however.

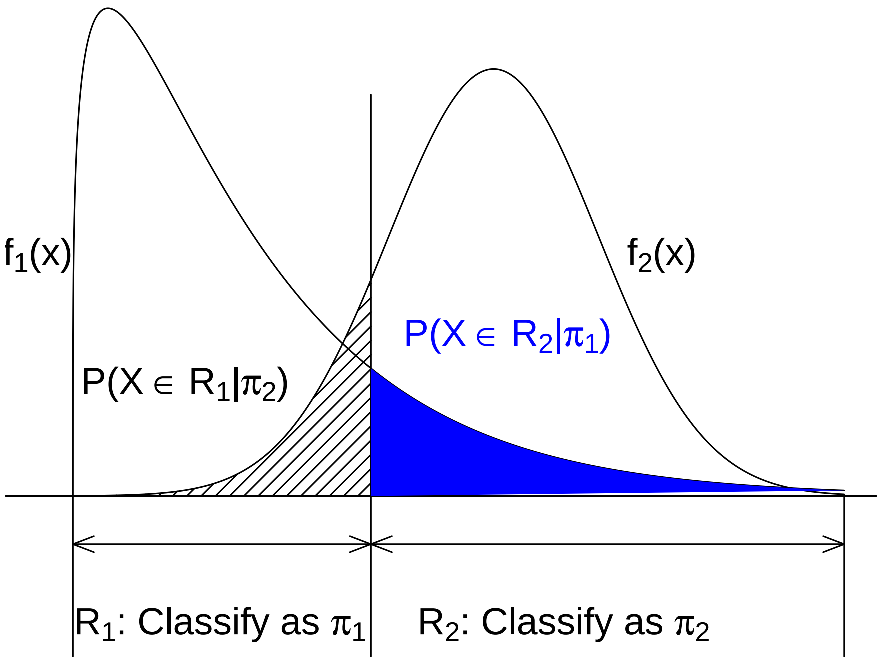
Another important consideration is the cost of misclassification. It is highly conceivable that the error of misclassifying an observation from population $\pi_1$ is much more severe than the error of misclassifying an observation from population $\pi_2$. For example, falsely rejecting a steel beam for having too low stress tolerance means that just that beam needs to be scrapped. However, if a steel beam has been falsely passed and is used in a construction, it could potentially cause the whole edifice to come crashing down, which costs thousands, perhaps millions times as much as rejecting a beam when it's not necessary. As such, we expect a good classification rule to take these costs into account as well.

To obtain this classification rule, or the regions $R_1$ and $R_2$, first compute the conditional probabilities of misclassifying an observation from population $\pi_1$ and of population $\pi_2$. Misclassifying an observation $\boldsymbol{x}$ from population $\pi_1$ simply means that $\boldsymbol{x} \in R_2$, hence

$$P(X \in R_2 \mid \pi_1) = \int_{R_2 = \Omega \setminus R_1} f_1(\boldsymbol{x}) d\boldsymbol{x},$$

and analogously, the probability of misclassifying an observation drawn from $\pi_2$ is

$$P(X \in R_1 \mid \pi_2) = \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}.$$


$f_1(x)$
$f_2(x)$
$P(X \in R_2|\pi_1)$
$P(X \in R_1|\pi_2)$
$R_1$: Classify as $\pi_1$
$R_2$: Classify as $\pi_2$

Using these expressions, the overall probabilities of incorrectly classifying observations are obtained easily:

$$
\begin{aligned}
P(\text{observation is misclassified as } \pi_2) &= P(\boldsymbol{X} \in R_2 \mid \pi_1)P(\pi_1) \\
&= p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} \text{ and} \\
P(\text{observation is misclassified as } \pi_1) &= P(\boldsymbol{X} \in R_1 \mid \pi_2)P(\pi_2) \\
&= p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}
$$

The *total probability of misclassification* (TPM) then becomes

$$
\begin{aligned}
TPM &= P(\text{observation is misclassified as } \pi_2) \\
&\qquad + P(\text{observation is misclassified as } \pi_1) \\
&= P(\boldsymbol{X} \in R_2 \mid \pi_1)p_1 + P(\boldsymbol{X} \in R_1 \mid \pi_2)p_2. \\
&= p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}
\tag{10.1.1}
$$

Classification rules are often evaluated based on their misclassification probabilities, but this ignores the misclassification cost. As illustrated above, even a seemingly small probability $P(X \in R_1 \mid \pi_2)$, e.g. falsely passing a steel beam and using it in a construction, may be too large if the cost of making an incorrect assignment to $\pi_1$ is extremely high (the collapse of a building in the example). As such, ignoring the costs of misclassification may cause problems. Define the costs of misclassification by a cost matrix:

| | | Classify as: | |
|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ |
| True population: | $\pi_1$ | 0 | $c(\boldsymbol{x} \in R_2 \mid \pi_1)$ |
| | $\pi_2$ | $c(\boldsymbol{x} \in R_1 \mid \pi_2)$ | 0 |

This allows us to define the *expected cost of misclassification* (ECM) of a classification rule in the following way:

$$
\begin{aligned}
ECM = c(\boldsymbol{x} \in R_2 \mid \pi_1)&P(\boldsymbol{X} \in R_2 \mid \pi_1)p_1 \\
&+ c(\boldsymbol{x} \in R_1 \mid \pi_2)P(\boldsymbol{X} \in R_1 \mid \pi_2)p_2.
\end{aligned}
$$

This ECM should be (nearly) as small as possible for a reasonable classification rule. Using the expressions for the misclassification probabilities, we can obtain the following result:

**Result 9.** The regions $R_1$ and $R_2$ that minimise the ECM are defined by the values $\boldsymbol{x}$ for which the following inequalities hold:

$$\begin{aligned} R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &\geqslant \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1} \\ R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &< \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}, \end{aligned} \tag{10.1.2}$$

where $f_1$ and $f_2$ are the probability density functions of $\pi_1$ and $\pi_2$ respectively, $p_1$ and $p_2$ the prior probabilities of belonging to $\pi_1$ and $\pi_2$ respectively, and $c(\boldsymbol{x} \in R_2 \mid \pi_1)$ and $c(\boldsymbol{x} \in R_1 \mid \pi_2)$ the misclassification costs.

Observe that the classification rule can be written in terms of ratios of probability density functions, prior probabilities and misclassification costs. This has a significant advantage, because it is often much easier to specify these ratios than their component parts.

For example, consider a college admission test. If the test accidentally passes a person who is not college mature, will drop out after a year or two, that incurs a certain cost to the taxpayers, which can probably be roughly assessed. On the other hand, estimating the cost to the university and to society of not admitting a student who is capable of graduating is much harder. However, finding a realistic number for the ratio of these misclassification costs is more feasible. For example, not admitting a capable student may be roughly five times as costly, over a suitable time horizon, than admitting an eventual dropout. As such, the cost ratio is five in this example (or 1/5).

In some special cases, the classification regions reduce to simpler expressions.

1. Consider the case where the prior probabilities are equal, $p_1 = p_2$. The classification rule then reduces to

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \qquad R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)}.$$

2. The second case is the case where the misclassification costs are equal, $c(\boldsymbol{x} \in R_2 \mid \pi_1) = c(\boldsymbol{x} \in R_1 \mid \pi_2)$. In that case, the classification rule

simplifies to

$$R_1: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1}.$$

Note that this is equivalent to minimising the TPM in (10.1.1).

3. In the final case, we assume both the prior probabilities and the misclassification costs to be equal, or that the misclassification cost ratio is the inverse of the prior probability ratio, leading to

$$R_1: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant 1 \qquad R_2: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1$$

as classification rule.

When the prior probabilities are unknown, we often take them to be equal, and the minimum ECM rule involves a comparison between the ratio of the population densities and the ratio of appropriate misclassification costs. If the misclassification cost is indeterminate, it is usually assumed to be unity, and the ratio of population densities is compared with the ratio of prior probabilities (in reverse order!). Finally, when both the prior probabilities and the misclassification cost ratios are unknown, researchers often assume that the ratio of misclassification costs is the reciprocal of the ratio of prior probabilities, leading to the rule of comparing the ratio of population densities with one.

**Example: classifying a new observation into one of two populations**

Assume that a researcher has sufficient data available to estimate the density functions $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ associated with populations $\pi_1$ and $\pi_2$, respectively. Assume that the costs of misclassifying observations are $c(\boldsymbol{x} \in R_2 \mid \pi_1) = 5$ units and $c(\boldsymbol{x} \in R_1 \mid \pi_2) = 10$ units. In addition, the researcher knows that 20% of *all* objects belong to population $\pi_2$, so the prior probabilities are $p_1 = 0.8$ and $p_2 = 0.2$.

Given this information, we can use the general classification rule (10.1.2) to derive the classification regions $R_2$ and $R_1$. Specifically, we obtain

$$R_1: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{10}{5} \times \frac{0.2}{0.8} = 0.5 \qquad R_2: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{10}{5} \times \frac{0.2}{0.8} = 0.5.$$

Now assume that, for a new observation $\boldsymbol{x}_0$, the density functions evaluate to $f_1(\boldsymbol{x}_0) = 0.3$ and $f_2(\boldsymbol{x}_0) = 0.4$. Will the researcher classify this observation as belonging to $\pi_1$ or $\pi_2$? Since

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{0.3}{0.4} = 0.75 \geqslant 0.5,$$

$x_0 \in R_1$ and the observation is classified as belonging to $\pi_1$.

Alternatively, we can adopt a more Bayesian approach, and allocate a new observation $\boldsymbol{x}_0$ to the population with the highest *posterior* probability, $P(\pi_i \mid \boldsymbol{X} = \boldsymbol{x}_0)$. By Bayes's rule, the posterior probabilities are

$$\begin{aligned} P(\pi_1 \mid \boldsymbol{X} = \boldsymbol{x}_0) &= \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}, \text{ and} \\ P(\pi_2 \mid \boldsymbol{X} = \boldsymbol{x}_0) &= \frac{p_2 f_2(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}. \end{aligned}$$

Classifying an observation $\boldsymbol{x}_0$ as $\pi_1$ when $P(\pi_1 \mid \boldsymbol{X} = \boldsymbol{x}_0) > P(\pi_2 \mid \boldsymbol{X} = \boldsymbol{x}_0)$ is, once more, equivalent to using the ECM rule with equal misclassification costs. However, computing these posterior probabilities is frequently useful for purposes of identifying the less clear-cut assignments.

## 10.2 Classification with two multivariate normal populations

In this section, we examine the ECM classification rule with the additional assumption that both populations $\pi_1$ and $\pi_2$ are normally distributed. This is a common assumption in statistical practice, due to the simplicity and reasonably high efficiency of the resulting classification rule across a wide range of population models.

For the remainder of this section, we assume that $\pi_1$ is normally distributed with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\Sigma_1$, and that $\pi_2$ is normally distributed with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\Sigma_2$.

Note that this assumption will not be required for logistic regression which will be considered later in Chapter 20!

### 10.2.1 Equal covariance matrices

Assume that $\Sigma_1 = \Sigma_2 = \Sigma$, and that the population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\Sigma$ are known. As such, the joint densities of $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ for populations $\pi_1$ and $\pi_2$ are given by

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$$

for $i = 1, 2$. This implies, after cancellation of the factors $(2\pi)^{p/2}|\Sigma|^{1/2}$, that the classification regions based on the minimum ECM rule (10.1.2) become

$$\begin{aligned}
R_1 : \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)& \\
\geqslant \frac{c(\boldsymbol{x}\in R_1 \mid \pi_2)}{c(\boldsymbol{x}\in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}& \\
R_2 : \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)& \\
< \frac{c(\boldsymbol{x}\in R_1 \mid \pi_2)}{c(\boldsymbol{x}\in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}&
\end{aligned} \tag{10.2.1}$$

Given these regions $R_1$ and $R_2$, we obtain the following result for the classification rule.

**Result 10.** Let the populations $\pi_1$ and $\pi_2$ be distributed as a multivariate, normal distribution with population means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and covariance matrices equal to $\Sigma$. The allocation rule that minimises the ECM is the following:
Allocate a new observation $\boldsymbol{x}_0$ to $\pi_1$ if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geqslant \log \left[ \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1} \right], \tag{10.2.2}$$

and allocate $\boldsymbol{x}_0$ to $\pi_2$ otherwise.

The proof is left to the reader.

Observe that the classification rule reduces to determining whether a linear function of $\boldsymbol{x}_0$ is larger than a certain threshold value or not. Hence, this technique for establishing a classification rule is often called *linear discriminant analysis.*

In most practical situations however, the population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\Sigma$ are unknown, so a modification of the classification rule (10.2.2) is necessary. One way of obtaining a modified classification rule is by replacing the unknown population parameters by their sample counterparts.

Assume that we have a sample consisting of $n_1$ observation of $\boldsymbol{X}$ from population $\pi_1$, denoted $\boldsymbol{x}_{1j}$ for $j = 1, \ldots, n_1$, and $n_2$ observations of $\boldsymbol{X}$ from population $\pi_2$, denoted $\boldsymbol{x}_{2j}$ for $j = 1, \ldots, n_2$. Furthermore, assume that $n_1 + n_2 - 2 \geqslant p$. For each population, we can obtain the sample averages and sample covariance matrices as

$$\bar{\boldsymbol{x}}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} \boldsymbol{x}_{1j} \qquad \hat{\Sigma}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)^t (\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)$$

$$\bar{\boldsymbol{x}}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \boldsymbol{x}_{2j} \qquad \hat{\Sigma}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)^t (\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2).$$

Because we assume that the population covariance matrices are the same and equal to $\Sigma$, we can combine the sample covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ to derive a single, unbiased estimate of $\Sigma$. In particular, the pooled covariance matrix

$$\hat{\Sigma}_p = \left( \frac{n_1 - 1}{n_2 + n_1 - 2} \right) \hat{\Sigma}_1 + \left( \frac{n_2 - 1}{n_2 + n_1 - 2} \right) \hat{\Sigma}_2$$

is an unbiased estimate of $\Sigma$ assuming that the samples drawn from populations $\pi_1$ and $\pi_2$ are random.

Substituting $\bar{\boldsymbol{x}}_1$ for $\boldsymbol{\mu}_1$, $\bar{\boldsymbol{x}}_2$ for $\boldsymbol{\mu}_2$, and $\hat{\Sigma}_p$ for $\Sigma$ in (10.2.2) gives the sample classification rule:

**Result 11.** Allocate a new observation $\boldsymbol{x}_0$ to $\pi_1$ if

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geqslant \log\left[\frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}\right],$$

and allocate $\boldsymbol{x}_0$ to $\pi_2$ otherwise.

In practical applications, the prior probabilities $p_1$ and $p_2$ are often unknown. In these cases, the researcher can either assume equality of the prior probabilities, as described in Section 10.1, or estimate these probabilities by the respective sample proportions

$$\hat{p}_1 = \frac{n_1}{n_1 + n_2} \text{ and } \hat{p}_2 = \frac{n_2}{n_1 + n_2}.$$

**Example: hemophilia test**

To develop a test for potential hemophilia carriers, blood samples were taken from two groups of patients. The two variables measured are AHF activity and AHF-like antigen where AHF means AntiHemophilic Factor. Both variables are analyzed on the logarithmic (base 10) scale. The dataset consists of 75 observations. The first group of $n_1 = 30$ patients did not carry the hemophilia gene. The second group consisted of known hemophilia carriers. The first 5 patients carrying the hemophilia gene are taken aside as test observations.

We perform a linear discriminant analysis on the remaining 70 observations using the software R, and obtain the following results.

```
library(rrcov)
data(hemophilia)
library(MASS)
output.lda=lda(gr~AHFactivity+AHFantigen,data=hemophilia[-(31:35),])
output.lda

Call:
lda(gr ~ AHFactivity + AHFantigen, data = hemophilia[-(31:35),
    ])

Prior probabilities of groups:
  carrier    normal
0.5714286 0.4285714

Group means:
        AHFactivity  AHFantigen
carrier  -0.3003825  0.00477000
normal   -0.1348700 -0.07785667

Coefficients of linear discriminants:
                  LD1
AHFactivity  9.017436
AHFantigen  -8.387345
```

The output shows the sample proportions of patients carrying the hemophilia gene and not carrying the gene; in this case $40/70 \approx 57\%$ of the patients carries the hemophilia gene. The group means are also displayed; from these, we can already infer that patients with more AHF activity and more AHF-like antigen are more inclined to carry the hemophilia gene, as can be expected. The coefficients of linear discriminants are obtained in the following way:

Define $\Delta^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, then the coefficients are equal to

$$\frac{1}{\Delta}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1}.$$

The "allocation cutoff" value, which determines when an observation is allocated to $\pi_1$ or $\pi_2$, is similarly equal to

$$\hat{m} = \frac{1}{2\Delta}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + \frac{1}{\Delta} \log \left[ \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1} \right],$$

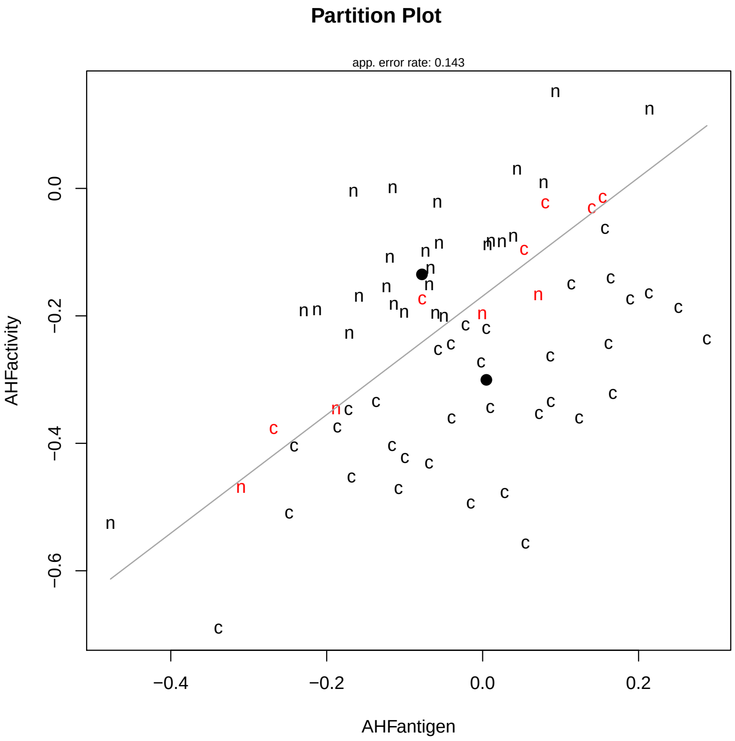
here equal to $\hat{m} = -1.52$. If the linear classification function,

$$\frac{1}{\Delta}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \hat{\Sigma}_p^{-1} \boldsymbol{x}_0$$

evaluates to a number higher than this for a new observation $\boldsymbol{x}_0$, allocate the observation to the population of carriers; otherwise, allocate it to the population of noncarriers.

We can visualise the classification boundary in the following way:

```r
library(klaR)
partimat(as.factor(gr) ~ AHFactivity+AHFantigen, data = hemophilia[-(31:35),],
         method = "lda", imageplot = FALSE)
```



For more than two variables, add the option `plot.matrix=TRUE`, and the plots show the classification boundaries for classifiers built with each pair of variables. In this case we observe that 4 of the noncarriers ('normal') are misclassified as carriers ('carrier') of the gene, and that 6 of the carriers are misclassified as being noncarriers. We can also see this if we produce the confusion matrix or classification table:

```
ldapred <- predict(output.lda, hemophilia[-(31:35),])$class
table(hemophilia[-(31:35),]$gr, ldapred)

         ldapred
          carrier normal
  carrier      34      6
  normal        4     26
```

The top row corresponds to the patients carrying the hemophilia gene, the bottom row to the patients not carrying the gene, and the columns correspond to the patients being predicted as carrying and not carrying the gene (from left to right).

Now assume that we wish to make a prediction for the first five patients that were taken aside as test observations.

```
predict(output.lda, hemophilia[31:35,])

$class
[1] carrier carrier carrier carrier normal
Levels: carrier normal

$posterior
     carrier      normal
31 0.9964344 0.003565637
32 0.5294873 0.470512703
33 0.9927427 0.007257256
34 0.7469073 0.253092740
35 0.3681125 0.631887465

$x
          LD1
31 -2.2896101
32  0.2337082
33 -1.9627442
34 -0.2074226
35  0.5349683
```

We see that for the first 4 patients the model correctly predicts that they carry the hemophilia gene, while the fifth patients is incorrectly classified as a noncarrier. The posterior probabilities reflect for each of the patients the probability that the patient is a carrier or noncarrier according to the model. We can see that the classification uncertainty is very small for patients 1 and 3, but much higher for the other three patients which are much closer to the classification boundary.

The values for LD1 are the values of the linear classification function for each of the 5 observations. In R the groups are re-centered before calculating these values. The corresponding allocation cutoff for the re-centered data is $\hat{m} = 0.37$. If the value of an observation falls below this cutoff, we allocate the observation to the population of carriers and otherwise the observation is allocated to the population of noncarriers.

### 10.2.2 Unequal covariance matrices

As expected, the classification rule becomes more complicated when the covariance matrices $\Sigma_1$ and $\Sigma_2$ are not equal.

Once again, assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is drawn from a multivariate normal distribution, with mean $\boldsymbol{\mu}_1$ and covariance matrix $\Sigma_1$ if it belongs to population $\pi_1$, and with mean $\boldsymbol{\mu}_2$ and covariance matrix $\Sigma_2$ otherwise. As we have seen, the classification rule from the ECM depends on the ratio of densities $f_1(\boldsymbol{x})/f_2(\boldsymbol{x})$, and that, in the case of equal covariance matrices, the factor $(2\pi)^{p/2}|\Sigma|^{1/2}$ cancels out, which is not the case in the more general case of unequal covariance matrices. Furthermore, the quadratic forms $-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^t\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)$ in the exponents of the density functions will no longer reduce to a linear expression of $\boldsymbol{x}$.

After taking the natural logarithm of the general classification rule (10.1.2), and simplifying the expressions, we obtain the following classification regions:

$$\begin{aligned} R_1 : -\frac{1}{2}\boldsymbol{x}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^t\Sigma_1^{-1} - \boldsymbol{\mu}_2^t\Sigma_2^{-1})\boldsymbol{x} - k \\ \geqslant \log\left[\frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}\right] \\ R_2 : -\frac{1}{2}\boldsymbol{x}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^t\Sigma_1^{-1} - \boldsymbol{\mu}_2^t\Sigma_2^{-1})\boldsymbol{x} - k \\ < \log\left[\frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1}\right] \end{aligned} \tag{10.2.3}$$

where

$$k = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^t\Sigma_2^{-1}\boldsymbol{\mu}_2).$$

As such, we observe that the classification regions are defined by quadratic functions of $\boldsymbol{x}$, and that, when $\Sigma_2 = \Sigma_1$, the quadratic term $-\frac{1}{2}\boldsymbol{x}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x}$ disappears, yielding the simplified classification regions of (10.2.1).

The classification rule for general multivariate populations follows directly from the obtained regions.

**Result 12.** Let the populations $\pi_1$ and $\pi_2$ be distributed as a multivariate, normal distribution with population means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and covariance matrices equal to $\Sigma_1$ and $\Sigma_2$ respectively. The allocation rule that minimises the ECM is the following:

Allocate a new observation $\boldsymbol{x}_0$ to $\pi_1$ if

$$-\frac{1}{2}\boldsymbol{x}_0^t(\Sigma_1^{-1}-\Sigma_2^{-1})\boldsymbol{x}_0+(\boldsymbol{\mu}_1^t\Sigma_1^{-1}-\boldsymbol{\mu}_2^t\Sigma_2^{-1})\boldsymbol{x}_0-k\geqslant\log\left[\frac{c(\boldsymbol{x}\in R_1\mid\pi_2)}{c(\boldsymbol{x}\in R_2\mid\pi_1)}\times\frac{p_2}{p_1}\right],$$

and allocate $\boldsymbol{x}_0$ to $\pi_2$ otherwise, where

$$k=\frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)+\frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^t\Sigma_2^{-1}\boldsymbol{\mu}_2).$$

We refer to this approach as *quadratic discriminant analysis*.

In practice, we substitute the (generally) unknown population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\Sigma_1$, and $\Sigma_2$ by their equivalent sample quantities $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\hat{\Sigma}_1$, $\hat{\Sigma}_2$, and obtain this result.

**Result 13.** Allocate a new observation $\boldsymbol{x}_0$ to $\pi_1$ if

$$-\frac{1}{2}\boldsymbol{x}_0^t(\hat{\Sigma}_1^{-1}-\hat{\Sigma}_2^{-1})\boldsymbol{x}_0+(\bar{\boldsymbol{x}}_1^t\hat{\Sigma}_1^{-1}-\bar{\boldsymbol{x}}_2^t\hat{\Sigma}_2^{-1})\boldsymbol{x}_0-k\geqslant\log\left[\frac{c(\boldsymbol{x}\in R_1\mid\pi_2)}{c(\boldsymbol{x}\in R_2\mid\pi_1)}\times\frac{p_2}{p_1}\right],$$

and allocate $\boldsymbol{x}_0$ to $\pi_2$ otherwise, where

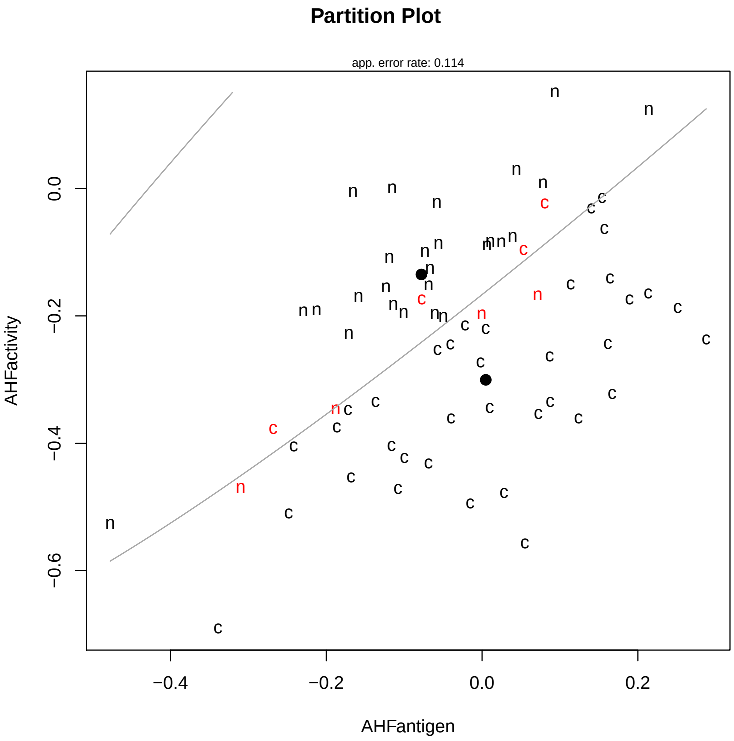
$$k=\frac{1}{2}\log\left(\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|}\right)+\frac{1}{2}(\bar{\boldsymbol{x}}_1^t\hat{\Sigma}_1^{-1}\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2^t\hat{\Sigma}_2^{-1}\bar{\boldsymbol{x}}_2).$$

### Example: hemophilia data

We return to the hemophilia example. Instead of a linear discriminant analysis, we now perform a quadratic discriminant analysis on these data to obtain the results.

```
output.qda <- qda(gr~AHFactivity+AHFantigen,data=hemophilia[-(31:35),])
output.qda

Call:
```

```
qda(gr ~ AHFactivity + AHFantigen, data = hemophilia[-(31:35),
    ])

Prior probabilities of groups:
  carrier    normal
0.5714286 0.4285714

Group means:
        AHFactivity  AHFantigen
carrier  -0.3003825  0.00477000
normal   -0.1348700 -0.07785667

qdapred <- predict(output.qda, hemophilia[-(31:35),])$class
table(hemophilia[-(31:35),]$gr, qdapred)

         qdapred
          carrier normal
  carrier      36      4
  normal        4     26

partimat(as.factor(gr) ~ AHFactivity+AHFantigen, data = hemophilia[-(31:35),],
         method = "qda", imageplot = FALSE)
```



```
predict(output.qda, hemophilia[31:35,])

$class
[1] carrier carrier carrier carrier normal
Levels: carrier normal


$posterior
     carrier       normal
31 0.9999435 5.646517e-05
32 0.5345803 4.654197e-01
33 0.9994472 5.528084e-04
34 0.7843529 2.156471e-01
35 0.3663978 6.336022e-01
```

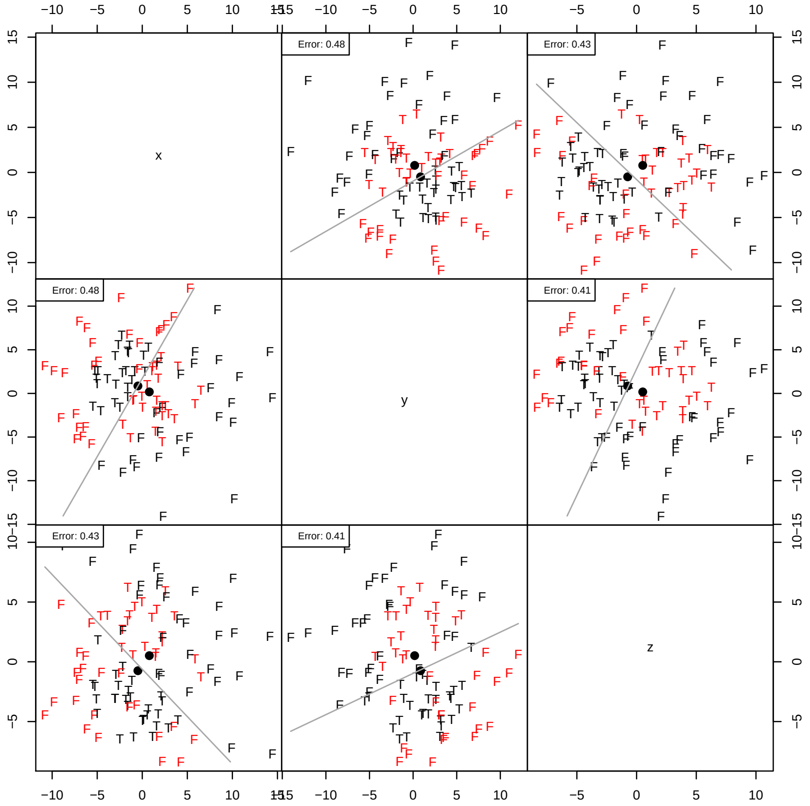
Performing a quadratic discriminant analysis returns similar output, minus the coefficients of the linear discriminant, so we will not address that again. A visualisation of the data and the classification rule shows that the classification boundary, the line on both graphs, is quite similar for both analyses. Only near the boundary of our data we observe a clear deviation due to the different assumptions that we made which allows two of the carriers at the boundary to lie on the correct side of the boundary now while they were lying on the wrong side of the LDA boundary.

If we make predictions for the 5 left-out observations, then the quadratic model predicts the same classes as the LDA, with similar posterior probabilities in all cases.

### Example 2

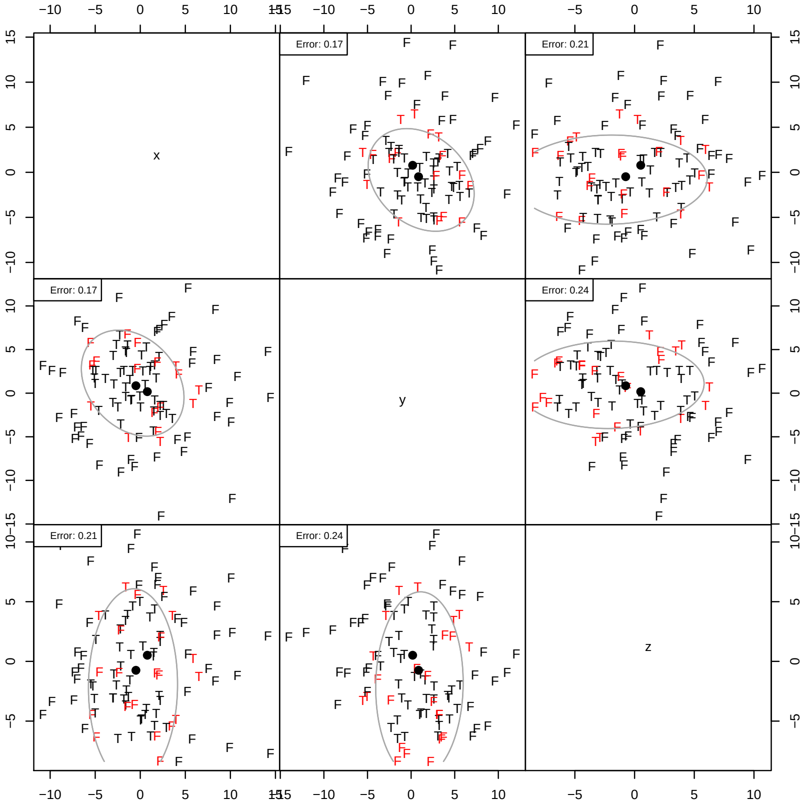
For this example, we use an artificial dataset consisting of 100 observations, 3 variables `x`, `y`, and `z`, and binary response `posit`. Once again, we shall perform both a linear and a quadratic discriminant analysis.

```
output.lda <- lda(posit ~ x + y + z, data = datamat)
partimat(posit ~ x + y + z, data = datamat, method = "lda",
         imageplot = FALSE, plot.matrix = TRUE)
ldapred <- predict(output.lda, datamat)$class
table(posit, ldapred)

       ldapred
posit   FALSE TRUE
  FALSE    29   22
  TRUE     21   28
```



If we examine the classification in the scatterplot matrix and look at the classification table, we see that a linear classifier performs very poorly at classifying these observations correctly. In fact, nearly half of the observations are allocated to the wrong class! Closer inspection of the plots reveals that the cases labelled TRUE seem to be centered in the middle of the dataset, while the cases labelled FALSE are closer to the edge. Perhaps a quadratic classifier will be able to capture this behavior better.

```
output.qda <- qda(posit ~ x + y + z, data = datamat)
partimat(posit ~ x + y + z, data = datamat, method = "qda",
         imageplot = FALSE, plot.matrix = TRUE)
qdapred <- predict(output.qda, datamat)$class
table(posit, qdapred)

       qdapred
posit   FALSE TRUE
  FALSE    47    4
  TRUE      4   45
```



Observe that, due to the quadratic nature of the classifier, the classification rule manages to define the classification regions a lot better than the linear classification rule, with only 8 out of 100 observations still misclassified. Note that the error rates mentioned on the plots are higher, but this is due to the fact that each subplot is the result of a discriminant analysis using only two of the three variables.

## 10.3 Evaluating classification rules

One way of judging the performance of any classification rule is by determining its "error rates", or misclassification probabilities. In this section, we illustrate three different ways of computing an error rate, along with their main advantages and disadvantages.

When the forms of the populations are completely known, we can easily compute the misclassification probability of the classification rule, which is the total probability of misclassification

$$TPM = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}.$$

The smallest value of this quantity, which is achieved by a good choice of $R_2$ and $R_1$, is called the optimum error rate (OER). This quantity can be expressed as

$$OER = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

where $R_2$ and $R_1$ are determined by special case (2) of the general rule (10.1.2), since equal misclassification costs are implied here. Thus, the OER is the error rate for the minimum TPM classification rule.

**Example: normal populations with equal covariance matrices**

Assume that the prior probabilities of belonging to populations $\pi_1$ and $\pi_2$ are equal, so $p_1 = p_2 = \frac{1}{2}$. Since the OER is based on the minimum TPM classification rule, where it is assumed that the misclassification costs are equal, $c(\boldsymbol{x} \in R_2 \mid \pi_1) = c(\boldsymbol{x} \in R_1 \mid \pi_2)$. We apply classification rule (10.2.2) with

$$\log \left[ \frac{c(\boldsymbol{x} \in R_1 \mid \pi_2)}{c(\boldsymbol{x} \in R_2 \mid \pi_1)} \times \frac{p_2}{p_1} \right] = 0,$$

and find that

$$\begin{aligned} R_1 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geqslant 0 \\ R_2 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0. \end{aligned}$$

Denote $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \boldsymbol{x} = \boldsymbol{a}^t \boldsymbol{x}$, then

$$\begin{aligned} R_1 &: \boldsymbol{a}^t \boldsymbol{x} \geqslant \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\ R_2 &: \boldsymbol{a}^t \boldsymbol{x} < \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \end{aligned}$$

Because the corresponding random variable $Y = \boldsymbol{a}^t\boldsymbol{X}$ is a linear combination of normal random variables for each population $\pi_1$ and $\pi_2$, $Y$ also has a univariate normal distribution for each population, with

$$\begin{aligned}
\mu_{1Y} &= \boldsymbol{a}^t\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\Sigma^{-1}\mu_1, \\
\mu_{2Y} &= \boldsymbol{a}^t\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\Sigma^{-1}\mu_2, \text{ and} \\
\sigma_Y^2 &= \boldsymbol{a}^t\Sigma\boldsymbol{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t = \Delta^2.
\end{aligned}$$

With this simplification we can easily determine (try it!) that the misclassification probabilities are

$$\begin{aligned}
&P(\text{observation is misclassified as } \pi_1) \\
&= P(\text{observation is misclassified as } \pi_2) = \Phi\left(-\frac{\Delta}{2}\right),
\end{aligned}$$

with $\Phi(\cdot)$ the cumulative distribution function of a standard normal variable. From this, we obtain that

$$OER = \text{minimum } TPM = \Phi\left(-\frac{\Delta}{2}\right).$$

In general, the population density functions aren't fully known and certain population parameters must be estimated from the sample. This of course makes evaluating the error rates less straightforward.

One possible approach is to calculate the actual error rate (AER),

$$AER = p_1\int_{\hat{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2\int_{\hat{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x},$$

where $\hat{R}_1$ and $\hat{R}_2$ represent the classification regions determined by samples of size $n_1$ and $n_2$ respectively. However, like the optimal error rate, the AER still depends on the unknown density functions, so it cannot be exactly calculated. This quantity can be estimated by the apparent error rate (APER), which is the fraction of observations in the training sample that are misclassified by the sample classification function. This measure of performance does not depend on the form of the population distributions and can, in fact, be computed for any classification procedure.

To determine the APER, first construct the *confusion matrix*, which shows the actual population memberships versus the predicted memberships.

|  |  | Predicted membership |  |  |
|---|---|---|---|---|
|  |  | $\pi_1$ | $\pi_2$ |  |
| Actual | $\pi_1$ | $n_{1c}$ | $n_{1m} = n_1 - n_{1c}$ | $n_1$ |
| membership | $\pi_2$ | $n_{2m} = n_2 - n_{2c}$ | $n_{2c}$ | $n_2$ |

In this table $n_{ic}$ is the number of observations from population $\pi_i$ that are correctly classified, and $n_{im}$ is the number of observations from population $\pi_i$ that are misclassified, for $i = 1, 2$. The APER is then equal to

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2}.$$

**Example: hemophilia data**

```
output.lda=lda(gr~AHFactivity+AHFantigen,data=hemophilia[-(31:35),])
ldapred <- predict(output.lda, hemophilia[-(31:35),])$class
table(hemophilia[-(31:35),]$gr, ldapred)

         ldapred
          carrier normal
  carrier      34      6
  normal        4     26
```

Coming back to the hemophilia example, we see that 6 of the patients carrying the hemophilia gene are incorrectly classified by the linear discriminant method as noncarriers, and that, conversely, 4 of the patients not carrying the gene are incorrectly classified as carriers. The apparent error rate for this classification rule and sample is then

$$APER = \frac{6+4}{40+30} = \frac{10}{70} = 14.3\%.$$

Similarly, for the quadratic classification rule obtained earlier we find an APER of 11.4%.

Though the APER is intuitively appealing and easy to calculate, it tends to underestimate the AER because it evaluates the classification rule based on the sample that was used to build the rule. The classification rule is thus optimally adapted to the data which we use to evaluate its performance ,which will not be the case for new observations. This remains an issue unless $n_1$ and $n_2$ are very large.

A workaround to this problem is to split the sample in a training sample and a validation sample, build the classification rule using the training sample, and determine which fraction of the observations of the validation sample are misclassified. Though this method bypasses the issue that the APER has, the classification rule will suffer from reduced accuracy because not all available observations are used to build the classification rule.

Another approach that works well is the "holdout" or "leave-one-out" procedure:

1. From the observations of population $\pi_1$, omit one observation, and build a classification rule based on the remaining $n_1 - 1$, $n_2$ observations.
2. Use this classification rule to classify the "holdout" observation.
3. Repeat steps 1 and 2 for each observation in $\pi_1$, and denote $n_{1m}^{(H)}$ the number of holdout observations in this group that are misclassified.
4. Repeat steps 1 through 3 for the observations of $\pi_2$, and denote $n_{2m}^{(H)}$ the number of holdout observations in this group that are misclassified.

We can then obtain estimates of the misclassification probabilities by

$$\begin{aligned}\hat{P}(\text{observation from } \pi_1 \text{ is misclassified as } \pi_2) &= \frac{n_{1m}^H}{n_1}\\ \hat{P}(\text{observation from } \pi_2 \text{ is misclassified as } \pi_1) &= \frac{n_{2m}^H}{n_2},\end{aligned}$$

and, for moderate samples, a nearly unbiased estimate of the expected actual error rate, $\mathrm{E}[AER]$, as

$$\hat{\mathrm{E}}[AER] = \frac{n_{1m}^H + n_{2m}^H}{n_1 + n_2}.$$

Though a new classification rule has to be built for each "holdout" observation, this approach is still computationally feasible with the linear and quadratic discriminant rules (10.2.1) and (10.2.3), respectively.

**Example: hemophilia data**

We revisit the hemophilia example again.

```
hemo.lda.cv=lda(gr~AHFactivity+AHFantigen,data=hemophilia[-(31:35),],CV=T)
table(hemophilia[-(31:35),]$gr,hemo.lda.cv$class)

          carrier normal
  carrier      34      6
  normal        4     26
```

With the additional option `CV = TRUE` to the function `lda`, a linear discriminant analysis is performed with each observation held out in turn, and predictions can be obtained. We observe that, with the holdout procedure applied to the linear classification rule, 6 of the patients carrying the hemophilia gene are misclassified, as well as 4 of the patients not carrying the gene. We find the following estimates for the misclassification probabilities,

$$\hat{P}(\text{observation from } \pi_1 \text{ is misclassified as } \pi_2) = \frac{6}{40} = 15\%$$
$$\hat{P}(\text{observation from } \pi_2 \text{ is misclassified as } \pi_1) = \frac{4}{30} = 13.3\%,$$

and an expected actual error rate of

$$\hat{\mathrm{E}}[AER] = \frac{6+4}{40+30} = \frac{10}{70} = 14.3\%,$$

which equals the APER of 14.3% for this example.

We also revisit the quadratic case,

```
hemo.qda.cv=qda(gr~AHFactivity+AHFantigen,data=hemophilia[-(31:35),],CV=T)
table(hemophilia[-(31:35),]$gr,hemo.qda.cv$class)

          carrier normal
  carrier      34      6
  normal        5     25
```

which yields an expected actual error rate of

$$\hat{\mathrm{E}}[AER] = \frac{6+5}{40+30} = \frac{11}{70} = 15.7\%,$$

a significant increase of the APER which was 11.4% in this case. This expected actual error rate of QDA is also higher than the expected actual error rate of LDA on these data, a clear indication that a linear classifier suffices to classify patients as carriers of the hemophilia gene or not, based on their AHF activity and AHF antigen levels.

## 10.4 Additional classification techniques

### 10.4.1 Fisher's discriminant function

Assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is a $p$-variate random variable, which belongs to either of populations $\pi_1$ or $\pi_2$, with prior probabilities $p_1$ and $p_2$ respectively. Furthermore, we will also assume that the population covariance matrices $\Sigma_1$ and $\Sigma_2$, of $\pi_1$ and $\pi_2$ respectively, are equal, but make no further assumptions on the distribution of $\boldsymbol{X}$ for either population.

Instead of finding regions $R_1$ and $R_2$ which minimise the cost of misclassification, or the total probability of misclassification, we can try find a transformation of the multivariate observations $\boldsymbol{x}_i$ to univariate observations $y_i$, such that the transformed observations of either populations are separated as much as possible. Fisher suggested to use a linear transformation because of its simplicity and ease of handling.

To derive *Fisher's discriminant function*, start with samples $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}$ of population $\pi_1$ and $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}$ of population $\pi_2$, and take a fixed linear combination $y = \boldsymbol{a}^t\boldsymbol{x}$ of the observations in the samples. The separation of the transformed observations $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$ is the difference between their sample means, $\bar{y}_1$ and $\bar{y}_2$ respectively, expressed in standard deviation units. In other words,

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \text{ where } s_y^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the variance of $Y$. Our objective is to select the linear combination which maximises the separation of the sample means.

**Result 14.** The linear combination $\hat{y} = \hat{\boldsymbol{a}}^t\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t\hat{\Sigma}_p^{-1}\boldsymbol{x}$, with $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ the sample means of the untransformed samples, maximises the ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}^t\boldsymbol{\delta})^2}{\hat{\boldsymbol{a}}^t\hat{\Sigma}_p\hat{\boldsymbol{a}}}$$

over all possible coefficient vectors $\boldsymbol{a}$, where $\boldsymbol{\delta} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$. The maximum of this ratio is $\Delta^2 = \boldsymbol{\delta}^t\hat{\Sigma}_p^{-1}\boldsymbol{\delta}$.

The proof of this result is a direct application of the maximisation lemma, which states that, for $B \in \mathbb{R}^{p\times p}$ a positive definite matrix, and $\boldsymbol{\delta} \in \mathbb{R}^p$, the maximum of

$$\frac{(\boldsymbol{x}^t\boldsymbol{\delta})^2}{\boldsymbol{x}^t B\boldsymbol{x}},$$

taken over all nonzero vectors $\boldsymbol{x} \in \mathbb{R}^p$, is attained for $\boldsymbol{x} = cB^{-1}\boldsymbol{\delta}$, and equals $\boldsymbol{\delta}^t B^{-1}\boldsymbol{\delta}$ for any nonzero constant $c$.

Based on this result, we obtain the following allocation rule, as illustrated in Figure 11.8.

**Result 15.** Allocate a new observation $\boldsymbol{x}_0$ to $\pi_1$ if

$$\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t\hat{\Sigma}_p^{-1}\boldsymbol{x}_0 \geqslant \hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t\hat{\Sigma}_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2),$$

and to $\pi_2$ otherwise.

Note that Fisher's discriminant function is actually a special case of the linear classification rule for two normal populations, where we assume that the ratio of the misclassification costs, $c(\boldsymbol{x} \in R_1 \mid \pi_2)/c(\boldsymbol{x} \in R_2 \mid \pi_1)$, is the reciprocal of the ratio of prior probabilities, $p_2/p_1$.

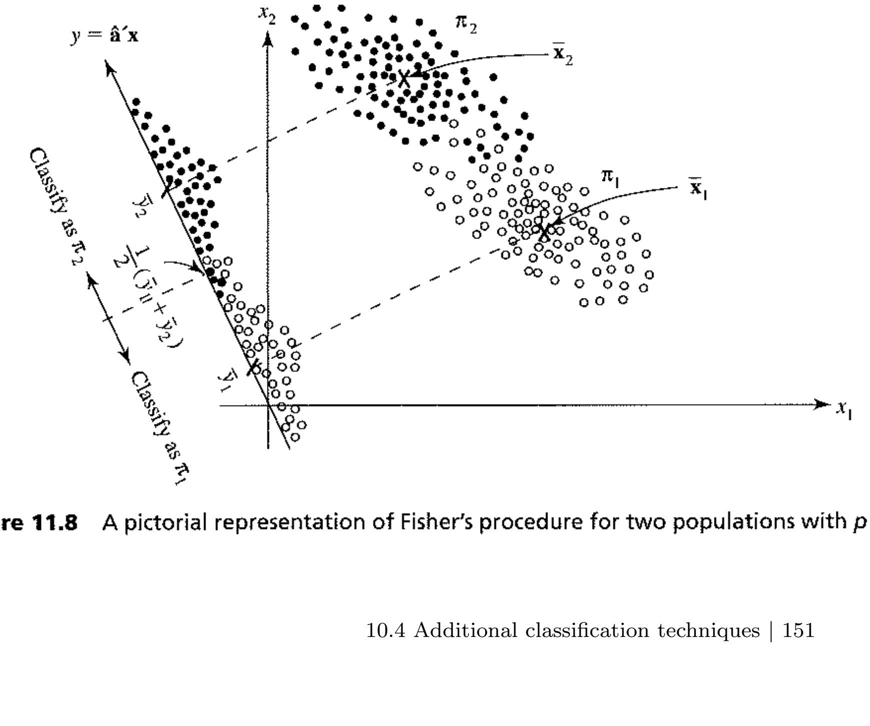


**Figure 11.8** A pictorial representation of Fisher's procedure for two populations with $p = 2$.

### 10.4.2 ECM for multiple populations

The expected cost of misclassification (ECM) rule can be extended, in a natural way, to cover situations in which there are more than two distinct populations. Assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is a $p$-variate random variable, which belongs to either of populations $\pi_1, \ldots, \pi_g$, with prior probabilities $p_1, \ldots, p_g$ respectively, and $g > 2$. Our goal is to find regions $R_1, \ldots, R_g$ such that the resulting allocation rules assign an observation $\boldsymbol{x}$ to $\pi_j$ if $\boldsymbol{x} \in R_j$, is optimal in some sense.

Denote the misclassification costs of classifying an observation from $\pi_i$ to $\pi_j$ as $c(\boldsymbol{x} \in R_j \mid \pi_i)$, with $c(\boldsymbol{x} \in R_i \mid \pi_i) = 0$, for $i, j = 1, \ldots, g$. The associated probabilities are

$$P(\boldsymbol{X} \in R_j \mid \pi_i) = \int_{R_j} f_i(\boldsymbol{x}) d\boldsymbol{x}$$

for $i, j = 1, \ldots, g$, where $f_i(\boldsymbol{x})$ is the density function of population $\pi_i$.

The conditional expected cost of misclassifying an observation $\boldsymbol{x}$ from population $\pi_i$ into a different population is

$$ECM(i) = \sum_{j=1, j\neq i}^{g} P(\boldsymbol{X} \in R_j \mid \pi_i) c(\boldsymbol{x} \in R_j \mid \pi_i).$$

After multiplying these expressions with their respective prior $p_i$ and summing over $i$, we obtain the expected cost of misclassification

$$ECM = \sum_{i=1}^{g} p_i \left( \sum_{j=1, j\neq i}^{g} P(\boldsymbol{X} \in R_j \mid \pi_i) c(\boldsymbol{x} \in R_j \mid \pi_i) \right). \qquad (10.4.1)$$

Building an optimal classification rule then amounts to determining the regions $R_1, \ldots, R_g$, mutually exclusive and exhaustive, such that (10.4.1) is minimal.

**Result 16.** The classification regions that minimise the ECM (10.4.1) are defined by allocating $\boldsymbol{x}_0$ to that population $\pi_i$, $i = 1, \ldots, g$, for which

$$\sum_{j=1, j\neq i}^{g} p_j f_j(\boldsymbol{x}_0) c(\boldsymbol{x} \in R_i \mid \pi_j)$$

is smallest. If a tie occurs, $\boldsymbol{x}$ can de assigned to any of the tied populations.

If all the costs of misclassification are equal, the ECM rule reduces to minimising the total probability of misclassification (TPM), and we would allocate $\boldsymbol{x}_0$ to that population $\pi_i$ for which

$$\sum_{j=1, j\neq i}^{g} p_j f_j(\boldsymbol{x}_0)$$

is smallest. This expression is smallest when the omitted term, $p_i f_i(\boldsymbol{x}_0)$, is largest. Consequently, the TPM classification rule for multiple populations has the following rather simple form:

**Result 17.** The classification rule which minimises the TPM, allocates $\boldsymbol{x}_0$ to population $\pi_i$ if

$$p_i f_i(\boldsymbol{x}_0) > p_j f_j(\boldsymbol{x}_0) \text{ for all } j \neq i.$$

Note that this result is identical to the classification rule that maximizes the posterior probability $P(\boldsymbol{X} \in \pi_i \mid \boldsymbol{X} = \boldsymbol{x}_0)$, where

$$P(\boldsymbol{X} \in \pi_i \mid \boldsymbol{X} = \boldsymbol{x}_0) = \frac{p_i f_i(\boldsymbol{x}_0)}{\sum_{j=1}^{g} p_j f_j(\boldsymbol{x}_0)}.$$

If the distribution of $X$ is normal with mean $\mu_i$ and covariance matrix $\Sigma_i$ for every population $\pi_i$, and if $\Sigma_i = \Sigma$ for all $i = 1, \ldots, g$, the TPM rule reduces to the following linear classification rule:

**Result 18.** The classification rule which minimises the TPM for normal populations with equal covariance matrices, allocates $\boldsymbol{x}_0$ to population $\pi_i$ if

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1} (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) > \log\left[\frac{p_j}{p_i}\right],$$

for all $j \neq i$.

This result is very similar to the result obtained in Section 10.2.1, where we derived the linear discriminant function for two populations. In practice however we do not have the values of the population parameters $\boldsymbol{\mu}_i$, $i = 1, \ldots, g$, and $\Sigma$, so we must estimate these quantities. Assume that, for each $i = 1, \ldots, g$, we have a sample of size $n_i$ from population $\pi_i$. The population means and covariance matrices are estimated in the usual way, and we obtain a pooled estimator of the covariance matrix $\Sigma$ as

$$\hat{\Sigma}_p = \sum_{i=1}^{g} \left( \frac{n_i - 1}{N - g} \right) \hat{\Sigma}_i,$$

where $N = \sum_{i=1}^{g} n_i$. This leads to the sample classification rule below.

**Result 19.** Allocate $\boldsymbol{x}_0$ to population $\pi_i$ if

$$(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j)^t \hat{\Sigma}_p^{-1} \boldsymbol{x}_0 - \frac{1}{2} (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j)^t \hat{\Sigma}_p^{-1} (\bar{\boldsymbol{x}}_i + \bar{\boldsymbol{x}}_j) > \log \left[ \frac{\hat{p}_j}{\hat{p}_i} \right],$$

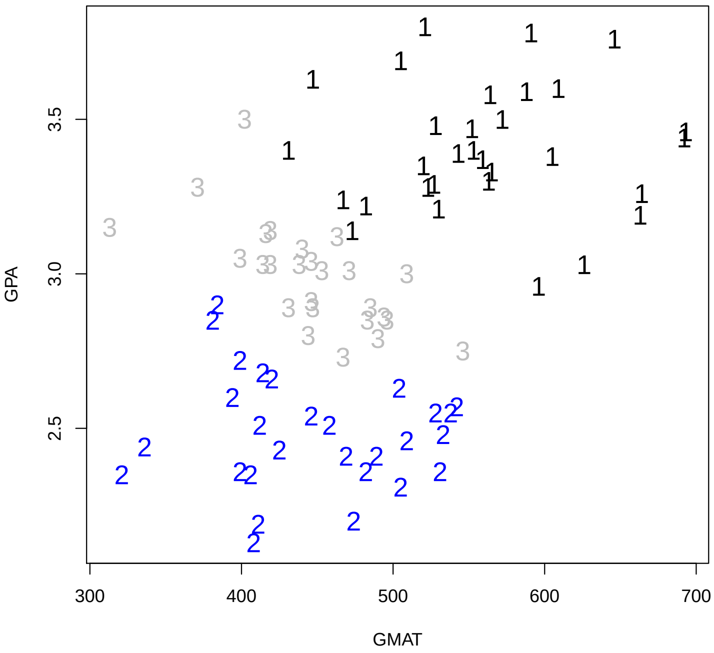
for all $j \neq i$, where $\hat{p}_i = \dfrac{n_i}{N}$ and $\hat{p}_j = \dfrac{n_j}{N}$ are the sample proportions of populations $\pi_i$ and $\pi_j$ respectively.

**Example: Business school admission**

The admission board of a business school uses two measures to decide on admittance of applicants:

- GPA= undergraduate grade point average
- GMAT=graduate management aptitude test score

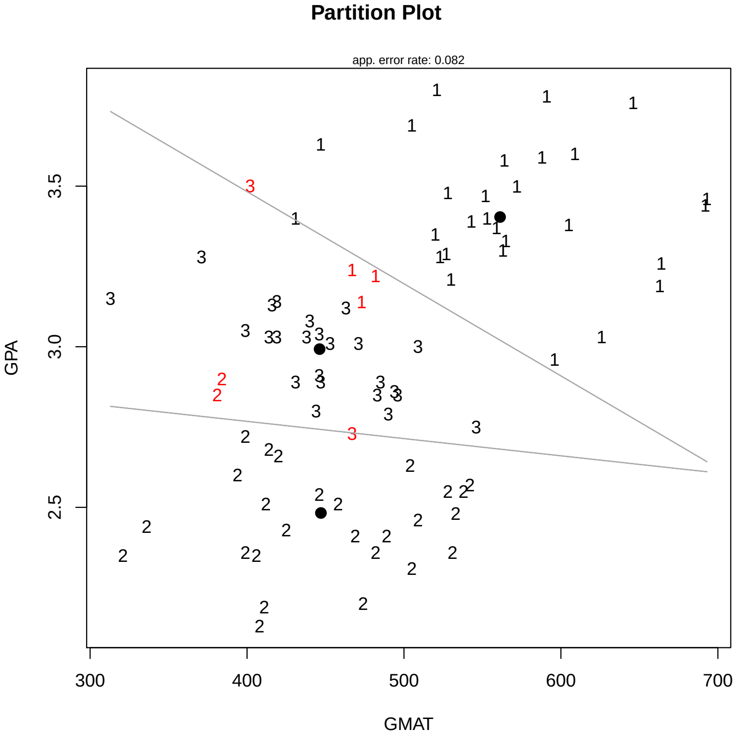
Based on these measures applicants are categorized as: admit ($\pi_1$), do not admit ($\pi_2$), and borderline ($\pi_3$). The available data are shown in the scatterplot.



From the plot, we can see that applicants with a high GPA and GMAT score are admitted, while applicants which score low in both cases are not admitted. The borderline cases lie in between both groups. Next, we perform a linear discriminant analysis on these data.

```
admit.lda=lda(class~GPA+GMAT,data=admit)
ldapred <- predict(admit.lda, admit)$class
table(admit$class, ldapred)

   ldapred
     1  2  3
  1 28  0  3
  2  0 26  2
  3  1  1 24

partimat(class~GPA+GMAT, data = admit, method = "lda",
         imageplot = FALSE)
output.ldacv <- lda(class~GPA+GMAT, data = admit, CV = TRUE)
table(admit$class, output.ldacv$class)

     1  2  3
```

```
1 27  0  4
2  0 26  2
3  1  1 24
```



After building the linear classifier and analysing its performance, we observe that, for example, of the students that were not admitted, 1 was misclassified as a borderline case, and 1 as admitted . In general, $0+3+0+2+1+1=7$ of the 85 students were misclassified, and as such, the apparent error rate is equal to

$$APER = \frac{7}{85} = 8.2\%.$$

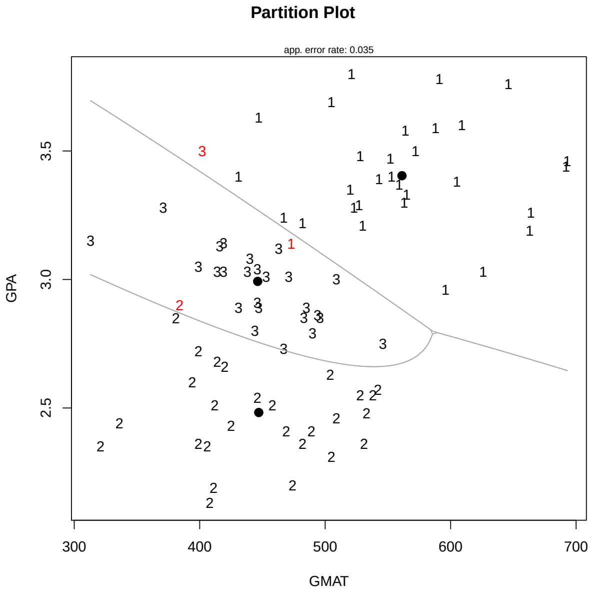
We can also estimate the expected actual error rate ($\mathrm{E}[AER]$) using the holdout method and find

$$\hat{\mathrm{E}}[AER] = \frac{0+4+0+2+1+1}{85} = \frac{8}{85} = 9.4\%.$$

The visual representation of the classification rule shows that our first intuition regarding the admissions is valid. Note that, in general, there are three classification boundaries which meet in a single point, assuming $g=3$ populations and $p=2$ variables. In this case, however, the third classification boundary and the intersection point aren't visible on the graph.

We can compare these results with the results obtained by a quadratic classification rule.

```
admit.qda <- qda(class~GPA+GMAT, data = admit)
qdapred <- predict(admit.qda, admit)$class
table(admit$class, qdapred)

   qdapred
     1  2  3
  1 30  0  1
  2  0 27  1
  3  1  0 25

partimat(class~GPA+GMAT, data = admit, method = "qda",
         imageplot = FALSE)
output.qdacv <- qda(class~GPA+GMAT, data = admit, CV = TRUE)
table(admit$class, output.qdacv$class)

     1  2  3
  1 30  0  1
  2  0 27  1
  3  1  1 24
```



We observe that the quadratic classification rule shows the biggest performance improvement for classifying admitted students, compared to the linear classification rule. In general, $0 + 1 + 0 + 1 + 1 + 0 = 3$ of the 85 students were misclassified, and as such, the apparent error rate is equal to

$$APER = \frac{3}{85} = 3.5\%.$$

This is roughly 5% below the APER of the linear classification rule, indicating that a quadratic classification rule is more suitable. However, this can be due to overfitting of the actual sample. To see whether this is the case, we also estimate the expected actual error rate ($\mathrm{E}[AER]$). If we do this, we find an error rate of

$$\hat{\mathrm{E}}[AER] = \frac{4}{85} = 4.7\%.$$

In this case, we observe that the estimated $\mathrm{E}[AER]$ of the quadratic classification rule decreases to about 4.7% when compared to the linear classification rule, giving a stronger indication that a quadratic rule works better in this case. From the visual representation of the quadratic classification rule we see that the largest difference with the linear classification rule is between the borderline cases and not admitted students.

### 10.4.3 $k$-nearest neighbours

Assume that we have a sample of $n$ observations of a random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, where each observation can be drawn from a number of distinct populations $\pi_i$, $i = 1, \ldots, g$. Furthermore, assume that there is a new observation $\boldsymbol{x}_0$ of $\boldsymbol{X}$, then the following allocation procedure can be applied:

1. Compute the (Euclidean) distance of $\boldsymbol{x}_0$ to each observation $\boldsymbol{x}$ in the sample, and rank the observations in the sample from smallest to largest distance to $\boldsymbol{x}_0$.

2. Define the set $S$ as the set of the first $k$ of the ordered observations. In the case that the $(k+j)$-th ordered observation has the exact same distance to $\boldsymbol{x}_0$ as the $k$-th ordered observation, add that observation to the set $S$ as well.

3. Define $n_i(\boldsymbol{x}_0)$, $i = 1, \ldots, g$, the number of observations in the set $S$ which are drawn from population $\pi_i$.

4. Allocate $\boldsymbol{x}_0$ to the population $\pi_i$ for which $n_i(\boldsymbol{x}_0)$ is largest. In case of a tie, allocate $\boldsymbol{x}_0$ to any of the tied populations.

This procedure for building a classification rule is called the *$k$-nearest neighbour method.*

Crucial in this procedure is the choice of $k$: if $k$ is chosen too small, then the classification rule will suffer from a high actual error rate (AER) even though the apparent error rate (APER) will be small (what happens to the APER if $k = 1$?). On the other hand, if $k$ is too large, the classification rule will not be able to capture all the information which is present in the sample, once again leading to a high AER and APER (what are the AER and APER if $k = n$?).

**Example: Business school admission**

We revisit the example of the admissions to a business school, but this time we will not make any assumptions on the distributions of the exam scores. Hence, we will perform the classification using the $k$-nearest neighbour algorithm, where we choose $k = 3$ and $k = 5$. Since we use euclidean distances, it is advisable to standardize the variables.

```
library(class)
admit.stand=scale(admit[1:2])
output.knn3 <- knn(admit.stand, admit.stand, admit$class, k = 3)
table(admit$class, output.knn3)

   output.knn3
     1  2  3
  1 31  0  0
  2  0 28  0
  3  1  1 24

output.knn3cv <- knn.cv(admit.stand, admit$class, k = 3)

   output.knn3cv
     1  2  3
  1 31  0  0
  2  0 28  0
  3  1  1 24
```

If we set $k = 3$, we see that the $k$-nearest neighbour procedure only misclassifies 2 out of 85 observations, for an apparent error rate of 2.4%, which is an improvement of about 1.1% compared to the quadratic classifier obtained in Section 10.4.2. This may indicate that the 3-nearest neighbour classifier outperforms the quadratic classifier. To obtain a more well-founded conclusion about the quality of the classification rule, we also estimate the expected actual error rate ($\mathrm{E}[AER]$) with the 'leave-one-out' method. We see that 2 observations are misclassified using this method, leading to $\hat{\mathrm{E}}[AER] = 2.4\%$, which is the same as the APER for this method and lower than the error rate obtained by the quadratic classification rule.

We now consider the results for k=5.

```
output.knn5 <- knn(admit.stand, admit.stand, admit$class, k = 5)
table(admit$class, output.knn5)

   output.knn5
     1  2  3
```

```
  1 31  0  0
  2  0 28  0
  3  0  1 25
```

```
output.knn5cv <- knn.cv(admit.stand, admit$class, k = 5)
table(admit$class, output.knn5cv)
```

```
   output.knn5cv
     1  2  3
  1 28  0  3
  2  0 27  1
  3  0  1 25
```

If we use the 5-nearest neighbour method, we observe that even only 1 out of 85 (1.2%) is misclassified. However, if we compute the $\hat{\mathrm{E}}[AER]$, we obtain 5.9%, which is higher than for $k = 3$. Therefore we can conclude that the performance of the $k$-nearest neighbour is better when we set $k = 3$.

# Part II

# Regression Analysis

# Introduction

In its simplest form regression aims to model the relation between an input variable $X$ and an output or *response* variable $Y$. Contrary to a correlation analysis, the regression model is *asymmetric*. It models the **influence** or **effect** of the input or *predictor* variable $X$ on the response variable $Y$. The regression model allows us to evaluate to what extent the outcome $Y$ changes due to a change in the value of $X$. The regression model can then be used to **predict** $Y$ from $X$. Therefore, the input variable $X$ is also called the *independent* variable, or *regressor*, whereas the response variable $Y$ is also called the *dependent* variable.

More generally, regression analysis models the relationship between a set of predictor variables $X_1, X_2, \ldots, X_{l-1}$ and a response variable $Y$ that are measured on $n$ observations. The goal is now to find a **relation** between the $X_j$ $(j = 1, \ldots, l-1)$ and $Y$, which reveals the joint influence of the $X$-variables on $Y$. This model can then also be used to predict the dependent variable $Y$ from the independent variables $X_1, \ldots, X_{l-1}$. In a very general form, we seek real functions $g, f$ and a parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^t$ such that $g(Y)$ can be well described by $f(X_1, \ldots, X_{l-1}, \boldsymbol{\beta})$. Unless otherwise stated, we will assume that the response variable is *continuous*.

Since the observations will in general not satisfy this functional relation exactly, the regression model will also include a stochastic component $\epsilon$ which expresses the variation of the data points around the regression curve. A regression model thus postulates that:

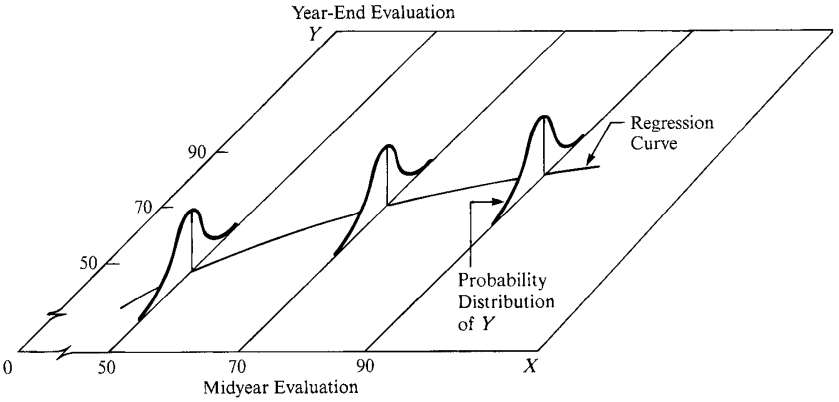
1. There is a probability distribution of $Y$ for each level of

$X = (X_1, \ldots, X_{l-1})$.

2. The means of these probability distributions vary in some systematic fashion with $X$.

This is illustrated in Figure 1.4.

**FIGURE 1.4 Pictorial Representation of Regression Model.**



We will especially study the **general linear** regression model which is defined as

$$g(y_i) = \beta_0 + \beta_1 f_1(x_{i1}, \ldots, x_{i,l-1}) + \ldots + \beta_{p-1} f_{p-1}(x_{i1}, \ldots, x_{i,l-1}) + \epsilon_i$$

for $i = 1, \ldots, n$ and for certain choices of $g, f_1, \ldots, f_{p-1}$. The error terms $\epsilon_i$ represent the random variation of the data points around the regression curve. We assume that the standard Gauss-Markov conditions are satisfied:

$$\begin{aligned} \mathrm{E}[\epsilon_i] &= 0 \\ \mathrm{Var}[\epsilon_i] &= \sigma^2 \\ \mathrm{E}[\epsilon_i \epsilon_j] &= 0 \text{ for all } i \neq j. \end{aligned}$$

The first condition expresses that at each level of $(X_1, \ldots, X_{l-1})$ the regression curve represents the mean of the corresponding probability distribution of $Y$. The second condition states that the probability distribution of $Y$ at each level of $(X_1, \ldots, X_{l-1})$ has the same variance, namely $\sigma^2$. The last condition implies that the error terms are uncorrelated.

This general linear model includes:

1. the first-order regression model: $l = p, g(y_i) = y_i, f_j(x_{i1}, \ldots, x_{i,p-1}) = x_{ij} \ \ (j = 1, \ldots, p-1)$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

2. simple regression: the first-order regression model with $p = 2$.

3. polynomial regression: $g, f_1, \ldots, f_{l-1}$ as in the first-order regression model, and e.g. additionally $f_l = X_1^2, f_{l+1} = X_3^2, f_{l+2} = X_1 X_2$.

4. variable selection: $g, f_1, \ldots, f_{l-3}$ as in the first-order regression model, all other $f_j = 0$.

5. transformations in $X$ or $Y$: $g(Y) = \log(Y), g(Y) = \frac{y^\lambda - 1}{\lambda}, f_j = \log(X_j)$.

Note that this model is linear in $\boldsymbol{\beta}$ and not necessarily in the independent variables $X_j$. An example of a nonlinear model is

$$y_i = \beta_0 + \beta_1 e^{\beta_2 x_i} + \epsilon_i.$$

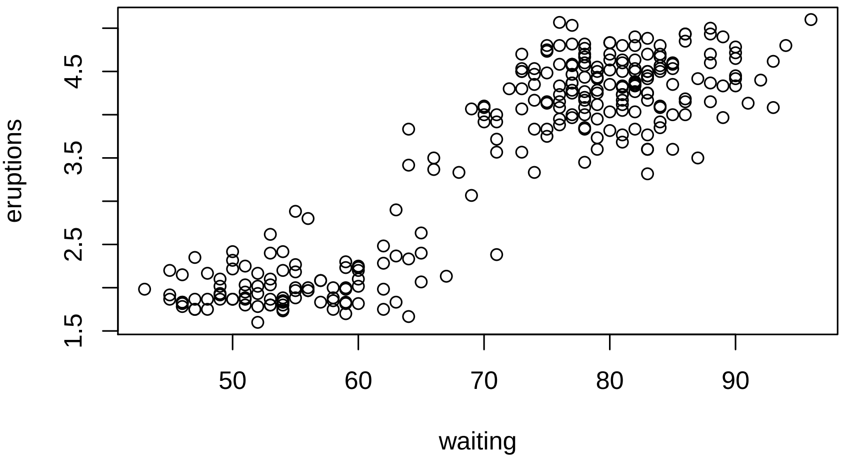
# Chapter 11

# The simple regression model

## 11.1 Examples

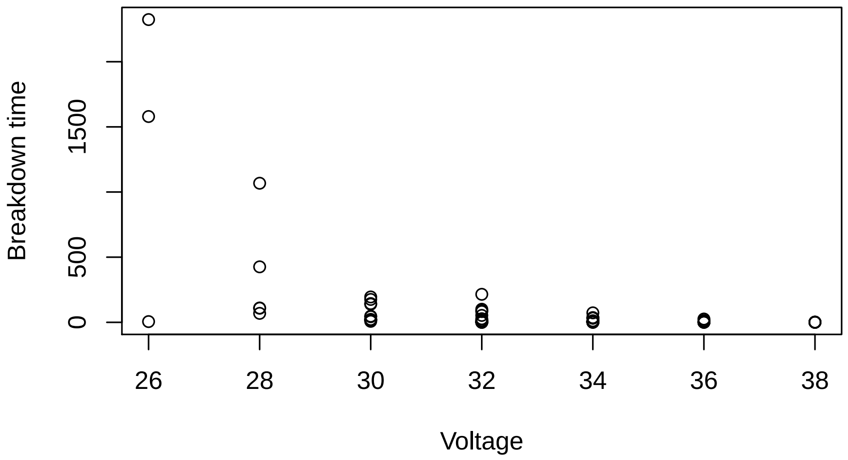
The 'old faithful geyser' is the most famous geyser in the Yellowstone National Park (Wyoming, USA). Eruptions occur in intervals with length between 45 minutes and 125 minutes. An eruption lasts 1.5 to 5 minutes during which 14,000 to 32,000 liters of boiling water is shot in the air to a height of 32 to 56 meters. It has been observed that there is a relation between the waiting time until an eruption and the duration of that eruption. To examine this relation both times (in minutes) have been recorded for 272 eruptions.

Questions that can be examined based on these data are: Is there indeed a strong influence of waiting time on the following eruption time? Can the waiting time be used to predict the length of the subsequent eruption? To answer these questions we model the relation between the waiting and eruption times. First, we graphically explore this relationship by making a scatterplot of the data.

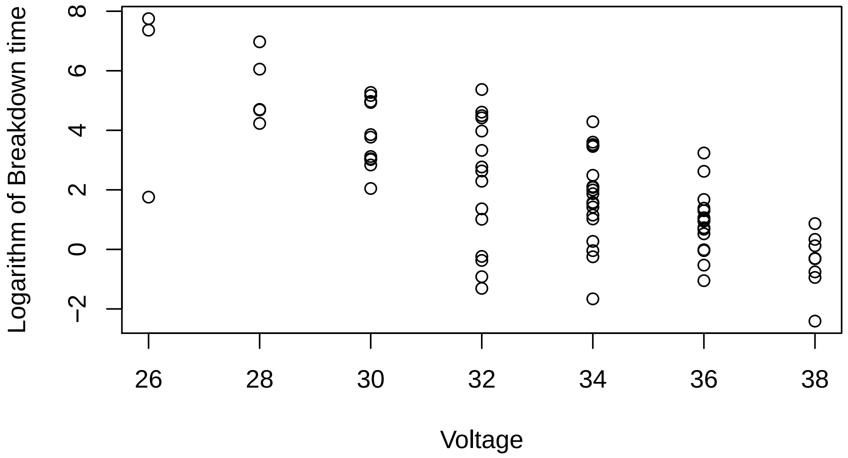


Note that the predictor variable 'waiting time' is plotted horizontally, while the response variable 'eruption time' is plotted vertically. The scatterplot clearly reveals that longer waiting times result in longer eruption times. The main pattern can at least approximately be represented by a line. However, it is also clear that the relation between both times is far from perfect. We will see how we can model the effect of the waiting time on the eruption time.

The second example is the result of an industrial laboratory experiment. Under uniform conditions, batches of electrical insulating fluid have been subjected to a constant voltage dose (in kV) until the insulating property of the fluid broke down. The process was repeatedly executed for seven different voltage levels and in each experiment the time until breakdown (in minutes) of the insulating property of the fluid is recorded. In total 76 experiments were executed. The main question of interest is to understand how the breakdown time depends on the administered voltage.



A scatterplot of the data is again used to explore the relationship. From the scatterplot we clearly see that the breakdown time decreases rapidly as the voltage increases. However, the type of relation between the two variables is difficult to see from this plot. This is caused by the skewness in the response variable. To improve the graphical representation of the data, we apply a logarithmic transformation on the response variable.



This scatterplot provides more insight in the data. We see a decrease of the breakdown time (logarithmic scale) with increasing dosage of voltage, a pattern that can be represented by a decreasing line. For every dose of voltage we can also see that there is considerable variation in the logarithmic breakdown time. We are now ready to model this relationship.

Note that there is an important difference between both examples. In the first example, the recorded values for both the waiting and eruption times are observed values. In this example both the $X$ and $Y$ variable are thus random variables. The observed measurement pairs can even be considered to be a random sample from the joint distribution of the two variables. In the second example, the voltage dose is chosen by the experimenters and the breakdown time is recorded for these fixed doses. In this example, the $Y$ variable is still random, but the $X$ variable is not. Consequently, in this case the paired dose-time measurements can also not be a random sample. Regression modeling as discussed next can be used for both types of data under suitable conditions.

## 11.2 The simple linear model

The simple linear model is given by

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i} \tag{11.2.1}$$

for $i = 1, \dots, n$. The parameter $\beta_0$ is called the *intercept*, whereas $\beta_1$ is called the regression *slope*.

In this model the values $x_i$ are not necessarily values of the observed predictor variable $X$, but can be values for any suitable function $f(X)$. We assume that $X$ does not contain any random effect or measurement error. Note that this assumption is naturally satisfied in an experimental setting as in the second example where the values of the predictor are chosen and fixed by the experimenter. In the case of an observational study where the values of $X$ are observed, as is the case for the response $Y$, it is far more difficult to satisfy this assumption. In this case, it is up to the statistician and/or data collectors to judge whether the $X$ variable is observed with sufficient accuracy so that the assumption is (approximately) satisfied. If there is considerable randomness in the observation of $X$, then more complex models such as measurement error models are needed.

Similarly as for $X$, the $y_i$ values are not necessarily values of the observed response variable $Y$, but can be values for any suitable function $g(Y)$. For the error term, we assume that the Gauss-Markov conditions are satisfied which means that

$$\mathrm{E}[\epsilon_i] = 0 \tag{11.2.2}$$

$$\mathrm{Var}[\epsilon_i] = \sigma^2 \tag{11.2.3}$$

$$\mathrm{E}[\epsilon_i \epsilon_j] = 0 \text{ for all } i \neq j \tag{11.2.4}$$

for $i = 1, \dots, n$.

In the case that $X$ is random, it is also assumed that the errors $\epsilon_i$ are independent of $X$.

As the $\epsilon_i$ are random variables with zero mean, also $Y$ is a random variable that satisfies:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X.$$

Here, $\mathrm{E}[Y|X]$ is a function of $X$ that for each value $X = x$ yields the mean of the corresponding distribution of the response variable $Y$ at $X = x$. Conditionally on the observed values for $X$, this can also be written as:

$$\mathrm{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i \tag{11.2.5}$$

At the first-order regression model where the $X$ and $Y$ variables in (11.2.1) correspond to the observed predictor variable and response variable respectively, this linear relation geometrically implies that we try to estimate a regression line

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X$$

in the $(X, Y)$-space.

For $X = 0$ we immediately obtain from (11.2.5) that $\mathrm{E}[Y|X = 0] = \beta_0$, hence the intercept of the model can be interpreted as the expected response when $X$ equals zero. That is, $\beta_0$ is the mean of the distribution of $Y$ at $X = 0$. If we increase the value of $X$ from an arbitrary value $x$ to $x + 1$, then the expected response increases from $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$ to $\mathrm{E}[Y|X = x + 1] = \beta_0 + \beta_1(x + 1)$. Therefore, we find that

$$\beta_1 = \mathrm{E}[Y|X = x + 1] - \mathrm{E}[Y|X = x],$$

so the slope $\beta_1$ can be interpreted as the change in the expected response $Y$ if $X$ increases by one unit. That is, if $X$ increases one unit, then $Y$ is expected to change by $\beta_1$ units on average.

## 11.3 Estimation of the regression parameters

### 11.3.1 The least squares estimator

The simple linear model in (11.2.1) contains three parameters $\beta_0$, $\beta_1$ and $\sigma$. Note that the two regression parameters $\beta_0$ and $\beta_1$ are inherent in the model (11.2.1) while the scale parameter $\sigma$ is a consequence of the second Gauss-Markov condition (11.2.3). These model parameters are unknown and need to be estimated from the available data. A natural strategy is to estimate the regression parameters such that the corresponding linear function fits the available data points as well as possible. Otherwise stated, the estimation method should aim to keep the errors as small as possible. Here, the errors corresponding to any parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$e_i(\hat{\beta}_0, \hat{\beta}_1) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i; \qquad i = 1, \ldots, n.$$

The first Gauss-Markov condition (11.2.2) implies that positive and negative errors occur. To avoid that large positive and large negative errors can cancel each other out in the estimation strategy, a function needs to be used that adds up all errors regardless of their sign. The two most common functions to achieve this goal are the absolute value and the square. Hence, the parameters $\beta_0$ and $\beta_1$ can be estimated by minimizing the sum of the absolute errors:

$$(\hat{\beta}_{0,\text{LAD}}, \hat{\beta}_{1,\text{LAD}}) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} |e_i(\beta_0, \beta_1)| = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|. \tag{11.3.1}$$

This estimator is called the *least absolute deviations estimator*. The other option is to estimate the parameters $\beta_0$ and $\beta_1$ by minimizing the sum of the squared errors:

$$(\hat{\beta}_{0,\text{LS}}, \hat{\beta}_{1,\text{LS}}) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} e_i^2(\beta_0, \beta_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2. \tag{11.3.2}$$

This estimator is called the **least squares estimator**. Both estimators have their merits, but the least squares estimator is the standard estimator for the regression parameters in linear models because it can be solved analytically and it has some good (optimal) statistical properties that will be discussed later.

The residual sum of squares $\sum_{i=1}^n e_i^2(\beta_0, \beta_1)$ is called the *objective function* or loss function of the least squares estimator. Differentiating this objective function $L(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$ with respect to $\beta_0$ and $\beta_1$ and setting these derivatives equal to zero, yields the normal equations

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] = 0 \qquad (11.3.3)$$

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]x_i = 0. \qquad (11.3.4)$$

The least squares estimators $\hat{\beta}_{0,\mathrm{LS}}$ and $\hat{\beta}_{1,\mathrm{LS}}$ for the simple regression model are the solution of this system of equations. From (11.3.3) we find that

$$\hat{\beta}_{0,\mathrm{LS}} = \bar{y}_n - \hat{\beta}_{1,\mathrm{LS}}\bar{x}_n. \qquad (11.3.5)$$

The second equation can be replaced by

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} - \bar{x}_n \frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} = 0$$

which leads to the equation

$$\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)](x_i - \bar{x}_n) = 0.$$

By substituting result (11.3.5) into this equation, we obtain that

$$\sum_{i=1}^n [(y_i - \bar{y}_n) - \hat{\beta}_{1,\mathrm{LS}}(x_i - \bar{x}_n)](x_i - \bar{x}_n) = 0.$$

Solving this equation yields

$$\hat{\beta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{\mathrm{cov}(X,Y)}{s_X^2} = \mathrm{cor}(X,Y)\frac{s_Y}{s_X}, \qquad (11.3.6)$$

where $s_X$ en $s_Y$ are the sample standard deviations of the variables $X$ and $Y$, and $\mathrm{cov}(X,Y)$ and $\mathrm{cor}(X,Y)$ are respectively the sample covariance and sample correlation between $X$ and $Y$.

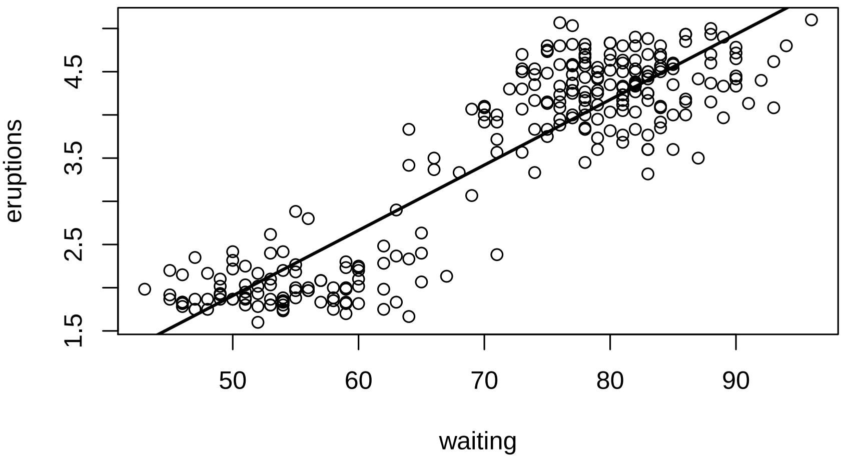
Note that for the existence of the least squares estimator it is required that $s_X^2 > 0$. This means that the values of the variable $X$ cannot be all the same, which is a natural condition. If all values of $X$ are equal to each other, then the data do not provide any information on how the value of the response $Y$ changes with changes in $X$.

### 11.3.2 Examples

In the old faithful geyser example, the simple regression model estimated based on the available data of 272 eruptions becomes

$$\text{Average eruption time} = -1.87 + 0.076 * \text{Waiting time}.$$

The graphical representation of this regression fit shows that the estimated regression line represents well the main trend in the data.

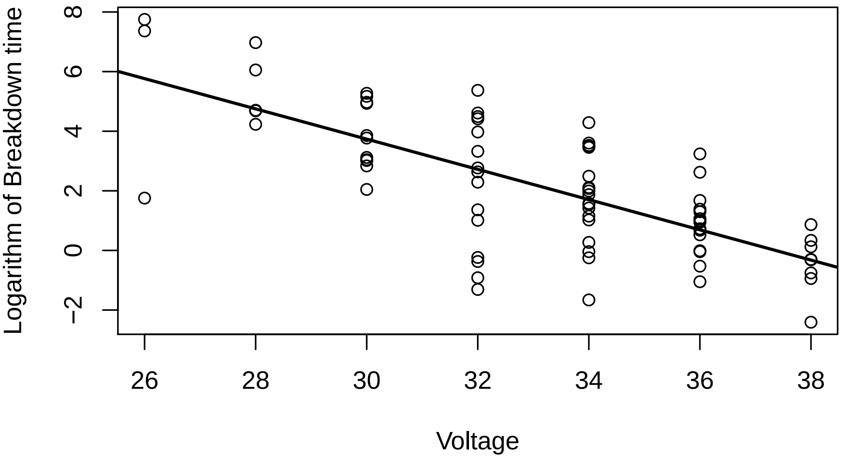


Note that the intercept has a negative sign which is physically not a meaningful value because an eruption time cannot be negative. This is an illustration of the danger of extrapolation from a regression model. The data do not contain any information about the length of eruption for very short waiting times (because short waiting times do not occur in reality). Hence, the model cannot be used to reliably predict what would happen with the eruption time after such small waiting times. There is no reason why the model would still be valid in this case, and in fact the unrealistic values predicted by the model indicate that it is not valid for such events beyond the range of information.

For the insulating fluid experiments, the simple regression model estimated from the available data is

$$\text{Average log(Breakdown time)} = 18.96 - 0.51 * \text{Voltage}.$$

The graphical representation of this regression fit shows that the estimated regression line again represents well the main trend in the data.



The interpretation of the regression coefficients in terms of the transformed response variable remains as before. For instance, the slope yields the change in average logarithmic breakdown time when the voltage dose is increased by one unit. However, this cannot easily be transformed into an interpretation in terms of the originally measured response variable. The reason is that the expected logarithmic response cannot be related to some expectation of the response variable on its original scale (why not?).

In case of a logarithmic transformation the regression coefficients can be interpreted in terms of the original response if we assume that the error distribution is symmetric around its mean zero. This is an acceptable assumption because it implies that at each value of $X$ the corresponding distribution of $Y$ is centered around the value of the regression line with equal positive and negative deviations. The symmetry assumption implies that $\mathrm{E}[Y|X = x] = \mathrm{med}[Y|X = x]$, that is, the mean coincides with the median of $Y$ at each $X = x$. Hence the regression line also models

$$\text{med}[Y|X] = \beta_0 + \beta_1 X$$

in this case. Now, if the response $Y$ in the simple linear model is the logarithm of an actual observed response $\tilde{Y}$, i.e. $Y = \log(\tilde{Y})$, then we obtain that

$$\text{med}[Y|X] = \text{med}[\log(\tilde{Y})|X] = \log(\text{med}[\tilde{Y}|X]) = \beta_0 + \beta_1 X,$$

or equivalently,

$$\text{med}[\tilde{Y}|X]) = \exp(\beta_0)\exp(\beta_1 X).$$

Moreover, we easily find that

$$\frac{\text{med}[\tilde{Y}|X = x+1])}{\text{med}[\tilde{Y}|X = x])} = \exp(\beta_1),$$

or

$$\text{med}[\tilde{Y}|X = x+1]) = \exp(\beta_1)\,\text{med}[\tilde{Y}|X = x])$$

Hence, $\exp(\beta_1)$ is the multiplicative change of the median of the measured response $\tilde{Y}$ if the predictor $X$ increases by one unit. A similar interpretation holds for the intercept. In the insulating fluid example, we find that $\exp(-0.51) = 0.60$ so with every unit increase in voltage the median breakdown point is only 60% of what it was before. Otherwise stated, the median breakdown point decreases by 40% for every unit increase in $X$. For example, the median breakdown point at $X = 32$ kV is estimated at 15.2 minutes. The median breakdown point at $X = 33$ kV then becomes $15.2 * 0.6 = 9.1$ minutes.

Similarly, if the predictor is included in the linear model in a transformed manner, then one should also think carefully about the interpretation of the regression coefficients in terms of a change in the value of the originally measured $X$. For example, consider a linear model of the form

$$E[Y|X] = \beta_0 + \beta_1 \log(X).$$

Hence, the predictor $X$ has been logarithmically transformed before inclusion in the simple regression model. The slope $\beta_1$ can of course be interpreted as the change in the expected response $Y$ if $\log(X)$ increases by one unit. However, what can we say about $Y$ if the originally measured $X$ is changed? By using properties of the logarithm, we find for any $c > 0$ that

$$E[Y|X = cx] = \beta_0 + \beta_1 \log(cx) = \beta_0 + \beta_1 \log(c) + \beta_1 \log(x),$$

such that

$$E[Y|X = cx] - E[Y|X = x] = \beta_1 \log(c).$$

For $c = 2$, this result implies that with every doubling of the value of $X$, the expected response $Y$ changes by the value $\beta_1 \log(2)$. Similarly, a ten-fold increase of $X$ induces a $\beta_1 \log(10)$ change in the expected value of $Y$.

Being able to estimate the parameters of a linear model is not sufficient. We should be able to check model adequacy and measure precision of parameter estimates. In particular, we are often interested in testing whether $X$ does have an effect on the response $Y$. Also of interest may be the prediction of the expected or individual response value for given values of $X$. Such questions can be answered by confidence intervals and hypothesis tests. However, the construction of such confidence intervals and hypothesis tests requires assumptions about the distribution of the errors $\epsilon_i$. Usually, it is assumed that these errors follow a normal distribution. It then becomes important to investigate whether this assumption is sufficiently reasonable to validate the resulting inference. These aspects will be discussed in the next chapters in the more general context of linear models with multiple regressors together with solutions if certain assumptions are not satisfied.

# Chapter 12

# The general linear model

## 12.1 The linear model

To simplify the notations we will write the general linear model as

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i} \tag{12.1.1}$$

for $i = 1, \ldots, n$. The parameter $\beta_0$ is called the *intercept*, whereas the $\beta_j$ $(j = 1, \ldots, p-1)$ are the regression *slopes*. In this model the $X_j$ do not necessarily stand for the observed predictor variables, but for any function $f_j$ of them. We also assume that the $X_j$ do not contain any random effect or measurement error. Moreover we assume that the Gauss-Markov conditions are satisfied. For all $i = 1, \ldots, n$ they state:

$$\mathrm{E}[\epsilon_i] = 0 \tag{12.1.2}$$

$$\mathrm{Var}[\epsilon_i] = \sigma^2 \tag{12.1.3}$$

$$\mathrm{E}[\epsilon_i \epsilon_j] = 0 \text{ for all } i \neq j. \tag{12.1.4}$$

As the $\epsilon_i$ are random with zero mean, also $Y$ is a random variable that satisfies:

$$\mathrm{E}[Y|X_1, \ldots, X_{p-1}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}.$$

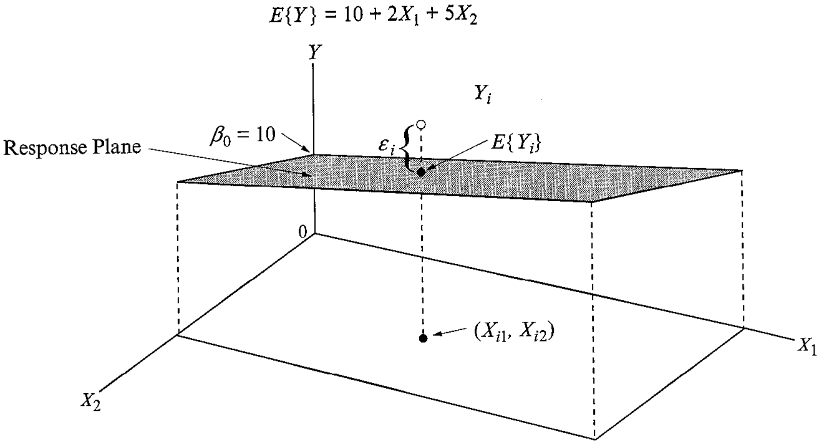
Conditionally on the observed values for $X_1, \ldots, X_{p-1}$, this can also be written as:

$$\mathrm{E}[Y|\boldsymbol{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} \tag{12.1.5}$$

with $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^t$. Note that the first element of the $\boldsymbol{x}$-vector is 1, which is the $x$-value for the intercept. At the first-order regression model (where the $X_j$ in (12.1.1) correspond with the observed predictor variables), this linear relation geometrically implies that we try to estimate a hyperplane in the $(X, Y)$-space. With $p = 2$ we recover simple regression as a special case and thus fit the regression line

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X.$$

**FIGURE 6.1 Response Function is a Plane—Sales Promotion Example.**



The intercept $\beta_0$ is the expected response value at $\boldsymbol{x}_i = (1, 0, \ldots, 0)^t$, that is when all predictors take the value 0. The slope parameter $\beta_j$ now indicates the change in the expected value of the response $Y$ due to a unit increase in the variable $X_j$ *when all other predictor variables are held constant.* Let $\boldsymbol{x}_{i(j)} = (1, x_{i1}, \ldots, x_{ij}, \ldots, x_{i,p-1})^t$ and $\boldsymbol{x}_{i(j+1)} = (1, x_{i1}, \ldots, x_{ij} + 1, \ldots, x_{i,p-1})^t$, then from (12.1.5) it follows that

$$\begin{aligned}\mathrm{E}(Y|\boldsymbol{x}_{i(j)}) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_{p-1} x_{i,p-1} \\ \mathrm{E}(Y|\boldsymbol{x}_{i(j+1)}) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j (x_{ij} + 1) + \ldots + \beta_{p-1} x_{i,p-1}\end{aligned}$$

hence indeed

$$\beta_j = \mathrm{E}(Y|\boldsymbol{x}_{i(j+1)}) - \mathrm{E}(Y|\boldsymbol{x}_{i(j)}).$$

Often it is very convenient to write the general linear model (12.1.1) in matrix form. Let the vectors $\boldsymbol{y} = (y_1, \ldots, y_n)^t, \boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)^t$ and the matrix $X =$

$(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^t$, then (12.1.1) is equivalent to

$$\boxed{\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}} \tag{12.1.6}$$

whereas (12.1.2), (12.1.3) and (12.1.4) correspond with

$$\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \tag{12.1.7}$$

$$\Sigma(\boldsymbol{\varepsilon}) = \sigma^2 I_n. \tag{12.1.8}$$

Here, $\Sigma(\boldsymbol{\varepsilon})$ stands for the variance-covariance matrix of the errors, and $I_n$ for the $n \times n$ identity matrix.

## 12.2 Estimation of the regression parameters

### 12.2.1 The least squares estimator

Any parameter estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1})^t$ yields fitted values $\hat{y}_i$ and residuals $e_i$:

$$\begin{aligned} e_i(\hat{\boldsymbol{\beta}}) &= y_i - \hat{y}_i \\ &= y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}. \end{aligned}$$

The **least squares estimator** $\hat{\boldsymbol{\beta}}_{LS}$ is defined as the $\hat{\boldsymbol{\beta}}$ for which the sum of the squared residuals is minimal, or

$$\boxed{\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} e_i^2(\boldsymbol{\beta}).} \tag{12.2.1}$$

The residual sum of squares $\sum_{i=1}^n e_i^2(\boldsymbol{\beta})$ is called the objective function. Differentiating this objective function with respect to each $\beta_j \quad (j = 0, \ldots, p-1)$ and setting the derivatives equal to zero, yields the normal equations

$$X^t X\boldsymbol{\beta} = X^t\boldsymbol{y}.$$

If $\mathrm{rank}(X) = p \leqslant n$, the solution of this linear system is given by:

$$\boxed{\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (X^tX)^{-1}X^t\boldsymbol{y}.} \tag{12.2.2}$$

Note that $X^tX$ is the matrix of cross-products:

$$(X^tX)_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik} \tag{12.2.3}$$

$$(X^tX)_{jj} = \sum_{i=1}^{n} x_{ij}^2 \tag{12.2.4}$$

The condition $\operatorname{rank}(X) = p \leqslant n$ is necessary to ensure that $X^tX$ is nonsingular. Indeed, assume that $X^tX$ is singular. Then $\exists\ \boldsymbol{a} \in \mathbb{R}^p, \boldsymbol{a} \neq \mathbf{0}$ such that $X^tX\boldsymbol{a} = \mathbf{0}$ $(\in \mathbb{R}^p)$. Consequently, $0 = \boldsymbol{a}^tX^tX\boldsymbol{a} = \|X\boldsymbol{a}\|^2$, and thus $X\boldsymbol{a} = \mathbf{0}$ $(\in \mathbb{R}^n)$. This implies that there exists a linear relation between the columns of $X$, or $\operatorname{rank}(X) < p$.

Figure 13.3 shows the LS objective function $\sum_1^n e_i^2(\boldsymbol{\beta})$ for varying values of $\boldsymbol{\beta}$ (here, for two regressors). If the rank of $X$ is exactly $p$, as in Figure 13.3(a), we see that this objective function is convex and hence yields a unique minimum which can be derived analytically. If $\operatorname{rank}(X) < p$ as in Figure 13.3(b), there are an infinite number of LS solutions. In practice, such a perfect linear relationship between the $X$-variables is not often encountered, but the $X$-variables might be strongly correlated. This situation is known as *multicollinearity*. In such a case, the LS fit is uniquely defined, but many other parameter estimates $\hat{\boldsymbol{\beta}}$ attain a residual sum of squares which is close to the minimal value of $\hat{\boldsymbol{\beta}}_{\text{LS}}$ (see Figure 13.3(c)). Consequently, small changes in the data set may cause a large change in the parameter estimates. Figure 13.2 illustrates these effects in the data space.

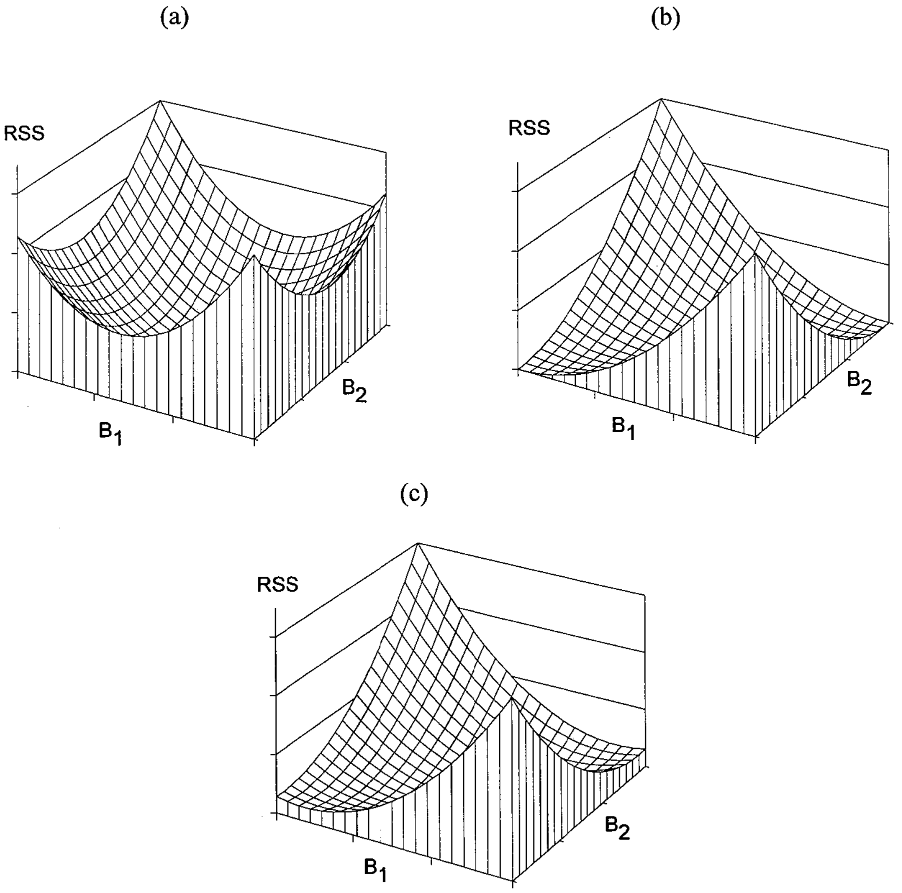


**Figure 13.3.** The residual sum of squares as a function of the slope coefficients $B_1$ and $B_2$. In each graph, the vertical axis is scaled so that the least-squares value of RSS is at the bottom of the axis. When, as in (*a*), the correlation between the independent variables $X_1$ and $X_2$ is small, the residual sum of squares has a well-defined minimum, much like a deep bowl. When there is a perfect linear relationship between $X_1$ and $X_2$, as in (*b*), the residual sum of squares is flat at its minimum, above a line in the $B_1$, $B_2$ plane: The least-squares values of $B_1$ and $B_2$ are not unique. When, as in (*c*), there is a strong, but less-than-perfect, linear relationship between $X_1$ and $X_2$, the residual sum of squares is nearly flat at its minimum, so values of $B_1$ and $B_2$ quite different from the least-squares values are associated with residual sums of squares near the minimum.

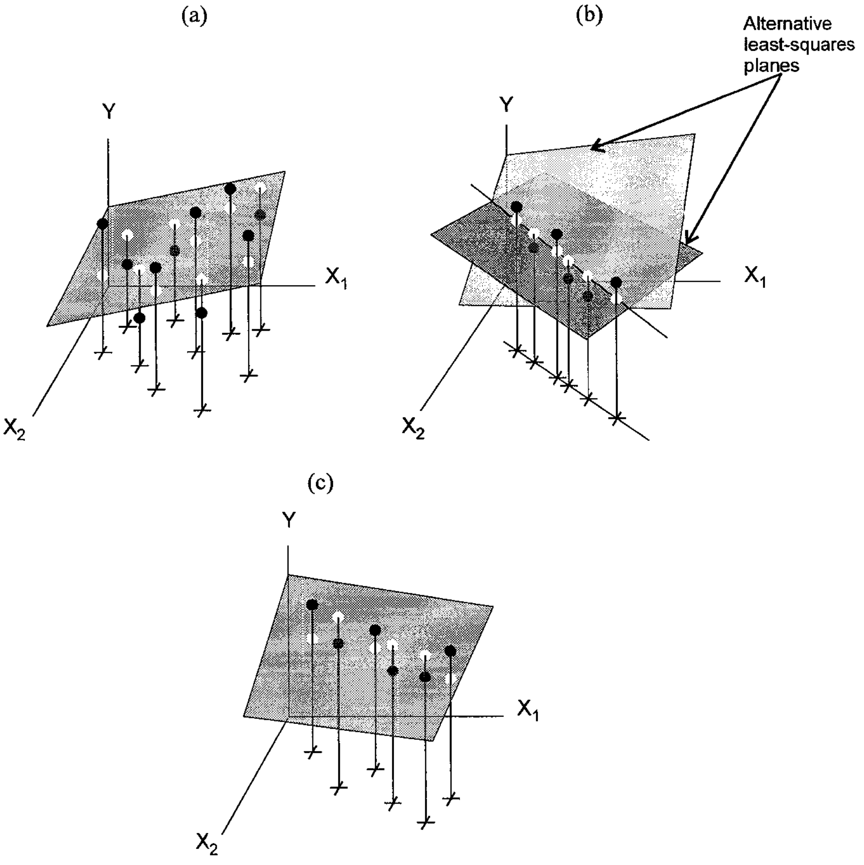


**Figure 13.2.** The impact of collinearity on the stability of the least-squares regression plane. In (*a*), the correlation between $X_1$ and $X_2$ is small, and the regression plane therefore has a broad base of support. In (*b*), $X_1$ and $X_2$ are perfectly correlated; the least-squares plane is not uniquely defined. In (*c*), there is a strong, but less-than-perfect, linear relationship between $X_1$ and $X_2$; the least-squares plane is uniquely defined, but it is not well supported by the data.

### 12.2.2 Properties and geometrical interpretation

If no confusion is possible, we simply denote $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ as $\hat{\boldsymbol{\beta}}$. Let

$$\boxed{H = X(X^tX)^{-1}X^t} \tag{12.2.5}$$

and

$$M = I_n - H$$

then the following relations hold for $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$ and $\boldsymbol{e} = (e_1, \ldots, e_n)^t$:

$$\hat{\mathbf{y}} = H\boldsymbol{y} \tag{12.2.6}$$

$$\boldsymbol{e} = M\boldsymbol{y} \tag{12.2.7}$$

$$\boldsymbol{e} = M\boldsymbol{\varepsilon} \tag{12.2.8}$$

$$\Sigma(\boldsymbol{e}) = \sigma^2(I_n - H) = \sigma^2 M. \tag{12.2.9}$$

Equations (12.2.6) and (12.2.7) are trivial and explain why the matrix $H$ is called the **hat matrix**. This hat matrix $H$ is symmetric $H^t = H$ and idempotent: $H^2 = HH = H$. From (12.2.6) we also derive that $\Sigma(\hat{\mathbf{y}}) = H\Sigma(\boldsymbol{y})H^t = \sigma^2 H$, hence the diagonal elements of the hat matrix $h_{ii}$ are always positive. Other properties of the hat matrix will be derived in Section 19.2.3.

Equation (12.2.8) follows from $\boldsymbol{e} = M\boldsymbol{y} = M(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = M\boldsymbol{\varepsilon}$ because $MX = X - HX = X - X(X^tX)^{-1}X^tX = 0_{n,p}$. Finally, $\Sigma(\boldsymbol{e}) = M\Sigma(\boldsymbol{\varepsilon})M^t = \sigma^2 MM^t = \sigma^2 M$ because also $M$ is symmetric and idempotent.

Moreover, the least squares residuals satisfy:

$$\sum_{i=1}^{n} e_i = 0 \tag{12.2.10}$$

$$\sum_{i=1}^{n} x_{ij} e_i = 0 \text{ for all } j = 1, \ldots, p-1 \tag{12.2.11}$$

$$\sum_{i=1}^{n} e_i \hat{y}_i = 0 \tag{12.2.12}$$

The first two equations (12.2.10) and (12.2.11) follow from $X^t\boldsymbol{e} = X^t M\boldsymbol{\varepsilon} = \mathbf{0}_{p,n}\boldsymbol{\varepsilon} = \mathbf{0}_p$. Moreover, $\hat{\boldsymbol{y}}^t\boldsymbol{e} = \hat{\boldsymbol{\beta}}^t X^t\boldsymbol{e} = 0$. These equations thus imply that the mean of the least squares residuals is zero, and that the residuals are orthogonal to the design matrix $X$ as well as to the predicted values.

From (12.2.10) we can also deduce that the LS hyperplane passes through the mean of the data points. Indeed, as $\frac{1}{n}\sum_i (y_i - \hat{y}_i) = 0$ we have that

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_i \hat{y}_i \\ &= \frac{1}{n}\sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{p-1} x_{i,p-1}) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_{p-1}\bar{x}_{p-1}. \end{aligned} \tag{12.2.13}$$

As a result, the intercept of the LS fit will be zero if we first mean-center the data, by setting $y_i^c = y_i - \bar{y}$ and $x_{ij}^c = x_{ij} - \bar{x}_j$ for each $i = 1, \ldots, n$ and $j = 1 \ldots, p-1$. From (12.2.13) we see indeed that the intercept $\hat{\beta}_0^c$ of the LS fit through the transformed data equals

$$\hat{\beta}_0^c = \bar{y}^c - \hat{\beta}_1^c \bar{x}_1^c - \ldots - \hat{\beta}_{p-1}^c \bar{x}_{p-1}^c = 0.$$

Relations (12.2.11) and (12.2.12) can also be derived from the **geometrical interpretation** of the least squares estimator. If we consider the observed $y$-vector and the column vectors of $X$ as points in $\mathbb{R}^n$, the least squares estimate $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is defined as the vector $\boldsymbol{\beta}$ which minimizes the Euclidean norm $\|\boldsymbol{y} - X\boldsymbol{\beta}\|$. This is because for any vector $\boldsymbol{z} \in \mathbb{R}^n$, it holds that

$$\|\boldsymbol{z}\| = \Big(\sum_{i=1}^{n} z_i^2\Big)^{1/2} = \sqrt{\boldsymbol{z}^t\boldsymbol{z}}.$$

Therefore, $\hat{\mathbf{y}}$ is the orthogonal projection of $\boldsymbol{y}$ on the linear subspace of $\mathbb{R}^n$ spanned by the columns of $X$. Consequently, $\boldsymbol{e} = \boldsymbol{y} - \hat{\mathbf{y}}$ is orthogonal to both $\hat{\mathbf{y}}$ and each column of $X$ (see Figure 10.6).

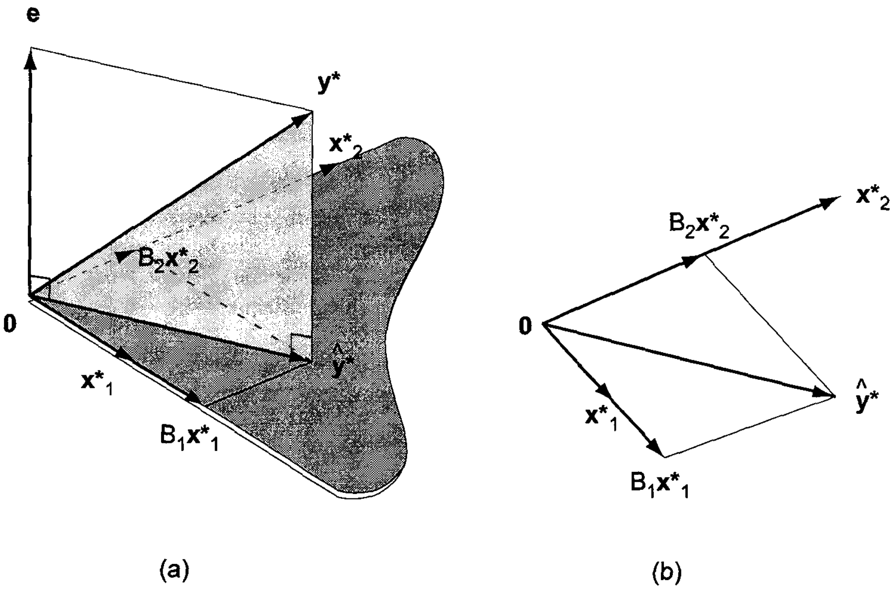


**Figure 10.6.** The vector geometry of least-squares fit in multiple regression, with the variables in mean-deviation form. The vectors y*, $\mathbf{x}_1^*$, and $\mathbf{x}_2^*$ span a three-dimensional subspace, shown in (*a*). The fitted *Y* vector, ŷ*, is the orthogonal projection of y* onto the plane spanned by $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$. The $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane is shown in (*b*).

The variance of the errors $\sigma^2$ can be estimated by the mean squared error (MSE):

$$\boxed{\hat{\sigma}^2 = s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2.}$$

Following (12.2.9) the variance-covariance matrix of the residuals is then estimated by

$$\hat{\Sigma}(\boldsymbol{e}) = s^2(I_n - H) = \text{MSE}(I_n - H). \qquad (12.2.14)$$

### 12.2.3 Statistical properties of the LS estimator

Under the Gauss-Markov conditions (12.1.2), (12.1.3) and (12.1.4), the following properties hold:

**Property 1.** *The least squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ is an unbiased and consistent estimator of $\boldsymbol{\beta}$ (we only need* (12.1.2) *to obtain the unbiasedness).*

The unbiasedness follows directly from

$$E[\hat{\boldsymbol{\beta}}_{\mathrm{LS}}] = (X^tX)^{-1}X^tE[\boldsymbol{y}] = (X^tX)^{-1}X^tX\boldsymbol{\beta} = \boldsymbol{\beta}.$$

**Property 2.** *The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{LS}$ is given by:*

$$\Sigma(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2(X^tX)^{-1}. \tag{12.2.15}$$

This follows from

$$\Sigma(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = (X^tX)^{-1}X^t\Sigma(\boldsymbol{y})X(X^tX)^{-1} = \sigma^2(X^tX)^{-1}.$$

**Property 3** (Gauss-Markov theorem)**.** *$\hat{\boldsymbol{\beta}}_{LS}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, i.e. any other linear and unbiased estimator of the form $A\boldsymbol{y}$ has a larger variance than $\hat{\boldsymbol{\beta}}_{LS}$.*

**Property 4.** *The MSE $s^2$ is an unbiased and consistent estimator of $\sigma^2$.*

To show the unbiasedness of $s^2$ we compute

$$\begin{aligned} E[\sum e_i^2] &= E[\boldsymbol{e}^t\boldsymbol{e}] = E[\boldsymbol{\varepsilon}^tM^tM\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon}^tM\boldsymbol{\varepsilon}] \\ &= E[\mathrm{trace}(\boldsymbol{\varepsilon}^tM\boldsymbol{\varepsilon})] = E[\mathrm{trace}(M\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)] \\ &= \mathrm{trace}(E[M\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]) = \mathrm{trace}(ME[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]) = \sigma^2\mathrm{trace}(M) \end{aligned}$$

Further it holds that

$$\begin{aligned} \mathrm{trace}(M) &= \mathrm{trace}(I_n - H) = n - \mathrm{trace}(H) \\ &= n - \mathrm{trace}(X(X^tX)^{-1}X^t) = n - \mathrm{trace}((X^tX)^{-1}X^tX) \\ &= n - \mathrm{trace}(I_p) = n - p \end{aligned}$$

**Property 5.** *$s^2(X^tX)^{-1}$ is an unbiased and consistent estimator of $\sigma^2(X^tX)^{-1}$.*

We thus estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ as:

$$\boxed{\hat{\Sigma}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = s^2(X^tX)^{-1}.} \tag{12.2.16}$$

**Property 6.** *If the errors $\boldsymbol{\varepsilon}$ are normally distributed, $\hat{\boldsymbol{\beta}}_{LS}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. The maximum likelihood estimator of $\sigma^2$ is given by*

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{12.2.17}$$

If $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ the likelihood function can be expressed as

$$L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) = \prod_{i=1}^{n} \phi\Big(\frac{y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}}{\sigma}\Big)$$

with $\phi$ the density of the standard normal distribution. Hence

$$L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2\Big).$$

The log-likelihood function $l = \log(L)$ is then equal to

$$l(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2. \tag{12.2.18}$$

Consequently, the log-likelihood function is maximized over $\boldsymbol{\beta}$ by minimizing the last term, of equivalently by taking $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{LS}}$. Finally it is easy to show that $l(\hat{\boldsymbol{\beta}}_{\text{LS}}, \sigma^2 | \boldsymbol{y})$ is maximized (over all $\sigma^2$) by the expression (12.2.17).

### 12.2.4 An example in R

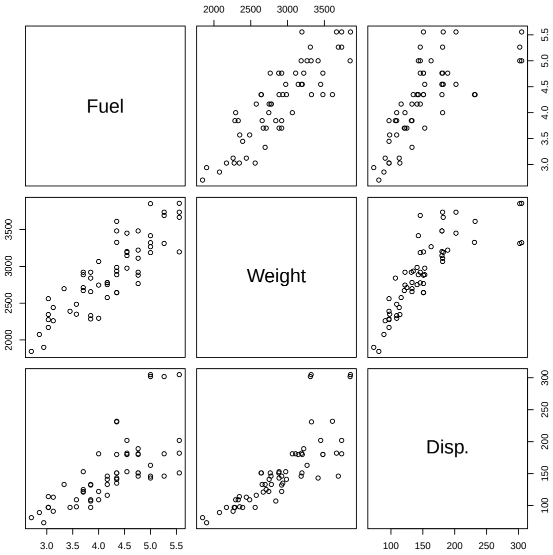
The data frame `fuel.frame` (from the '`SemiPar`' library) contains information of 60 cars. This data set contains 5 variables: `Weight` (the weight of the car in pounds), `Disp.` (the engine displacement in liters), `Mileage` (gas mileage in miles/gallon), `Fuel` (fuel consumption in gallons per 100 miles, it thus is equal to 100/Mileage), and `Type` (a factor giving the general type of car, with levels: Small, Sporty, Compact, Medium, Large, Van).

We want to predict the fuel consumption of a car by its weight and engine displacement. The postulated model is:

$$\text{Fuel}_i = \beta_0 + \beta_1 \ \text{Weight}_i + \beta_2 \ \text{Disp}_i + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma^2)$.

```
data(fuel.frame)
attach(fuel.frame)
## help(fuel.frame)
names(fuel.frame)
pairs(~Fuel+Weight+Disp.)
```



The pairwise plots suggest that both `Weight` and `Disp.` are linearly related to `Fuel`. The analysis yields:

```
Fuelfit <- lm(Fuel~Weight+Disp.)
Fuelsum <- summary(Fuelfit)
Fuelsum

Call:
lm(formula = Fuel ~ Weight + Disp.)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81089 -0.25586  0.01971  0.26734  0.98124 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.4789731  0.3417877   1.401    0.167    
Weight      0.0012414  0.0001720   7.220 1.37e-09 ***
Disp.       0.0008544  0.0015743   0.543    0.589    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3901 on 57 degrees of freedom
Multiple R-squared:  0.7438,Adjusted R-squared:  0.7348 
F-statistic: 82.75 on 2 and 57 DF,  p-value: < 2.2e-16
```

The fitted model is thus:

$$\hat{\text{Fuel}}_i = 0.48 + 0.0012 \text{ Weight}_i + 0.00085 \text{ Disp}_i$$

with $\hat{\sigma} = 0.39$.

## 12.3 Analysis of variance

### 12.3.1 The decomposition of the total sum of squares

For an individual observation we have the identity

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

which is illustrated for simple regression in Figure 5.4.

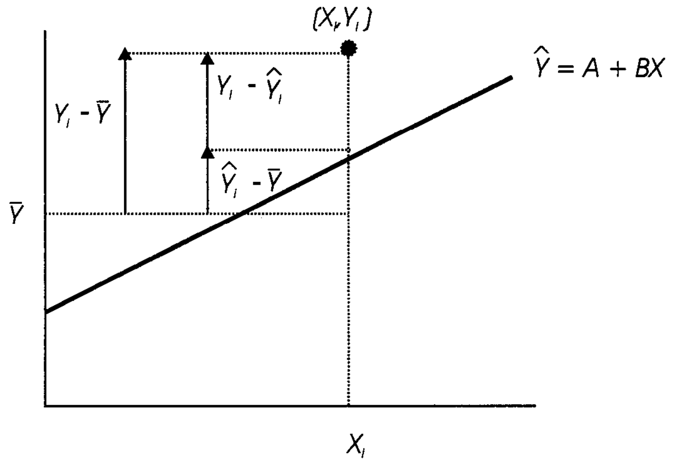


**Figure 5.4.** Decomposition of the total deviation $Y_i - \overline{Y}$ into components $Y_i - \hat{Y}_i$ and $\hat{Y}_i - \overline{Y}$.

Squaring both sides of the equation and summing over all observations gives

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{12.3.1}$$

The cross-product term vanishes because of the Pythagorean theorem (see Figure 10.6). It can also be deduced from (12.2.10) and (12.2.12) by noting that

$$\sum_i(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_i(\hat{y}_i - \bar{y})e_i = \sum_i \hat{y}_i e_i - \bar{y}\sum_i e_i = 0.$$

Relation (12.3.1) is the ANOVA decomposition which says that the total variation (SST) in the response $\boldsymbol{y}$ can be decomposed into an 'explained' component due to the regression (SSR) and an 'unexplained' component due to the errors (SSE). We thus have:

$$\text{SST} = \text{SSR} + \text{SSE}$$

with degrees of freedom $n-1, p-1$ and $n-p$. The mean squares are defined as the sum of squares divided by their degrees of freedom:

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{p-1}$$
$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-p}$$

They are typically written in an ANOVA table as in Table 6.1.

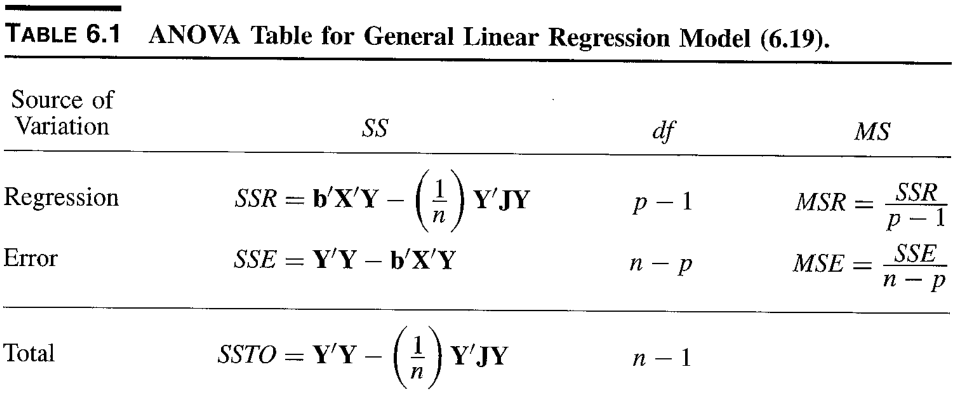


**TABLE 6.1 ANOVA Table for General Linear Regression Model (6.19).**

| Source of Variation | *SS* | *df* | *MS* |
|---|---|---|---|
| Regression | $SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$ | $p-1$ | $MSR = \frac{SSR}{p-1}$ |
| Error | $SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$ | $n-p$ | $MSE = \frac{SSE}{n-p}$ |
| Total | $SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$ | $n-1$ | |

### 12.3.2 The coefficient of multiple determination

The *coefficient of multiple determination* is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{12.3.2}$$

$$= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{12.3.3}$$

It measures the proportion of the *total variation in the response* $\boldsymbol{y}$ that is explained by the linear model (12.1.1) which includes the variables $X_1, \ldots, X_{p-1}$. By construction $0 \leqslant R^2 \leqslant 1$. The minimum value 0 is attained when all $\hat{y}_i = \bar{y}$, i.e. when all $\hat{\beta}_j = 0$ for $j = 1, \ldots, p-1$. The maximum value 1 is attained when all the observations fall exactly on the fitted regression surface, i.e. when $y_i = \hat{y}_i$ for all cases $i$.

Remarks:

1. In simple regression $R^2$ coincides with the squared correlation coefficient $r^2$ between $X = X_1$ and $Y$.

2. A high value of $R^2$ does not necessarily imply that the fitted model is useful to make predictions.

3. One can always increase $R^2$ by adding variables to the model. Therefore the *adjusted coefficient of determination* $R_a^2$ corrects for the number of variables:
$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}. \tag{12.3.4}$$
However, $R_a^2$ can take negative values.

4. If the general model (12.1.1) does not contain an intercept term, that is, when $\beta_0 = 0$, the ANOVA decomposition becomes:
$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$
The $R^2$ coefficient is then defined by:
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n y_i^2}. \tag{12.3.5}$$
As the denominators of (12.3.3) and (12.3.5) are different, it is dangerous to compare the $R^2$-value of a model with intercept to the $R^2$-value of the same model without intercept.

5. The positive square root of $R^2$, i.e. $R = \sqrt{R^2}$ is called the *coefficient of multiple correlation.* From Figure 10.8 is it clear that

$$R = \cos(\boldsymbol{y}, \hat{\boldsymbol{y}}).$$

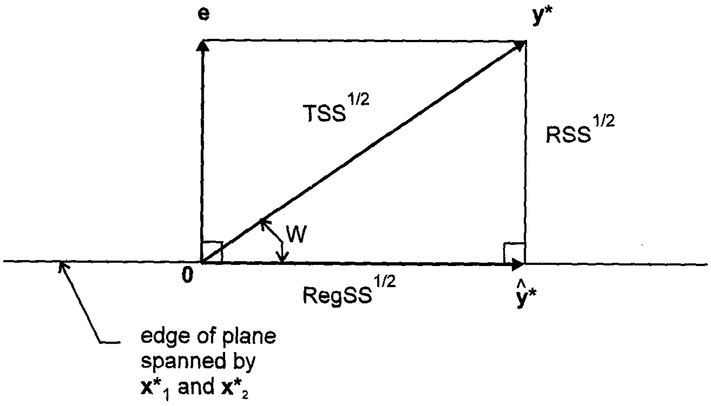


**Figure 10.8.** The analysis of variance for multiple regression appears in the plane spanned by y* and ŷ*. The multiple correlation is the cosine of the angle *W* separating y* and ŷ*.

### 12.3.3 The extra sum of squares

The extra sum of squares measures the marginal *reduction in the error sum of squares* (SSE) when one or several predictor variables are added to the regression model, given that other variables are already in the model. Equivalently, it measures the marginal *increase in the regression sum of squares* (SSR) when one or several regressors are added to the model.

Consider as an example a regression model with three predictors: $X_1, X_2$ and $X_3$. If we only include $X_1$ in the model, we know that

$$\text{SST} = \text{SSR}(X_1) + \text{SSE}(X_1). \tag{12.3.6}$$

Now, add variable $X_2$, then

$$\text{SST} = \text{SSR}(X_1, X_2) + \text{SSE}(X_1, X_2) \tag{12.3.7}$$

because the total sum of squares did not change. Note that

$$\text{SSE}(X_1, X_2) \leqslant \text{SSE}(X_1).$$

We now define the *extra sum of squares* as:

$$\boxed{\text{SSR}(X_2|X_1) = \text{SSE}(X_1) - \text{SSE}(X_1, X_2)} \tag{12.3.8}$$

or equivalently

$$\boxed{\text{SSR}(X_2|X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1).} \tag{12.3.9}$$

Analogously, if we include $X_3$ as well:

$$\begin{aligned}\text{SSR}(X_3|X_1, X_2) &= \text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)\\ &= \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1, X_2)\end{aligned}$$

These definitions and formulas also hold when we include several regressors at once. For example,

$$\begin{aligned}\text{SSR}(X_2, X_3|X_1) &= \text{SSE}(X_1) - \text{SSE}(X_1, X_2, X_3)\\ &= \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1).\end{aligned}$$

When we combine (12.3.6) and (12.3.8) we see that the total sum of squares (SST) can be written as

$$\text{SST} = \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSE}(X_1, X_2)$$

or as

$$\text{SST} = \text{SSR}(X_2) + \text{SSR}(X_1|X_2) + \text{SSE}(X_1, X_2)$$

if we include $X_2$ first in the regression model.

Equation (12.3.9) is equivalent to:

$$\boxed{\text{SSR}(X_1, X_2) = \text{SSR}(X_1) + \text{SSR}(X_2|X_1).} \tag{12.3.10}$$

We can thus decompose the SSR of the full model (here, with all 3 predictors) into several extra sum of squares, as in Table 7.3. Note that the degrees of freedom associated with each sum of squares is equal to the number of variables that are added to the model.

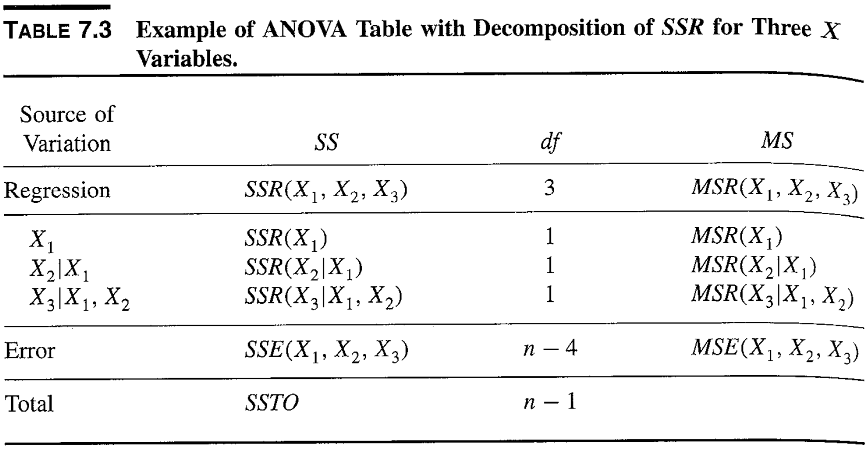


**TABLE 7.3 Example of ANOVA Table with Decomposition of *SSR* for Three $X$ Variables.**

| Source of Variation | *SS* | *df* | *MS* |
|---|---|---|---|
| Regression | $SSR(X_1, X_2, X_3)$ | 3 | $MSR(X_1, X_2, X_3)$ |
| $X_1$ | $SSR(X_1)$ | 1 | $MSR(X_1)$ |
| $X_2 \vert X_1$ | $SSR(X_2 \vert X_1)$ | 1 | $MSR(X_2 \vert X_1)$ |
| $X_3 \vert X_1, X_2$ | $SSR(X_3 \vert X_1, X_2)$ | 1 | $MSR(X_3 \vert X_1, X_2)$ |
| Error | $SSE(X_1, X_2, X_3)$ | $n-4$ | $MSE(X_1, X_2, X_3)$ |
| Total | $SSTO$ | $n-1$ | |

The ANOVA analysis of the `fuel.frame` data set yields:

```
summary(aov(Fuelfit))

            Df Sum Sq Mean Sq F value Pr(>F)
Weight       1 25.139  25.139 165.209 <2e-16 ***
Disp.        1  0.045   0.045   0.295  0.589
Residuals   57  8.673   0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Note the difference
Fuelfit2<- lm(Fuel~Disp.+Weight)
summary(aov(Fuelfit2))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
Disp.        1 17.253  17.253  113.38 3.58e-15 ***
Weight       1  7.931   7.931   52.12 1.37e-09 ***
Residuals   57  8.673   0.152                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 12.4 Equivariance properties

The least squares estimator satisfies several equivariance properties. If the model

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \epsilon_i \tag{12.4.1}$$

holds, then also

$$\begin{aligned} y_i + \boldsymbol{x}_i^t \boldsymbol{v} &= \boldsymbol{x}_i^t \boldsymbol{\beta} + \boldsymbol{x}_i^t \boldsymbol{v} + \epsilon_i \\ &= \boldsymbol{x}_i^t (\boldsymbol{\beta} + \boldsymbol{v}) + \epsilon_i \end{aligned}$$

for any vector $\boldsymbol{v}$. Hence, if a regression estimator applied to the $(\boldsymbol{x}_i, y_i)$ yields $\hat{\boldsymbol{\beta}}$, then it is desirable that the estimator applied to the $(\boldsymbol{x}_i, y_i + \boldsymbol{x}_i^t \boldsymbol{v})$ yields $\hat{\boldsymbol{\beta}} + \boldsymbol{v}$.

**Property 7.** $\hat{\boldsymbol{\beta}}_{LS}$ *is regression equivariant:*

$$\hat{\boldsymbol{\beta}}_{LS}(\boldsymbol{x}_i, y_i + \boldsymbol{x}_i^t \boldsymbol{v}) = \hat{\boldsymbol{\beta}}_{LS}(\boldsymbol{x}_i, \boldsymbol{y}_i) + \boldsymbol{v}$$

*for any vector* $\boldsymbol{v}$*.*

This follows from

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\boldsymbol{x}_i, y_i + \boldsymbol{x}_i^t \boldsymbol{v}) &= \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i + \boldsymbol{x}_i^t \boldsymbol{v} - \boldsymbol{x}_i^t \boldsymbol{\beta})^2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^t (\boldsymbol{\beta} - \boldsymbol{v}))^2 \end{aligned}$$

Thus $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\boldsymbol{x}_i, y_i + \boldsymbol{x}_i^t \boldsymbol{v}) - \boldsymbol{v} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\boldsymbol{x}_i, y_i)$.

Model (12.4.1) also yields the equalities

$$c y_i = \boldsymbol{x}_i^t c \boldsymbol{\beta} + c \epsilon_i$$

for any constant $c$ and

$$y_i = (A \boldsymbol{x}_i)^t (A^t)^{-1} \boldsymbol{\beta} + \epsilon_i$$

for any non-singular $p \times p$ matrix.

**Property 8.** $\hat{\boldsymbol{\beta}}_{LS}$ *is scale equivariant:*

$$\hat{\boldsymbol{\beta}}_{LS}(\boldsymbol{x}_i, c y_i) = c \hat{\boldsymbol{\beta}}_{LS}(\boldsymbol{x}_i, y_i)$$

*for any constant* $c$ *and*

$$\hat{\sigma}^2_{LS}(\boldsymbol{x}_i, c y_i) = c^2 \hat{\sigma}^2_{LS}(\boldsymbol{x}_i, y_i).$$

**Property 9.** $\hat{\boldsymbol{\beta}}_{LS}$ *is affine equivariant:*

$$\hat{\boldsymbol{\beta}}_{LS}(A\boldsymbol{x}_i, y_i) = (A^t)^{-1}\hat{\boldsymbol{\beta}}_{LS}(\boldsymbol{x}_i, y_i)$$

*for any non-singular* $p \times p$ *matrix.*

This implies that the fit is essentially independent of the choice of measurement unit for the response variable $y$. Also, if we apply e.g. a logarithmic transformation to the $y$, it does not really make a difference whether we use the natural logarithm $\log(y) = \ln(y)$ or $\log_{10}(y)$ as they only differ up to a constant factor.

The affine equivariance allows linear transformations of the regressors, including changes in the measurement units.

## 12.5 The standardized regression model

The computation of the least squares parameters involves the inverse of $X^tX$. This matrix operation will be sensitive to roundoff errors, when

1. the determinant of $X^tX$ is close to zero (multicollinearity!)

2. the elements of $X^tX$ differ significantly in order of magnitude, which occurs when the predictor variables have substantially different magnitudes.

The standardized regression model is obtained by transforming the $X$ (and $Y$) variables such that the new $X^tX$ matrix corresponds with the correlation matrix of the original $X$-variables. Consequently its entries (in absolute value) are bounded by 1 and thus are less sensitive to roundoff errors.

This transformation is also used when we want to compare the regression coefficients in common units. Consider e.g. the estimated regression plane:

$$\hat{Y} = 200 + 20000X_1 + 0.2X_2$$

with $Y$ measured in dollars, $X_1$ in thousand dollars and $X_2$ in cents. Then a 1-unit increase of $X_1$ (i.e. a \$1000 dollar increase) when $X_2$ is constant corresponds with $\hat{\beta}_1 = \$20000$. The \$1000 increase is equal to a 100000-unit increase of $X_2$ with $X_1$ constant, which equals $100000\hat{\beta}_2 = \$20000$. Both regressors thus have the same effect on $Y$ although the regression coefficients suggest the inverse.

The *correlation transformation* is defined for each observation $i = 1, \ldots, n$ and for each variable $j = 1, \ldots, p-1$ as:

$$x'_{ij} = \frac{1}{\sqrt{n-1}} \left( \frac{x_{ij} - \bar{x}_j}{s_j} \right) \tag{12.5.1}$$

$$y'_i = \frac{1}{\sqrt{n-1}} \left( \frac{y_i - \bar{y}}{s_Y} \right) \tag{12.5.2}$$

with $s_j$ resp. $s_Y$ the standard deviation of $X_j$ resp. $Y$. Using (12.2.3) and (12.2.4) we then obtain for the transformed variables:

$$\begin{aligned}((X')^tX')_{jk} &= \sum_{i=1}^{n} x'_{ij}x'_{ik}\\ &= \frac{1}{n-1}\frac{\sum(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{s_js_k}\\ &= \frac{\text{cov}(X_j,X_k)}{s_js_k} = r_{jk}\\ ((X')^tX')_{jj} &= \frac{s_j^2}{s_js_j} = 1\\ ((X')^ty')_j &= \frac{\text{cov}(X_j,Y)}{s_js_Y} = r_{jy}\end{aligned}$$

with $r_{jk}$ the simple correlation between $X_j$ and $X_k$, and $r_{jy}$ the correlation between $X_j$ and $Y$.

In terms of the transformed variables, the general linear model (12.1.1)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1}x_{i,p-1} + \epsilon_i$$

now becomes

$$y_i - \bar{y} = \beta_1(x_{i1}-\bar{x}_1) + \ldots + \beta_{p-1}(x_{i,p-1}-\bar{x}_{p-1}) + \epsilon_i$$

and thus

$$\frac{y_i-\bar{y}}{s_Y} = \beta_1\frac{s_1}{s_Y}\left(\frac{x_{i1}-\bar{x}_1}{s_1}\right) + \ldots + \beta_{p-1}\frac{s_{p-1}}{s_Y}\left(\frac{x_{i,p-1}-\bar{x}_{p-1}}{s_{p-1}}\right) + \frac{\epsilon_i}{s_Y}.$$

We could drop the intercept term from the model because the observations are mean-centered! If finally we divide each term by $\sqrt{n-1}$, we obtain the *standardized regression model*

$$y'_i = \beta'_1x'_{i1} + \beta'_2x'_{i2} + \ldots + \beta'_{p-1}x'_{i,p-1} + \epsilon'_i \qquad (12.5.3)$$

for $i = 1,\ldots,n$ with

$$\epsilon'_i = \frac{\epsilon_i}{\sqrt{n-1}s_Y} \qquad (12.5.4)$$

$$\beta'_j = (\frac{s_j}{s_Y})\beta_j. \qquad (12.5.5)$$

The regression coefficients $\beta'_j$ are often called the *standardized regression coefficients.* Because of the correlation transformation their least squares estimates satisfy:

$$\hat{\boldsymbol{\beta}}' = R_{XX}^{-1}r_{XY}. \qquad (12.5.6)$$

Here, $R_{XX}$ is the correlation matrix of $X$, and $r_{XY} = (r_{1y}, \ldots, r_{p-1,y})^t$ contains the correlations between each predictor variable and the response variable.

To return to the estimates with respect to the original variables we use (12.5.5), (12.2.13) and the equivariance properties of the least squares estimator:

$$\hat{\beta}_j = (\frac{s_Y}{s_j})\hat{\beta}'_j \tag{12.5.7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \ldots - \hat{\beta}_{p-1}\bar{x}_{p-1}. \tag{12.5.8}$$

**Example.**

Dwaine Studios Inc. operates portrait studios in 21 cities of medium size. They are specialized in portraits of children. The company wants to investigate whether sales in a community ($Y$, expressed in \$1000) can be predicted from the number of persons aged 16 or younger in that community ($X_1$ in thousands of persons) and the per capita personal income ($X_2$ in \$1000). Some data and results are shown in Table 7.5. The standardized regression model yields:

$$\hat{y}'_i = 0.75x'_{i1} + 0.25x'_{i2}$$

whereas the model in the original variables yields estimated values:

$$\hat{y}_i = -68.86 + 1.45x_{i1} + 9.36x_{i2}.$$

In the latter model, $\hat{\beta}_1$ and $\hat{\beta}_2$ can not be compared directly because both variables $X_1$ and $X_2$ are measured in other units. The standardized coefficients tell us that an increase of one standard deviation of $X_1$ when $X_2$ is fixed leads to a much larger increase in expected sales than if we fix $X_1$ and increase $X_2$ by one standard deviation. We should however be cautious about this interpretation as also the correlation between $X_1$ and $X_2$ has an effect on the regression coefficients.

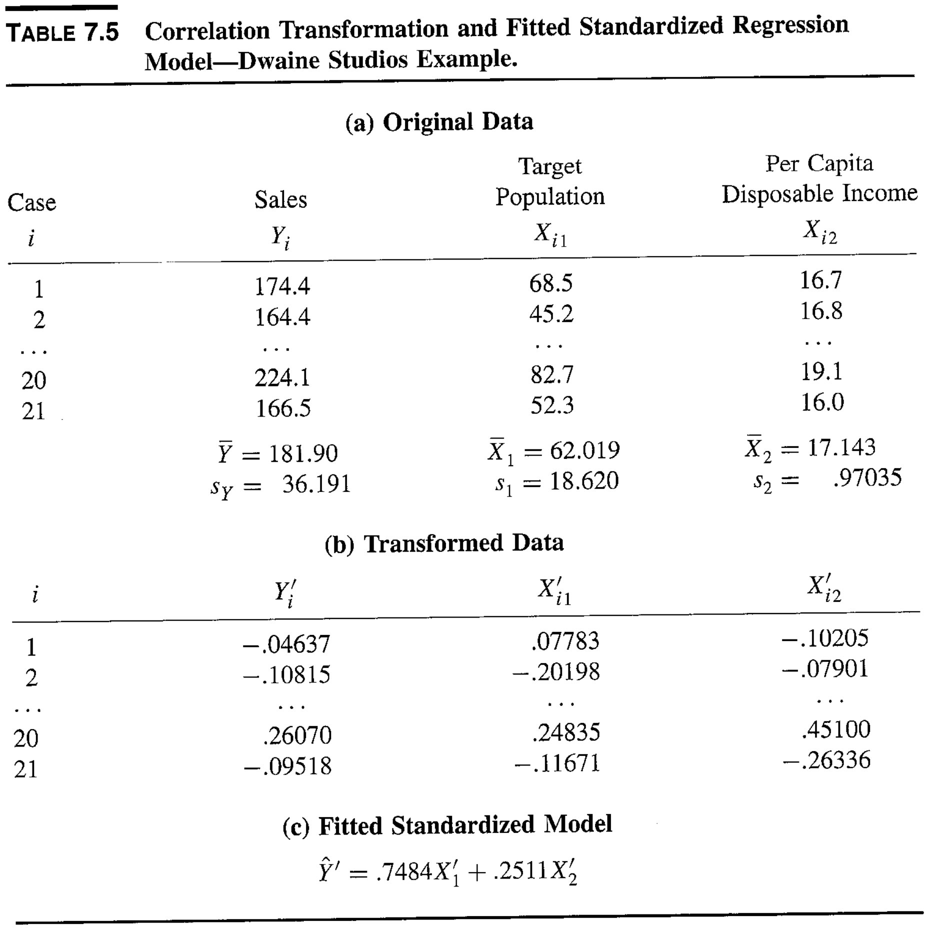


**TABLE 7.5 Correlation Transformation and Fitted Standardized Regression Model—Dwaine Studios Example.**

**(a) Original Data**

| Case<br>$i$ | Sales<br>$Y_i$ | Target<br>Population<br>$X_{i1}$ | Per Capita<br>Disposable Income<br>$X_{i2}$ |
|---|---|---|---|
| 1 | 174.4 | 68.5 | 16.7 |
| 2 | 164.4 | 45.2 | 16.8 |
| ... | ... | ... | ... |
| 20 | 224.1 | 82.7 | 19.1 |
| 21 | 166.5 | 52.3 | 16.0 |
|  | $\bar{Y} = 181.90$ | $\bar{X}_1 = 62.019$ | $\bar{X}_2 = 17.143$ |
|  | $s_Y = 36.191$ | $s_1 = 18.620$ | $s_2 = .97035$ |

**(b) Transformed Data**

| $i$ | $Y_i'$ | $X_{i1}'$ | $X_{i2}'$ |
|---|---|---|---|
| 1 | −.04637 | .07783 | −.10205 |
| 2 | −.10815 | −.20198 | −.07901 |
| ... | ... | ... | ... |
| 20 | .26070 | .24835 | .45100 |
| 21 | −.09518 | −.11671 | −.26336 |

**(c) Fitted Standardized Model**

$$\hat{Y}' = .7484X_1' + .2511X_2'$$

# Chapter 13

# Statistical inference

When we want to make inferences about $\boldsymbol{\beta}$ we assume that the errors are independent and *normally distributed*, i.e.

$$\boxed{\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n).} \tag{13.0.1}$$

Under this condition, the general linear model satisfies:

$$\boldsymbol{y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n) \tag{13.0.2}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^t X)^{-1}). \tag{13.0.3}$$

Note that (13.0.2) does not tell that the $\{y_i,\ i = 1, \ldots, n\}$follow a common univariate normal distribution. It says that at a certain $\boldsymbol{x}$, the corresponding response variable $y$ is normally distributed. In particular, $y_i \sim N(\boldsymbol{x}_i^t \boldsymbol{\beta}, \sigma^2)$ for $i = 1, \ldots, n$. This is in general difficult to check as we often only have one measurement for each $\boldsymbol{x}_i$. Normality of the residuals on the other hand can be verified using residual plots (see Section 13.7).

## 13.1 Inference for individual parameters

From (13.0.3), we obtain

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^t X)^{-1}_{jj})$$

and its estimated standard error

$$\mathrm{s}(\hat{\beta}_j) = s\sqrt{(X^t X)^{-1}_{jj}}.$$

Moreover it can be shown that

$$(n-p)\frac{s^2}{\sigma^2} \sim \chi^2_{n-p}$$

and that $\hat{\beta}_j$ and $s^2$ are independent. Consequently,

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{s}(\hat{\beta}_j)} \sim t_{n-p}$$

Under the null hypothesis

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

it then holds that

$$\boxed{t = \frac{\hat{\beta}_j}{\mathrm{s}(\hat{\beta}_j)} \sim_{H_0} t_{n-p}} \tag{13.1.1}$$

These $t$-values and their corresponding $p$-values are usually reported in the output of an analysis with a statistical software package. If the $p$-value is smaller than $\alpha$, we reject the $H_0$ hypothesis in favor of the alternative.

Equivalently, we can construct a $(1-\alpha)100\%$ confidence interval for $\beta_j$:

$$\mathrm{CI}(\beta_j, \alpha) = [\hat{\beta}_j - t_{n-p,\frac{\alpha}{2}}\mathrm{s}(\hat{\beta}_j), \hat{\beta}_j + t_{n-p,\frac{\alpha}{2}}\mathrm{s}(\hat{\beta}_j)]$$

and reject $H_0$ if 0 does not belong to $\mathrm{CI}(\beta_j, \alpha)$. Note that the quantile $t_{n-p,\alpha/2}$ satisfies

$$P(T > t_{n-p,\frac{\alpha}{2}}) = \frac{\alpha}{2} \text{ with } T \sim t_{n-p}.$$

Remember that $\alpha$ is the probability of a type I error, i.e.

$$\boxed{\alpha = P(H_0 \text{ is rejected } | H_0 \text{ is correct}).}$$

**Example: Fuel data**

Let us look again at the fuel consumption data. Based on the output of the linear model fit in Section 12.2.4 we can now examine the relevance of both regressors in the model

The hypothesis $H_0 : \beta_1(\text{Weight}) = 0$ is rejected because the corresponding $p$-value is essentially 0 (1.37e-09). On the other hand, the hypothesis $H_0 : \beta_2(\text{Disp.}) = 0$ cannot be rejected at the 5% significance level because its p-value $0.59 > 0.05$.

## 13.2 Inference for several parameters

When we want to test whether a group of parameters is significant we state the null and alternative hypothesis as:

$$\begin{aligned} &H_0 : \beta_{p-q} = \beta_{p-q+1} = \ldots = \beta_{p-1} = 0 \\ &H_1 : \text{not all } \beta_j \text{ equal zero } (j = p-q, \ldots, p-1) \end{aligned}$$

(we assume that we want to make a test on the last $q$ parameters).

Example:

Suppose we fit a regression model with three slope parameters:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

and we want to test whether

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0.$$

We could then accept $H_0$ at the $\alpha 100\%$ significance level if $0 \in \text{CI}(\beta_2, \alpha)$ and $0 \in \text{CI}(\beta_3, \alpha)$. However, the probability of a type I error then increases:

$$\begin{aligned} &P(H_0 \text{ is accepted } | H_0 \text{ is correct}) \\ &\quad = P(0 \in \text{CI}(\beta_2, \alpha) \text{ and } 0 \in \text{CI}(\beta_3, \alpha)) \\ &\quad = 1 - P(0 \notin \text{CI}(\beta_2, \alpha) \text{ or } 0 \notin \text{CI}(\beta_3, \alpha)) \\ &\quad = 1 - P(0 \notin \text{CI}(\beta_2, \alpha)) - P(0 \notin \text{CI}(\beta_3, \alpha)) + P(0 \notin \text{CI}(\beta_2, \alpha) \text{ and } 0 \notin \text{CI}(\beta_3, \alpha)) \\ &\quad \geqslant 1 - P(0 \notin \text{CI}(\beta_2, \alpha)) - P(0 \notin \text{CI}(\beta_3, \alpha)) \\ &\quad = 1 - \alpha - \alpha = 1 - 2\alpha. \end{aligned}$$

Hence,

$$P(H_0 \text{ is rejected } | H_0 \text{ is correct}) \leqslant 1 - (1 - 2\alpha) = 2\alpha.$$

If we want to be sure that

$$P(H_0 \text{ is rejected } | H_0 \text{ is correct}) \leqslant \alpha$$

we can apply the Bonferroni correction. For this, we construct *simultaneous confidence intervals* which are wider than the individual confidence intervals:

$$\text{SCI}(\beta_j, \alpha) = \text{CI}(\beta_j, \frac{\alpha}{2}).$$

In general simultaneous confidence intervals for testing $g \leqslant p$ parameters with a confidence of at least $1 - \alpha$ are given by

$$[\hat{\beta}_j - t_{n-p,\frac{\alpha}{2g}}\ \mathrm{s}(\hat{\beta}_j), \hat{\beta}_j + t_{n-p,\frac{\alpha}{2g}}\ \mathrm{s}(\hat{\beta}_j)].$$

Because the simultaneous confidence intervals are wider than the individual confidence intervals, they define a much larger region in $\mathbb{R}^g$. Therefore they yield a larger type II error, i.e. the probability that the $H_0$ hypothesis will not be rejected although the alternative is true, will be larger. Equivalently, the probability to detect that $H_1$ is correct, will be small.

$$\boxed{\beta = P(\text{type II error}) = P(H_0 \text{ is accepted}|H_1 \text{ is correct}).}$$

Therefore, the Bonferroni method is to be preferred when the number of hypotheses that we want to test is limited. Otherwise, other simultaneous confidence intervals (e.g. Scheffé) will have a coverage which is closer to $(1-\alpha)100\%$. They will give a better balance between the probabilities of type I and type II error.

Another disadvantage of this procedure is that the correlation between the parameter estimates is not taken into account. A *partial* F-test, which is based on $\hat{\Sigma}(\hat{\boldsymbol{\beta}})$ attains the correct significance level. Under $H_0$ we obtain the reduced model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-q-1} x_{i,p-q-1} + \epsilon_i.$$

Let $\mathrm{SSE}_{p-q}$ denote the error sum of squares under this reduced model, i.e. $\mathrm{SSE}_{p-q} = \mathrm{SSE}(X_1, \ldots, X_{p-q-1})$ and let $\mathrm{SSE}_p$ be the error sum of squares under the full model. Thus $\mathrm{SSE}_p = \mathrm{SSE}(X_1, \ldots, X_{p-1})$. Under condition (13.0.1) it can be shown that

$$\boxed{F = \frac{(\mathrm{SSE}_{p-q} - \mathrm{SSE}_p)/q}{\mathrm{SSE}_p/(n-p)} \sim_{H_0} F_{q,n-p}} \tag{13.2.1}$$

This test statistic can as well be described using the extra sum of squares. By (12.3.8) we have that

$$\mathrm{SSE}_{p-q} - \mathrm{SSE}_p = \mathrm{SSR}(X_{p-q}, \ldots, X_{p-1}|X_1, \ldots, X_{p-q-1}).$$

Thus, (13.2.1) becomes:

$$\boxed{F = \frac{\mathrm{MSR}(X_{p-q}, \ldots, X_{p-1}|X_1, \ldots, X_{p-q-1})}{\mathrm{MSE}(X_1, \ldots, X_{p-1})}} \tag{13.2.2}$$

Moreover,

$$\mathrm{MSR}(X_{p-q},\ldots,X_{p-1}|X_1,\ldots,X_{p-q-1}) = \frac{1}{q}\big(\mathrm{SSR}(X_{p-q}|X_1,\ldots,X_{p-q-1})+$$
$$\mathrm{SSR}(X_{p-q+1}|X_1,\ldots,X_{p-q}) + \ldots + \mathrm{SSR}(X_{p-1}|X_1,\ldots,X_{p-2})\big).$$

Therefore, this F-statistic can easily be computed from a (detailed) ANOVA table as in Table 7.3.

**Example: Body Fat Data.**

The Body Fat data study the relation of amount of body fat ($Y$) to three possible predictor variables: triceps skinfold thickness ($X_1$), thigh circumference ($X_2$) and midarm circumference ($X_3$). Measurements are taken on 20 healthy women between 25 and 34 years old. Assume we want to test:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero}$$

Using the ANOVA Table 7.4, we derive the partial F-statistic

$$F = \frac{(33.17 + 11.54)/2}{98.41/16} = 3.63.$$

As $F_{2,16,0.05} = 3.63$, the $p$-value of our test is 5% and we are at the boundary of the decision rule. At the 1% significance level e.g. we would not reject the $H_0$ hypothesis.

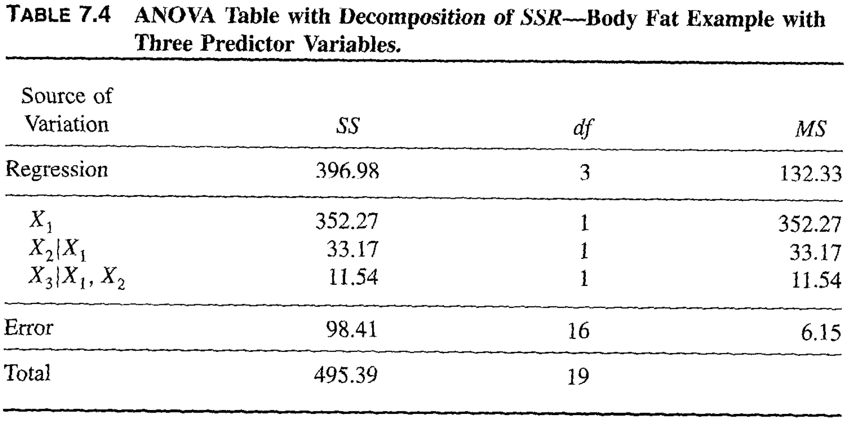


**TABLE 7.4** **ANOVA Table with *Decomposition of SSR*—Body Fat Example with Three Predictor Variables.**

| Source of Variation | *SS* | *df* | *MS* |
|---|---|---|---|
| Regression | 396.98 | 3 | 132.33 |
| $X_1$ | 352.27 | 1 | 352.27 |
| $X_2 \| X_1$ | 33.17 | 1 | 33.17 |
| $X_3 \| X_1, X_2$ | 11.54 | 1 | 11.54 |
| Error | 98.41 | 16 | 6.15 |
| Total | 495.39 | 19 | |

How is this F-statistic related to $\hat{\Sigma}(\hat{\boldsymbol{\beta}}) = s^2(X^tX)^{-1}$? Let $\mathbf{b}_q = (\hat{\beta}_{p-q}, \ldots, \hat{\beta}_{p-1})^t$ be the vector with the last $q$ components of the LS fit $\hat{\boldsymbol{\beta}}$ on the full model, and let $V_{qq}$ represent the square submatrix consisting of the last $q$ rows and columns of $(X^tX)^{-1}$. Then it can be shown that

$$\mathrm{SSE}_{p-q} - \mathrm{SSE}_p = \mathbf{b}_q^t V_{qq}^{-1} \mathbf{b}_q.$$

## 13.3 The overall F-test

The *overall* F-test is used to test whether there is a regression relation between the response variable $Y$ and the set of $X$-variables $X_1, \ldots, X_{p-1}$:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$
$$H_1 : \text{not all } \beta_j \text{ equal zero}$$

The test statistic is derived from the partial F-test (13.2.2) with $q = p - 1$:

$$\boxed{F = \frac{\mathrm{MSR}}{\mathrm{MSE}} \sim_{H_0} \quad F_{p-1,n-p}} \tag{13.3.1}$$

From (12.3.2) and (13.3.1) it can easily be derived that this F-statistic is equivalent to:

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}.$$

Its value is usually reported in a statistical package:

```
F-statistic: 82.75 on 2 and 57 DF,  p-value: < 2.2e-16
```

## 13.4 Test for all parameters

A $(1-\alpha)100\%$ joint confidence region for the unknown $\boldsymbol{\beta} \in \mathbb{R}^p$ is given by an ellipsoid with center $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$:

$$E_\alpha = \{\boldsymbol{x} \in \mathbb{R}^p | \frac{(\boldsymbol{x} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})^t (X^tX)(\boldsymbol{x} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})}{ps^2} \leqslant F_{p,n-p,\alpha}\}.$$

This ellipsoid can be used to test hypotheses of the form:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$
$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

for some fixed vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. If $\boldsymbol{\beta}_0$ does not belong to $E_\alpha$, we reject the $H_0$ hypothesis at the $\alpha$ significance level.

Again, this procedure attains the correct significance level by employing the covariance matrix of $\hat{\boldsymbol{\beta}}$. A drawback is that if the $H_0$ hypothesis is rejected, we can not directly deduce statements about the individual parameters. This is why individual or simultaneous confidence intervals for $\beta_{0j}$ are still useful. The geometric differences between the two types of tests are illustrated in Figure 5.1.

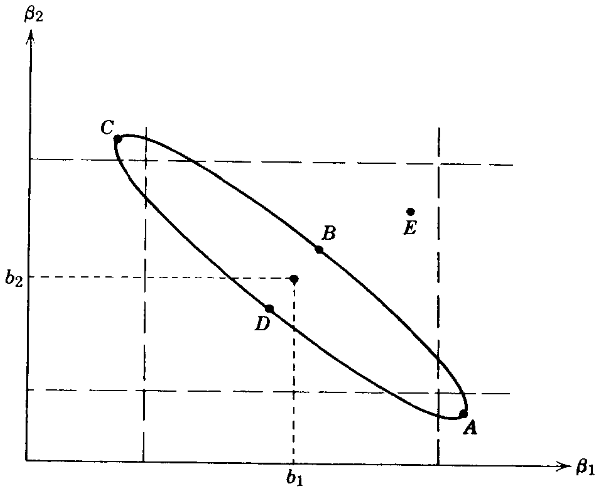


**Figure 5.1.** Joint and individual confidence statements. The point $(b_1, b_2)$ defined by the least squares estimates is at the center of both ellipse and rectangle.

If the correlation between the parameter estimates is large, the ellipsoid will be more elongated, and tests based on confidence intervals will be too conservative. So, it is very important to look at the correlation of the regression parameters.

### Example: Fuel data

```
summary(Fuelfit,correlation=TRUE)
```

The last part of the output now yields the correlation between the three parameter estimates for the fuel consumption data:

```
Correlation of Coefficients:
       (Intercept) Weight
Weight -0.90
Disp.   0.47       -0.80
```

Clearly, the correlation between `Weight` and `Disp.` is very high, as in Figure 5.1. Hence, the large *p*-value for `Disp.` is not very informative.
This high correlation is due to a high correlation between `Weight` and `Disp.`:

```
cor(Weight,Disp.)

[1] 0.8032804
```

Here, $\mathrm{cor}(\texttt{Weight},\texttt{Disp.}) = -\mathrm{cor}(\hat{\beta}_1, \hat{\beta}_2)$ but this equality is not satisfied in general.

## 13.5 A general linear hypothesis

All the above hypotheses belong to the class of linear hypotheses of the form:

$$\begin{aligned} H_0 &: C\boldsymbol{\beta} = \mathbf{0} \\ H_1 &: C\boldsymbol{\beta} \neq \mathbf{0} \end{aligned} \tag{13.5.1}$$

with $C$ a $(q \times p)$ matrix with $\text{rank}(C) = q \leqslant p$.

Example 1:

$$H_0 : \beta_1 = \beta_2, \beta_3 = 0$$

is equivalent to (13.5.1) with

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

Example 2:

$$H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

is equivalent to (13.5.1) with $C = (\mathbf{0} \ \ I_{p-1})$.

Example 3:

$$H_0 : \ \beta_{p-q} = \beta_{p-q+1} = \ldots = \beta_{p-1} = 0$$

is equivalent to (13.5.1) with $C = (0_{q,p-q} \ \ I_q)$.

The linear hypothesis $C\boldsymbol{\beta} = \mathbf{0}$ provides $q$ independent equations, so $q$ of the $\beta_j$'s can be expressed in terms of the other $p - q$. Under the null hypothesis we thus obtain a reduced model with only $p - q$ parameters. Let $\text{SSE}_{p-q}$ again denote the error sum of squares under this reduced model. Then we obtain the same partial F-statistic (13.2.1).

## 13.6 Mean response and prediction

### 13.6.1 Inference about the mean response

At a fixed point $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0,p-1})^t$, the (unknown) *mean response* is denoted as $E[Y_0|\boldsymbol{x}_0] = \boldsymbol{x}_0^t\boldsymbol{\beta}$. An unbiased estimator of the mean response is given by

$$\hat{y}_0 = \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}}$$

with

$$\text{Var}(\hat{y}_0) = \boldsymbol{x}_0^t\Sigma(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0$$

which is estimated by $s^2\boldsymbol{x}_0^t(X^tX)^{-1}\boldsymbol{x}_0$.
A $(1-\alpha)100\%$ confidence interval for the mean response $E[Y_0|\boldsymbol{x}_0]$ is then given by

$$\hat{y}_0 \pm t_{n-p,\frac{\alpha}{2}} s\sqrt{\boldsymbol{x}_0^t(X^tX)^{-1}\boldsymbol{x}_0}.$$

### 13.6.2 Inference about the unknown response

Consider now a new point $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0,p-1})^t$ which does not belong to the data set. A confidence interval for the unknown response

$$y_0 = \boldsymbol{x}_0^t\boldsymbol{\beta} + \varepsilon_0$$

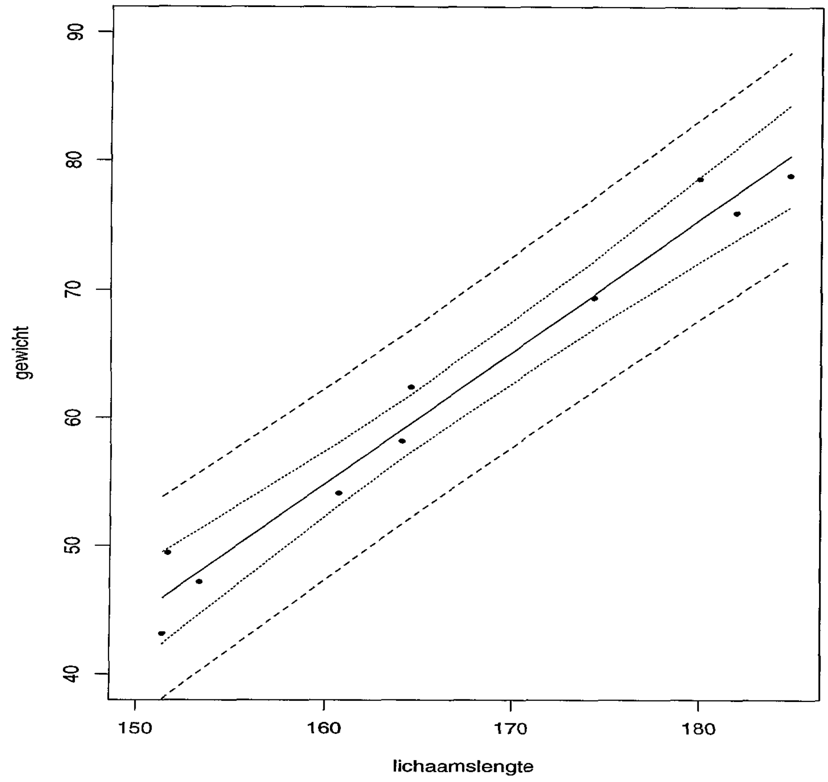
is constructed as follows. Consider the random variable $\hat{y}_0 - y_0 = \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^t\boldsymbol{\beta} - \epsilon_0$. It holds that

$$E[\hat{y}_0 - y_0] = \boldsymbol{x}_0^tE[\hat{\boldsymbol{\beta}}] - \boldsymbol{x}_0^t\boldsymbol{\beta} = 0$$
$$\text{Var}[\hat{y}_0 - y_0] = \sigma^2\boldsymbol{x}_0^t(X^tX)^{-1}\boldsymbol{x}_0 + \sigma^2$$

because $\hat{\boldsymbol{\beta}}$ and $\varepsilon_0$ are independent and $\epsilon_0 \sim N(0, \sigma^2)$.

A $(1-\alpha)100\%$ prediction interval for the unknown response $y_0$ is then given by

$$\hat{y}_0 \pm t_{n-p,\frac{\alpha}{2}} s\sqrt{\boldsymbol{x}_0^t(X^tX)^{-1}\boldsymbol{x}_0 + 1}.$$



This interval is larger than the confidence interval for the mean response because it also includes the uncertainty given by $\epsilon_0$.

**Example.**

Figure 13.6.2 contains the 95% confidence intervals for the average weight of 10 people based on their length (dotted curves). This is however not so interesting as we are merely interested in the prediction of someone's weight, based on its length. For instance, given an individual with length $x = 170$cm, what can we say about his/her weight? The prediction interval (dashed curve) is larger than the confidence interval because it takes into account two sources of variability: the variability of the fitted line, and the variability of the observation around the regression line.

We notice that the confidence and the prediction intervals become larger as we move away from the mean of the data. This illustrates how dangerous it is to draw conclusions about an observation with $\boldsymbol{x}$-values outside the range of the observed $\boldsymbol{x}_i$ values. This is called extrapolation.

## 13.7 Residual plots

Residual plots are very helpful to check the validity of our model assumptions. If our model is correct, we can consider the residuals $e_i$ as the observed errors. These residuals should exhibit tendencies which confirm the assumptions we have made, or at least do not exhibit a denial of the assumptions. Residual plots will be helpful to check for

1. non-normality

2. time effects (correlation)

3. nonconstant variance (and transformations of $Y$)

4. curvature (and transformations of the regressors)

5. outlier detection ...

### Normal quantile plot

We usually assume condition (13.0.1) which a.o. says that the errors are normally distributed with zero mean. Since the least squares residuals always have zero mean, remember (12.2.10), we do not have to check for it! Normality can be verified using a *normal quantile plot*. One can make such a qq-plot of the residuals themselves, but they have different standard errors, hence different distributions. Therefore, we prefer to make a normal quantile plot of the *standardized* residuals. They are defined by dividing each residual by its standard error (12.2.14):

$$\boxed{e_i^{(s)} = \frac{e_i}{s\sqrt{1-h_{ii}}}} \tag{13.7.1}$$

with $h_{ii}$ the $i$th diagonal element of the hat matrix $H$. Note that the normality assumption can also be accessed formally through the Shapiro-Wilk statistic or the Kolmogorov-Smirnov test.

### Plot of the residuals versus their index

This plot is useful if the index $i$ has a physical interpretation, such as ‘time’. When we make a plot of the $e_i$ versus $i$, we do not expect to see any pattern

**FIGURE 3.9** Normal Probability Plots when Error Term Distribution Is Not Normal.

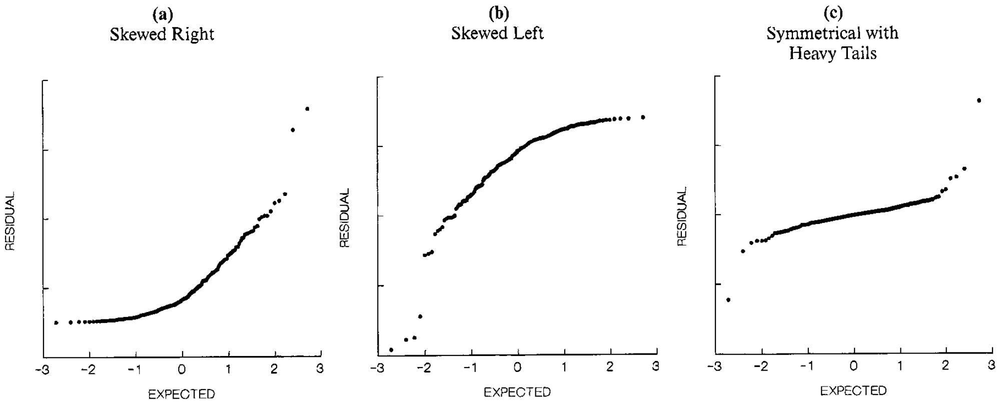
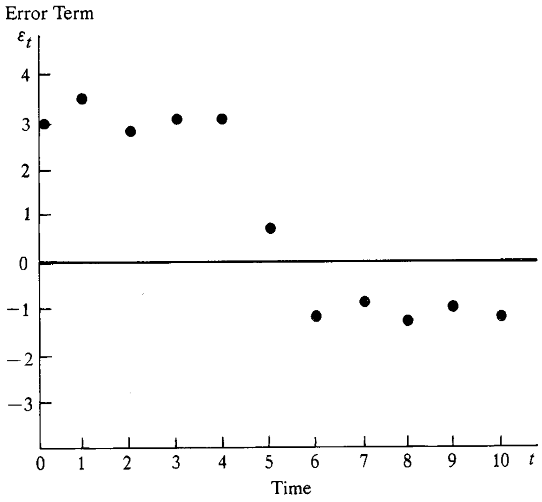


**FIGURE 12.1** Example of Positively Autocorrelated Error Terms.



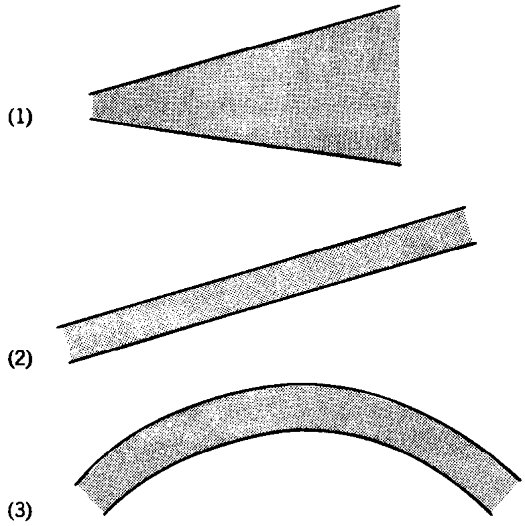


**Figure 2.6.** Examples of characteristics shown by unsatisfactory residuals behavior.

as the errors should be uncorrelated. An example of positively autocorrelated error terms is presented in Figure 12.1.

Moreover we can check whether the variance is constant over time (known as homoscedasticity). A pattern as in Figure 2.6 (1) gives evidence that the variance is increasing (heteroscedasticity). Patterns (2) and (3) show that the linear model should be refined by adding first-order or second order terms.

### Plot of the residuals versus fitted values

If the linear model is correct, it follows from equations (12.2.10) and (12.2.12) that the $e_i$ are uncorrelated with the $\hat{y}_i$. A non-horizontal or curved band in a $(\hat{y}_i, e_i)$ plot thus shows that the linear model is not appropriate. Sometimes the variance of the errors increases with the level of the dependent variable. This is again visible when we observe a funnel.

### Plot of the residuals versus independent variables

From equation (12.2.11) we can deduce that each independent variable $X_j$ is uncorrelated with the residuals $e_i$. Therefore, $(x_{ij}, e_i)$ plots (for $j = 1, \ldots, p-1$) again can indicate that the regression fit is defective.

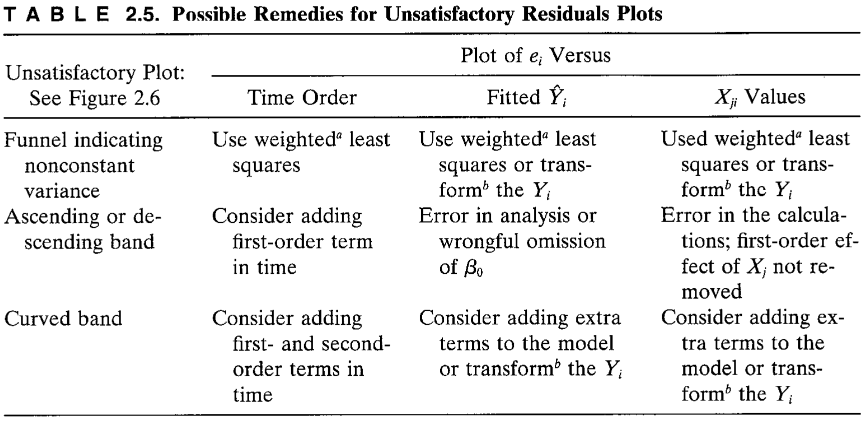


**TABLE 2.5. Possible Remedies for Unsatisfactory Residuals Plots**

| Unsatisfactory Plot: | Plot of $e_i$ Versus | | |
|---|---|---|---|
| See Figure 2.6 | Time Order | Fitted $\hat{Y}_i$ | $X_{ji}$ Values |
| Funnel indicating nonconstant variance | Use weighted$^a$ least squares | Use weighted$^a$ least squares or transform$^b$ the $Y_i$ | Used weighted$^a$ least squares or transform$^b$ thc $Y_i$ |
| Ascending or descending band | Consider adding first-order term in time | Error in analysis or wrongful omission of $\beta_0$ | Error in the calculations; first-order effect of $X_j$ not removed |
| Curved band | Consider adding first- and second-order terms in time | Consider adding extra terms to the model or transform$^b$ the $Y_i$ | Consider adding extra terms to the model or transform$^b$ the $Y_i$ |

## Plot of the standardized residuals

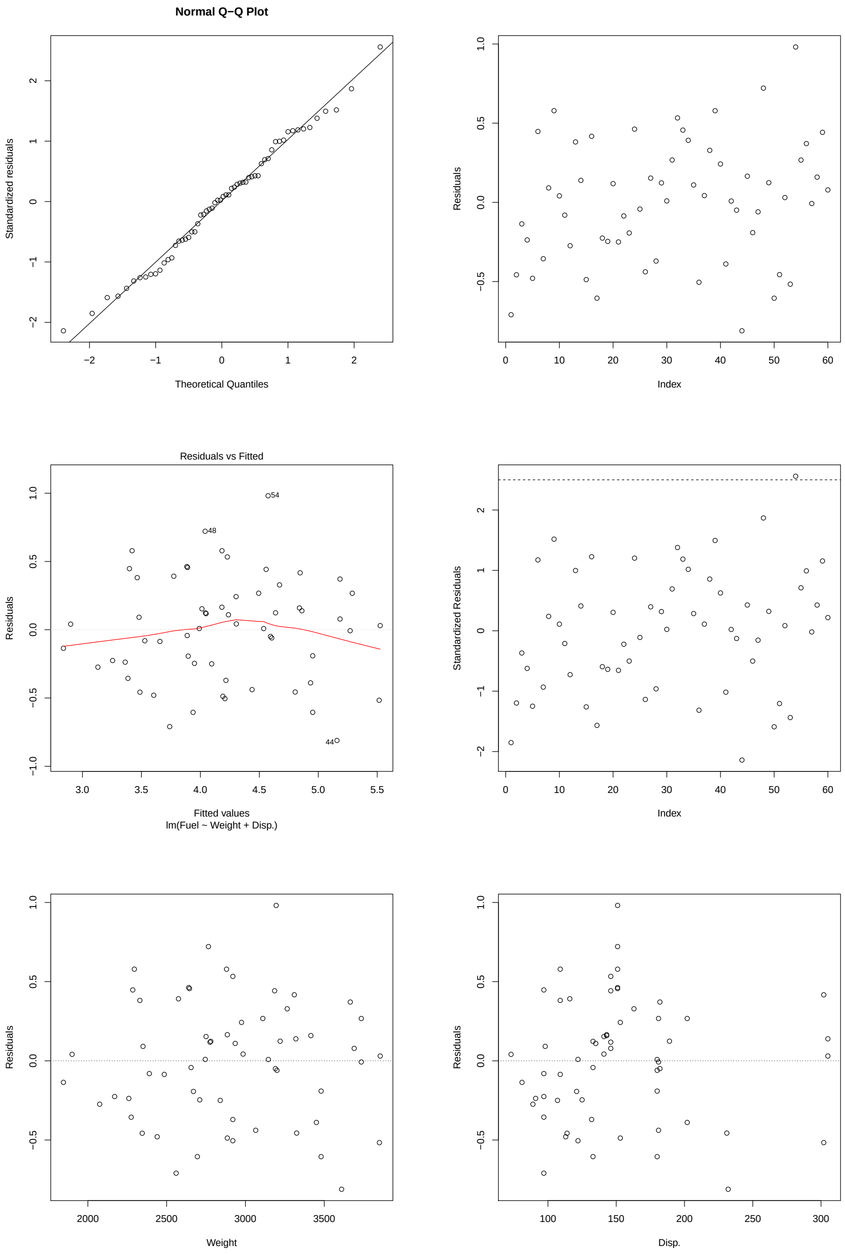
Diagnostics to detect influential observations and outliers will be discussed in Chapter 19. Now, we can already make a plot of the standardized residuals (13.7.1). If they are a good approximation of the true errors, they should be approximately gaussian distributed. Hence observations whose absolute standardized residual is larger than, say 2.5, can be pinpointed as outliers.

In R the standardized residuals can be retrieved from the `lm.influence` function, or using the `library MASS`:

```r
Fuelinf <- lm.influence(Fuelfit)
e <- residuals(Fuelfit)
h <- Fuelinf$hat
s <- Fuelsum$sigma
es <- e/(s*(1-h)^.5)
library(MASS)
es2 <- stdres(Fuelfit)
qqnorm(es,ylab="Standardized residuals")
qqline(es)
plot(e,xlab="Index",ylab="Residuals")
plot(Fuelfit) # first graph yields residuals versus fitted values
plot(es,xlab="Index",ylab="Standardized Residuals")
abline(h=-2.5,lty=2)
abline(h=2.5,lty=2)
```

```r
plot(Weight,e,ylab="Residuals")
abline(h=0,lty=3)
plot(Disp.,e,ylab="Residuals")
abline(h=0,lty=3)
```

This yields the following plots.



# Chapter 14

# Categorical predictors

Contrary to *continuous* or *quantitative* regressors, *categorical* or *qualitative* predictor variables only take on a finite number of values. Typical examples are gender (male/female), age (child, adult, senior), profession, type of production machine, time period (1941-1960, 1961-1980, 1981-2000), ...

## 14.1 One dichotomous predictor variable

### 14.1.1 Constructing the model

In the simplest case, we have one continuous regressor and one categorical predictor that takes on only two different values.

**Example: Insurance Innovation data set.**
An economist wished to relate the speed with which a particular insurance innovation is adopted to the size of the insurance firm and the type of firm. The variables in this data set are:

$Y$ : the number of months elapsed between the time the first firm adopted the innovation and the time the given firm adopted the innovation

$X_1$ : the size of the firm, measured by the amount of total assets, in million \$

$T_2$ : the type of the firm: stock company or mutual company

```
firm

   Months Size   Type
1      17  151 Mutual
2      26   92 Mutual
3      21  175 Mutual
4      30   31 Mutual
5      22  104 Mutual
6       0  277 Mutual
7      12  210 Mutual
8      19  120 Mutual
9       4  290 Mutual
10     16  238 Mutual
11     28  164  Stock
12     15  272  Stock
13     11  295  Stock
14     38   68  Stock
15     31   85  Stock
16     21  224  Stock
17     20  166  Stock
18     13  305  Stock
19     30  124  Stock
20     14  246  Stock
```

To include the type of firm in a regression model, this dichotomous variable is usually converted into a binary variable:

$$X_2^{(1)} = \begin{cases} 1 & \text{if firm } i \text{ is a stock company} \\ 0 & \text{if firm } i \text{ is a mutual company} \end{cases} \tag{14.1.1}$$

The regression model then becomes

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^{(1)} + \epsilon_i.} \tag{14.1.2}$$

The interpretation of the regression coefficients is as follows:

- for a stock company, we have

$$\begin{aligned}E[Y|X_1] &= \beta_0 + \beta_1 X_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 X_1\end{aligned}$$

  which is a straight line with slope $\beta_1$ and intercept $\beta_0 + \beta_2$.

- for a mutual company, we obtain

$$\begin{aligned}E[Y|X_1] &= \beta_0 + \beta_1 X_1 + 0 \\ &= \beta_0 + \beta_1 X_1\end{aligned}$$

  which is a straight line with the same slope $\beta_1$ and intercept $\beta_0$.

In the $(X_1, Y)$ space, both lines are thus parallel. The parameter $\beta_0$ is the expected value for mutual companies at $x_1 = 0$, and $\beta_2$ indicates how much higher the elapsed time is for stock firms that for mutual firms, for any given size of firm, see Figure 11.1.

In general, $\beta_2$ shows how much higher or lower the mean response line is for the class coded 1 than the line for the class coded 0, for any given level of $X_1$. The class coded 0 thus serves as the reference group.

Other coding schemes could be used as well, e.g.

$$X_2^{(2)} = \begin{cases} 0 & \text{if firm } i \text{ is a stock company} \\ 1 & \text{if firm } i \text{ is a mutual company} \end{cases}$$

or

$$X_2^{(3)} = \begin{cases} 1 & \text{if firm } i \text{ is a stock company} \\ -1 & \text{if firm } i \text{ is a mutual company} \end{cases}$$

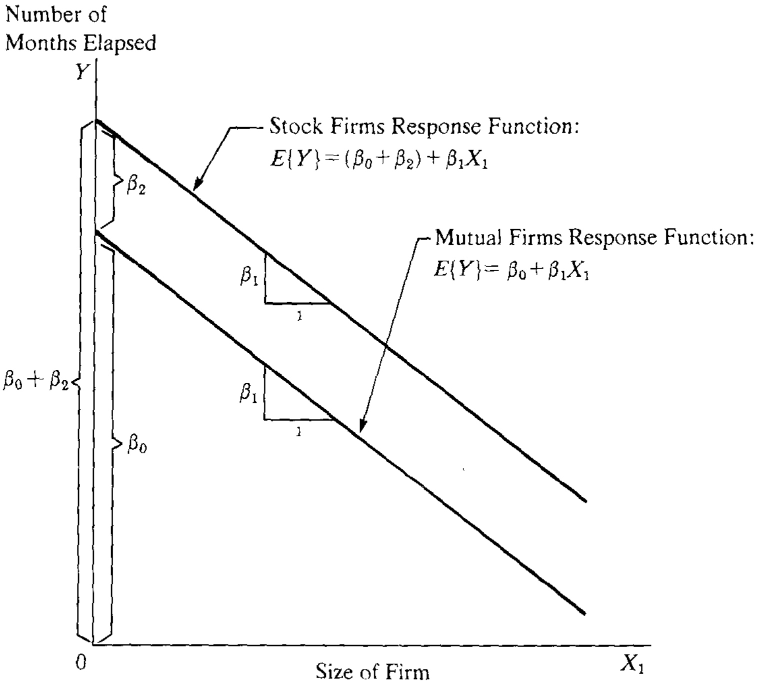
Then of course the interpretation of the regression coefficients changes! Consider e.g. the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^{(3)} + \epsilon_i. \qquad (14.1.3)$$

Then stock firms have

$$E[Y|X_1] = (\beta_0 + \beta_2) + \beta_1 X_1$$

FIGURE 11.1 **Illustration of Meaning of Regression Coefficients for Regression Model (11.4) with Indicator Variable $X_2$—Insurance Innovation Example.**



whereas mutual firms have

$$E[Y|X_1] = (\beta_0 - \beta_2) + \beta_1 X_1.$$

Now the difference between the expected time of a stock firm and the expected time of a mutual firm at a given $x_1$ is expressed by $2\beta_2$! The reference line $y = \beta_0 + \beta_1 x_1$ then lies in between the other two response functions.

Note that we only need one binary variable $X_2$ although $T_2$ has two levels. If we would e.g. define

$$X_2 = \begin{cases} 1 & \text{if firm } i \text{ is a stock company} \\ 0 & \text{if firm } i \text{ is a mutual company} \end{cases}$$

and

$$X_3 = \begin{cases} 0 & \text{if firm } i \text{ is a stock company} \\ 1 & \text{if firm } i \text{ is a mutual company} \end{cases}$$

then our design matrix would not have full rank, because $x_{i2} + x_{i3} = 1$ which is the intercept term. Consequently the LS estimator would not be unique.

### 14.1.2 Estimation and inference

The parameters in models (14.1.2) or (14.1.3) can be estimated as before, because these models belong to the class of general linear models (12.1.1). Which binary variable is used depends mainly on the tests we want to apply afterwards, or the confidence intervals we want to construct.

**Example.**
Assume that the economist is most interested in the effect of type of firm ($T_2$) on the elapsed time and wished to obtain a 95% confidence interval for the mean increase of the time of stock firms compared to mutual firms. Then it is recommend to work with the binary variable $X_2^{(1)}$. In R we obtain

```
attach(firm)
mst <- lm(Months ~ Size + Type)
coefficients(summary(mst))

              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 33.8740690 1.813858297  18.675146 9.145269e-13
Size        -0.1017421 0.008891218 -11.442990 2.074687e-09
TypeStock    8.0554692 1.459105700   5.520826 3.741874e-05
```

We find that the fitted model is

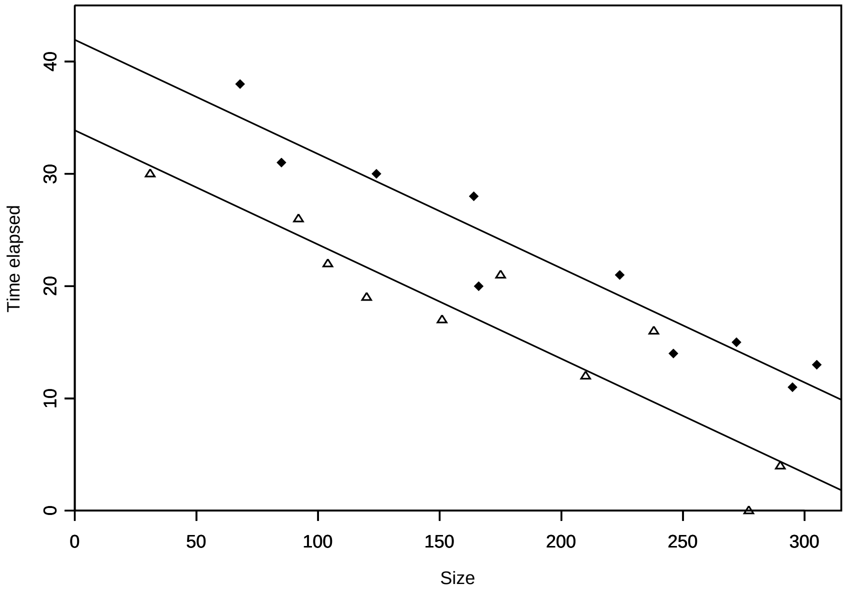
$$\hat{y}_i = 33.874 - 0.102x_{i1} + 8.055x_{i2}^{(1)}$$

and that a 95% confidence interval for $\beta_2$ is given by

$$8.055 \pm t_{17,0.025}1.459 = [4.98; 11.13].$$

So, with 95% confidence, we conclude that on the average, stock companies tend to adopt the innovation between 5 and 11 months later than mutual companies, for any given size of firm. A scatter plot of the data and the two regression lines are shown in Figure 14.1. This is the default coding scheme in R.

Figure 14.1: The Insurance Innovation data set, with the two regression lines superimposed.



The coding scheme can be checked by examining the design matrix

```
model.matrix(mst)
   (Intercept) Size TypeStock
1            1  151         0
2            1   92         0
3            1  175         0
...
18           1  305         1
19           1  124         1
20           1  246         1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$Type
[1] "contr.treatment"
```

To work with $X_2^{(3)}$ we set

```r
options(contrasts = c("contr.helmert","contr.poly"))
mst2 <- lm(Months ~ Size + Type)
summary(mst2)
```

which yields

```
Call:
lm(formula = Months ~ Size + Type)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6915 -1.7036 -0.4385  1.9210  6.3406

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.901804   1.770041  21.413 9.78e-14 ***
Size        -0.101742   0.008891 -11.443 2.07e-09 ***
Type1        4.027735   0.729553   5.521 3.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared:  0.8951,Adjusted R-squared:  0.8827
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

The design matrix now is

```r
model.matrix(mst2)
```

```
   (Intercept) Size Type1
1            1  151    -1
2            1   92    -1
3            1  175    -1
...
```

```
18           1  305     1
19           1  124     1
20           1  246     1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$Type
[1] "contr.helmert"
```

Note that the fitted values and consequently also the residuals are the same whether we use $X_2^{(1)}$ or $X_2^{(3)}$. Hence, the MSE and the overall F-test yield the same results.

### 14.1.3 Adding interaction terms

So far, we have assumed that the regression lines of both classes are parallel. If this assumption is not plausible, we can extend model (14.1.2) by adding an interaction term. This yields the model

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i} \tag{14.1.4}$$

with $X_2 = X_2^{(1)}$ defined in (14.1.1). The meaning of the regression coefficients now becomes:

- for stock firms:

  $$\begin{aligned} E[Y|X_1] &= \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \end{aligned}$$

  which is a line with intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$.

- for mutual firms:

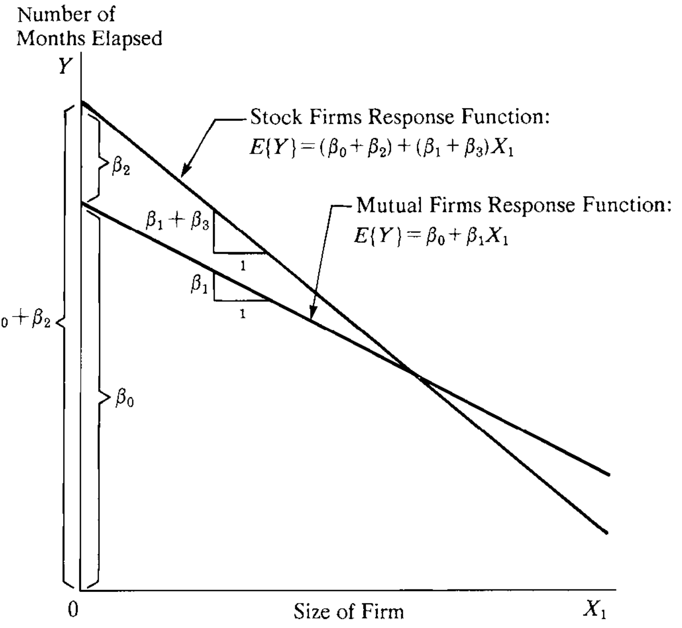
  $$E[Y|X_1] = \beta_0 + \beta_1 X_1$$

  which is a line with intercept $\beta_0$ and slope $\beta_1$ (see Figure 11.3).

The hypothesis test

$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$$

**FIGURE 11.3 Illustration of Meaning of Regression Coefficients for Regression Model (11.6) with Indicator Variable $X_2$ and Interaction Term — Insurance Innovation Example.**


Number of Months Elapsed
Y
Stock Firms Response Function:
$E\{Y\}=(\beta_0+\beta_2)+(\beta_1+\beta_3)X_1$
Mutual Firms Response Function:
$E\{Y\}=\beta_0+\beta_1X_1$
$\beta_2$
$\beta_1+\beta_3$
1
$\beta_1$
1
$\beta_0+\beta_2$
$\beta_0$
0
Size of Firm
$X_1$

is performed to see whether the interaction term is significant, or equivalently whether the two regression lines are parallel or not.

**Example.**

If we include an interaction term in the Insurance Innovation regression and use the default dummy encoding in R, then the result of the analysis is:

```
mst3 <- lm(Months ~ Size * Type)
summary(mst3)

Call:
lm(formula = Months ~ Size * Type)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7144 -1.7064 -0.4557  1.9311  6.3259 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    33.8383695  2.4406498  13.864 2.47e-10 ***
Size           -0.1015306  0.0130525  -7.779 7.97e-07 ***
TypeStock       8.1312501  3.6540517   2.225   0.0408 *  
Size:TypeStock -0.0004171  0.0183312  -0.023   0.9821    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.32 on 16 degrees of freedom
Multiple R-squared:  0.8951,Adjusted R-squared:  0.8754 
F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

The large $p$-value for $\beta_3$ confirms that the simplified regression model (14.1.2) without interaction term is appropriate for this data set.

Remark:

1. The model with the interaction term (14.1.4) is almost identical to the model in which we assume a separate regression line for both groups that are defined by the dichotomous predictor variable. The only difference is that model (14.1.4) assumes that the data points of both classes show the

same variability around their regression line. Doing so, tests about the equality of the slopes and the intercepts become very easy to apply.

2. The *principle of marginality* specifies that a model including a high-order term, such as an interaction, should normally also include the lower-order relatives of that term (the main effects that compose the interaction).

Suppose that we fit model (14.1.4) and conclude that $\beta_2 = 0$ but $\beta_3 \neq 0$, which would result in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i1} x_{i2} + \epsilon_i.$$

This model describes regression lines that have the same intercept but different slopes which is a peculiar specification (generally) of no substantive interest. Similarly, the model which retains $\beta_3$ but removes $\beta_1$:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

has a zero slope for the class coded $X_2 = 0$ which is usually too restrictive.

## 14.2 Extensions

There are two straightforward extensions of models (14.1.2) and (14.1.4). The first one includes a predictor variable with more than two classes, the second one includes more than one categorical variable.

### 14.2.1 One polytomous predictor variable

If a qualitative predictor variable $T_2$ has more than two levels, we need additional indicator variables in the regression model. Assume e.g. that $T_2$ indicates a tool model which can take on four different values: M1, M2, M3 and M4. Then we need three binary variables:

$$X_2 = \begin{cases} 1 & \text{if } T_2 = M1 \\ 0 & \text{if } T_2 \neq M1 \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if } T_2 = M2 \\ 0 & \text{if } T_2 \neq M2 \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if } T_2 = M3 \\ 0 & \text{if } T_2 \neq M3 \end{cases}$$

With $Y =$ tool wear, and $X_1 =$ tool speed, a first-order regression model is:

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i.} \qquad (14.2.1)$$

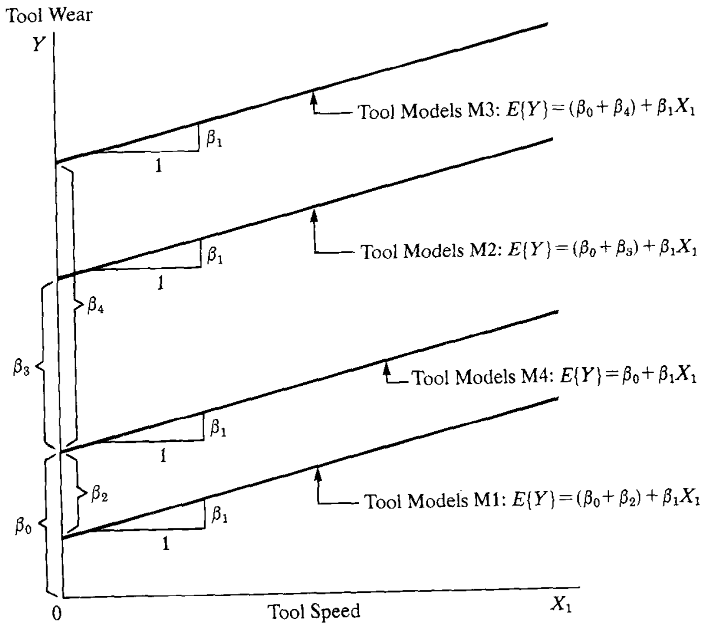
The response functions are again lines with the same slope $\beta_1$ for all values of $T_2$, see Figure 11.5. The coefficients $\beta_2, \beta_3$ and $\beta_4$ indicate how much higher (lower) the response functions are for tool models M1, M2 and M3 than for tool model M4, for any given level of tool speed.

An hypothesis test of the form $H_0 : \beta_j = 0$ for $j = 2, \ldots, 4$ is thus concerned with the differential effect of class $j$ compared with the reference class M4 for which $X_2 = X_3 = X_4 = 0$. If we want to compare the intercepts of e.g. class M3 with M2 we use the point estimator $\hat{\beta}_4 - \hat{\beta}_3$ with variance

$$\text{Var}(\hat{\beta}_4 - \hat{\beta}_3) = \text{Var}(\hat{\beta}_4) + \text{Var}(\hat{\beta}_3) - 2\text{cov}(\hat{\beta}_4, \hat{\beta}_3)$$

which can be estimated from $\hat{\Sigma}(\hat{\boldsymbol{\beta}})$.

**FIGURE 11.5 Illustration of Regression Model (11.9)—Tool Wear Example.**


Tool Wear
Y
Tool Models M3: E{Y} = (β0 + β4) + β1X1
Tool Models M2: E{Y} = (β0 + β3) + β1X1
Tool Models M4: E{Y} = β0 + β1X1
Tool Models M1: E{Y} = (β0 + β2) + β1X1
β1
1
β4
β3
β2
β0
0
Tool Speed
X1

If interaction effects are present, the regression model (14.2.1) becomes:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i2} + \beta_6 x_{i1} x_{i3} + \beta_7 x_{i1} x_{i4} + \epsilon_i.$$

This model again implies that each tool model has its own regression line, with different intercepts and slopes for the different tool models.

### 14.2.2 More than one categorical variable

If several predictor variables are qualitative, they should all be converted into indicator variables.

Example: one quantitative and two dichotomous qualitative regressors. The first-order model becomes:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

The response functions can be summarized as:

| | $X_2 = 0$ | $X_2 = 1$ |
|---|---|---|
| $X_3 = 0$ | $\beta_0 + \beta_1 X_1$ | $(\beta_0 + \beta_2) + \beta_1 X_1$ |
| $X_3 = 1$ | $(\beta_0 + \beta_3) + \beta_1 X_1$ | $(\beta_0 + \beta_2 + \beta_3) + \beta_1 X_1$ |

With interactions between each pair of the predictor variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \epsilon_i$$

with response functions

| | $X_2 = 0$ | $X_2 = 1$ |
|---|---|---|
| $X_3 = 0$ | $\beta_0 + \beta_1 X_1$ | $(\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$ |
| $X_3 = 1$ | $(\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$ | $(\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) X_1$ |

**Remarks.**

- If all the explanatory variables are qualitative, the models are called *analysis of variance* models.
- If the model contains qualitative and quantitative regressors, but the main variables of interest are the qualitative ones, it is called an *analysis of covariance* model.

## 14.3 Piecewise linear regression

Indicator variables can also be used when the regression of $Y$ on $X$ follows a certain linear relation in some range of $X$ but follows a different relation elsewhere.

**Example: lot size data set.**

The unit cost of a lot depends linearly on the lot size, but the slope changes once the lot size exceeds 500 (e.g. because the unit price of some raw materials decrease if larger amounts are purchased). The model, as illustrated in Figure 11.9, can be expressed as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 500) x_{i2} + \epsilon_i$$

where $X_1$ is the lot size and

$$X_2 = \begin{cases} 1 & \text{if } X_{i1} > 500 \\ 0 & \text{otherwise.} \end{cases}$$

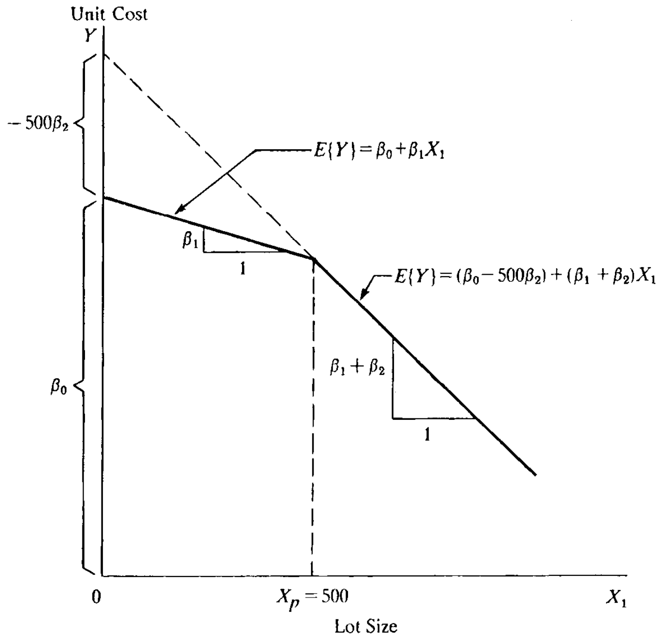
For $X_1 \leqslant 500, X_2 = 0$ we obtain the response function:

$$E[Y|X_1] = \beta_0 + \beta_1 X_1$$

whereas for $X_1 > 500, X_2 = 1$ we have

$$\begin{aligned} E[Y|X_1] &= \beta_0 + \beta_1 X_1 + \beta_2 (X_1 - 500) \\ &= (\beta_0 - 500\beta_2) + (\beta_1 + \beta_2) X_1. \end{aligned}$$

**FIGURE 11.9 Illustration of Meaning of Regression Coefficients of Piecewise Linear Response Function (11.25).**



When the regression function not only changes its slope at some value $X_p$ but also makes a jump there, then we need an additional term. The response functions in Figure 11.10 could be modelled as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 40) x_{i2} + \beta_3 x_{i2} + \epsilon_i$$

with $X_2 = I(X_1 > 40)$. Then for $X_1 \leqslant 40$ the response function becomes

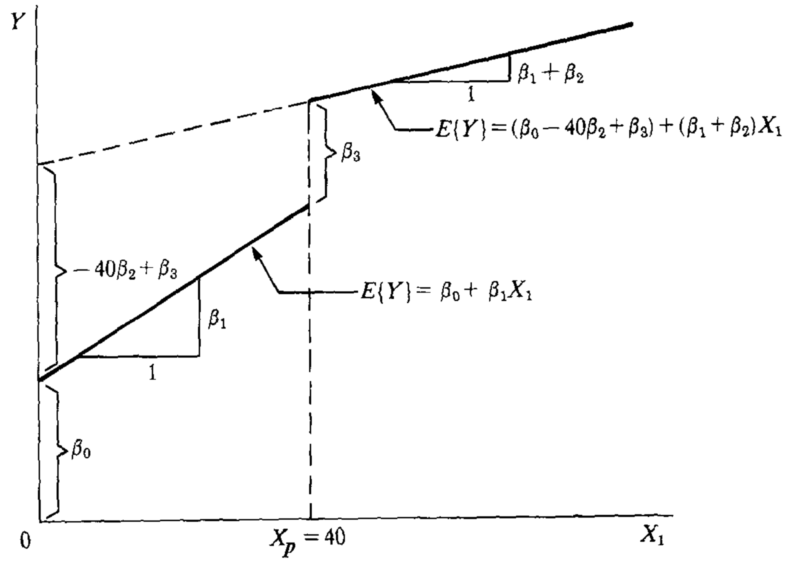
$$E[Y|X_1] = \beta_0 + \beta_1 X_1$$

and for $X_1 > 40$ we obtain

$$\begin{aligned} E[Y|X_1] &= \beta_0 + \beta_1 X_1 + \beta_2 (X_1 - 40) + \beta_3 \\ &= (\beta_0 - 40\beta_2 + \beta_3) + (\beta_1 + \beta_2) X_1 \end{aligned}$$

so $\beta_2$ represents the difference in the slopes, and $\beta_3$ the difference in the mean responses at $X_p = 40$.

**FIGURE 11.10** **Illustration of Response Function (11.28) for Discontinuous Piecewise Linear Regression.**



$Y$
$\beta_1 + \beta_2$
1
$E\{Y\} = (\beta_0 - 40\beta_2 + \beta_3) + (\beta_1 + \beta_2)X_1$
$\beta_3$
$-40\beta_2 + \beta_3$
$E\{Y\} = \beta_0 + \beta_1 X_1$
$\beta_1$
1
$\beta_0$
0
$X_p = 40$
$X_1$

# Chapter 15

# Comparisons among several groups

In the previous chapter we have illustrated the techniques to use when a categorical regressor is included in a linear regression model, alongside one or more continuous regressors. This was illustrated during the analysis of the dataset in Section 14.1.
In this chapter we shall build on this idea, but we will consider models with only categorical variables included as regressors.

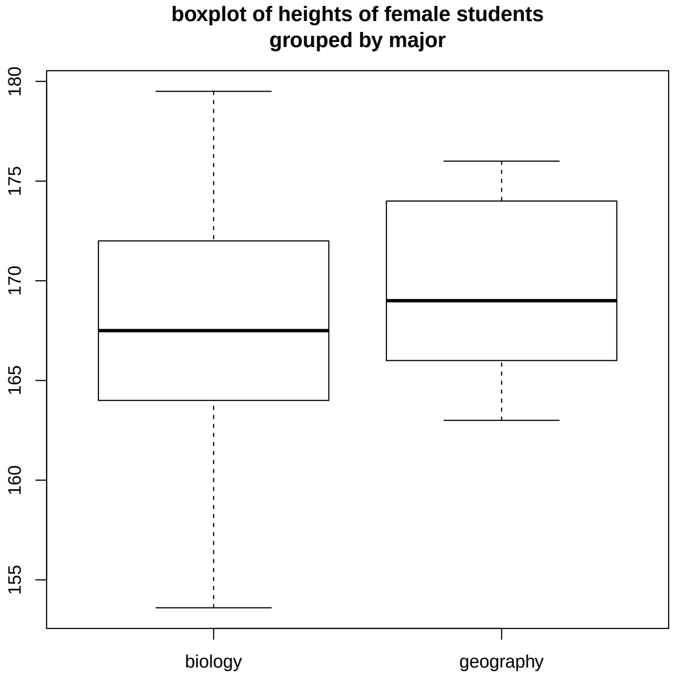
## 15.1 Two groups

In the simple case, we have a sample where a variable is recorded for two different groups. Hence, this can be modelled by a linear model with one categorical regressor that can take only two values.

**Example: Heights of female bachelor students**
We wish to find a model for the height of female bachelor students, differentiated by major (geography or biology), and use this model to determine whether geography and biology students, on average, have different heights.

```
   length     major
4     175 geography
34    165 geography
35    176 geography
```

```
36    164 geography
37    173 geography
38    175 geography
...
98     170   biology
99     166   biology
100    169   biology
107    174 geography
108    166 geography
109    167 geography
```



The boxplot of the data indicates that there is little difference between the medians of the heights of both groups of students, and that the heights of the biology students has a larger range than that of the geography students.

Similar to equation (14.1.2), we can model the relation between major and height through the regression model

$$y_i = \beta_0 + \beta_1 x_{i1}^{(1)} + \epsilon_i \tag{15.1.1}$$

where the binary variable $X_1^{(1)}$ denotes the student's gender and is encoded as

$$X_1^{(1)} = \begin{cases} 1 & \text{if the student is a geography major} \\ 0 & \text{if the student is a biology major} \end{cases} \tag{15.1.2}$$

and $Y$ denotes the student's height in centimetres. For inferential purposes we further assume that (12.1.7) and (12.1.8) are satisfied, together with normality of the error terms, i.e. that

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n). \tag{15.1.3}$$

The property that the variances are equal is referred to as *homoskedasticity*. The interpretation of the regression coefficients is then

- For geography students, we have

$$E[Y_g] = \beta_0 + \beta_1.$$

- For biology students, we have

$$E[Y_b] = \beta_0.$$

To test whether geography and biology students have, on average, the same height, we have to test

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

Fitting the linear model in R yields

```
lengths.lm1 <- lm(length ~ major, data=lengths)
summary(lengths.lm1)

Call:
lm(formula = length ~ major, data = lengths)


Residuals:
     Min       1Q   Median       3Q      Max
-14.3357  -3.6986  -0.4615   4.3014  11.5643
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     167.936      1.565 107.335   <2e-16 ***
majorgeography    1.526      2.255   0.677    0.505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.854 on 25 degrees of freedom
Multiple R-squared:  0.01799,Adjusted R-squared:  -0.02129
F-statistic: 0.4579 on 1 and 25 DF,  p-value: 0.5048
```

R encodes the categorical variable `major` as `majorgeography`, indicating that the value `biology` is the reference group and `geography` is the treatment group. It can be easily derived that the least-squares estimates correspond with $\hat{\beta}_0 = \bar{y}_b$ (the empirical average of the heights of the biology students) and $\hat{\beta}_1 = \bar{y}_g - \bar{y}_b$ (the difference between the empirical averages of the heights of the geography and biology students). We observe that on average, female biology students are about 168 cm tall, female geography students are about 1.5 cm shorter, and that this difference is not statistically significant ($p \approx 0.5$).

Alternatively, the major can be encoded by the variable $X_1^{(3)}$ in the following manner

$$X_1^{(3)} = \begin{cases} 1 & \text{if the student is a geography major} \\ -1 & \text{if the student is a biology major} \end{cases} \tag{15.1.4}$$

and the regression model then becomes

$$y_i = \beta_0' + \beta_1' x_{i1}^{(3)} + \epsilon_i', \tag{15.1.5}$$

with $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$ This yields

$$E[Y_g] = \beta_0' + \beta_1'$$

for geography students, and

$$E[Y_b] = \beta_0' - \beta_1'$$

for biology students, from which it easily follows that

$$\begin{aligned} \beta_0' &= \beta_0 + \tfrac{1}{2}\beta_1, \text{ and} \\ \beta_1' &= \tfrac{1}{2}\beta_1. \end{aligned}$$

We also see that the parameter $\beta_0'$ corresponds with the average of the group means, i.e. $\beta_0' = \frac{1}{2}\big(E[Y_g] + E[Y_b]\big)$, whereas $|\beta_1'|$ yields the deviation of each group mean to $\beta_0'$.

As a result of this, testing whether female biology and geography students, on average, have the same height corresponds with

$$H_0 : \beta_1' = 0 \text{ vs. } H_1 : \beta_1' \neq 0$$

when using the alternative parametrisation. In R this encoding can be obtained as follows.

```
options(contrasts = c("contr.helmert", "contr.poly"))
length.lm2 <- lm(length ~ major, data=lengths)
summary(length.lm2)

Call:
lm(formula = length ~ major, data = lengths)


Residuals:
     Min       1Q   Median       3Q      Max
-14.3357  -3.6986  -0.4615   4.3014  11.5643


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 168.6986     1.1274 149.634   <2e-16 ***
major1        0.7629     1.1274   0.677    0.505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 5.854 on 25 degrees of freedom
Multiple R-squared:  0.01799,Adjusted R-squared:  -0.02129
F-statistic: 0.4579 on 1 and 25 DF,  p-value: 0.5048
```

We observe that the average of both group means is about 168.7 cm, that the mean height of geography students is about 0.8 cm more than that (the variable `major1` equals 1 for geography students and $-1$ for biology students), and that this deviation is not statistically significant ($p \approx 0.5$). Note that the $t$-values and the $p$-values of $\beta_1$ and $\beta_1'$ in the models above are the same.

From (15.1.1) and (15.1.3), it follows that $Y_i \sim N(\beta_0 + \beta_1, \sigma^2) = N(\mu_1, \sigma^2)$ for geography students, and that $Y_i \sim N(\beta_0, \sigma^2) = N(\mu_2, \sigma^2)$ for biology students. As such, if we denote the heights of the biology students by $y_{i1}$, $i = 1, \ldots, n_1$, and the heights of the geography students by $y_{i2}$, $i = 1, \ldots, n_2$, the hypothesis test on whether both groups have the same height becomes a $t$-test on the mean for two groups with equal variance.

```
t.test(length ~ major, data=lengths, var.equal = TRUE)

Two Sample t-test


data:  length by major
t = -0.67669, df = 25, p-value = 0.5048
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.169710  3.118061
sample estimates:
  mean in group biology mean in group geography
               167.9357                169.4615
```

Once again, we observe that the difference in heights between the biology and geography students is not statistically significant. It is easy to see that this $t$-test always gives the exact same result as the hypothesis test on the linear model.

We also validate the model assumptions. Testing whether the variances of the heights of both groups are equal can be done using an $F$-test.

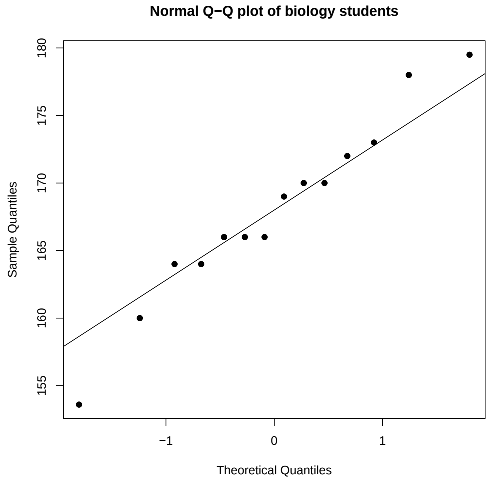
```
var.test(length ~ major, data=lengths)

F test to compare two variances
```

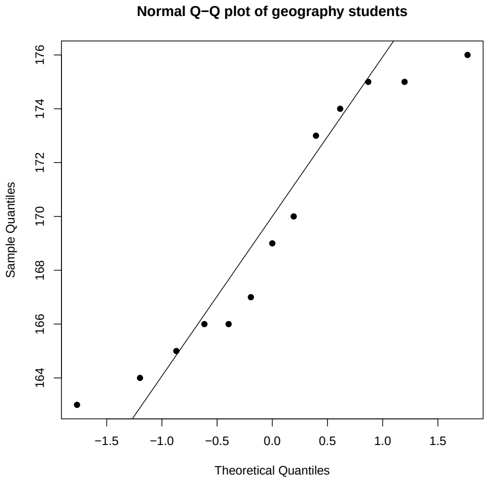
```
data:  length by major
F = 2.1278, num df = 13, denom df = 12, p-value = 0.201
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6568722 6.7092697
sample estimates:
ratio of variances
          2.127782
```

The $F$-test shows that, although the difference between the variances of both populations appears significant from the boxplot, they aren't statistically significant ($p \approx 0.2$). Hence, we can assume that the homoskedasticity assumption is valid.

Normality of the samples can be verified using normal QQ-plots.

```
qqnorm(lengths[lengths$major == "biology",1], main = "Normal Q-Q plot of biology st
       pch=19)
qqline(lengths[lengths$major == "biology",1])
qqnorm(lengths[lengths$major == "geography",1], main = "Normal Q-Q plot of geograph
       pch=19)
qqline(lengths[lengths$major == "geography",1])
```





These plots indicate that normality is a reasonable assumption, which is verified with the Shapiro-Wilk test ($p \approx 0.89$ for the biology students and $p \approx 0.17$ for the geology students).

```
shapiro.test(lengths[lengths$major == "biology",1])

Shapiro-Wilk normality test


data:  lengths[lengths$major == "biology", 1]
W = 0.97096, p-value = 0.8892

shapiro.test(lengths[lengths$major == "geography",1])

Shapiro-Wilk normality test


data:  lengths[lengths$major == "geography", 1]
W = 0.90693, p-value = 0.1665
```

Hence, the model assumptions have been verified, and the inference on the group means is valid.

## 15.2 Multiple groups

In this section we extend the ideas of the previous section towards the situation of comparing the means of three or more groups simultaneously.

Assume that there are $k$ distinct groups, and denote the observation of sample $j$ ($j = 1, \dots, k$) as $y_{ij}$ with $i = 1, \dots, n_j$. Hence, $n = \sum_{j=1}^{k} n_j$. As before, we will assume that the population distributions are normal and have the same variance, i.e. that $y_{ij} \sim N(\mu_j, \sigma^2)$. In addition to that, we assume independence among all observations.

Just as in the situation with two groups, we aim to test whether the population means of the $k$ groups are equal. In other words, we test

$$\begin{aligned} &H_0 : \mu_1 = \cdots = \mu_k \\ &H_1 : \exists i, j \in \{1, \dots, k\} : \mu_i \neq \mu_j. \end{aligned} \tag{15.2.1}$$

To perform this test, we make an ANOVA table:

| Sum of Squares | d.f. | Mean of Squares | F-stat |
|---|---|---|---|
| $SSR$ | $k-1$ | $MSR = \frac{SSR}{k-1}$ | $\frac{MSR}{MSE} \sim F_{k-1,n-k}$ |
| $SSE$ | $n-k$ | $MSE = \frac{SSE}{n-k}$ | |
| $SST$ | $n-1$ | $MST = \frac{SST}{n-1} = S_Y^2$ | |

where $S_Y^2$ is the total variance of all observations $y_{ij}$,

$$\begin{aligned} SSR &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\hat{y}_{ij} - \bar{y})^2 \\ &= \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 \\ &= \text{Between groups sum of squares,} \end{aligned}$$

and

$$\begin{aligned} SSE &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2 \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 \\ &= \sum_{j=1}^{k}(n_j - 1)S_j^2 \\ &= \text{Within groups sum of squares.} \end{aligned}$$

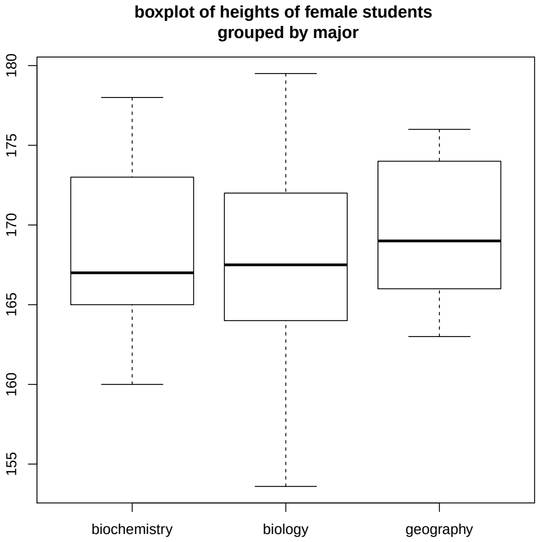
Here, $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ is the sample mean for group $j$, $\bar{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$ is the overall sample mean, and $S_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$ is the sample variance for group $j$.

We observe that, thanks to the specific nature of the problem, it is possible to compute the required sums of squares, and hence the test statistic $MSR/MSE$, without having to estimate a linear model. For this reason, this test is referred to as *one-way analysis of variance* (One-way ANOVA).

**Example: Heights of female bachelor students**

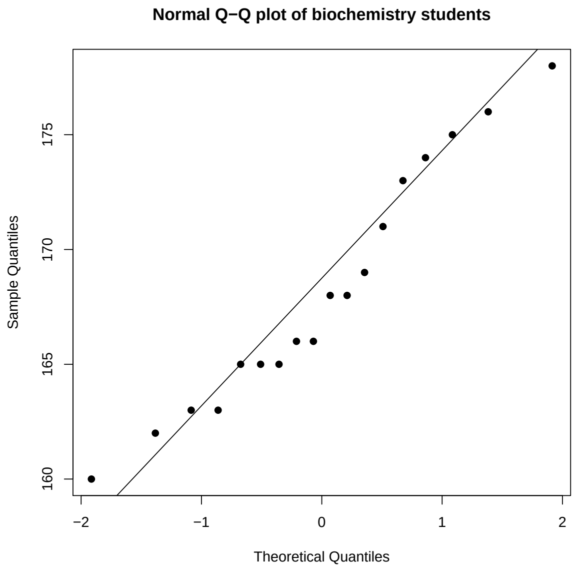
We revisit the bachelor student height example, but we now extend it to include the female biochemisty students in addition to the geography and biology students, and wish to test if there is a difference, on average, between the heights of the students over the various majors.

```
boxplot(length ~ major, data=lengths,
        main = "boxplot of heights of female students \n grouped by major")
```



The boxplot of these data shows that it is unlikely that there are significant differences between the average heights of the students, and that the spread of the heights of the biochemistry students is comparable to that of the geography students. We already know that the lengths of the biology and geography students are normally distributed, so we only have to verify this for the biochemistry majors.

```
qqnorm(lengths[lengths$major == "biochemistry",1],
       main = "Normal Q-Q plot of biochemistry students",pch=19)
qqline(lengths[lengths$major == "biochemistry",1])
```



Examining the normal QQ-plot does not give a strong indication that the lengths of the biochemistry students aren't normally distributed, and the Shapiro-Wilk test verifies this ($p \approx 0.42$).

```
shapiro.test(lengths[lengths$major == "biochemistry",1])

Shapiro-Wilk normality test


data:  lengths[lengths$major == "biochemistry", 1]
W = 0.94985, p-value = 0.4226
```

To test for homoskedasticity, we use Levene's test for equality of variances among several groups, which is available in the R package `car`. This test starts with defining $z_{ij} = |y_{ij} - \bar{y}_j|$ for each $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$, and the quantities $\bar{z}_j$ and $\bar{z}$ are defined analogously as $\bar{y}_j$ and $\bar{y}$. The test statistic is

$$W = \frac{n-k}{k-1} \frac{\sum_{j=1}^{k} n_j(\bar{z}_j - \bar{z})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (z_{ij} - \bar{z}_j)^2}$$

and has an $F$ distribution with $k-1$ and $n-k$ degrees of freedom. Note that this is in fact a one-way ANOVA performed on the observations $z_{ij}$. The test can also be performed with $z_{ij} = |y_{ij} - \text{med}_i(y_{ij})|$ and is then less sensitive to possible outliers in the data.

```
library(car)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   0.456 0.6369
      42
```

This test indicates that there is no significant issue with heteroskedasticity (variances differ between the groups) in our example ($p \approx 0.64$), so the model assumptions are valid. If a linear model without intercept is fitted, then the least squares slope estimates correspond to the sample averages for each group.

```
lengths.lm <- lm(length ~ 0 + major, data = lengths)
summary(lengths.lm)

Call:
lm(formula = length ~ 0 + major, data = lengths)


Residuals:
     Min       1Q   Median       3Q      Max
-14.3357  -3.4615  -0.4615   4.5385  11.5643


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
majorbiochemistry  168.167      1.324   127.0   <2e-16 ***
majorbiology       167.936      1.501   111.9   <2e-16 ***
majorgeography     169.462      1.558   108.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 5.617 on 42 degrees of freedom
Multiple R-squared:  0.999,Adjusted R-squared:  0.9989
F-statistic: 1.349e+04 on 3 and 42 DF,  p-value: < 2.2e-16
```

We see that the average height of a female biochemistry student is about 168.2 cm, of a biology student is 167.9 cm and of a geography student is 169.5 cm. Let us now test whether there is a significant difference in height between female students of the three majors, i.e. $H_0 : \mu_1 = \mu_2 = \mu_3$.

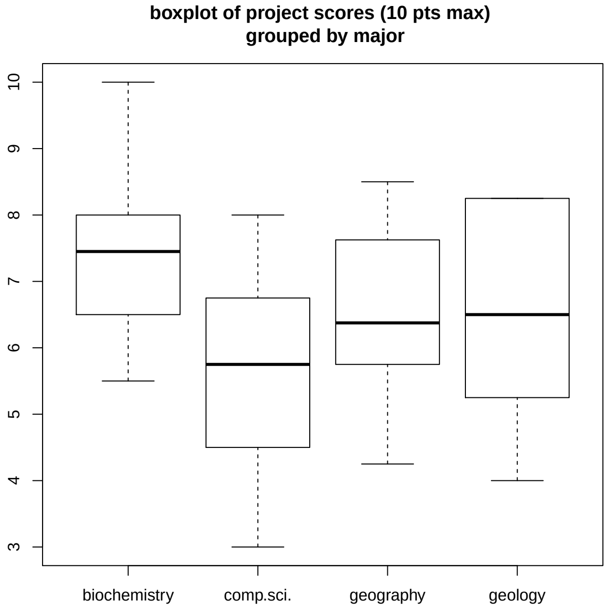
```
lengths.aov <- aov(length ~ major, data = lengths)
summary(lengths.aov)

            Df Sum Sq Mean Sq F value Pr(>F)
major        2   18.4   9.217   0.292  0.748
Residuals   42 1325.3  31.554
```

The ANOVA table shows that only a small portion of the total sum of squares can be explained by the differences between groups ($SSR \ll SSE$), and hence that there is no significant difference in heights, on average, between female students of different majors ($p \approx 0.75$).

### Another example: project results of bachelor students

The dataset ‘results’ contains the results of a statistics project done by bachelor students a few years ago. We wish to investigate whether there are significant differences, on average, in the results of students of different majors (Biochemistry, Computer Science, Geography, Geology).

```
boxplot(project10 ~ major, data=results,
        main = "boxplot of project scores (10 pts max) \n grouped by major")
```



The boxplot gives a first indication that it is not likely that the scores for all the groups are equal, on average. Also, the distributions of the data of the various groups seem to have more or less the same variance, with the project scores for the computer science major a bit more spread out, and they appear rather symmetric. Hence, the normality and homoskedasticity assumptions are likely.

```
library(car)
# Shapiro-Wilk tests omitted
leveneTest(project10 ~ major, data = results)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
```

```
group  3   1.071 0.3654
      91
```

The Shapiro-Wilk tests for normality verify the conclusions drawn from the boxplot, yielding $p$-values of 0.28, 0.37,0.14, and 0.12 for resp. the biochemistry, computer science, geography, and geology students. Hence, it is not unlikely that the observations originate from normally distributed populations. Levene's test for equality of variances indicates that there is little reason ($p \approx 0.37$) to doubt the homoskedasticity assumption. Hence, the model assumptions are validated and we can proceed with the analysis.

```
results.lm <- lm(project10 ~ 0 + major, data = results)
summary(results.lm)

Call:
lm(formula = project10 ~ 0 + major, data = results)


Residuals:
    Min      1Q  Median      3Q     Max 
-2.6471 -0.9485  0.1029  1.0202  2.7179 


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
majorbiochemistry   7.2821     0.2546   28.60   <2e-16 ***
majorcomp.sci.      5.6471     0.2311   24.44   <2e-16 ***
majorgeography      6.5625     0.2751   23.86   <2e-16 ***
majorgeology        6.2500     0.4492   13.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.347 on 91 degrees of freedom
Multiple R-squared:  0.9599,Adjusted R-squared:  0.9581 
F-statistic: 544.5 on 4 and 91 DF,  p-value: < 2.2e-16

results.aov <- aov(project10 ~ major, data = results)
```

```
summary(results.aov)

            Df Sum Sq Mean Sq F value   Pr(>F)    
major        3  41.87  13.957   7.687 0.000124 ***
Residuals   91 165.23   1.816                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The average scores for the various groups are: 7.3 for biochemistry students, 5.6 for computer science students, 6.6 for geography students and 6.2 for geology students. From the ANOVA table we see that these observed differences between the group means are (jointly) significant. Note however that this does not imply that ALL the pairwise differences between the groups are significant. To determine exactly which groups have significantly different results, we will have to perform a different hypothesis test. This will be covered in the next section.

## 15.3 Simultaneous estimation of contrasts

### 15.3.1 Contrasts

As mentioned in the previous section, testing the ANOVA null hypothesis $H_0 : \mu_1 = \ldots = \mu_k$ amounts to testing whether there is A significant difference between the means of the various groups, but it gives no indication as to which of the pairwise differences are significant and which aren't. This question can be expressed in terms of *contrasts*.

A contrast is a linear combination $\sum_{j=1}^{k} a_j\mu_j$ of the group (or treatment) means such that $\sum_{j=1}^{k} a_j = 0$.

In the example with the project scores, an example of a contrast would be the difference between the scores of the biochemistry major students and the computer science major, or between the computer science and geology major students.

Note that saying that the null hypotheses $H_0 : \mu_1 = \ldots = \mu_k$ holds (all group means are equal) is equivalent to saying that $\sum_{j=1}^{k} a_j\mu_j = 0$ for ALL linear combinations of the group means satisfying $\sum_{j=1}^{k} a_j = 0$. In short, it is equivalent to stating that all possible contrasts equal zero.

Since the sample group means $\bar{y}_j$ are unbiased estimators for the population means $\mu_j$, the linear combination $\sum_{j=1}^{k} a_j\bar{y}_j$ is an unbiased estimator for the contrast $\sum_{j=1}^{k} a_j\mu_j$. Furthermore, it is easily shown, under the ANOVA assumptions, that

$$\frac{\sum_{j=1}^{k} a_j\bar{y}_j - \sum_{j=1}^{k} a_j\mu_j}{\sqrt{\hat{\sigma}^2 \sum_{j=1}^{k} a_j^2/n_j}} \sim t_{n-k}.$$

Here, $\hat{\sigma}^2$ is the extension of the pooled variance to $k$ groups, i.e.

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 \tag{15.3.1}$$

Hence, we can find a $1-\alpha$ confidence interval for $\sum_{j=1}^{k} a_j\mu_j$ or perform the hypothesis test with $H_0 : \sum_{j=1}^{k} a_j\mu_j = 0$ in a similar manner as finding confidence intervals or performing hypothesis tests on a single mean.

In particular we can construct a $1-\alpha$ confidence interval for a pairwise difference between two group means, let's say $\mu_j - \mu_{j'}$ with $1 \leqslant j \neq j' \leqslant k$. It is given by

$$\bar{y}_j - \bar{y}_{j'} \pm t_{n-k,\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}.$$

### 15.3.2 The Bonferroni and Scheffé method

It is often the case in an ANOVA setting that we wish to draw conclusions about more than one contrast. However, if we use the confidence intervals at level $1-\alpha$ for each contrast, this will not lead to a simultaneous confidence level of $1-\alpha$.

One way to obtain the desired $1-\alpha$ simultaneous confidence level is by using the Bonferroni method. This method starts from the observation that, for $m$ events $A_j$, the inequality

$$P\left(\bigcap_{j=1}^{m} A_j\right) \geqslant 1 - \sum_{j=1}^{m} P(A_i^c)$$

holds. This implies that, if a confidence level of $1-\alpha$ is desired, and the probability of the $m$ events separately is at least $\gamma$, that $\gamma$ must satisfy $1-\alpha = 1 - m(1-\gamma)$, or that $\gamma = 1 - \frac{\alpha}{m}$.

If we want to determine which of the differences between the project scores are significant in the ANOVA setting, we would have to construct $m = k(k-1)/2$ simultaneous confidence intervals for each contrast $\mu_j - \mu_{j'}$, $1 \leqslant j \neq j' \leqslant k$, leading to a set of confidence intervals with individual confidence level $1 - \frac{2\alpha}{k(k-1)}$ if we want to achieve a simultaneous confidence level of $1-\alpha$. Hence, the confidence intervals will be wider compared to the non-simultaneous confidence intervals, especially when the number of groups $k$ becomes large.

A more elegant approach to this problem is to construct simultaneous confidence intervals for ALL contrasts, the Scheffé method. In short, we need a constant $M$ such that, with probability $1-\alpha$,

$$\sum_{j=1}^{k} a_j \bar{y}_j - M\hat{\sigma}\sqrt{\sum_{j=1}^{k} a_j^2/n_j} \leqslant \sum_{j=1}^{k} a_j \mu_j \leqslant \sum_{j=1}^{k} a_j \bar{y}_j + M\hat{\sigma}\sqrt{\sum_{j=1}^{k} a_j^2/n_j}$$

holds simultaneously for all $\boldsymbol{a} = (a_1, \ldots, a_k)^t \in \mathbb{R}^k$ satisfying $\sum_{j=1}^k a_j = 0$. First, denote

$$T_{\boldsymbol{a}}^2 = \frac{\left(\sum_{j=1}^k a_j \bar{y}_j - \sum_{j=1}^k a_j \mu_j\right)^2}{\hat{\sigma}^2 \sum_{j=1}^k a_j^2/n_j}$$

and $\mathcal{A} = \{\boldsymbol{a} \mid \sum_{j=1} a_j = 0\}$, then $M > 0$ has to satisfy

$$P\left(\sup_{\boldsymbol{a}\in\mathcal{A}} T_{\boldsymbol{a}}^2 \leqslant M^2\right) = 1 - \alpha.$$

Using the lemma below, with $c_j = n_j/\hat{\sigma}^2$ and $v_j = \bar{y}_j - \mu_j$, we find that

$$\sup_{\boldsymbol{a}\in\mathcal{A}} T_{\boldsymbol{a}}^2 = \frac{\sum_{j=1}^k n_j\big((\bar{y}_j - \bar{y}) - (\mu_j - \bar{\mu})\big)^2}{\hat{\sigma}^2}$$

with $\bar{\mu} = \frac{\sum n_j \mu_j}{\sum n_j}$. Under ANOVA assumptions it follows that $\sup_{\boldsymbol{a}\in\mathcal{A}} T_{\boldsymbol{a}}^2 \sim (k-1)F_{k-1,n-k}$, and hence $M^2 = (k-1)F_{k-1,n-k,\alpha}$.

**Lemma 10.** *Let $v_j \in \mathbb{R}$ and $c_j > 0$, $j = 1, \ldots, k$, then*

$$\max_{\boldsymbol{a}\in\mathcal{A}} \left\{ \frac{\left(\sum_{j=1}^k a_j v_j\right)^2}{\sum_{j=1}^k a_j^2/c_j} \right\} = \sum_{j=1}^k c_j (v_j - \bar{v}_c)^2$$

*where $\bar{v}_c = \sum c_j v_j / \sum c_j$. This upper bound is attained for $a_j = K c_j (v_j - \bar{v}_c)$ where $K \neq 0$ is constant.*

**Proof.**

Denote

$$b_j = \frac{a_j}{\sqrt{\sum_{j=1}^k a_j^2/c_j}}.$$

Then, we have to maximise $\left(\sum_{j=1}^k b_j v_j\right)^2$ over all $\boldsymbol{b} = (b_1, \ldots, b_k)^t$ satisfying $\sum_{j=1}^k b_j = 0$ and $\sum_{j=1}^k b_j^2/c_j = 1$. Using the Cauchy-Schwarz inequality, we find that

$$\begin{aligned}
\left(\sum_{j=1}^k b_j v_j\right)^2 &= \left(\sum_{j=1}^k b_j (v_j - \bar{v}_c)\right)^2 \\
&= \left(\sum_{j=1}^k \left(\frac{b_j}{\sqrt{c_j}}\right)\left(\sqrt{c_j}(v_j - \bar{v}_c)\right)\right)^2 \\
&\leqslant \sum_{j=1}^k (b_j^2/c_j) \sum_{j=1}^k c_j (v_j - \bar{v}_c)^2 \\
&= \sum_{j=1}^k c_j (v_j - \bar{v}_c)^2.
\end{aligned}$$

It is also easy to verify that this upper bound is attained at $a_j = Kc_j(v_j - \bar{v}_c)$.

□

### 15.3.3 Tukey's HSD method

The Scheffé method yields the confidence intervals for all possible contrasts. However, we are often interested in just a small subset of this, in the family of pairwise differences between the group means. Hence, the Scheffé method will often yield intervals with a confidence level greater than $1-\alpha$. By exploiting the structure of the specific multiple comparisons problem for the pairwise differences of means, it is often possible to find more narrow confidence intervals.

The Tukey multiple comparison method leads to confidence intervals for all pairwise comparisons of means with an exact $1-\alpha$ confidence level when all group sizes are equal, or $n_1 = n_2 = \ldots = n_k$. The procedure uses the *studentized range distribution.* When $X_1, \ldots, X_k$ are $k$ independent observations from $N(\mu, \sigma^2)$, then the standardised range $(\max_j(X_j) - \min_j(X_j))/\hat{\sigma}$ follows a studentized range distribution $q(k, v)$ with $v$ the degrees of freedom used to estimate $\sigma$. Here, we apply it to $X_j = \bar{Y}_j$ which under the ANOVA assumptions have a normal distribution with mean $\mu_j$ and common variance $\sigma^2/n_1$. Thus

$$\frac{\max(\bar{Y}_j - \mu_j) - \min(\bar{Y}_j - \mu_j)}{\hat{\sigma}/\sqrt{n_1}} \sim q(k, n_1k - k)$$

where $\hat{\sigma}$ is the pooled variance, as defined in (15.3.1). Consequently,

$$P\left(\frac{\max(\bar{Y}_j - \mu_j) - \min(\bar{Y}_j - \mu_j)}{\hat{\sigma}/\sqrt{n_1}} \leqslant q(k, n_1k - k, \alpha)\right) = 1 - \alpha.$$

It then follows that

$$P\left(\forall j, j' \quad \left|\frac{(\bar{Y}_j - \mu_j) - (\bar{Y}_{j'} - \mu_{j'})}{\hat{\sigma}/\sqrt{n_1}}\right| \leqslant q(k, n_1k - k, \alpha)\right) = 1 - \alpha$$

or equivalently

$$P\left(\forall j, j' \quad \left|\frac{(\bar{Y}_j - \mu_j) - (\bar{Y}_{j'} - \mu_{j'})}{\hat{\sigma}\sqrt{2/n_1}}\right| \leqslant \frac{q(k, n_1k - k, \alpha)}{\sqrt{2}}\right) = 1 - \alpha.$$

The Tukey HSD (Honest Significant Difference) confidence intervals with $1-\alpha$ confidence level are thus given by

$$\bar{y}_j - \bar{y}_{j'} \pm \frac{q(k, n_1k - k, \alpha)}{\sqrt{2}}\hat{\sigma}\sqrt{\frac{2}{n_1}}$$

for equal group sizes. When group sizes are unequal and do not differ much, the intervals

$$\bar{y}_j - \bar{y}_{j'} \pm \frac{q(k, n_1k - k, \alpha)}{\sqrt{2}}\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$$

can be used but typically have a confidence level of at least $1 - \alpha$. They are known as the *Tukey-Kramer intervals*.

In our example with the scores for the projects, we have found that there is a significant difference between the scores for the various majors, but not which majors differ significantly from each other. Let us examine now the pairwise differences.

```
summary(results.lm)

Call:
lm(formula = project10 ~ 0 + major, data = results)


Residuals:
    Min      1Q  Median      3Q     Max
-2.6471 -0.9485  0.1029  1.0202  2.7179


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
majorbiochemistry   7.2821     0.2546   28.60   <2e-16 ***
majorcomp.sci.      5.6471     0.2311   24.44   <2e-16 ***
majorgeography      6.5625     0.2751   23.86   <2e-16 ***
majorgeology        6.2500     0.4492   13.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.347 on 91 degrees of freedom
Multiple R-squared:  0.9599,Adjusted R-squared:  0.9581
F-statistic: 544.5 on 4 and 91 DF,  p-value: < 2.2e-16

print(TukeyHSD(results.aov),digits=5)

  Tukey multiple comparisons of means
    95% family-wise confidence level


Fit: aov(formula = project10 ~ major, data = results)
```

```
$major
                             diff       lwr      upr   p adj
comp.sci.-biochemistry -1.63508 -2.535048 -0.73512 0.00004
geography-biochemistry -0.71964 -1.700633  0.26135 0.22699
geology-biochemistry   -1.03214 -2.383432  0.31915 0.19595
geography-comp.sci.     0.91544 -0.024751  1.85563 0.05929
geology-comp.sci.       0.60294 -0.719027  1.92491 0.63241
geology-geography      -0.31250 -1.690908  1.06591 0.93385
```

From the results of the analysis we observe that the biochemistry students performed significantly better, on average, than the computer science students ($p \approx 4 \times 10^{-5}$). The other pairwise differences are not significant. However, the difference between the scores of the geography and the computer science students is close to being significant ($p \approx 0.06$). These results can be shown graphically as follows:

| Major: | comp.sci. | geol. | geog. | biochem. |
|---|---|---|---|---|
| Average: | 5.6 | 6.2 | 6.6 | 7.3 |

Reading this graph shows that among the computer science, geology students, and geography students, there are no significant differences, on average, between their project scores, and that there are also no significant differences, on average, between the scores of the geology, geography, and biochemistry students. Since there is no line directly connecting ‘biochem.’ to ‘comp.sci.’, this indicates the significant difference between the average scores of these groups.

# Chapter 16

# Transformations

In regression both transformations of the regressors and of the response variable are often useful to obtain a model that better fits the assumptions of the general linear model. First we will study the effect of a transformation on some variable $X$.

## 16.1 The family of power and root transformations

If $X$ is strictly positive, a useful group of transformations is the family of power and roots:

$$X \to X^{\lambda} \tag{16.1.1}$$

If $\lambda$ is negative, this transformation is an inverse power, e.g. $X^{-1} = 1/X, X^{-2} = 1/X^2$. If $\lambda$ is a fraction, the transformation represents a root, e.g. $X^{1/3} = \sqrt[3]{X}, X^{-1/2} = 1/\sqrt{X}$. A more convenient family of power transformations is defined as:

$$f(x) = x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0, x > 0 \\ \log(x) & \lambda = 0, x > 0 \end{cases} \tag{16.1.2}$$

Figure 4.1 shows several of those power transformations. Note that this family of transformations is monotone increasing in $X$, whereas the simple form (16.1.1) is decreasing if $\lambda < 0$. Moreover, for $\lambda = 0$, transformation (16.1.1) would be useless ($X^0 = 1$), whereas the logarithmic transformation is often very appropriate.

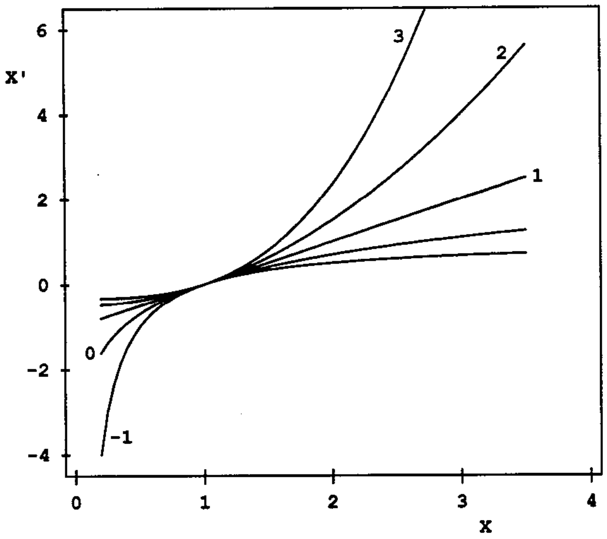


**Figure 4.1.** The family of power transformations $X'$ of $X$. The curve labeled $p$ is the transformation $X^{(p)}$, that is, $(X^p - 1)/p$; $X^{(0)}$ is $\log_e X$.

The effect of a power transformation is the following:

1. if $\lambda < 1$ large values of $X$ are compressed, whereas small values are more spread out.
2. if $\lambda > 1$ the inverse effect takes place: large values are more dispersed, whereas small values are compressed.

The first property makes the power transformation interesting if the distribution of $X$ is skewed to the right, the second if $X$ is skewed to the left (which occurs less in practice). Consider e.g. the distribution of income in Figure 4.2. The non-parametric density estimate clearly shows a right-tailed distribution. The log of income on the other hand is more symmetric, as illustrated in Figure 4.3.

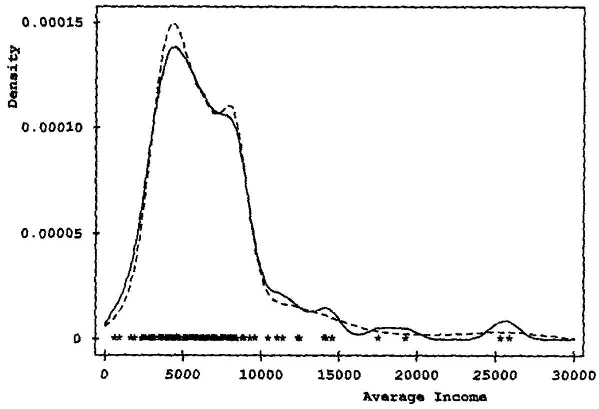


**Figure 4.2.** The distribution of income in the Canadian occupational prestige data. The solid line shows a kernel density estimate, the broken line an adaptive-kernel density estimate. The income values are displayed in the one-dimensional scatterplot at the bottom of the figure.

Remarks:

- If $X$ also contains negative values, we can first add a positive constant to each observation to make them all strictly positive. Example: $X = \{-3, -1, 4, 7, 9\}$ then $X + 4 = \{1, 3, 7, 11, 13\} > 0$.

- The power transformation is not very effective when the ratio between the largest and the smallest value is small. Example:

| | | | | | |
|---|---|---|---|---|---|
| $X$ : | 2001 | 2002 | 2003 | 2004 | 2005 |
| $\log(X)$ : | 7.6014 | 7.6019 | 7.6024 | 7.6029 | 7.6034 |

  Here, $2005/2001 = 1.002 \approx 1$. When we first subtract 2000 from the data, we obtain a ratio of $5/1 = 5$:

| | | | | | |
|---|---|---|---|---|---|
| $X - 2000$ : | 1 | 2 | 3 | 4 | 5 |
| $\log(X - 2000)$ : | 0 | 0.6931 | 1.0986 | 1.3863 | 1.6094 |

  and then the logarithmic transformation has more effect.

- An adequate power transformation can often be found in the range $-2.5 \leqslant \lambda \leqslant 3$. We usually select integer values, or simple fractions such as 1/2 or 1/3.

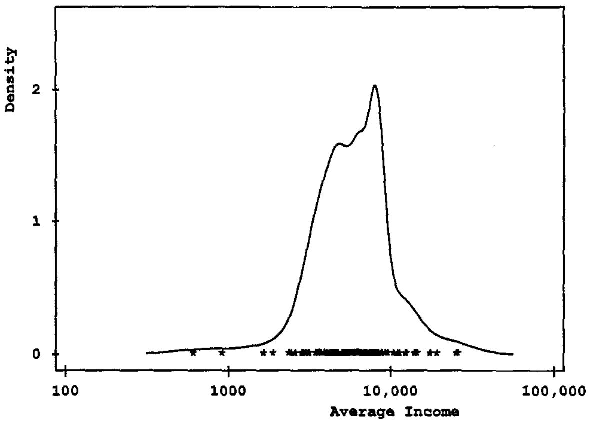


Figure 4.3. Adaptive-kernel density estimate for $\log_{10}$ average income in the Canadian occupational prestige data. The window width is 0.05 (on the log-income scale). A one-dimensional scatterplot of the data values appears at the bottom of the graph.

## 16.2 Transforming proportions

Power transformations are often not helpful for proportions, because these quantities are bounded below by 0 and above by 1. Also many sorts of rates (e.g. infant mortality rate per 1000 live births) are rescaled proportions. Or "the number of questions correct on an exam of fixed length" is essentially a proportion.

Common transformations for proportions are:

1. The logit transformation:

$$\text{logit}(P) = \log\Big(\frac{P}{1-P}\Big).$$

   The logit transformation (Figure 4.15) removes the boundaries of the scale, spreads out the tails of the distribution and makes the transformed variable symmetric around zero. It is the inverse of the cumulative distribution function of the logistic distribution and is essential in logistic regression.

2. The probit transformation uses the inverse of the standard normal distribution:

$$\text{probit}(P) = \Phi^{-1}(P)$$

   as is done in probit regression.

3. Also the arcsine-square-root $\arcsin(\sqrt{P})$ transformation has a similar shape.

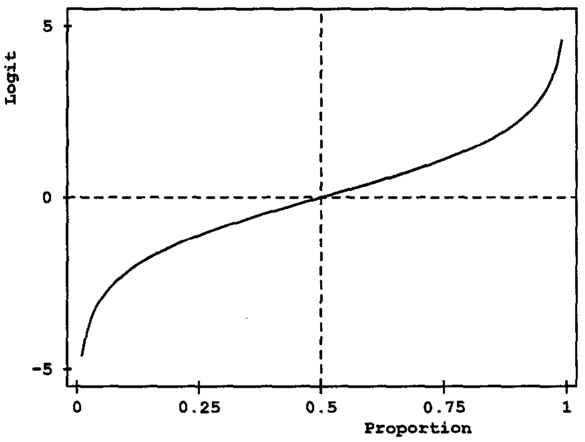


Figure 4.15. The logit transformation $\log[P/(1 - P)]$ of a proportion $P$.

## 16.3 Transformations in regression

### 16.3.1 Power transformation

When a normal quantile plot of the (standardized) residuals shows that they are not normally distributed, a power transformation of the response variable can sometimes correct this non-normality. A positive skew in the residuals e.g. will often be corrected by a log or a square-root transformation.

A power transformation can also be used to make a nonlinear relationship more nearly linear. Assume e.g. that $Y \approx 4X^2$. An $(x, y)$ plot will show a parabola. If we transform $Y$ into $Y' = \sqrt{Y}$ we obtain approximately a line since then $Y' \approx 2X$. The same happens when we transform $X$ into $X' = X^2$, yielding $Y \approx 4X'$. To select a good transformation, we can use Tukey and Mosteller's 'bulge rule', illustrated in Figure 4.6. Note that a transformation of one or more of the regressors is often preferred over a transformation of the response variable. Transforming $Y$ not only changes the shape of the error distribution (which is often the primary goal of the transformation), but it also has an effect on the relationship between $Y$ and the $X$'s.

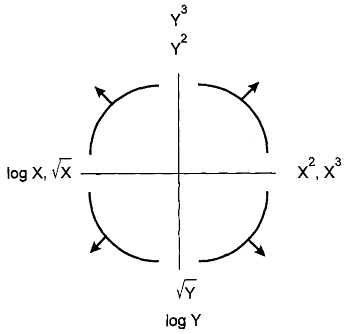


**Figure 4.6.** Tukey and Mosteller's "bulging rule": The direction of the "bulge" indicates the direction of the power transformation of $Y$ and/or $X$ to straighten the relationship between them.

### 16.3.2 Box-Cox transformation

A more sophisticated approach to transform the response variable is the *Box-Cox transformation.* The object of this transformation is to normalize the error distribution, to stabilize the error variance and to straighten the relation between $Y$ and $X$.

The general Box-Cox model assumes that a certain power transformation of $Y$ yields a general linear model with normal homoscedastic errors:

$$y_i^{(\lambda)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

with $\epsilon_i$ i.i.d. $N(0, \sigma^2)$, and $y_i^{(\lambda)}$ defined by (16.1.2). Note that all the $y_i$ must be positive, otherwise a constant should first be added. For a particular choice of $\lambda$, the maximized log-likelihood (profile log-likelihood) is

$$\log L(\lambda) = \text{const} - \frac{n}{2} \log \hat{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

where $\hat{\sigma}^2(\lambda) = \sum e_i^2(\lambda)/n$ and $e_i^2(\lambda)$ are the least squares residuals from the regression of $y^{(\lambda)}$ on the $X$'s. To find the maximum-likelihood estimator $\hat{\lambda}$ the profile log-likelihood is evaluated over a range of $\lambda$ values, say, between -2 and 2.

An approximate $(1-\alpha)100\%$ confidence interval for $\lambda$ is given by the interval of all $\lambda$ that satisfy

$$2(\log L(\hat{\lambda}) - \log L(\lambda)) \leqslant \chi^2_{1,\alpha}.$$

This stems from the fact that the likelihood-ratio statistic $G = -2(\log L(\lambda) - \log L(\hat{\lambda}))$ is asymptotically distributed as a $\chi^2_1$ distribution. Usually a plot of the log-likelihood $\log L(\lambda)$ versus $\lambda$ is made, with the 95% confidence interval for $\lambda$ indicated, as in Figure 12.8. Then a value of $\hat{\lambda}$ is selected which belongs to this confidence interval and which coincides with rounded numbers such as -1.5, -1, -0.5, 0, 0.5, 1, ....

Remark that the Box-Cox methodology assumes that there exists a $\lambda$ such that on the transformed scale the assumptions of the general linear model are fulfilled. It is thus important to verify (using residual plots and normal quantile

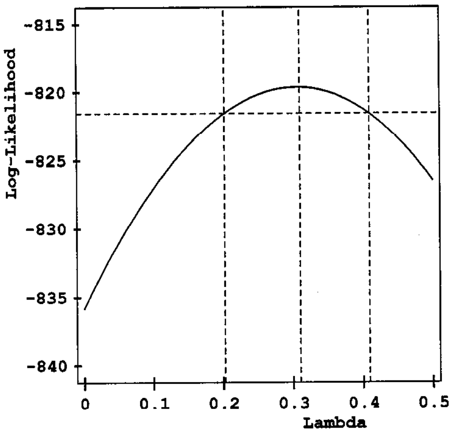


**Figure 12.8.** Box-Cox transformations for Ornstein's interlocking-directorate regression. The maximized log likelihood is plotted against the transformation parameter $\lambda$. The intersection of the line near the top of the graph with the log likelihood curve marks off a 95% confidence interval for $\lambda$. The maximum of the log likelihood corresponds to the MLE of $\lambda$.

plots) whether the proposed transformation indeed improved the appropriateness of the model assumptions.

## 16.4 Nonconstant variance

### 16.4.1 Detecting heteroscedasticity

The second Gauss-Markov condition (12.1.3) states that the error variance is everywhere the same around the regression surface. Nonconstant error variance is called *heteroscedasticity.* In that case the least squares estimator is still unbiased and consistent, but the variances of the parameter estimates tend to be large and thus affect the tests of hypothesis substantially. Also $s^2(X^tX)^{-1}$ need no longer be an unbiased estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{LS}}$.

Heteroscedasticity can sometimes be detected if we understand the underlying situation. For example, if the responses are counts, then it is likely that the response variable is approximately Poisson distributed, for which the variance is equal to its expected value, i.e. $\text{Var}[Y|\boldsymbol{x}_i] = E[Y|\boldsymbol{x}_i]$. Hence it cannot be expected that $\sigma_i^2 := \text{Var}[Y|\boldsymbol{x}_i]$ will be constant. Or consider house prices or income. Less expensive houses or low incomes usually show less variation than more expensive houses or higher incomes.

Heteroscedasticity can often be seen through residual plots. If $\sigma_i^2$ varies with $E[Y|\boldsymbol{x}_i]$, then a plot of the residuals, which are estimates of $\epsilon_i$, against the fitted values $\hat{y}_i$, which are estimates of $E[Y|\boldsymbol{x}_i]$, might reveal that the residuals are more spread out for some values of $\hat{y}_i$ than for others. A typical example is shown in Figure 6.2. Because the least-squares residuals have unequal variances even in the homoscedastic case, it is preferable to use the standardized residuals.

Also plots of the *absolute* (standardized) residuals or the *squared* residuals versus $\hat{y}_i$ are often used. Since

$$\sigma_i^2 = E[\epsilon_i^2] - (E[\epsilon_i])^2 = E[\epsilon_i^2]$$

we notice that the squared residual $e_i^2$ is an estimator of the variance $\sigma_i^2$, and that the absolute residual $|e_i|$ is an estimate of the standard deviation $\sigma_i$.

Sometimes the variance of the errors varies with one or more of the regressors $X$. Therefore residual plots versus each of the independent variables might also

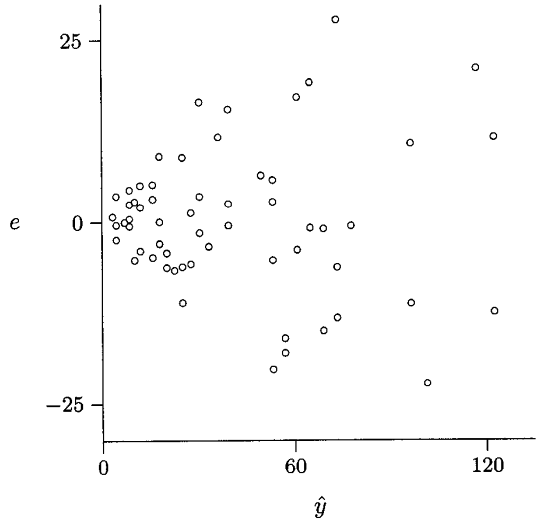


EXHIBIT 6.2: Plot of Residuals against Predicted for Speed-Braking Distance Data

be appropriate.

Note that heteroscedasticity can also be due to the omission of an important variable or interaction term from the model. This can however be very difficult to detect.

### 16.4.2 Variance-stabilizing transformations

Variance-stabilizing transformations can be used when the variance of the response variable depends on its expected value, e.g. when $Y_x := Y|\boldsymbol{x}_x$ has a Poisson or an exponential distribution.

Assume that $Y_x$ has mean $\mu_x$ and variance $\sigma_x^2$ and that $g$ is a function of $Y_x$ such that $E[g(Y_x)]$ can be well approximated with $g(E[Y_x]) = g(\mu_x)$. The Taylor expansion of $g(Y_x)$ around $\mu_x$ gives:

$$g(Y_x) \approx g(\mu_x) + (Y_x - \mu_x)g'(\mu_x).$$

Consequently

$$\begin{aligned}\mathrm{Var}(g(Y_x)) = E[(g(Y_x) - E[g(Y_x)])^2] &\approx E[(g(Y_x) - g(\mu_x))^2] \\ &\approx E[((Y_x - \mu_x)g'(\mu_x))^2] \\ &= g'(\mu_x)^2 E[(Y_x - \mu_x)^2] \\ &= g'(\mu_x)^2\sigma_x^2.\end{aligned}$$

If we want to apply a transformation such that $\mathrm{Var}(g(Y_x)) = c^2$ for a certain constant $c$, we thus have to find a function $g$ such that $g'(\mu_x)^2\sigma_x^2 = c^2$, or

$$g(\mu_x) = c\int \frac{d\mu_x}{\sigma_x}.$$

**Examples**.

1. Assume that $Y_x$ is Poisson distributed with $E(Y_x) = \lambda_x = \mathrm{Var}(Y_x)$. Then

   $$g(\lambda_x) = \int \frac{d\lambda_x}{\sqrt{\lambda_x}} = 2\sqrt{\lambda_x}.$$

   Hence taking the square root of $Y$ will lead to a more constant variance.

2. If the response is a proportion, $E(Y_x) = p_x$ and $\mathrm{Var}(Y_x) \sim p_x(1-p_x)$, thus

   $$g(p_x) = \int \frac{dp_x}{\sqrt{p_x(1-p_x)}} = \arcsin\sqrt{p_x}$$

   which gives us again the arcsine-square-root transformation.

### 16.4.3 Weighted least squares regression

Transformations in $Y$, such as the Box-Cox transformation, might be helpful to reduce unequal error variances, but they have the disadvantage of changing the relation between the response and the independent variables. If the linear relationship seems appropriate, but one is left with unequal error variances, the **Weighted Least Squares** procedure is recommended.

For this we consider the *generalized linear regression model*:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \tag{16.4.1}$$

for $i = 1, \ldots, n$, with

$$\boxed{\epsilon_i \text{ independent } N(0, \sigma_i^2).}$$

Moreover we assume that the variances $\sigma_i^2$ are known up to a constant of proportionality:

$$\boxed{\sigma_i^2 = \sigma^2 / w_i}$$

with $w_i$ known weights, and $\sigma^2$ unknown. The ratio of two variances $\sigma_j^2$ and $\sigma_k^2$ is then indeed known, since $\sigma_j^2/\sigma_k^2 = w_k/w_j$. Let $W = \operatorname{diag}(w_1, \ldots, w_n)$ then

$$\Sigma(\boldsymbol{\varepsilon}) = \sigma^2 W^{-1}. \tag{16.4.2}$$

This generalized linear model (16.4.1) is equivalent to the general linear model:

$$\sqrt{w_i} y_i = \beta_0 \sqrt{w_i} + \beta_1 \sqrt{w_i} x_{i1} + \beta_2 \sqrt{w_i} x_{i2} + \ldots + \beta_{p-1} \sqrt{w_i} x_{i,p-1} + \sqrt{w_i}\epsilon_i$$

since $\sqrt{w_i}\epsilon_i$ are independent $N(0, \sigma^2)$. Or

$$\boldsymbol{y}^{(W)} = X^{(W)} \boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(W)}$$

with $\boldsymbol{y}^{(W)} = W^{1/2}\boldsymbol{y}, X^{(W)} = W^{1/2}X, \boldsymbol{\varepsilon}^{(W)} = W^{1/2}\boldsymbol{\varepsilon}, \Sigma(\boldsymbol{\varepsilon}^{(W)}) = W^{1/2}\Sigma(\boldsymbol{\varepsilon})W^{1/2} = \sigma^2 I_n$.

The least-squares estimator thus becomes

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= ((X^{(W)})^t X^{(W)})^{-1} (X^{(W)})^t \boldsymbol{y}^{(W)} \\
&= (X^t W^{1/2} W^{1/2} X)^{-1} X^t W^{1/2} W^{1/2} \boldsymbol{y} \\
&= (X^t W X)^{-1} X^t W \boldsymbol{y}
\end{aligned}$$

Since for any $\boldsymbol{\beta}$, $\hat{y}^{(W)}(\boldsymbol{\beta}) = X^{(W)}\boldsymbol{\beta} = W^{1/2}X\boldsymbol{\beta}$ we see that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i^{(W)} - \hat{y}_i^{(W)})^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i} \boldsymbol{x}_i^t \boldsymbol{\beta})^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2.\end{aligned}$$

We thus apply the ordinary least squares estimator (OLS) on the weighted variables, where observations with a large variance get a small weight and those with a small variance a large weight ($w_i \sim 1/\sigma_i^2$).

It can be shown that $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is the BLUE estimator of $\boldsymbol{\beta}$ in the generalized linear model (16.4.1) that satisfies (16.4.2). The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is given by

$$\begin{aligned}\Sigma(\hat{\boldsymbol{\beta}}_{\text{WLS}}) &= (X^t W X)^{-1} X^t W \sigma^2 W^{-1} W^t X (X^t W X)^{-1} \\ &= \sigma^2 (X^t W X)^{-1}.\end{aligned}$$

The residual vector is

$$\begin{aligned}\boldsymbol{e}^{(W)} &= \boldsymbol{y}^{(W)} - \hat{\boldsymbol{y}}^{(W)} \\ &= W^{1/2}\boldsymbol{y} - W^{1/2} X \hat{\boldsymbol{\beta}}_{\text{WLS}} \\ &= W^{1/2}(\boldsymbol{y} - \hat{\boldsymbol{y}}) \qquad \text{with } \hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}_{\text{WLS}}. \qquad (16.4.3)\end{aligned}$$

An unbiased estimate of $\sigma^2$ is given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^{n} (e_i^{(W)})^2 \\ &= \frac{1}{n-p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2\end{aligned}$$

from which

$$\hat{\Sigma}(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = \hat{\sigma}^2 (X^t W X)^{-1} \qquad (16.4.4)$$

follows.

**Estimation of the variance function.**

Because the error variances $\sigma_i^2$ or the weights $w_i$ are in general not known, we are forced to estimate them. As the $\sigma_i^2$ often vary with one or several predictor variables or with the mean response $E(y_i)$, the following procedure might be helpful:

1. Fit the regression model by ordinary least squares (OLS) and analyze the residuals.
2. Regress the squared residuals or the absolute residuals on the fitted values or one or several independent variables. This makes sense because the squared residuals $e_i^2$ estimate the variances $\sigma_i^2$, while the absolute residuals $|e_i|$ are estimates of the standard deviations $\sigma_i$.
3. Use the fitted values from the estimated variance or standard deviation to obtain the weights $w_i$.
4. Fit the regression model by WLS and analyze the residuals.

If $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ differs significantly from $\hat{\boldsymbol{\beta}}_{\text{LS}}$, it is advisable to iterate the WLS process by using the residuals from the WLS fit to reestimate the variance or standard deviation function and then obtain revised weights. This iterative process is called *iteratively reweighted least squares.*

Note that the estimated standard deviations of the coefficients, derived from (16.4.4), are now only approximate, because the estimation of the variances $\sigma_i^2$ has introduced another source of variability. The approximation will often be quite good when the sample size is not too small.

**Example: Blood data set.**

Consider the Blood data set containing the age and the diastolic blood pressure of 54 healthy women.

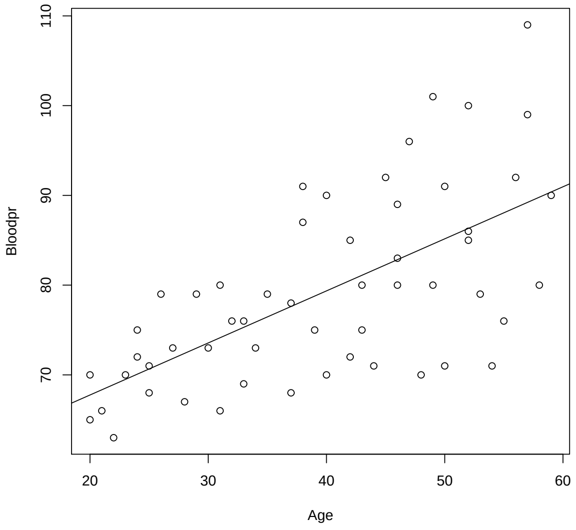
```
lmba <- lm(Bloodpr~Age)
summary(lmba)

Call:
lm(formula = Bloodpr ~ Age)


Residuals:
     Min       1Q   Median       3Q      Max 
-16.4786  -5.7877  -0.0784   5.6117  19.7813 


Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 56.15693    3.99367  14.061  < 2e-16 ***
Age          0.58003    0.09695   5.983 2.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-squared:  0.4077,Adjusted R-squared:  0.3963 
F-statistic: 35.79 on 1 and 52 DF,  p-value: 2.05e-07
```

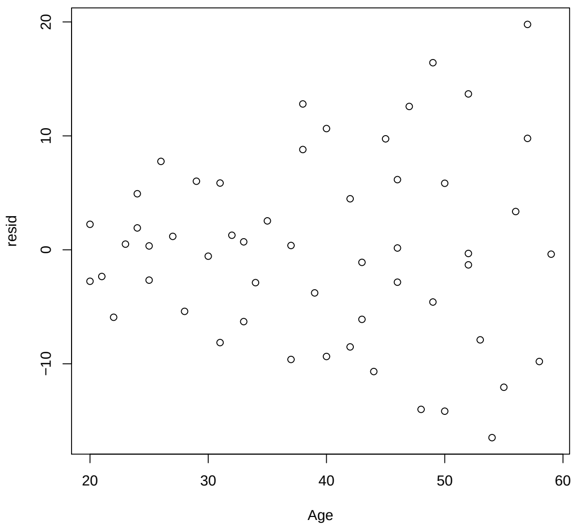
The OLS analysis yields the fitted model:

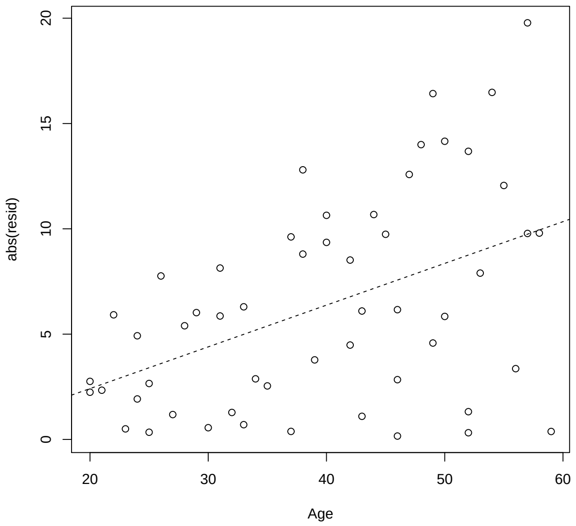
$$\texttt{Bloodpressure} = 56.16 + 0.58\texttt{Age}$$

The scatter plot of the data, the plot of residuals versus age, and absolute residuals versus age using OLS clearly demonstrate heteroscedasticity.

```
plot(Age,Bloodpr)
abline(lm(Bloodpr~Age))
resid <- residuals(lmba)
plot(Age,resid)
plot(Age,abs(resid))
```







When we regress the absolute residuals versus `Age` we obtain the estimated expected standard deviation:

$$\hat{\sigma}_i = s_i = -1.55 + 0.198\texttt{Age}$$

```
stdev <- lm(abs(resid)~Age)
abline(lm(abs(resid)~Age),lty=2)
summary(stdev)

Call:
lm(formula = abs(resid) ~ Age)


Residuals:
    Min      1Q  Median      3Q     Max
-9.7639 -2.7882 -0.1587  3.0757 10.0350


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.54948    2.18692  -0.709  0.48179
Age          0.19817    0.05309   3.733  0.00047 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.461 on 52 degrees of freedom
Multiple R-squared:  0.2113,Adjusted R-squared:  0.1962
F-statistic: 13.93 on 1 and 52 DF,  p-value: 0.0004705
```

Finally we apply WLS with weights $w_i = 1/s_i^2$.

```
weightblood <- 1/stdev$fitted^2
wlmba <- lm(Bloodpr~Age, weights=weightblood)
summary(wlmba)

Call:
lm(formula = Bloodpr ~ Age, weights = weightblood)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-2.0230 -0.9939 -0.0327  0.9250  2.2008 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 55.56577    2.52092  22.042  < 2e-16 ***
Age          0.59634    0.07924   7.526 7.19e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-squared:  0.5214,Adjusted R-squared:  0.5122
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

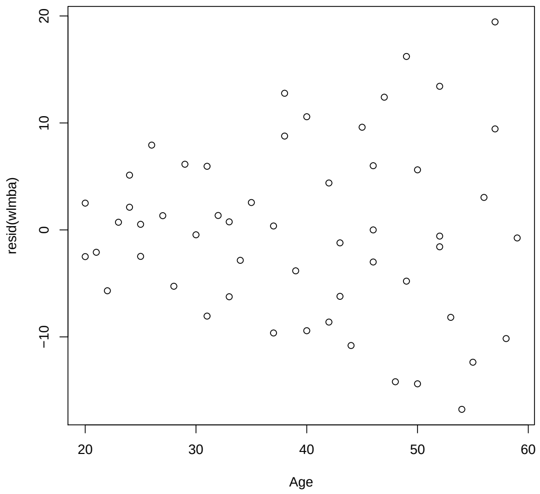
The final estimated regression line is thus:

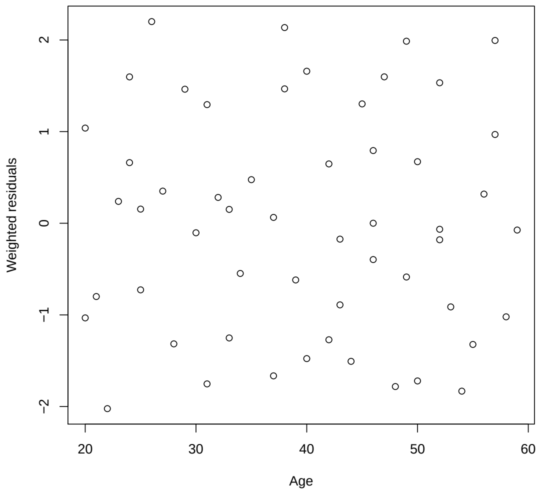
$$\texttt{Bloodpressure} = 55.56 + 0.596\texttt{Age}$$

which is not so different from the OLS line. Therefore, an extra reweighting should not be considered. We see that the standard error of $\hat{\beta}_1$ has decreased from 0.097 in the OLS analysis to 0.079 in the WLS analysis. Consequently the (approximate) 95% confidence interval for $\beta_1$, which is $0.596 \pm t_{53,0.025}0.079$, is also smaller than the confidence interval based on the ordinary analysis.

Finally we test whether the heteroscedasticity has gone now. Note that R computes the residuals as $e_i(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = y_i - \hat{y}_i(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}})$. They will still show the heteroscedasticity. By (16.4.3) it is the residuals $e_i^{(W)}(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = \sqrt{w_i}(y_i - \hat{y}_i(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}))$ who should have constant variance.

```
plot(Age,resid(wlmba))
plot(Age,resid(wlmba)*sqrt(weightblood),ylab="Weighted residuals")
```





# Chapter 17

# Variable selection methods

## 17.1 Reduction of explanatory variables

Variable reduction is often recommended because:

- a regression model with many predictors may be difficult and expensive to maintain
- a model with a limited number of regressors is easier to work with and to understand
- the presence of highly intercorrelated explanatory variables may increase the variance of the regression coefficients, and increase the problem of roundoff errors.
- the presence of explanatory variables that are not related to the response variable increase the variance of the predicted values and hence, decrease the model's predictive ability.

On the other hand, omitting important variables (or *latent explanatory variables*) leads to biased estimates of the regression coefficients, the error variance, the mean responses and predictions of new observations.

This is illustrated in Figure 4.3. If too many variables are selected, too much of the redundancy in the $x$-variables is used and the solution becomes overfitted. The regression equation will be very data dependent and gives poor prediction results. If too few variables are retained, it is called underfitting which means that the model is not large enough to capture the important variability in the data. The optimal number of variables is usually found in between the two extremes. It is therefore often a good idea to consider several 'good' subsets of explanatory variables.

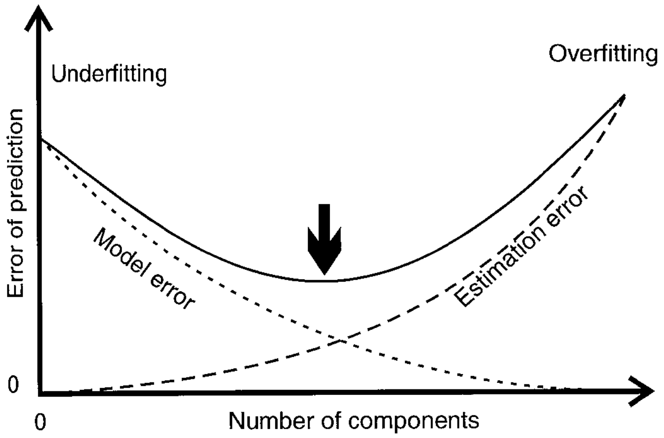


**Figure 4.3. The estimation error (– – –) and model error (· · ·) contribute to the prediction error (———). If too large a model is used, an overfitted solution is the result. The opposite is called underfitting. Usually a medium size model is to be preferred. Reproduced with permission from H. Martens and T. Næs, *Multivariate Calibration* (1989). © John Wiley & Sons Limited.**

### 17.1.1 Surgical unit example

We consider the surgical unit example to illustrate the model-building process and some variable reduction methods.

The data set consists of $n = 108$ patients undergoing a liver operation. A hospital surgical unit was interested to predict the survival time of patients. Available explanatory variables are:

| | |
|---|---|
| $X_1$: | blood clotting score |
| $X_2$: | prognostic index (including the age of the patient) |
| $X_3$: | enzyme function test score |
| $X_4$: | liver function test score |

An exploratory data analysis shows that a transformation of the response variable into $Y' = \log_{10} Y$ makes the distribution of the residuals in the first order linear model more normal (symmetric), eliminates the need for interaction

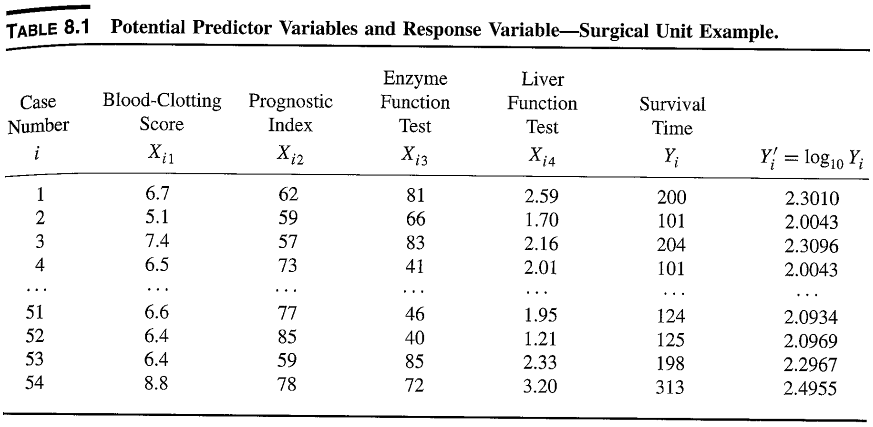
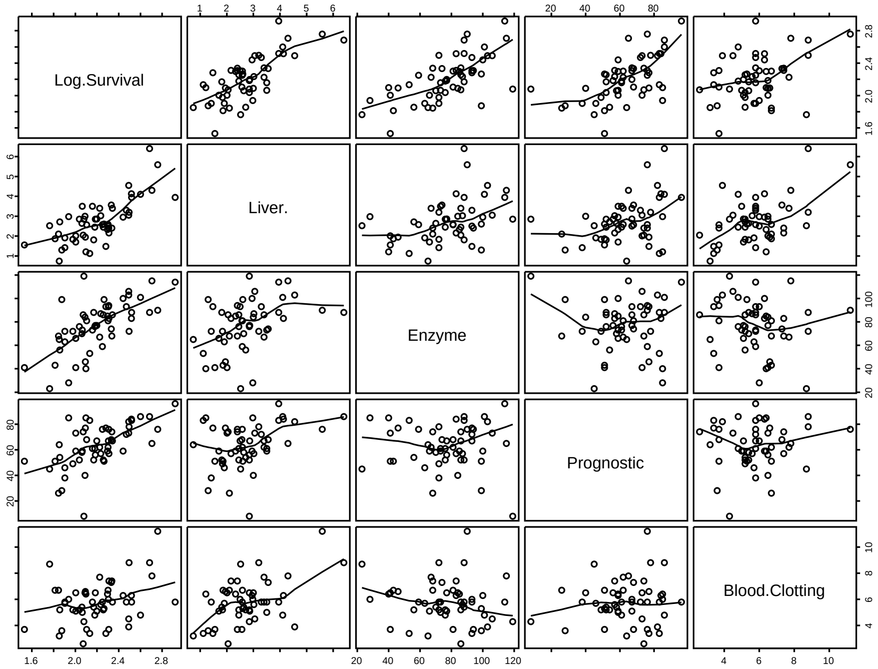


**TABLE 8.1** **Potential Predictor Variables and Response Variable—Surgical Unit Example.**

| Case Number $i$ | Blood-Clotting Score $X_{i1}$ | Prognostic Index $X_{i2}$ | Enzyme Function Test $X_{i3}$ | Liver Function Test $X_{i4}$ | Survival Time $Y_i$ | $Y_i' = \log_{10} Y_i$ |
|---|---|---|---|---|---|---|
| 1 | 6.7 | 62 | 81 | 2.59 | 200 | 2.3010 |
| 2 | 5.1 | 59 | 66 | 1.70 | 101 | 2.0043 |
| 3 | 7.4 | 57 | 83 | 2.16 | 204 | 2.3096 |
| 4 | 6.5 | 73 | 41 | 2.01 | 101 | 2.0043 |
| ... | ... | ... | ... | ... | ... | ... |
| 51 | 6.6 | 77 | 46 | 1.95 | 124 | 2.0934 |
| 52 | 6.4 | 85 | 40 | 1.21 | 125 | 2.0969 |
| 53 | 6.4 | 59 | 85 | 2.33 | 198 | 2.2967 |
| 54 | 8.8 | 78 | 72 | 3.20 | 313 | 2.4955 |

terms and makes the relation between every regressor and the response variable more linear.



We take the first half of the data as training data to build a regression model. Table 8.1 shows part of these data which are graphically shown in the scatterplot matrix The first order linear model (full model) for the data yields:

```
surgicalunit <- surgicalunit[,-(5:8)]
surgicalunit1 <- surgicalunit[1:54,]
attach(surgicalunit1)
```

```
surg.full <- lm(Log.Survival ~ Blood.Clotting + Prognostic + Enzyme + Liver)

Call:
lm(formula = Log.Survival ~ Blood.Clotting + Prognostic + Enzyme +
    Liver)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43500 -0.17591 -0.02091  0.18400  0.56192 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.851948   0.266258  14.467  < 2e-16 ***
Blood.Clotting 0.083684   0.028833   2.902  0.00554 ** 
Prognostic     0.012665   0.002315   5.471 1.51e-06 ***
Enzyme         0.015632   0.002100   7.443 1.37e-09 ***
Liver          0.032161   0.051465   0.625  0.53493    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2509 on 49 degrees of freedom
Multiple R-squared:  0.7592,Adjusted R-squared:  0.7396 
F-statistic: 38.62 on 4 and 49 DF,  p-value: 1.388e-14
```

## 17.2 All-possible-regressions procedure for variable reduction

The first procedure considers all possible subsets of the pool of potential predictors, including products of observed variables for interaction or higher-order terms. A regression is then performed for each subset and evaluated through some criterion. Note that the number of subsets grows exponentially with the number of predictors. For example there are $2^{10} = 1024$ regressions to be inspected for a data set with 10 possible predictors. Since it would not be possible to examine each model carefully, it is useful to compute one simple criterion for each model and then to select a few good subsets.

### 17.2.1 $R_p^2$ criterion

The $R_p^2$ coefficient (of multiple determination) reflects the proportion of the variance of the response variable that is explained by the linear model with $p$ coefficients. Resuming (12.3.2)

$$\boxed{R_p^2 = 1 - \frac{\mathrm{SSE}_p}{\mathrm{SST}}}$$

Subsets with a high $R_p^2$ coefficient (or equivalently with a low $\mathrm{SSE}_p$) are considered good. We know that $R_p^2$ always increases if we include additional variables in the model. Therefore it makes no sense to maximize $R_p^2$, but we should find the point where adding more variables is not worthwhile because it leads to a very small increase in $R_p^2$.

Figure 8.4 contains a plot of the $R_p^2$ values versus $p$, and its maximum for each number of parameters $p$ in the model. From this plot it can be seen that the inclusion of the fourth variable `Liver` does not lead to a large increase in the explained variance. This might be surprising because the correlation between `Liver` and `Log.Survival` is the largest among all the pairwise correlations with the response variable. This indicates that $X_1, X_2$ and $X_3$ contain much of the information presented by $X_4$.

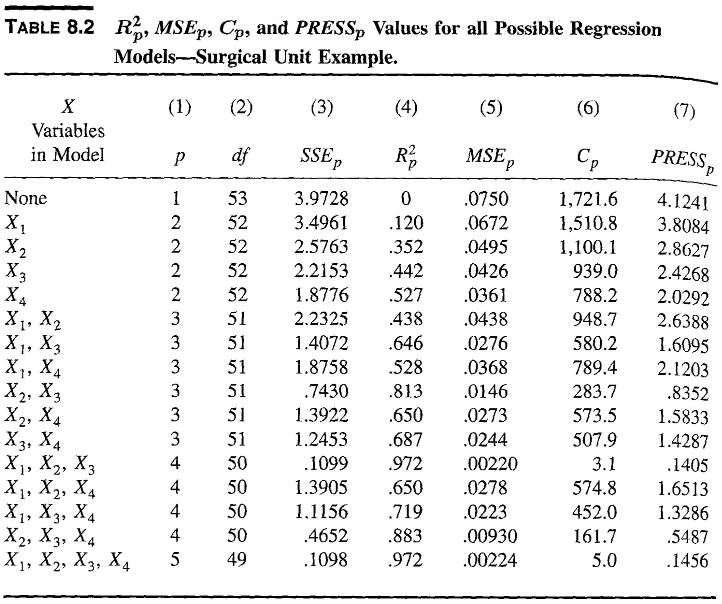
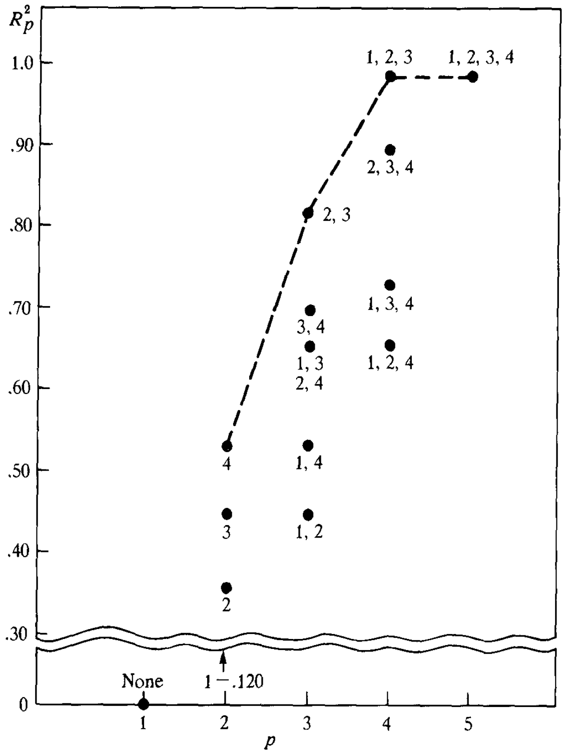


**TABLE 8.2** $R^2_p$, $MSE_p$, $C_p$, and $PRESS_p$ Values for all Possible Regression Models—Surgical Unit Example.

| $X$<br>Variables<br>in Model | (1)<br>$p$ | (2)<br>$df$ | (3)<br>$SSE_p$ | (4)<br>$R^2_p$ | (5)<br>$MSE_p$ | (6)<br>$C_p$ | (7)<br>$PRESS_p$ |
|---|---|---|---|---|---|---|---|
| None | 1 | 53 | 3.9728 | 0 | .0750 | 1,721.6 | 4.1241 |
| $X_1$ | 2 | 52 | 3.4961 | .120 | .0672 | 1,510.8 | 3.8084 |
| $X_2$ | 2 | 52 | 2.5763 | .352 | .0495 | 1,100.1 | 2.8627 |
| $X_3$ | 2 | 52 | 2.2153 | .442 | .0426 | 939.0 | 2.4268 |
| $X_4$ | 2 | 52 | 1.8776 | .527 | .0361 | 788.2 | 2.0292 |
| $X_1$, $X_2$ | 3 | 51 | 2.2325 | .438 | .0438 | 948.7 | 2.6388 |
| $X_1$, $X_3$ | 3 | 51 | 1.4072 | .646 | .0276 | 580.2 | 1.6095 |
| $X_1$, $X_4$ | 3 | 51 | 1.8758 | .528 | .0368 | 789.4 | 2.1203 |
| $X_2$, $X_3$ | 3 | 51 | .7430 | .813 | .0146 | 283.7 | .8352 |
| $X_2$, $X_4$ | 3 | 51 | 1.3922 | .650 | .0273 | 573.5 | 1.5833 |
| $X_3$, $X_4$ | 3 | 51 | 1.2453 | .687 | .0244 | 507.9 | 1.4287 |
| $X_1$, $X_2$, $X_3$ | 4 | 50 | .1099 | .972 | .00220 | 3.1 | .1405 |
| $X_1$, $X_2$, $X_4$ | 4 | 50 | 1.3905 | .650 | .0278 | 574.8 | 1.6513 |
| $X_1$, $X_3$, $X_4$ | 4 | 50 | 1.1156 | .719 | .0223 | 452.0 | 1.3286 |
| $X_2$, $X_3$, $X_4$ | 4 | 50 | .4652 | .883 | .00930 | 161.7 | .5487 |
| $X_1$, $X_2$, $X_3$, $X_4$ | 5 | 49 | .1098 | .972 | .00224 | 5.0 | .1456 |

```
print(cor(surgicalunit1[, -5]), digits = 2)

               Blood.Clotting Prognostic Enzyme Liver
Blood.Clotting           1.00      0.090 -0.150  0.50
Prognostic               0.09      1.000 -0.024  0.37
Enzyme                  -0.15     -0.024  1.000  0.42
Liver                    0.50      0.369  0.416  1.00
Log.Survival             0.25      0.470  0.654  0.65
               Log.Survival
Blood.Clotting         0.25
Prognostic             0.47
Enzyme                 0.65
Liver                  0.65
Log.Survival           1.00
```

FIGURE 8.4 $R_p^2$ Plot for Surgical Unit Example.



## 17.2.2 MSE$_p$ criterion

Because $R_p^2$ does not take account of the number of parameters in the regression model and because it can never decrease as $p$ increases, the adjusted $R_a^2$, defined in (12.3.4), can be used as an alternative criterion:

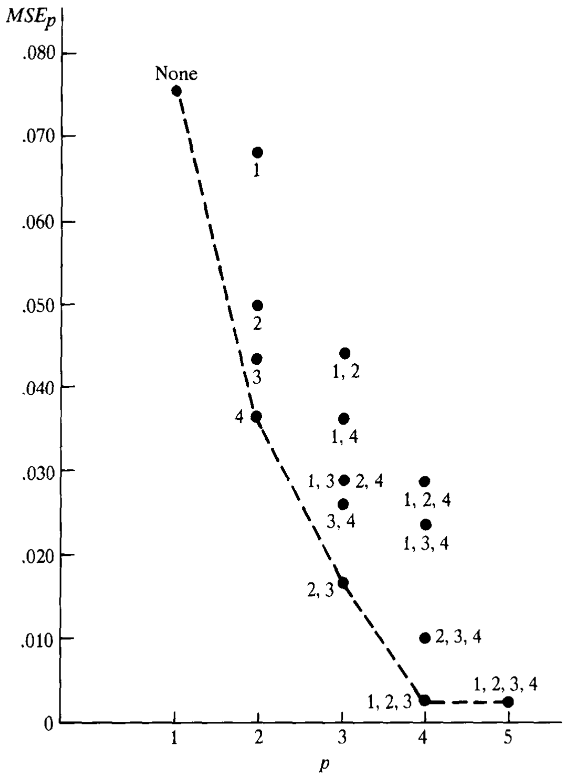
$$\begin{aligned} R_a^2 &= 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} \\ &= 1 - \frac{\text{MSE}_p}{\text{SST}/(n-1)}. \end{aligned}$$

Since SST remains constant over all regression models, considering the adjusted $R_a^2$ is equivalent to looking at the mean squared error $\text{MSE}_p = \hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n e_{i,p}^2$ with $e_{i,p}$ the residuals from a model with $p$ regressors. Although

$$\text{SSE}_{p+1} = \sum e_{i,p+1}^2 \leqslant \text{SSE}_p = \sum e_{i,p}^2,$$

the denominator $n-(p+1)$ of the larger model is smaller than the denominator $n-p$ of the smaller model. Hence, if the decrease in SSE is very small, the loss of one degree of freedom can result in an $\text{MSE}_{p+1} > \text{MSE}_p$.

**FIGURE 8.5** *$MSE_p$* **Plot for Surgical Unit Example.**



When we consider the MSE criterion we can thus look for the subset(s) with minimal MSE, or whose MSE is very close to the minimum. Figure 8.5 shows the $\mathrm{MSE}_p$ plot for the surgical unit data. Again, the fourth explanatory variable `Liver` appears not to be needed in the model.

### 17.2.3 Mallows' $C_p$

The $C_p$ statistic, suggested by C.L. Mallows, has the form

$$\boxed{C_p = \frac{\text{SSE}_p}{s^2} - (n - 2p)} \tag{17.2.1}$$

with $s^2$ the MSE from the 'largest' model, presumed to be a reliable unbiased estimate of the error variance $\sigma^2$.

It is an estimator of the standardized total mean squared error of the fitted values:

$$\begin{aligned}\gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{MSE}(\hat{y}_i)\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}E((\hat{y}_i - E(y_i|\boldsymbol{x}_i))^2)\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\text{bias}^2(\hat{y}_i) + \text{Var}(\hat{y}_i)).\end{aligned}$$

It can be shown that $\gamma_p$ can be expressed as

$$\gamma_p = \frac{E(\text{SSE}_p)}{\sigma^2} - (n - 2p)$$

from which (17.2.1) follows. In order to minimize the total mean squared error we prefer models with small $C_p$ value.

Note that there are several $C_p$ statistics for each $p$, except for the $C_p$ value when all variables are included, let's say $C_P$. At this model

$$C_P = \frac{(n-P)\text{MSE}_P}{s^2} - n + 2P = P$$

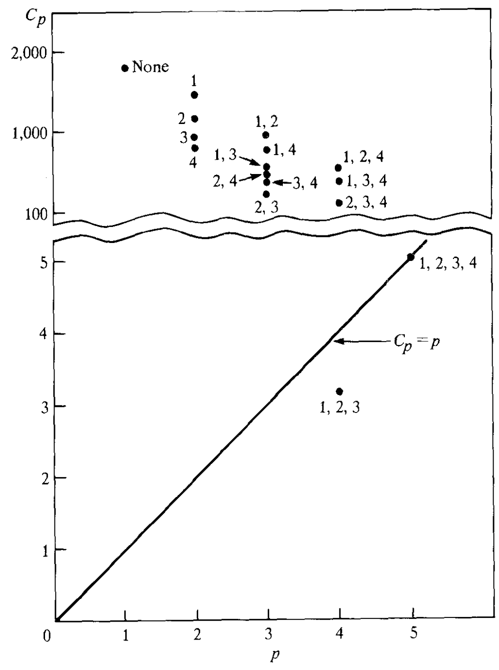
since $s^2 = \text{MSE}_P$.

If a model with $p$ parameters is adequate, $E(\text{SSE}_p) \approx (n-p)\sigma^2$. Since we assume that also $E(s^2) = \sigma^2$, approximately $E(\text{SSE}_p/s^2) \approx n - p$. Hence

$$\boxed{E(C_p) \approx p}$$

for an adequate model. When the $C_p$ values for all regression models are thus plotted against $p$, the models with little bias will be close to the line $C_p = p$.

**FIGURE 8.6** $C_p$ **Plot for Surgical Unit Example.**



Models with substantial bias (due to the omission of some predictors) will fall above this line.

The $C_p$ plot of the surgical unit data clearly shows that only the subset with the first three variables has little bias. Here, the $C_p$ value falls below the line $C_p = p$, because $\text{SSE}_{1,2,3} = 0.1099$ is only slightly larger than $\text{SSE}_{1,2,3,4} = 0.1098$ and consequently $\text{MSE}_{1,2,3} = 0.00220 < \text{MSE}_{1,2,3,4} = 0.00224$.

### 17.2.4 Akaike's Information Criterion

Akaike's information criterion (AIC) is an asymptotically unbiased estimator of the Kullback-Leibler information number (or discrepancy):

$$\text{K-L} = E_P\Big[\log\Big(\frac{f_P(x)}{f_p(x)}\Big)\Big]$$

with $f_P$ the likelihood of the full model, and $f_p$ the likelihood of the reduced model.

As derived in (12.2.18), for a regression model with $n$ observations, $p$ parameters and normally distributed errors, the log-likelihood function is given by:

$$\log L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2.$$

The least squares estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{LS}}$ maximizes this log-likelihood function:

$$\log L(\hat{\boldsymbol{\beta}}, \sigma^2|\boldsymbol{y}) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\text{SSE}_p.$$

Akaike's Information Criterion is now defined as

$$\boxed{\text{AIC}_p = -2\max\log L + 2p.}$$

If $\sigma^2$ is known, this corresponds to

$$\text{AIC}_p = \frac{\text{SSE}_p}{\sigma^2} + 2p + \text{const}.$$

From (17.2.1) we see that, if $\sigma^2$ is known,

$$C_p = \frac{\text{SSE}_p}{\sigma^2} - (n - 2p)$$

and thus $C_p$ and $\text{AIC}_p$ only differ by a constant.
If $\sigma^2$ is unknown, we estimate it as $\hat{\sigma}^2_{\text{ML}} = \text{SSE}_p/n$ (see (12.2.17)) and thus

$$\max\log L = \log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\text{ML}}|\boldsymbol{y}) = \text{const} - \frac{n}{2}\log\hat{\sigma}^2_{\text{ML}} - \frac{n}{2}$$

from which

$$\boxed{\text{AIC}_p = n\log(\text{SSE}_p/n) + 2p + \text{const}}$$

follows.

The AIC criterion is used in R to perform stepwise regression (see Section 17.3). In a forward search strategy, one starts with a model of e.g. $p-1$ explanatory variables, and then includes the variable that yields the largest reduction in the AIC.

### 17.2.5 PRESS$_p$ Criterion

The goal of the PRESS$_p$ criterion (predicted sum of squares) is to estimate the mean squared error of prediction:

$$\text{MSEP} = E[(y_0 - \hat{y}_0)^2]$$

with $y_0$ a new observation, independent of the original $n$ observations.

To obtain independent observations in a artificial way, every observed data point $(\boldsymbol{x}_i, y_i)$ in turn is withheld from the data set. This leaves a new data set with $n-1$ observations. Under the independent errors assumption, we know that $y_i$ is independent of this new data set. The remaining $n-1$ observations yield the least squares fit $\hat{\boldsymbol{\beta}}_{(i)}$ and for observation $i$ the fitted value $\hat{y}_{i(i)} = \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}_{(i)}$.

The PRESS$_p$ value is then obtained by summing the prediction errors $d_i = y_i - \hat{y}_{i(i)}$ over all $i = 1, \ldots, n$:

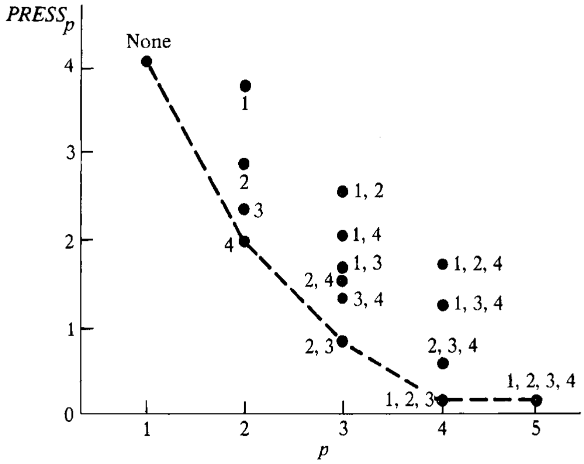
$$\boxed{\text{PRESS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2} \tag{17.2.2}$$

Models with small PRESS$_p$ values (or PRESS$_p/n$) are considered good candidate models. The prediction error $d_i = y_i - \hat{y}_{i(i)}$ is also called the *deleted residual* for the $i$th observation. It can be shown to be equal to

$$d_i = \frac{e_i}{1 - h_{ii}} \tag{17.2.3}$$

and thus can be computed without recomputing the regression function.

**FIGURE 8.7** ***PRESS$_p$* Plot for Surgical Unit Example.**



## 17.3 Stepwise regression

Stepwise regression procedures develop a sequence of regression models, such that at each step an explanatory variable is added to or removed from the model. They are automatic search methods that avoid fitting all subset regressions.

### 17.3.1 Backward elimination

The backward elimination method starts with the regression model that contains all $P$ explanatory variables. Then it computes for each variable $j = 1, \ldots, P$ in the model its partial F-value. Following (13.2.2):

$$F_j^* = \frac{\text{MSR}(X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_P)}{\text{MSE}(X_1, \ldots, X_P)}$$

represents the decrease in the SSE when $X_j$ is added to the model that does not yet contain $X_j$. Equivalently

$$F_j^* = \left( \frac{\hat{\beta}_j}{s(\hat{\beta}_j)} \right)^2$$

and thus equals the squared $t$-value for the parameter test $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.

The variable for which this $F_j^*$ is smallest is the candidate for deletion. If this $F_j^*$ value falls below a predetermined limit (e.g. $F_{1,n-P,\alpha}$ or the corresponding $p$-value is larger than $\alpha$), then this variable is deleted. Otherwise the process is stopped. If not, the procedure starts over with the $P - 1$ remaining variables. A drawback of this method is that a variable can never come back in the model once it is deleted.
To illustrate the stepwise procedures we use a random halfsample of the surgical unit data.

```
set.seed(1)
surgicalunit2 <- surgicalunit[sample(1:108,54),]
attach(surgicalunit2)
surg.full <- lm(Log.Survival ~ Blood.Clotting + Prognostic + Enzyme + Liver)
summary(surg.full)

Call:
lm(formula = Log.Survival ~ Blood.Clotting + Prognostic + Enzyme +
    Liver)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58486 -0.15028  0.01233  0.13650  0.64934

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.655732   0.272491  13.416  < 2e-16 ***
Blood.Clotting 0.089647   0.029422   3.047  0.00372 **
Prognostic     0.015938   0.002433   6.552 3.28e-08 ***
Enzyme         0.015687   0.002578   6.085 1.73e-07 ***
Liver          0.003911   0.059692   0.066  0.94802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.252 on 49 degrees of freedom
Multiple R-squared:  0.7583,Adjusted R-squared:  0.7386
F-statistic: 38.43 on 4 and 49 DF,  p-value: 1.52e-14
```

Because $F_4^* = 0.066^2 = 0.0044 < F_{1,54-5,0.05} = 4.04$ (or equivalently because the corresponding $p$-value $0.948 > 0.05$) the fourth variable `Liver` will be removed.

```
surg.bpe <- update(surg.full, .~. - Liver)
coefficients(summary(surg.bpe))

                 Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)    3.64634732 0.229497996 15.888362 3.516836e-21
Blood.Clotting 0.09075961 0.023785524  3.815750 3.747351e-04
Prognostic     0.01599824 0.002230012  7.174059 3.206410e-09
Enzyme         0.01581636 0.001633439  9.682863 4.701198e-13
```

The procedure stops because all p-values are smaller than the limit.
In R the function `stepAIC` performs backward selection automatically based on the AIC criterion. The variable whose removal results in the largest decrease in AIC is dropped from the model. The procedure stops when AIC cannot decrease anymore.

```
library(MASS)
surg.full <- lm(Log.Survival~Blood.Clotting + Prognostic + Enzyme + Liver)
surg.stepb <- stepAIC(surg.full, list(upper = ~ Blood.Clotting +
          Prognostic + Enzyme + Liver, lower = ~ 1), direction = "back")
```

```
Start:  AIC=-144.09
Log.Survival ~ Blood.Clotting + Prognostic + Enzyme + Liver

                 Df Sum of Sq    RSS     AIC
- Liver           1   0.00027 3.1129 -146.09
<none>                        3.1126 -144.09
- Blood.Clotting  1   0.58976 3.7024 -136.72
- Enzyme          1   2.35244 5.4651 -115.69
- Prognostic      1   2.72697 5.8396 -112.11

Step:  AIC=-146.08
Log.Survival ~ Blood.Clotting + Prognostic + Enzyme

                 Df Sum of Sq    RSS      AIC
<none>                        3.1129 -146.085
- Blood.Clotting  1    0.9065 4.0194 -134.284
- Prognostic      1    3.2043 6.3172 -109.869
- Enzyme          1    5.8372 8.9501  -91.055
```

Note that the column ”Sum of Sq” indicates the extra sum of squares $\mathrm{SSR}(X_j|X_1,\dots,X_{j-1},X_{j+1},\dots,X_p)$ whereas $\mathrm{RSS} = \mathrm{SSE}_{p-1}$.

### 17.3.2 Forward selection

The forward selection procedure starts with the model that only contains the intercept. Then a simple regression model is fitted for each of the $P - 1$ explanatory variables. Also here the partial $F^*$-value is computed:

$$F_j^* = \frac{\mathrm{MSR}(X_j)}{\mathrm{MSE}(X_j)}$$

and the variable with the largest $F_j^*$ value is the candidate for the first addition. If this $F_j^*$ value exceeds a predetermined level (or the $p$-value is lower than $\alpha$), the $X_j$ variable is added. Otherwise none of the regressors are considered to be helpful in the prediction of the response variable.
Assume variable $X_7$ is entered, then the next partial $F_j^*$ values are

$$F_j^* = \frac{\mathrm{MSR}(X_j|X_7)}{\mathrm{MSE}(X_j, X_7)}$$

and again the variable with the largest $F_j^*$-value is included (if it is large enough). As with the backward selection procedure, none of the variables can be removed once they are entered in the model.

```
surg.initial <- lm(Log.Survival~1)
surg.stepf <- stepAIC(surg.initial, list(upper = ~ Blood.Clotting +
      Prognostic + Enzyme + Liver, lower = ~ 1), direction = "forward")

Start:  AIC=-75.41
Log.Survival ~ 1


                 Df Sum of Sq     RSS      AIC
+ Liver           1    5.8972  6.9810 -106.473
+ Enzyme          1    5.5028  7.3755 -103.505
+ Prognostic      1    2.9983  9.8800  -87.718
+ Blood.Clotting  1    1.0741 11.8041  -78.109
<none>                        12.8782  -75.406

Step:  AIC=-106.47
Log.Survival ~ Liver


                 Df Sum of Sq    RSS     AIC
+ Prognostic      1   1.50547 5.4755 -117.59
+ Enzyme          1   0.92131 6.0597 -112.11
<none>                        6.9810 -106.47
+ Blood.Clotting  1   0.00119 6.9798 -104.48
```

```
Step:  AIC=-117.59
Log.Survival ~ Liver + Prognostic

                 Df Sum of Sq    RSS     AIC
+ Enzyme          1   1.77315 3.7024 -136.72
<none>                        5.4755 -117.59
+ Blood.Clotting  1   0.01047 5.4651 -115.69

Step:  AIC=-136.72
Log.Survival ~ Liver + Prognostic + Enzyme

                 Df Sum of Sq    RSS     AIC
+ Blood.Clotting  1   0.58976 3.1126 -144.09
<none>                        3.7024 -136.72

Step:  AIC=-144.09
Log.Survival ~ Liver + Prognostic + Enzyme + Blood.Clotting
```

### 17.3.3 Stepwise regression

Stepwise regression combines the forward and backward selection procedure. At each stage of the procedure, it is checked whether a new variable should come in or should be removed from the model. Let us consider the forward stepwise procedure. Then, starting with the intercept term, a variable is added as in the forward selection method. Assume it to be $X_3$. Next, a second variable is selected, e.g. $X_4$. But before a third one is added, it is checked whether $X_3$ should be dropped from the actual model or not. For this the partial $F_j^*$-value

$$F_j^* = \frac{\text{MSR}(X_3|X_4)}{\text{MSE}(X_3, X_4)}$$

is computed. Assume that both variables remain in the model, then the partial $F_j^*$-values

$$F_j^* = \frac{\text{MSR}(X_j|X_3, X_4)}{\text{MSE}(X_j, X_3, X_4)}$$

have to be computed and so on.

The stepwise regression thus allows a variable that is brought into the model at a certain stage, to be dropped consequently if it is no longer helpful in conjunction with other variables added at a later stage.

Stepwise regression can also be performed by looking at the increase and decrease of the AIC when removing and adding variables:

```
surg.stepfbi <- stepAIC(surg.initial, list(upper = ~ Blood.Clotting +
          Prognostic + Enzyme + Liver, lower = ~ 1), direction = "both")

Start:  AIC=-75.41
Log.Survival ~ 1

                 Df Sum of Sq     RSS      AIC
+ Liver           1    5.8972  6.9810 -106.473
+ Enzyme          1    5.5028  7.3755 -103.505
+ Prognostic      1    2.9983  9.8800  -87.718
+ Blood.Clotting  1    1.0741 11.8041  -78.109
<none>                        12.8782  -75.406

Step:  AIC=-106.47
Log.Survival ~ Liver

                 Df Sum of Sq     RSS      AIC
+ Prognostic      1    1.5055  5.4755 -117.589
+ Enzyme          1    0.9213  6.0597 -112.115
<none>                         6.9810 -106.473
+ Blood.Clotting  1    0.0012  6.9798 -104.482
- Liver           1    5.8972 12.8782  -75.406

Step:  AIC=-117.59
Log.Survival ~ Liver + Prognostic

                 Df Sum of Sq    RSS      AIC
+ Enzyme          1    1.7732 3.7024 -136.720
<none>                        5.4755 -117.589
+ Blood.Clotting  1    0.0105 5.4651 -115.693
- Prognostic      1    1.5055 6.9810 -106.473
- Liver           1    4.4044 9.8800  -87.718

Step:  AIC=-136.72
Log.Survival ~ Liver + Prognostic + Enzyme

                 Df Sum of Sq    RSS     AIC
+ Blood.Clotting  1   0.58976 3.1126 -144.09
```

```
<none>                       3.7024 -136.72
- Liver           1   0.31699 4.0194 -134.28
- Enzyme          1   1.77315 5.4755 -117.59
- Prognostic      1   2.35731 6.0597 -112.11

Step:  AIC=-144.09
Log.Survival ~ Liver + Prognostic + Enzyme + Blood.Clotting

                 Df Sum of Sq    RSS     AIC
- Liver           1   0.00027 3.1129 -146.09
<none>                       3.1126 -144.09
- Blood.Clotting  1   0.58976 3.7024 -136.72
- Enzyme          1   2.35244 5.4651 -115.69
- Prognostic      1   2.72697 5.8396 -112.11

Step:  AIC=-146.08
Log.Survival ~ Prognostic + Enzyme + Blood.Clotting

                 Df Sum of Sq    RSS      AIC
<none>                       3.1129 -146.085
+ Liver           1    0.0003 3.1126 -144.090
- Blood.Clotting  1    0.9065 4.0194 -134.284
- Prognostic      1    3.2043 6.3172 -109.869
- Enzyme          1    5.8372 8.9501  -91.055

surg.stepfbf <- stepAIC(surg.full, list(upper = ~ Blood.Clotting +
          Prognostic + Enzyme + Liver, lower = ~ 1), direction = "both")

Start:  AIC=-144.09
Log.Survival ~ Blood.Clotting + Prognostic + Enzyme + Liver

                 Df Sum of Sq    RSS     AIC
- Liver           1   0.00027 3.1129 -146.09
<none>                       3.1126 -144.09
- Blood.Clotting  1   0.58976 3.7024 -136.72
- Enzyme          1   2.35244 5.4651 -115.69
- Prognostic      1   2.72697 5.8396 -112.11

Step:  AIC=-146.08
Log.Survival ~ Blood.Clotting + Prognostic + Enzyme

                 Df Sum of Sq    RSS      AIC
<none>                       3.1129 -146.085
```

```
+ Liver           1    0.0003 3.1126 -144.090
- Blood.Clotting  1    0.9065 4.0194 -134.284
- Prognostic      1    3.2043 6.3172 -109.869
- Enzyme          1    5.8372 8.9501  -91.055
```

## 17.4 Model validation

Model validation involves checking the model against independent data. Three basic ways of validation are

- collection of new data to check the model and its predictive ability
- comparison of results with theoretical expectations, earlier empirical results, simulation results
- use of a holdout sample to check the model and its predictive ability

### 17.4.1 Collection of new data

The purpose of collecting new data is to able to examine whether the regression model developed from the earlier data is still applicable for the new data. This is in particular of interest for exploratory observational studies, as they also involve model building.

Validity checking can be performed

- by re-estimating the final model using the new data and comparing the estimated regression coefficients and other characteristics of the fitted model
- by re-estimating from the new data all the 'good' subset models that had been considered to see whether the selected regression model is still the preferred one.
- by measuring the predictive ability of the regression model. Since the selected model is chosen to fit the original data, the MSE often underestimates the true variance. The *mean squared prediction error*

  $$\text{MSEP} = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}$$

  computes the mean of the squared prediction errors of the new data (of size $m$), and should be compared with MSE. If MSEP is much larger than MSE, one should rely on the MSEP as an indicator of how well the selected regression model will predict in the future.

### 17.4.2 Data splitting

Since it is often very difficult to collect new data, an alternative is to split the data into two sets: a *training set* used to develop the model and a *validation set* which is used to evaluate the reasonableness and predictive capability of the selected model. This validation procedure is often called *cross-validation.* The validation set then plays the role of the 'new data' in the previous section.

To obtain reliable results, the training set should be large enough (remember, $n > 5p$) otherwise the variances of the regression coefficients will be too large. If data splitting is impractical at small data sets, the PRESS criterion (17.2.2)

$$\text{PRESS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2$$

can also be employed as a form of data splitting.

If a data set for an explanatory observational study is very large, it can even be split into three parts: one for developing the regression model, the second for estimating the parameters and the third for validation. This approach avoids bias resulting from estimating the regression parameters from the same data set used for developing the model. On the other hand this approach yields larger variances of the parameter estimates.

In any case, once the model has been validated, it is customary to use the entire data set for estimating the final regression model.

# Chapter 18

# Multicollinearity

## 18.1 The effects of multicollinearity

We say that there exists multicollinearity among the predictors if there exists a nontrivial linear combination of the regressors which is (almost) zero:

$$\exists \{c_j\} : \quad c_0 + \sum_{j=1}^{p-1} c_j X_j \approx 0.$$

As illustrated in Figures 13.2 and 13.3 of Chapter 12, there is a large effect on the regression parameter estimates when the predictor variables are correlated among themselves. Before exploring these difficulties in detail, we first examine the situation when all the predictor variables are uncorrelated.

### 18.1.1 Uncorrelated predictor variables

Let us consider the crew productivity example, investigating the effect of work crew size ($X_1$) and the level of bonus pay ($X_2$) on crew productivity ($Y$).

```
attach(crew)
crew

  crew.size bonus.pay crew.productivity
1         4         2                42
2         4         2                39
3         4         3                48
4         4         3                51
5         6         2                49
6         6         2                53
7         6         3                61
```

```
8          6          3                 60
```

Both predictors are uncorrelated, $r_{12} = 0$. We compare the regression when both $X_1$ and $X_2$ are included in the model, to the simple regression models containing only $X_1$ and only $X_2$.

```
crew.lm12 <- lm(crew.productivity ~ crew.size + bonus.pay)
summary(crew.lm12)

Call:
lm(formula = crew.productivity ~ crew.size + bonus.pay)

Residuals:
     1      2      3      4      5      6      7      8 
 1.625 -1.375 -1.625  1.375 -2.125  1.875  0.625 -0.375 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.3750     4.7405   0.079 0.940016    
crew.size     5.3750     0.6638   8.097 0.000466 ***
bonus.pay     9.2500     1.3276   6.968 0.000937 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.877 on 5 degrees of freedom
Multiple R-squared:  0.958,Adjusted R-squared:  0.9412 
F-statistic: 57.06 on 2 and 5 DF,  p-value: 0.000361

anova(crew.lm12)

Analysis of Variance Table

Response: crew.productivity
          Df  Sum Sq Mean Sq F value    Pr(>F)    
crew.size  1 231.125 231.125  65.567 0.0004657 ***
bonus.pay  1 171.125 171.125  48.546 0.0009366 ***
Residuals  5  17.625   3.525                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

crew.lm1 <- lm(crew.productivity ~ crew.size)
summary(crew.lm1)

Call:
lm(formula = crew.productivity ~ crew.size)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-6.750 -3.750  0.125  4.500  6.000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   23.500     10.111   2.324   0.0591 .
crew.size      5.375      1.983   2.711   0.0351 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.609 on 6 degrees of freedom
Multiple R-squared:  0.5505,Adjusted R-squared:  0.4755 
F-statistic: 7.347 on 1 and 6 DF,  p-value: 0.03508

anova(crew.lm1)

Analysis of Variance Table

Response: crew.productivity
          Df Sum Sq Mean Sq F value  Pr(>F)  
crew.size  1 231.12 231.125   7.347 0.03508 *
Residuals  6 188.75  31.458                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

crew.lm2 <- lm(crew.productivity ~ bonus.pay)
summary(crew.lm2)

Call:
lm(formula = crew.productivity ~ bonus.pay)

Residuals:
   Min     1Q Median     3Q    Max 
-7.000 -4.688 -0.250  5.250  7.250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   27.250     11.608   2.348   0.0572 .
bonus.pay      9.250      4.553   2.032   0.0885 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.439 on 6 degrees of freedom
Multiple R-squared:  0.4076,Adjusted R-squared:  0.3088
F-statistic: 4.128 on 1 and 6 DF,  p-value: 0.08846

anova(crew.lm2)

Analysis of Variance Table


Response: crew.productivity
          Df Sum Sq Mean Sq F value  Pr(>F)
bonus.pay  1 171.12 171.125  4.1276 0.08846 .
Residuals  6 248.75  41.458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The fitted response functions are:

$$\begin{aligned} \hat{Y} &= 0.375 + 5.375X_1 + 9.250X_2 \\ \hat{Y} &= 23.50 + 5.375X_1 \\ \hat{Y} &= 27.25 + 9.250X_2 \end{aligned}$$

We see that $\hat{\beta}_1 = 5.375$, the regression coefficient for $X_1$, is the same whether or not $X_2$ is also included in the model. The same holds for $\hat{\beta}_2 = 9.250$. This is a general result, which can be most easily deduced from the estimate of $\boldsymbol{\beta}$ in the standardized regression model (12.5.3):

$$y_i' = \beta_1' x_{i1}' + \beta_2' x_{i2}' + \ldots + \beta_{p-1}' x_{i,p-1}' + \epsilon_i' \tag{18.1.1}$$

with parameter estimates (12.5.6)

$$\hat{\boldsymbol{\beta}}' = R_{XX}^{-1} r_{XY}.$$

If all the $X$ variables are uncorrelated, $R_{XX} = I_{p-1}$, and thus $\hat{\beta}_j' = r_{jy}$ only depends on $X_j$ and $Y$. This remains true for the original coefficients

$$\hat{\beta}_j = (\frac{s_Y}{s_j})\hat{\beta}_j' = \frac{s_{jY}}{s_j^2}.$$

A geometrical illustration is provided in Figure 10.11 for uncorrelated regressors and in Figure 10.10 for correlated predictor variables.

Moreover we observe in our example that

$$\text{SSR}(X_2|X_1) = 171.125 = \text{SSR}(X_2)$$

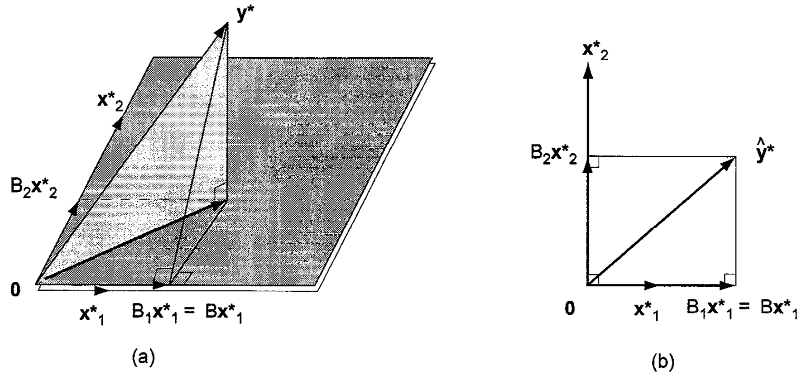


**Figure 10.11.** When the $X$'s are uncorrelated, the simple-regression slope $B$ and the multiple-regression slope $B_1$ are the same. The least-squares fit is in (*a*), the regressor plane in (*b*).

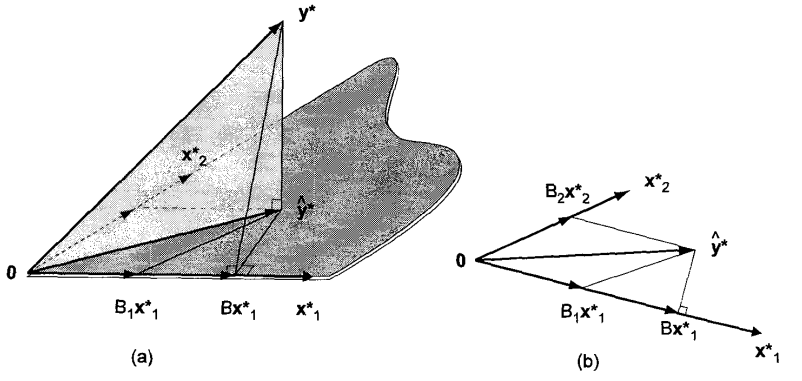


**Figure 10.10.** When the $X$'s are correlated (here positively), the slope $B$ for the simple regression of $Y$ on $X_1$ differs from the slope $B_1$ in the multiple regression of $Y$ on both $X_1$ and $X_2$. The least-squares fit is in (*a*), the regressor plane in (*b*).

and equivalently $\text{SSR}(X_1|X_2) = \text{SSR}(X_1) = 231.125$. Since

$$\text{SSR}(X_2|X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1) = \text{SSR}(X_2)$$

we see that the regression sum of squares due to $X_1$ and $X_2$ together can be split into the SSR due to $X_1$ alone and the SSR due to $X_2$ alone, when $X_1$ and $X_2$ are uncorrelated.

### 18.1.2 Perfectly or highly correlated predictors

Now, consider an example where two predictor variables are perfectly correlated, as in Table 7.8.

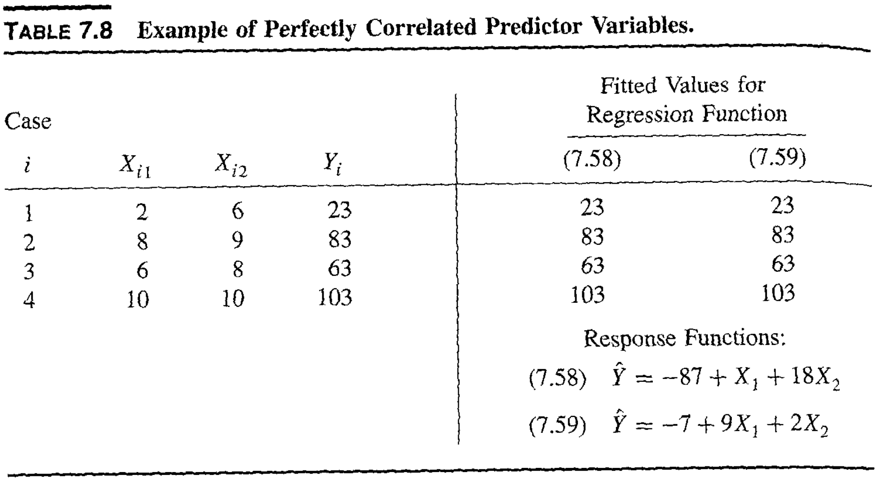


**TABLE 7.8 Example of Perfectly Correlated Predictor Variables.**

| Case |  |  |  | Fitted Values for Regression Function |  |
|---|---|---|---|---|---|
| $i$ | $X_{i1}$ | $X_{i2}$ | $Y_i$ | (7.58) | (7.59) |
| 1 | 2 | 6 | 23 | 23 | 23 |
| 2 | 8 | 9 | 83 | 83 | 83 |
| 3 | 6 | 8 | 63 | 63 | 63 |
| 4 | 10 | 10 | 103 | 103 | 103 |

Response Functions:

(7.58) $\hat{Y} = -87 + X_1 + 18X_2$

(7.59) $\hat{Y} = -7 + 9X_1 + 2X_2$

Here, the response function is not unique. Both $\hat{Y} = -87 + X_1 + 18X_2$ and $\hat{Y} = -7 + 9X_1 + 2X_2$ yield the same fitted values and (zero) residuals.

As argued in Section 12.2, the matrix of cross-products $X^tX$ is singular when $\text{rank}(X) < p$, hence the least-squares solution is not unique. This could also be seen in Figure 13.2(b) of Chapter 12.

At real data sets, predictor variables are rarely perfectly correlated and the least squares solution will thus still be unique. But other problems occur when some (or all) of the regressors are highly correlated:

- because $(X^tX)$ is close to a singular matrix, the variance of the estimated coefficients

  $$\Sigma(\hat{\boldsymbol{\beta}}) = \sigma^2(X^tX)^{-1}$$

  will be large. Consequently

  - a new sample can yield very different estimates
  - many of the regression coefficients may be statistically non significant (large confidence intervals), even though a statistical relation exists between the response variable and the set of predictor variables. The

$t$ and $p$-values corresponding to the univariate tests $\beta_j = 0$ are thus not very informative (see also the discussion in Section 13.4).

- the interpretation of a regression coefficient is not fully applicable. When regressors are strongly correlated, it might be hard to vary one predictor while holding the others constant (e.g. $X_1$ = amount of rainfall, $X_2$ = hours of sunshine).

We will illustrate some of these effects on the Body Fat data, relating the amount of body fat ($Y$) to triceps skinfold thickness ($X_1$), thigh circumference ($X_2$) and midarm circumference ($X_3$), measured at 20 healthy women of 25-34 years old.

```
attach(bodyfat)
bodyfat

   triceps thigh midarm body.fat
1     19.5  43.1   29.1     11.9
2     24.7  49.8   28.2     22.8
3     30.7  51.9   37.0     18.7
4     29.8  54.3   31.1     20.1
5     19.1  42.2   30.9     12.9
6     25.6  53.9   23.7     21.7
7     31.4  58.5   27.6     27.1
8     27.9  52.1   30.6     25.4
9     22.1  49.9   23.2     21.3
10    25.5  53.5   24.8     19.3
11    31.1  56.6   30.0     25.4
12    30.4  56.7   28.3     27.2
13    18.7  46.5   23.0     11.7
14    19.7  44.2   28.6     17.8
15    14.6  42.7   21.3     12.8
16    29.5  54.4   30.1     23.9
17    27.7  55.3   25.7     22.6
18    30.2  58.6   24.6     25.4
19    22.7  48.2   27.1     14.8
20    25.2  51.0   27.5     21.1
```

From the correlation matrix $R_{XX}$ we deduce that $X_1$ and $X_2$ are highly correlated.

```
print(cor(bodyfat),digits=2)

         triceps thigh midarm body.fat
triceps      1.00 0.924  0.458     0.84
thigh        0.92 1.000  0.085     0.88
midarm       0.46 0.085  1.000     0.14
body.fat     0.84 0.878  0.142     1.00
```

The third variable $X_3$ is not highly correlated with $X_1$ and $X_2$ but if we regress $X_3$ on $X_1$ and $X_2$ we obtain $R^2 = 0.98$. From the table below, we see that the regression coefficient for a certain predictor varies a lot depending on the presence of some or all of the other predictors, and even can change sign as for $\hat{\beta}_2$. Also the standard error of the parameter estimates increases considerably when more variables are added to the model.

| variables in model | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $s(\hat{\beta}_1)$ | $s(\hat{\beta}_2)$ |
|---|---|---|---|---|
| $X_1$ | 0.8572 | - | 0.1288 | - |
| $X_2$ | - | 0.8565 | - | 0.1100 |
| $X_1, X_2$ | 0.2224 | 0.6594 | 0.3034 | 0.2912 |
| $X_1, X_2, X_3$ | 4.3341 | -2.8568 | 3.0155 | 2.5820 |

This example illustrates again that a regression coefficient reflects the marginal or partial effect of a predictor on the response variable, given the other variables in the model!

If we compute the extra sums of squares for $X_1$, we see that $\text{SSR}(X_1) = 352.27$ is much larger that $\text{SSR}(X_1|X_2) = 3.47$. This is again due to the fact that $X_1$ and $X_2$ are highly correlated. When $X_2$ is already in the model, the marginal contribution of $X_1$ is small, because $X_1$ contains almost the same information as $X_2$.

Finally we illustrate the effect of the multicollinearity on hypothesis tests. Consider the linear model with only the first two predictors.

```
body.lm12 <- lm(body.fat ~ triceps + thigh)
summary(body.lm12, correlation = TRUE)

Call:
lm(formula = body.fat ~ triceps + thigh)


Residuals:
    Min       1Q  Median      3Q      Max
```

```
-3.9469 -1.8807  0.1678  1.3367  4.0147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
triceps       0.2224     0.3034   0.733   0.4737
thigh         0.6594     0.2912   2.265   0.0369 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared:  0.7781,Adjusted R-squared:  0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

Correlation of Coefficients:
        (Intercept) triceps
triceps  0.73
thigh   -0.93       -0.92
```

The overall F-test

$$H_0:\ \beta_1 = \beta_2 = 0$$

has a $p$-value 0, which leads us to conclude that $\beta_1$ or $\beta_2$ are significant. The individual $t$-tests for $\beta_1$ and $\beta_2$:

$$H_0 : \beta_1 = 0 \qquad \text{and} \qquad H_0 : \beta_2 = 0$$

are however both in favor of the $H_0$ hypothesis. If we use the Bonferroni method at the $\alpha = 5\%$ significance level, we see that both $p$-values (0.47 and 0.037) are larger than $0.025 = \alpha/2$. Reconsidering Figure 5.1 from Chapter 13, we notice that the absolute value of the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is very high (-0.92) and thus leads to a very tight confidence ellipse which excludes $(0, 0)$. But the two univariate confidence intervals contain 0.

## 18.2 Multicollinearity diagnostics

### 18.2.1 Informal methods

To summarize, multicollinearity can be informally detected by the following diagnostics:

1. large changes in the estimated regression coefficients when a predictor variable is added or deleted
2. nonsignificant results in individual tests on the regression coefficients for important predictor variables
3. estimated regression coefficients with an opposite sign as we would expect from theoretical considerations or prior experience
4. large simple correlations between pairs of predictor variables.

### 18.2.2 Variance inflation factors

A formal method to detect multicollinearity is by means of the **variance inflation factors**.

Let $R_j^2$ be the value of the $R^2$ coefficient of determination when $X_j$ is regressed against all the other independent variables. Then, for each $j = 1, \ldots, p-1$ the variance inflation factor is defined as

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R_j^2}}$$

- If $X_j$ is not related to the other $X$ variables, $R_j^2 = 0$ and hence, $\mathrm{VIF}_j = 1$.
- When there is a perfect linear association, $R_j^2 = 1$ and $\mathrm{VIF}_j$ is unbounded.
- In the general case $0 < R_j^2 < 1$, so $1 < \mathrm{VIF}_j < \infty$.

It can be shown that $\mathrm{VIF}_j$ equals the $j$th diagonal element of the inverse correlation matrix of $X$.

$$\boxed{\mathrm{VIF}_j = (R_{XX}^{-1})_{jj}} \tag{18.2.1}$$

For the Body Fat data, we have indeed large variance inflation factors for all three regressors: $\mathrm{VIF}_1 = 708.84, \mathrm{VIF}_2 = 564.34$ and $\mathrm{VIF}_3 = 104.61$.

```
solve(cor(bodyfat[, 1:3]))

          triceps     thigh    midarm
triceps  708.8429 -631.9152 -270.9894
thigh   -631.9152  564.3434  241.4948
midarm  -270.9894  241.4948  104.6060
```

The variance inflation factors measure how much the variances of the estimated regression coefficients are inflated compared to when the predictor variables are not linearly related. This can be seen as follows: at the standardized regression model (12.5.3):

$$\Sigma(\hat{\boldsymbol{\beta}}') = (\sigma')^2((X')^t X')^{-1} = (\sigma')^2 R_{XX}^{-1}$$

with $(\sigma')^2$ the error variance of the transformed data. Because of (18.2.1) we derive that $\mathrm{Var}(\hat{\beta}_j') = (\sigma')^2 \mathrm{VIF}_j$. In terms of the original variables, this yields

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_j) &= \Big(\frac{s_Y}{s_j}\Big)^2 \mathrm{Var}(\hat{\beta}_j') \\ &= \Big(\frac{s_Y}{s_j}\Big)^2 \frac{1}{(n-1)s_Y^2} \sigma^2 \mathrm{VIF}_j = \frac{\sigma^2}{(n-1)s_j^2} \mathrm{VIF}_j.\end{aligned}$$

using (12.5.5) and (12.5.4).

We speak about strong multicollinearity if

- the largest VIF is larger than 10, or if
- the mean of the VIF values is considerably larger than 1.

### 18.2.3 The eigenvalues of the correlation matrix

As another diagnostic tool we can look at the **eigenvalues of** $R_{XX}$. Because the correlation matrix is symmetric and semi-positive definite, it can be decomposed as

$$R_{XX} = PLP^t = \sum_{j=1}^{p-1} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^t \tag{18.2.2}$$

with the columns of $P = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p-1})$ containing the (normalized) eigenvectors of $R_{XX}$ and $L = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p-1})$ the corresponding eigenvalues. We assume from now on that the eigenvalues are sorted in descending order, so

$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_{p-1}$. If there is perfect multicollinearity, some of the eigenvalues are zero. Near collinearities are associated with small eigenvalues.

To judge the eigenvalues with respect to their size, we use the equality

$$\sum_{j=1}^{p-1} \lambda_j = \operatorname{tr}(R_{XX}) = p - 1$$

so we can

- compare the eigenvalues with $p - 1$ by computing $\lambda_j/(p-1)$
- compare them with the largest eigenvalue $\lambda_{\max} = \lambda_1$. The *condition number* is defined as

  $$\eta_j = \sqrt{\lambda_{\max}/\lambda_j}.$$

  A condition number $\eta_j > 30$ is an indication for multicollinearity.

## 18.3 Multicollinearity remedies

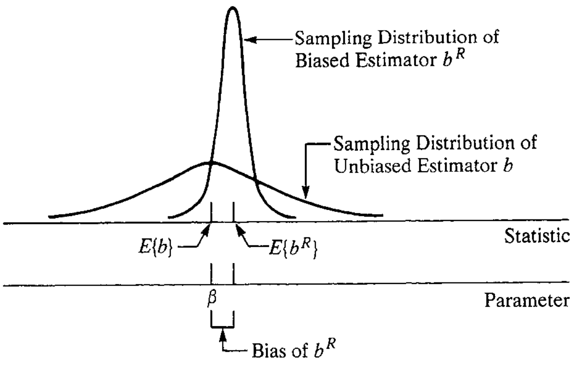
### 18.3.1 Specific solutions

To reduce multicollinearity, we can

- drop one or several predictor variables that are highly correlated to the remaining variables
- for polynomial regression: center the variables
- apply a biased regression method with a smaller variance.

In Figure 10.2 it is illustrated why a biased estimator with a small variance might be preferable over an unbiased estimator with large variance: the biased one will have a larger probability of being close to the true parameter value. Two popular biased estimators are: *Principal component regression* and *Ridge regression.*

**FIGURE 10.2 Biased Estimator with Small Variance May Be Preferable to Unbiased Estimator with Large Variance.**


Sampling Distribution of Biased Estimator $b^R$
Sampling Distribution of Unbiased Estimator $b$
$E\{b\}$
$E\{b^R\}$
Statistic
$\beta$
Parameter
Bias of $b^R$

### 18.3.2 Principal component regression

Principal Component Regression (PCR) combines Principal Component Analysis (PCA) and linear regression. In the first step, the largest principal components of the regressors are computed, yielding a new set of uncorrelated regressors. Secondly, the response variable is regressed onto these principal components.

This is illustrated in Figure 4.2.

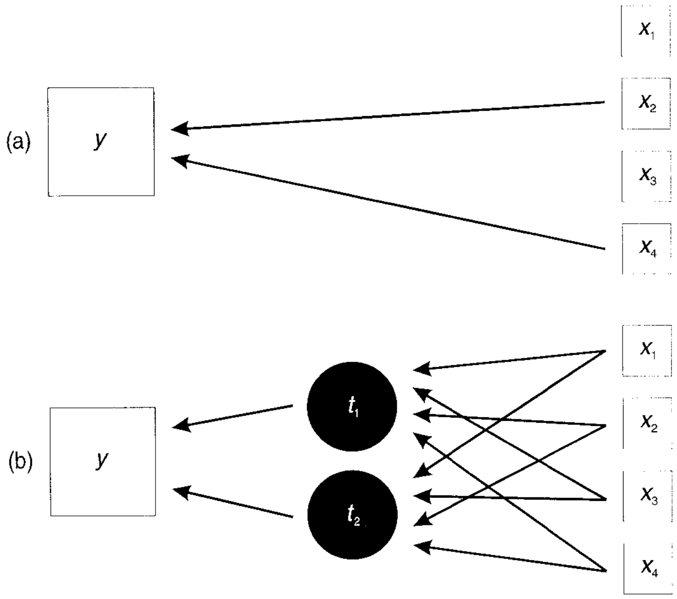


**Figure 4.2. Conceptual illustration of the difference between methods for solving the multicollinearity problem. (a) indicates that variable selection deletes some of the variables from the model. (b) shows how all *x*-variables are transformed into linear combinations $t_1$ and $t_2$, which are related to *y* in a regression equation.**

Because principal components are attracted by the variables that have the largest variance, it is common to start standardizing the variables by the correlation transformation (12.5.1) and (12.5.2). We thus consider the standardized linear model:

$$y'_i = \beta'_1 x'_{i1} + \ldots + \beta'_{p-1} x'_{i,p-1} + \epsilon'_i.$$

Denoting $Z_{n,p-1} = Z = X'$, the least squares estimator then satisfies:

$$\hat{\boldsymbol{\beta}}' = (\hat{\beta}'_1, \ldots, \hat{\beta}'_{p-1})^t = (Z^t Z)^{-1} Z^t \boldsymbol{y}' = R_{XX}^{-1} Z^t \boldsymbol{y}'$$

with variance $\Sigma(\hat{\boldsymbol{\beta}}') = \sigma'^2 (Z^t Z)^{-1}$.

From (18.2.2) it follows that

$$(Z^t Z)^{-1} = \sum_{j=1}^{p-1} \lambda_j^{-1} \boldsymbol{v}_j \boldsymbol{v}_j^t = P L^{-1} P^t. \tag{18.3.1}$$

Here, the loading vectors $\boldsymbol{v}_j$ define the principal components $T_j = v_{1j} Z_1 + v_{2j} Z_2 + \ldots + v_{p-1,j} Z_{p-1}$ of $Z$ satisfying the property that $\mathrm{Var}(T_j)$ is maximal under the constraints that $\|\boldsymbol{v}_j\| = 1$ and $\mathrm{cor}(T_j, T_l) = 0$ for all $j < l$.

Relation (18.3.1) illustrates again that the presence of small eigenvalues yields a large sampling variability. Hence, to reduce the variance of $\hat{\boldsymbol{\beta}}'$, we can decide to eliminate the eigenvectors for which the corresponding eigenvalue is too small. If $\lambda_{k+1}, \ldots, \lambda_{p-1}$ are sufficiently small, this corresponds to setting

$$(Z^t Z)^+ = \sum_{j=1}^{k} \lambda_j^{-1} \boldsymbol{v}_j \boldsymbol{v}_j^t.$$

and defining

$$\boxed{\hat{\boldsymbol{\beta}}^+ = (Z^t Z)^+ Z^t \boldsymbol{y}'.}$$

Equivalently we obtain $\hat{\boldsymbol{\beta}}^+$ by first applying a principal component analysis to the $z$'s and retaining the first $k$ principal components. They coincide with the $T_j$ for $j = 1, \ldots, k$. These principal components span a $k$-dimensional subspace of $\mathbb{R}^{p-1}$ with basis vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. The coordinates of the observations projected onto this subspace are given by the scores

$$\boldsymbol{t}_i = (\tilde{P}_{k,p-1})^t \boldsymbol{z}_i$$

with $\tilde{P}_{p-1,k} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$, or equivalently $T_{n,k} = Z_{n,p-1}\tilde{P}_{p-1,k}$ with $T_{n,k} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)^t$. In the remainder, we drop the subscripts and write $T = Z\tilde{P}$.

Next, the response variable $y'$ is regressed onto the scores. We thus consider the regression model

$$y_i' = \boldsymbol{t}_i^t \boldsymbol{\alpha} + \epsilon_i$$

with least squares estimates $\hat{\boldsymbol{\alpha}} = (T^tT)^{-1}T^t\boldsymbol{y}'$ where $T$ we dropped the subscripts in $T = T_{n,k}$. This estimate is not affected by multicollinearity problems as the scores are uncorrelated! Since $T = Z\tilde{P}$ (where $\tilde{P}$ is short for $\tilde{P}_{p-1,k}$), we find using (18.2.2) that

$$T^tT = \tilde{P}^tZ^tZ\tilde{P} = \tilde{P}^tR_{XX}\tilde{P} = \tilde{P}^t_{k,p-1}P_{p-1,p-1}L_{p-1,p-1}P^t_{p-1,p-1}\tilde{P}_{p-1,k} = \tilde{L}_{k,k}$$

with $\tilde{L}_{k,k} = \tilde{L}$ the upper left $k \times k$ submatrix of $L$. Thus, $\hat{\boldsymbol{\alpha}} = \tilde{L}^{-1}\tilde{P}^tZ^t\boldsymbol{y}'$. Finally we note that $y_i' = \boldsymbol{t}_i^t\boldsymbol{\alpha} + \epsilon_i = \boldsymbol{z}_i^t\tilde{P}\boldsymbol{\alpha} + \epsilon_i$, so

$$\hat{\boldsymbol{\beta}}^+ = \tilde{P}\hat{\boldsymbol{\alpha}} = \tilde{P}_{p-1,k}\tilde{L}^{-1}_{k,k}\tilde{P}^t_{k,p-1}Z^t\boldsymbol{y}' = (Z^tZ)^+Z^t\boldsymbol{y}'.$$

This approach is illustrated in Figure 5.1.

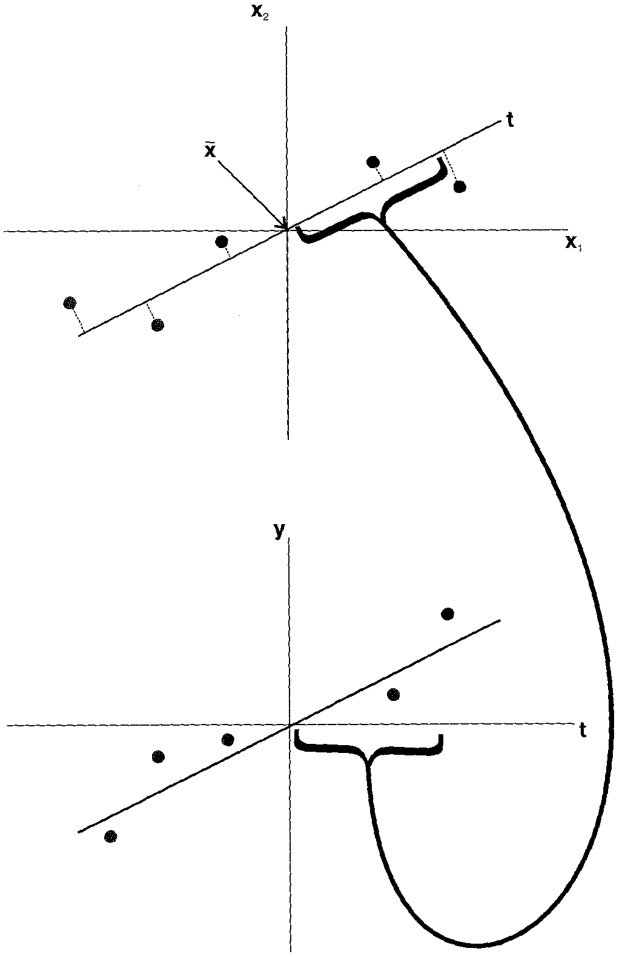


Figure 5.1. Geometrical illustration of the model structure used for the methods PCR and PLS. The information in X is first compressed down to a few components t before these components are used as independent variables in a regression equation with y as dependent variable. The two parts of the illustration correspond to the two equations in (5.1). The first is data compression in x-space, the other is regression based on the compressed components.

The PCR estimator $\hat{\boldsymbol{\beta}}^+$ is biased. Using the orthogonality of the eigenvectors, it follows that

$$\begin{aligned}(Z^tZ)^+(Z^tZ) &= \big(\sum_{j=1}^{k}\lambda_j^{-1}\boldsymbol{v}_j\boldsymbol{v}_j^t\big)\big(\sum_{j=1}^{k}\lambda_j\boldsymbol{v}_j\boldsymbol{v}_j^t + \sum_{j=k+1}^{p-1}\lambda_j\boldsymbol{v}_j\boldsymbol{v}_j^t\big)\\ &= \sum_{j=1}^{k}\boldsymbol{v}_j\boldsymbol{v}_j^t\\ &= I_{p-1} - \sum_{j=k+1}^{p-1}\boldsymbol{v}_j\boldsymbol{v}_j^t\end{aligned}$$

and that

$$(Z^tZ)^+(Z^tZ)(Z^tZ)^+ = (Z^tZ)^+.$$

Consequently

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}^+) &= (Z^tZ)^+Z^tE(Y')\\ &= (Z^tZ)^+Z^tZ\boldsymbol{\beta}'\\ &= \boldsymbol{\beta}' - \sum_{j=k+1}^{p-1}\boldsymbol{v}_j\boldsymbol{v}_j^t\boldsymbol{\beta}'.\end{aligned}$$

On the other hand, the variance of $\hat{\boldsymbol{\beta}}^+$ has decreased:

$$\begin{aligned}\Sigma(\hat{\boldsymbol{\beta}}^+) &= (Z^tZ)^+Z^t\Sigma(Y')Z(Z^tZ)^+\\ &= (Z^tZ)^+\sigma^2\end{aligned}$$

hence

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_l^+) = \sigma^2\sum_{j=1}^{k}\lambda_j^{-1}v_{lj}^2$$

whereas

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_l') = \sigma^2\sum_{j=1}^{p-1}\lambda_j^{-1}v_{lj}^2.$$

**Remarks.**

- The PCR method is in particular very useful when $n < p$. When there are more variables than observations, there is always perfect multicollinearity because $\mathrm{rank}(X) \leq \min(n, p) = n < p$.

- Another advantage of PCR is its ease of computation and its transparency.
- A drawback is that the principal components are not always easy to interpret. Moreover they only make sense if all the variables are measured in the same units.
- PCR selects components which contain most of the variation in the regressors. More sophisticated methods such as Partial Least Squares Regression (PLS) compute components that maximize their covariance with the response variable, with the goal of retaining components that are more informative with respect to the regression model.

A very important issue in PCR is the choice of $k$, the optimal number of principal components that are retained in the analysis. Some popular strategies are the following:

- to make a scree plot, which is a graph of the eigenvalues in decreasing order.
- to select $k$ such that the first $k$ components explain a prescribed percentage of the total variance of the $x$-variables, e.g. one could take $k$ as the smallest integer such that
$$(\sum_{j=1}^{k} \lambda_j)/(\sum_{j=1}^{p-1} \lambda_j) \geqslant 80\%$$
- to use variable selection techniques as discussed in Chapter 17.
- to compute the RMSEP value at a validation set (or by cross-validation):
$$\text{RMSEP}_k = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_{i,k})^2}$$
with $\hat{y}_{i,k}$ the fitted response value for the $i$th case based on a PCR regression with $k$ components, and $m$ the number of observations in the validation set. The RMSEP$_k$ curve for $k = 1, \dots, k_{\max}$ often has the shape of the upper curve of Figure 4.3 (in Chapter 17, Section 17.1). Its minimal value then determines the chosen number of components, see Figure 5.2.

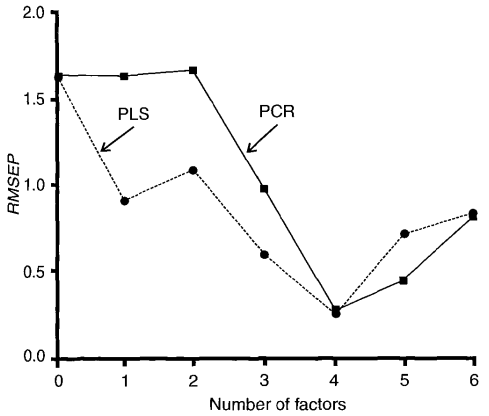


**Figure 5.2. Prediction ability of PCR and PLS for $A = 0, 1, \ldots, 6$. The *RMSEP* is the root mean square error of prediction (see Chapter 13), and measures the prediction ability. Small values of *RMSEP* are to be preferred. Note that the *RMSEP* follows a similar curve to the prediction error in the conceptual illustration in Figure 4.3.**

**Example: Police Height data.**

Measurements on height are taken for 33 female police department applicants together with 9 predictor variables: sitting height, upper arm length, forearm length, hand length, upper leg length, lower leg length, foot length, brachial index and Tibio-Femural index.

The smallest two eigenvalues of $Z^tZ$ are $\lambda_9 = 0.00047$ and $\lambda_8 = 0.00087$ whereas $\lambda_7 = 0.23145$, so 7 principal components are retained. This yields the parameter estimates and the variance inflation factors of Table 10.3. We see that the least squares and the PCR estimates differ very little for all the predictor variables that have a small VIF for least squares. These regressors are not intercorrelated and are only slightly affected by the deletion of $\boldsymbol{v}_8$ and $\boldsymbol{v}_9$. The correlated variables on the other hand have estimates that are greatly altered by the elimination of $\boldsymbol{v}_8$ and $\boldsymbol{v}_9$ and their VIF values are significantly reduced. Also the closeness of $\hat{\sigma}^2$ and $R^2$ show that the PCR model is appropriate here.

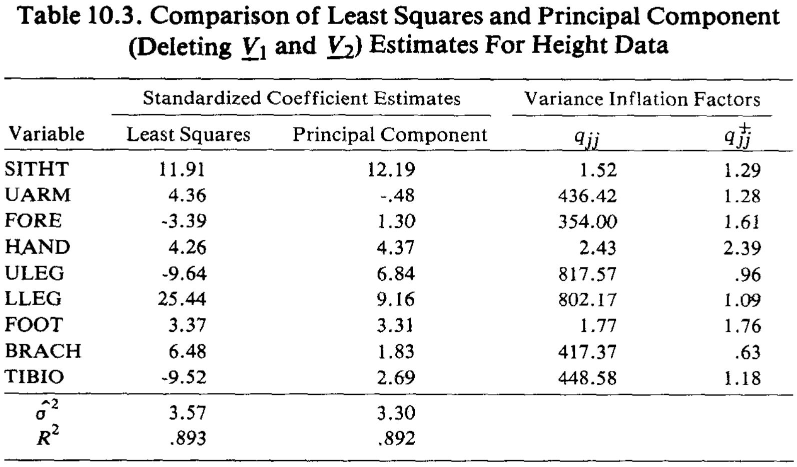


**Table 10.3. Comparison of Least Squares and Principal Component (Deleting $\underline{V}_1$ and $\underline{V}_2$) Estimates For Height Data**

| | Standardized Coefficient Estimates | | Variance Inflation Factors | |
|---|---|---|---|---|
| Variable | Least Squares | Principal Component | $q_{jj}$ | $q_{jj}^{+}$ |
| SITHT | 11.91 | 12.19 | 1.52 | 1.29 |
| UARM | 4.36 | -.48 | 436.42 | 1.28 |
| FORE | -3.39 | 1.30 | 354.00 | 1.61 |
| HAND | 4.26 | 4.37 | 2.43 | 2.39 |
| ULEG | -9.64 | 6.84 | 817.57 | .96 |
| LLEG | 25.44 | 9.16 | 802.17 | 1.09 |
| FOOT | 3.37 | 3.31 | 1.77 | 1.76 |
| BRACH | 6.48 | 1.83 | 417.37 | .63 |
| TIBIO | -9.52 | 2.69 | 448.58 | 1.18 |
| $\hat{\sigma}^2$ | 3.57 | 3.30 | | |
| $R^2$ | .893 | .892 | | |

### 18.3.3 Ridge regression

Ridge regression is a biased regression method which starts by transforming the variables by the correlation transformation, yielding the standardized regression model (12.5.3), whose least squares solution satisfies

$$R_{XX}\boldsymbol{\beta} = r_{XY}.$$

The ridge standardized regression estimators $\boldsymbol{\beta}^*$ are obtained by introducing a constant $c \geqslant 0$ to the diagonal elements of the correlation matrix of $X$:

$$(R_{XX} + cI_{p-1})\boldsymbol{\beta}^* = r_{XY} \tag{18.3.2}$$

With $c = 0$ the ridge and the least squares estimators coincide. When $c > 0$ the ridge estimator is biased, but has less variability.

It can be shown that the bias of $\boldsymbol{\beta}^*$ increases with $c$, whereas the variance (expressed as the trace of the variance-covariance matrix) decreases with $c$. The mean squared error combines the bias and the variance of an estimator. For an estimator of a univariate parameter $\beta$:

$$\begin{aligned} \text{MSE}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] &= (E[\hat{\beta}] - \beta)^2 + E[(\hat{\beta} - E(\hat{\beta}))^2] \\ &= \text{bias}(\hat{\beta})^2 + \text{Var}(\hat{\beta}). \end{aligned}$$

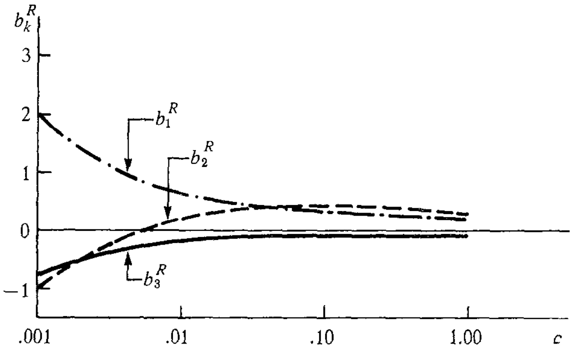
For a $p-1$-dimensional estimator of $\boldsymbol{\beta}$, the total mean squared error can be defined as

$$\text{TMSE}(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{p-1} E[(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^2] = \sum_{j=1}^{p-1} [\text{bias}(\hat{\boldsymbol{\beta}}_j)^2 + \text{Var}(\hat{\boldsymbol{\beta}}_j)].$$

It has been shown that for any data set there exists always a value of $c$ such that the ridge estimator $\boldsymbol{\beta}^*$ has a smaller TMSE than the least squares estimator $\hat{\boldsymbol{\beta}}_{\text{LS}}$.

To determine the constant $c$ we will consider the *ridge trace* method, and the *variance inflation factors.* The ridge trace plots the evolution of the ridge standardized regression coefficients $\beta_j^*$ for different values of $c$, usually between 0 and 1.

**FIGURE 10.3 Ridge Trace of Estimated Standardized Regression Coefficients —Body Fat Example with Three Predictor Variables.**



The VIF values for ridge regression are defined as for OLS: they measure for each coefficient how large the variance of $\hat{\beta}_j^*$ is relative to what the variance would be if the predictors were uncorrelated. It can be shown that $\text{VIF}_j$ for ridge regression equals the $j$-diagonal element of the matrix

$$(R_{XX} + cI)^{-1} R_{XX} (R_{XX} + cI)^{-1}.$$

In the Body fat example (Table 10.3) we see that the VIF's decrease rapidly as $c$ changes from 0 towards 1. The constant $c$ is then chosen as the smallest value where the plot and the VIF's become stable. Here, it was decided to employ $c = 0.02$ since the VIF values are then close to 1 and the regression coefficients are quite stable. The resulting fitted model for $c = 0.02$ is:

$$\hat{Y}' = 0.5463\,X_1' + 0.3774\,X_2' - 0.1369\,X_3'$$

or in terms of the original variables

$$\texttt{Bodyfat} = -7.4034 + 0.5554\,\texttt{triceps} + 0.3681\,\texttt{thigh} - 0.1916\,\texttt{midarm}$$

Also notice that the $R^2$ value only decreased slightly: from 0.8014 to 0.7818. Since the total sum of squares for the transformed variables

$$\text{SST} = \sum_{i=1}^{n}(y_i' - \bar{y}')^2 = 1$$

the coefficient of multiple determination for ridge regression equals

$$\begin{aligned} R^2 &= 1 - \text{SSE} \\ &= 1 - \sum_{i=1}^{n}(y_i' - \hat{y}_i')^2 \end{aligned}$$

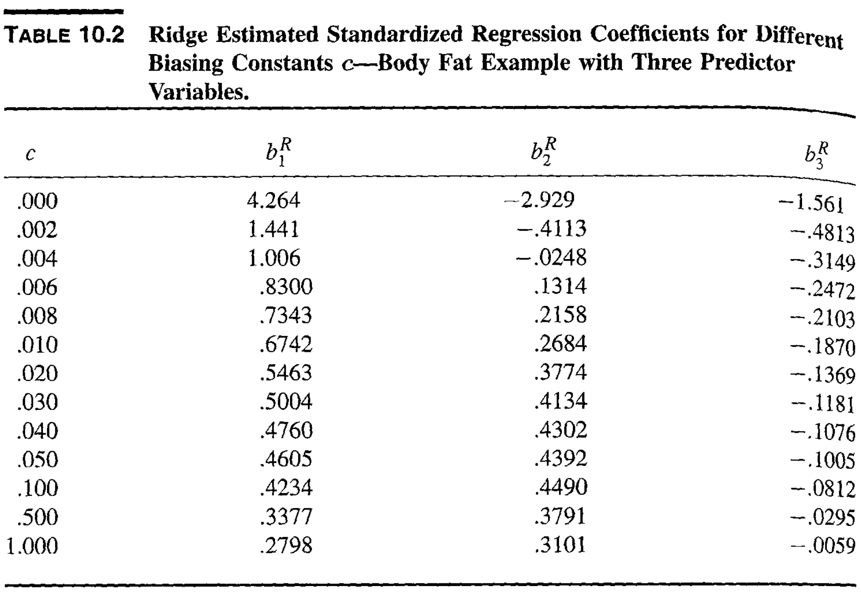


**Table 10.2** Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants $c$—Body Fat Example with Three Predictor Variables.

| $c$ | $b_1^R$ | $b_2^R$ | $b_3^R$ |
|---|---|---|---|
| .000 | 4.264 | −2.929 | −1.561 |
| .002 | 1.441 | −.4113 | −.4813 |
| .004 | 1.006 | −.0248 | −.3149 |
| .006 | .8300 | .1314 | −.2472 |
| .008 | .7343 | .2158 | −.2103 |
| .010 | .6742 | .2684 | −.1870 |
| .020 | .5463 | .3774 | −.1369 |
| .030 | .5004 | .4134 | −.1181 |
| .040 | .4760 | .4302 | −.1076 |
| .050 | .4605 | .4392 | −.1005 |
| .100 | .4234 | .4490 | −.0812 |
| .500 | .3377 | .3791 | −.0295 |
| 1.000 | .2798 | .3101 | −.0059 |

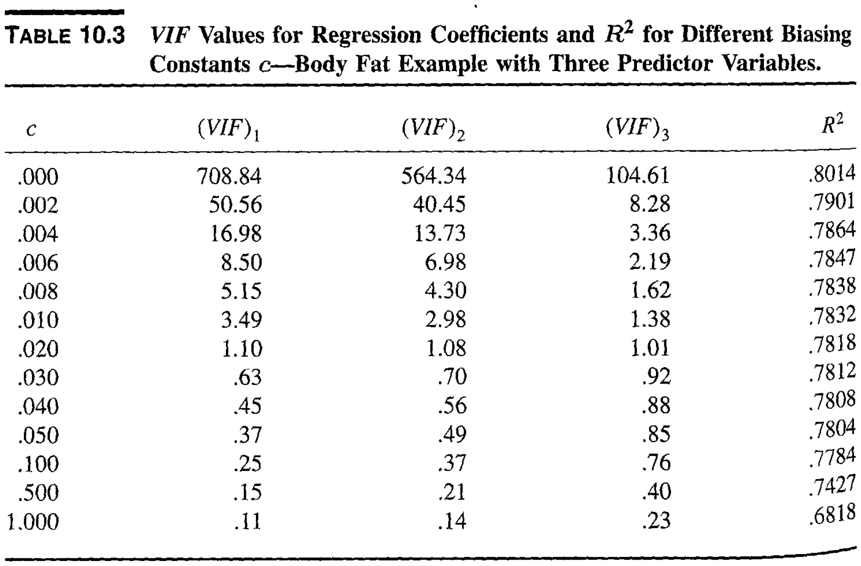


**Table 10.3** *VIF* Values for Regression Coefficients and $R^2$ for Different Biasing Constants $c$—Body Fat Example with Three Predictor Variables.

| $c$ | $(VIF)_1$ | $(VIF)_2$ | $(VIF)_3$ | $R^2$ |
|---|---|---|---|---|
| .000 | 708.84 | 564.34 | 104.61 | .8014 |
| .002 | 50.56 | 40.45 | 8.28 | .7901 |
| .004 | 16.98 | 13.73 | 3.36 | .7864 |
| .006 | 8.50 | 6.98 | 2.19 | .7847 |
| .008 | 5.15 | 4.30 | 1.62 | .7838 |
| .010 | 3.49 | 2.98 | 1.38 | .7832 |
| .020 | 1.10 | 1.08 | 1.01 | .7818 |
| .030 | .63 | .70 | .92 | .7812 |
| .040 | .45 | .56 | .88 | .7808 |
| .050 | .37 | .49 | .85 | .7804 |
| .100 | .25 | .37 | .76 | .7784 |
| .500 | .15 | .21 | .40 | .7427 |
| 1.000 | .11 | .14 | .23 | .6818 |

# Chapter 19

# Influential observations and outliers

Real data sets often contain outlying observations. Although a precise definition of outliers is hard to give, they are characterized as *the observations that do not follow the pattern of the majority of the data.* In regression, data points can be split into 4 types:

1. **regular observations** with internal $\boldsymbol{x}_i$ and well-fitting $y_i$
2. **vertical outliers**, with internal $\boldsymbol{x}_i$ and non-fitting $y_i$
3. **good leverage points**, with outlying $\boldsymbol{x}_i$ and well-fitting $y_i$
4. **bad leverage points**, with outlying $\boldsymbol{x}_i$ and non-fitting $y_i$

For simple regression, these different types of observations are illustrated in Figure 19.1. It is well-known that the least-squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is very sensitive to vertical outliers and bad leverage points.

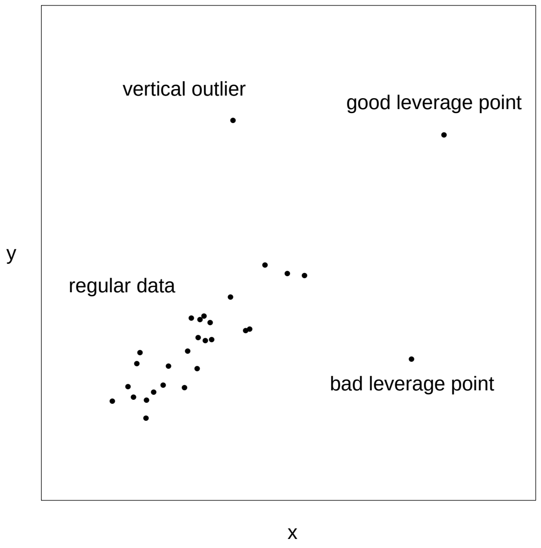


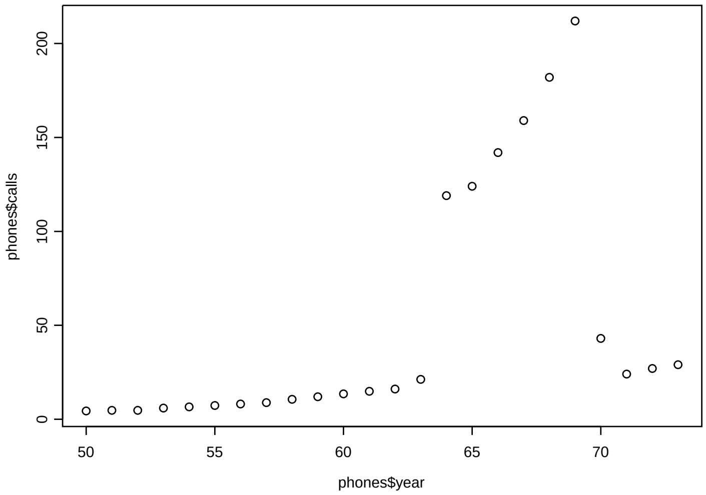
Figure 19.1: Different types of outliers in regression.

## 19.1 Vertical outliers

We consider the `Telephone` data set, which contains the number of international telephone calls (in millions) from Belgium in the years 1950-1973.

```
library(MASS)
phones

$year
 [1] 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
[21] 70 71 72 73


$calls
 [1]   4.4   4.7   4.7   5.9   6.6   7.3   8.1   8.8  10.6  12.0
[11]  13.5  14.9  16.1  21.2 119.0 124.0 142.0 159.0 182.0 212.0
[21]  43.0  24.0  27.0  29.0
```

This data set contains six remarkable vertical outliers. It turned out that from 1964 to 1969 another recording system was used, giving the total number of *minutes* of these calls. The LS fit has clearly been affected by the outlying $y$-values, as shown in Figure 19.2. The robust LTS method, which will be defined in Section 19.4.1, avoids the outliers and fits nicely the linear model of the majority of the data.



```
attach(phones)
plot(year,calls)
phones.lm <- lm(calls ~ year)
abline(phones.lm)
text(70,100,"LS")
library(robustbase)
phones.wlts <- ltsReg(calls~year,alpha=0.75)
abline(phones.wlts,lty=2)
text(67,30,"LTS")
```

To detect vertical outliers, we consider the **standardized robust residuals**, defined as

$$\boxed{e_{i,R}^{(s)} = \frac{y_i - \hat{y}_{i,R}}{s_R}} \tag{19.1.1}$$

with $\hat{y}_{i,R}$ the fitted values obtained by applying a robust regression method, and $s_R$ a robust measure of scale. If the majority of the data points follows the general linear model with normal errors, these standardized robust residuals approximately lie in [-2,2] with a confidence of 95% and in [-2.5,2.5] with a confidence of 99%. The robust LTS method nicely detects the outliers, as shown on the residual plot in Figure 19.3. On the other hand, none of the observations seems outlying if we look at the plot of the standardized LS residuals (Figure 19.4).

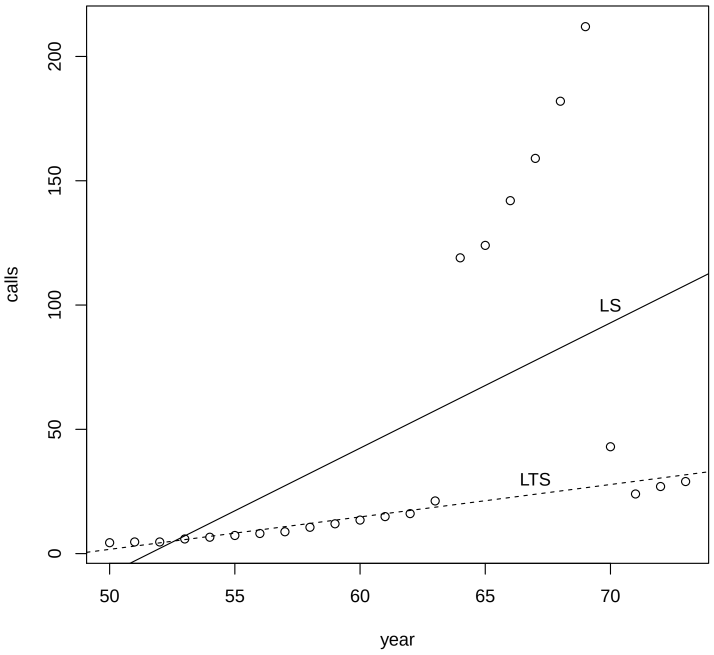


Figure 19.2: Telephone data set with LS and LTS fit superimposed.

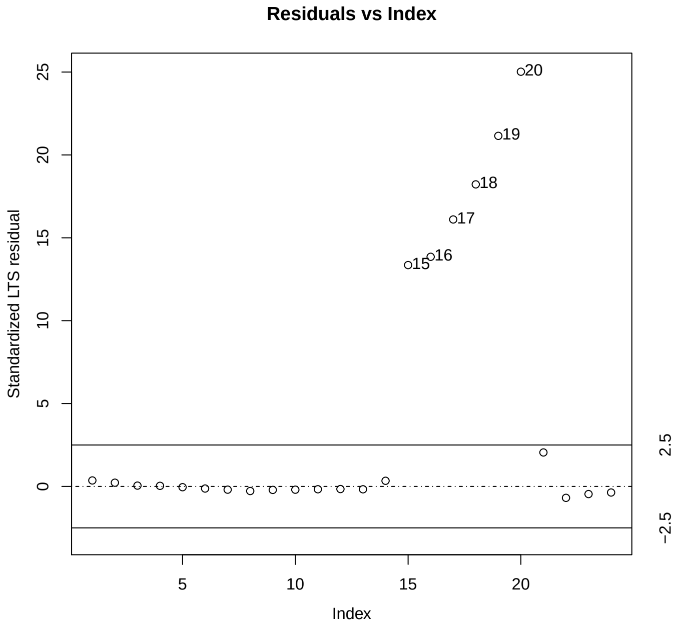


Figure 19.3: Telephone data set: Index plot of the standardized robust residuals.

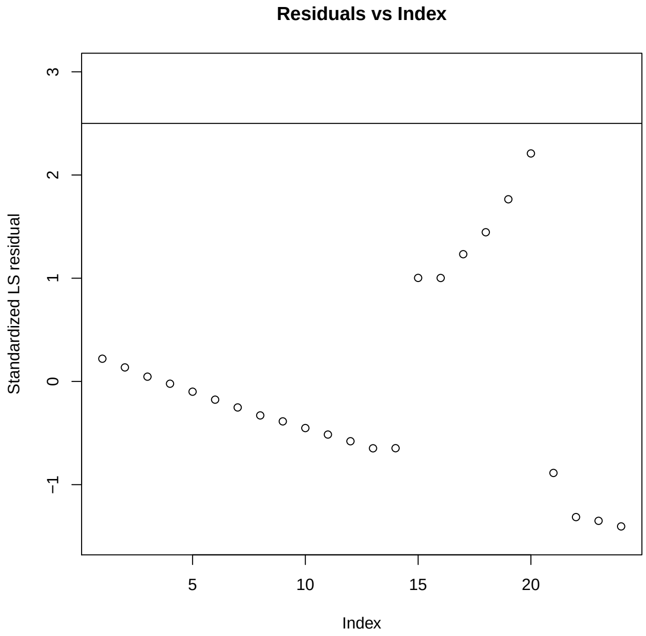


Figure 19.4: Telephone data set: Index plot of the standardized least squares residuals.

The LS residuals do not detect the vertical outliers because the LS fit itself is attracted to those outliers as it tries to make the (squared) residuals of all the cases as small as possible. A classical approach to find the vertical outliers, thus based on $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, consists of computing the deleted residual, introduced in Chapter 17, equation (17.2.3):

$$\begin{aligned} d_i &= y_i - \hat{y}_{i(i)} \\ &= \frac{e_i}{1 - h_{ii}} \end{aligned}$$

with $\hat{y}_{i(i)}$ the fitted value of case $i$, excluded from the data set to estimate the regression coefficients. It can be shown that

$$s(d_i) = \frac{s_{(i)}}{\sqrt{1 - h_{ii}}}$$

and that

$$\boxed{e_i^* = \frac{d_i}{s(d_i)} = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} \sim t_{n-p-1}.} \tag{19.1.2}$$

Hence, the $e_i^*$ are called the **studentized residuals**. They can be computed without refitting the model each time an observation is deleted, by using the relation

$$e_i^* = e_i \sqrt{\frac{n - p - 1}{\mathrm{SSE}(1 - h_{ii}) - e_i^2}}.$$

Here, also a plot of the studentized residuals of the Telephone data (Figure 19.5), does not pinpoint the outliers. This is due to the fact that the outliers are not isolated here. Deleting one of the outliers does not change the fit drastically!

```
phones.studres <- studres(phones.lm)
```

or alternatively

```
phones.lmi <- lm.influence(phones.lm)
si <- phones.lmi$sigma
h <- phones.lmi$hat
phones.studres <- residuals(phones.lm)/(si*(1-h)^0.5)
```

```
plot(phones.studres,ylim=c(-3,3))
abline(h=c(-2.5,2.5))
```

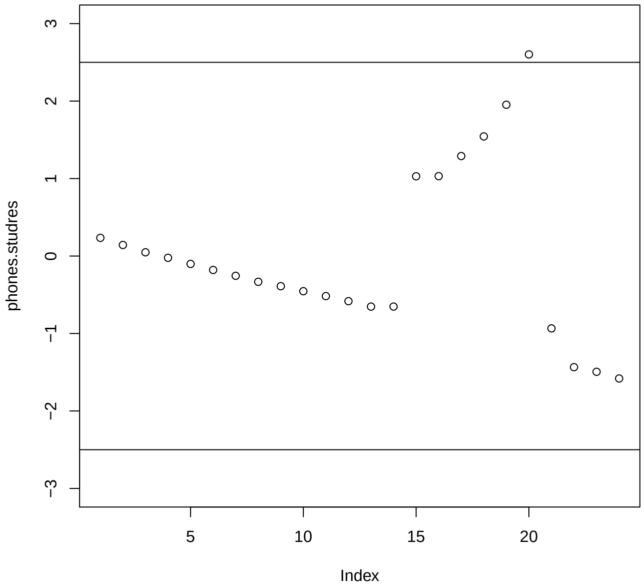


Figure 19.5: Telephone data set: Index plot of the studentized residuals.

## 19.2 Leverage points

### 19.2.1 Residuals

Bad leverage points, which are outlying observations in the predictor space that do not follow the linear model of the majority of the data points, also have a large influence on the classical LS estimator. Let us illustrate this effect on the `stars` data set. These data form the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. The regressor $X$ is the logarithm of the effective temperature at the surface of the star, and the response $Y$ is the logarithm of its light intensity.

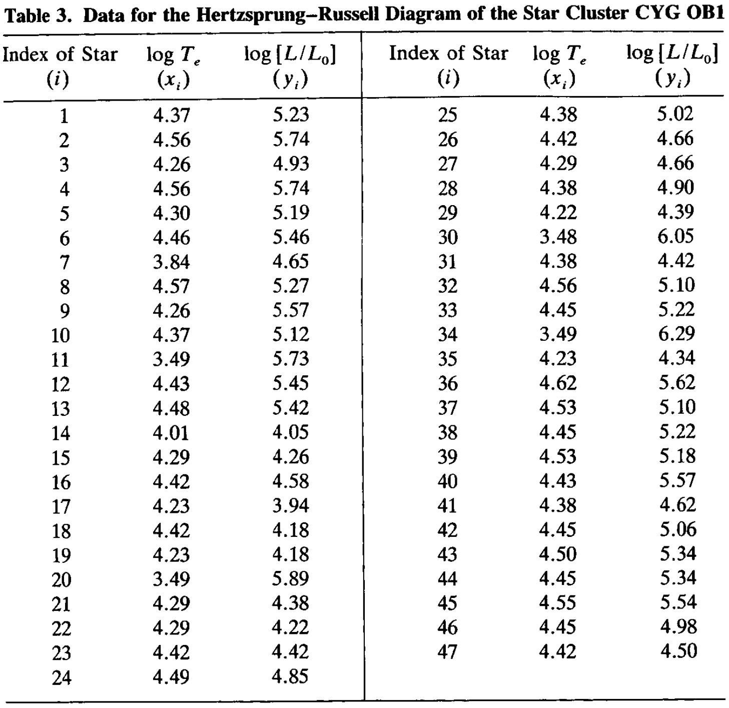


**Table 3. Data for the Hertzsprung–Russell Diagram of the Star Cluster CYG OB1**

| Index of Star ($i$) | $\log T_e$ ($x_i$) | $\log[L/L_0]$ ($y_i$) | Index of Star ($i$) | $\log T_e$ ($x_i$) | $\log[L/L_0]$ ($y_i$) |
|---|---|---|---|---|---|
| 1 | 4.37 | 5.23 | 25 | 4.38 | 5.02 |
| 2 | 4.56 | 5.74 | 26 | 4.42 | 4.66 |
| 3 | 4.26 | 4.93 | 27 | 4.29 | 4.66 |
| 4 | 4.56 | 5.74 | 28 | 4.38 | 4.90 |
| 5 | 4.30 | 5.19 | 29 | 4.22 | 4.39 |
| 6 | 4.46 | 5.46 | 30 | 3.48 | 6.05 |
| 7 | 3.84 | 4.65 | 31 | 4.38 | 4.42 |
| 8 | 4.57 | 5.27 | 32 | 4.56 | 5.10 |
| 9 | 4.26 | 5.57 | 33 | 4.45 | 5.22 |
| 10 | 4.37 | 5.12 | 34 | 3.49 | 6.29 |
| 11 | 3.49 | 5.73 | 35 | 4.23 | 4.34 |
| 12 | 4.43 | 5.45 | 36 | 4.62 | 5.62 |
| 13 | 4.48 | 5.42 | 37 | 4.53 | 5.10 |
| 14 | 4.01 | 4.05 | 38 | 4.45 | 5.22 |
| 15 | 4.29 | 4.26 | 39 | 4.53 | 5.18 |
| 16 | 4.42 | 4.58 | 40 | 4.43 | 5.57 |
| 17 | 4.23 | 3.94 | 41 | 4.38 | 4.62 |
| 18 | 4.42 | 4.18 | 42 | 4.45 | 5.06 |
| 19 | 4.23 | 4.18 | 43 | 4.50 | 5.34 |
| 20 | 3.49 | 5.89 | 44 | 4.45 | 5.34 |
| 21 | 4.29 | 4.38 | 45 | 4.55 | 5.54 |
| 22 | 4.29 | 4.22 | 46 | 4.45 | 4.98 |
| 23 | 4.42 | 4.42 | 47 | 4.42 | 4.50 |
| 24 | 4.49 | 4.85 | | | |

In the plot of the data in Figure 19.6 we see two groups of observations: the majority, following a steep band, and four stars in the upper left corner (with indices 11, 20, 30 and 34). The 43 'regular' observations lie on the main sequence, whereas the four outlying data points are giant stars. The LS fit is again highly attracted by the giant stars and does not at all reflect the linear trend of the majority of the data points, in contrast to the robust LTS fit.

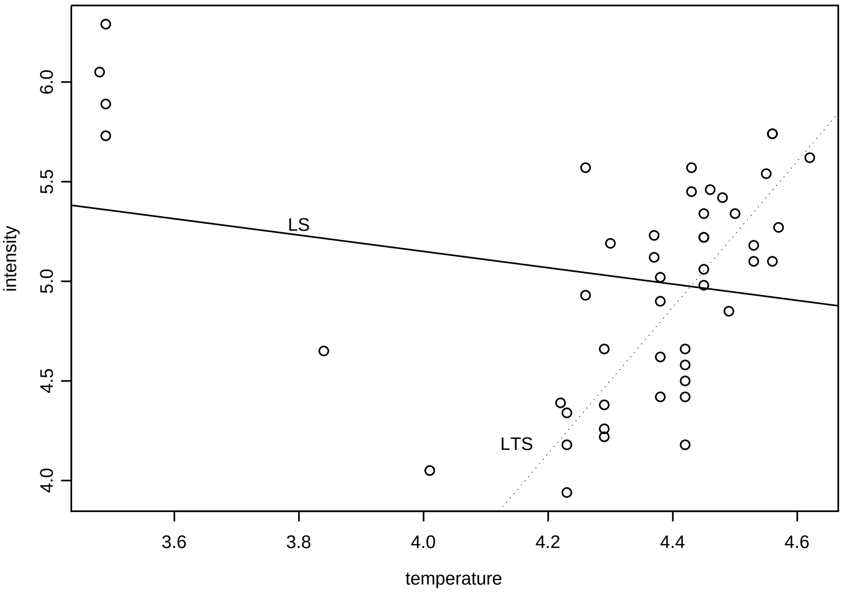


Figure 19.6: Stars data set with LS and LTS fit superimposed.

If we plot the studentized LS residuals (Figure 19.7) we can not detect any deviating observation, but the four outliers stand out in the plot of the standardized LTS residuals (Figure 19.8).

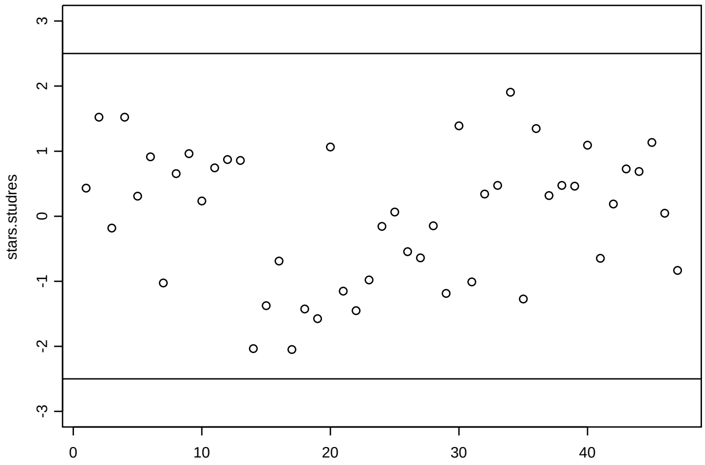


Figure 19.7: Stars data set: Index plot of studentized LS residuals.

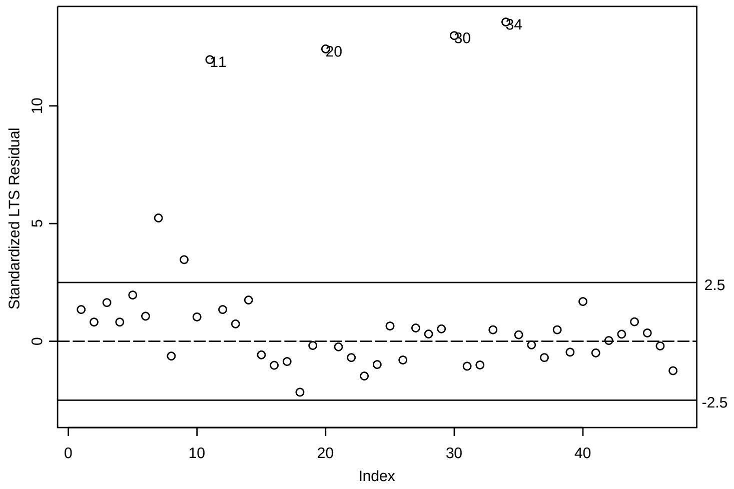


Figure 19.8: Stars data set: Index plot of standardized LTS residuals.

### 19.2.2 Diagnostic plot

In this example, the outliers are bad leverage points, hence they can be detected based on their (large) robust residuals. This residual plot however can not distinguish between bad leverage points and vertical outliers. Also good leverage points will not be highlighted on a residual plot, as they have a small residual.

Therefore we will need a metric within the $X$-space to compute the distance of each observation to the center of the data cloud. For $(p-1)$-dimensional vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i,p-1})^t$ the classical **Mahalanobis distance** is defined as:

$$\boxed{\mathrm{MD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}} \tag{19.2.1}$$

with

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

the sample mean, and

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t$$

the empirical covariance matrix of the $\boldsymbol{x}_i$. Both the sample mean and the sample covariance matrix are however non-robust: the mean will be shifted towards the outliers whereas the covariance matrix will be inflated to them. Consider e.g. Figure 19.9 with contains the logarithms of the body weight (in kilograms) and the brain weight (in grams) of 28 animals.

```
data(Animals)
Animals

                      body  brain
Mountain beaver      1.350    8.1
Cow                465.000  423.0
Grey wolf           36.330  119.5
Goat                27.660  115.0
Guinea pig           1.040    5.5
Dipliodocus      11700.000   50.0
Asian elephant    2547.000 4603.0
Donkey             187.100  419.0
Horse              521.000  655.0
Potar monkey        10.000  115.0
Cat                  3.300   25.6
Giraffe            529.000  680.0
Gorilla            207.000  406.0
Human               62.000 1320.0
African elephant  6654.000 5712.0
Triceratops       9400.000   70.0
Rhesus monkey        6.800  179.0
Kangaroo            35.000   56.0
```

```
Golden hamster      0.120    1.0
Mouse               0.023    0.4
Rabbit              2.500   12.1
Sheep              55.500  175.0
Jaguar            100.000  157.0
Chimpanzee         52.160  440.0
Rat                 0.280    1.9
Brachiosaurus   87000.000  154.5
Mole                0.122    3.0
Pig               192.000  180.0
```

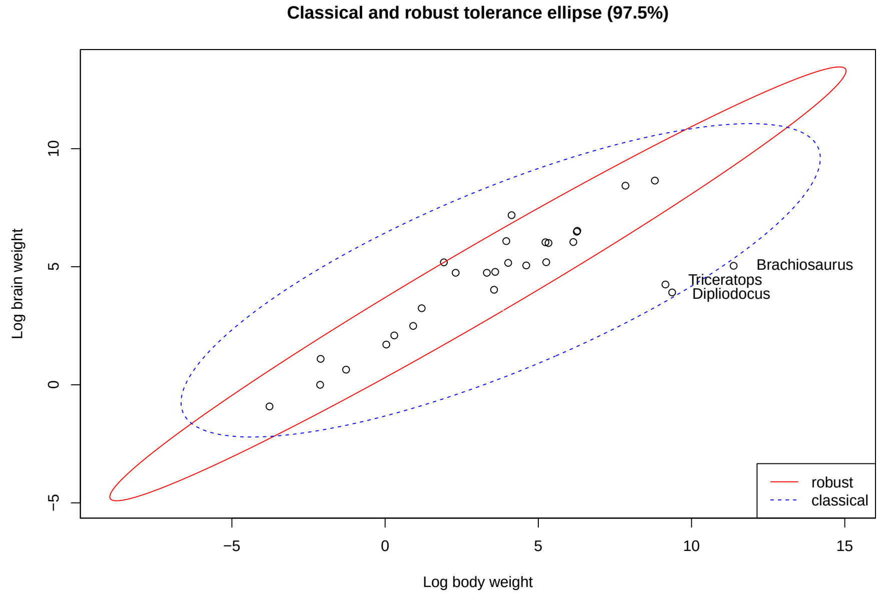


Figure 19.9: Body and brain weight for 28 animals with classical and robust tolerance ellipse superimposed.

Three animals are clearly outlying: these are dinosaurs, with a small brain as compared with a heavy body. We see that the classical mean, indicated by a plus sign, is shifted towards the outliers. The covariance matrix can be visualized through the classical *tolerance ellipsoid*, defined by

$$\{\boldsymbol{x}\ ;\ \mathrm{MD}(\boldsymbol{x}) = \sqrt{\chi^2_{p-1,0.025}}\}.$$

At a $(p-1)$-variate normal distribution this ellipsoid should contain approximately 97.5% of the data points, since the squared Mahalanobis distances are then $\chi^2_{p-1}$ distributed. We see that this ellipsoid is highly attracted to the outliers, and tries to engulf them. On the other hand, the robust tolerance ellipsoid, defined as

$$\{\boldsymbol{x}\ ;\ \mathrm{RD}(\boldsymbol{x}) = \sqrt{\chi^2_{p-1,0.025}}\}$$

is much smaller and essentially contains the majority of the data points. Here, the **robust distance** is defined analogously to the Mahalanobis distance:

$$\boxed{\mathrm{RD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_R)^t \hat{\Sigma}_R^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_R)}} \qquad (19.2.2)$$

where $\hat{\boldsymbol{\mu}}_R$ and $\hat{\Sigma}_R$ are robust estimates of the center $\boldsymbol{\mu}$ and shape $\Sigma$ of the $\boldsymbol{x}$-part of the data points. In Section 19.5 we will discuss the MCD estimator as a highly robust estimator of location and shape.

If we compare the Mahalanobis distances of the `Animals` data set with the robust distances as in Figure 19.10 we see that the MD distances are all but one smaller than $\sqrt{\chi^2_{2,0.025}} = 2.72$, whereas the robust distances of the dinosaurs are much larger than this cut-off value.

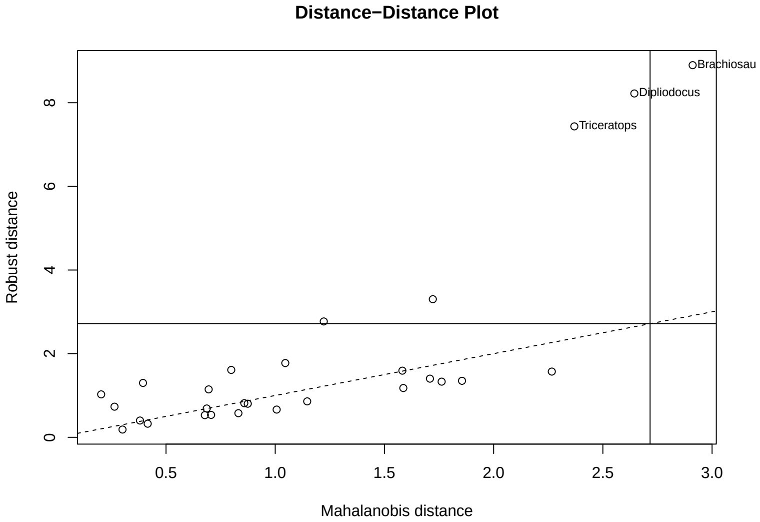


Figure 19.10: Animals data set: Robust distances versus Mahalanobis distances.

Leverage points will thus be characterized as having a large robust distance. If we now plot for each observation its standardized robust residual versus its robust distance, we obtain the **diagnostic plot**, on which the four types of observations can be distinguished as in Figure 19.11.

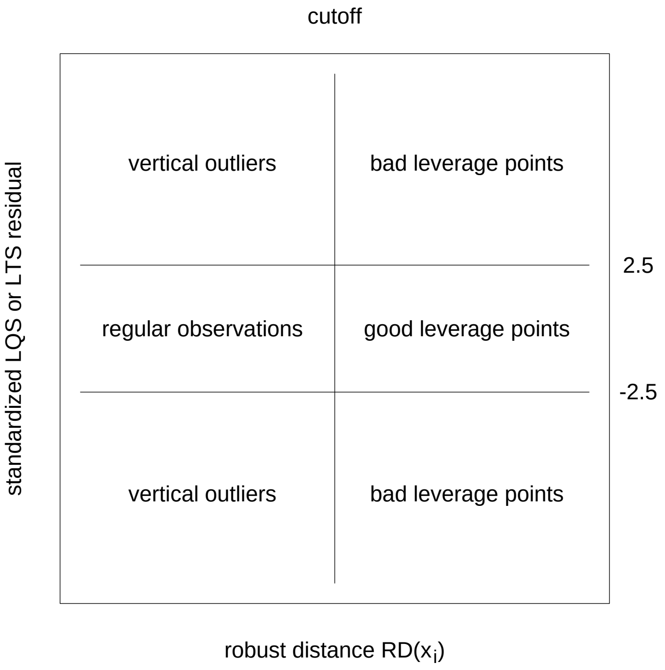


Figure 19.11: Diagnostic plot with 4 types of observations.

For the stars data set, this yields Figure 19.12, on which we clearly see the giant stars and star 7 as bad leverage points. Star 14 is a good leverage point, whereas star 9 is found to be a vertical outlier.

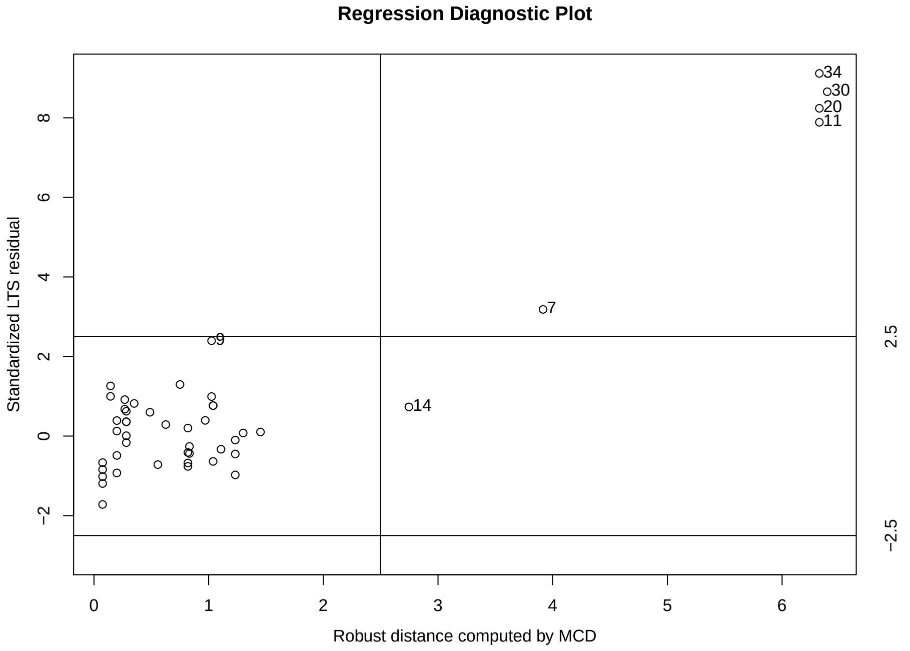


Figure 19.12: Stars data set: diagnostic plot.

### 19.2.3 The Hat matrix

Classical diagnostics to detect leverage points are based on the hat matrix

$$H = X(X^tX)^{-1}X^t$$

as defined in (12.2.5) which transforms the observed response vector $\boldsymbol{y}$ into its LS estimate

$$\boxed{\hat{\mathbf{y}} = H\boldsymbol{y}}$$

or equivalently

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n.$$

(Note that the $X$ matrix here includes a constant column of ones for the intercept term.) The element $h_{ij}$ of $H$ thus measures the effect of the $j$th observation on $\hat{y}_i$, and the diagonal element $h_{ii}$ the effect of the $i$th observation on its own prediction. A diagonal element $h_{ii} = 0$ indicates a point with no influence on the fit. Since

$$\mathrm{tr}(H) = \mathrm{tr}(X(X^tX)^{-1}X^t) = \mathrm{tr}(X^tX(X^tX)^{-1}) = \mathrm{tr}(I_p) = p$$

we have

$$\sum_{i=1}^{n} h_{ii} = p$$

and consequently

$$\bar{h}_{ii} = p/n.$$

Moreover, since $H$ is symmetric $H^t = H$ and idempotent $HH = H$, we see that

$$\begin{aligned} h_{ii} = (HH)_{ii} &= \sum_{j=1}^{n} h_{ij}h_{ji} \\ &= h_{ii}h_{ii} + \sum_{j\neq i} h_{ij}h_{ji} \\ &= h_{ii}^2 + \sum_{j\neq i} h_{ij}^2 \end{aligned}$$

and thus $0 \leqslant h_{ii}$ and $h_{ii} \geqslant h_{ii}^2$ for all $i = 1, \ldots, n$. Finally, this implies

$$0 \leqslant h_{ii} \leqslant 1.$$

These limits do not yet tell us when $h_{ii}$ is large. Some authors suggest to use

$$h_{ii} > \frac{2p}{n}$$

as cut-off value. Note that, when $h_{ii} = 1$, $h_{ij} = 0$ for all $j \neq i$, and consequently $\hat{y}_i = 1y_i$ and $r_i = y_i - \hat{y}_i = 0$. The $i$th observation is thus so influential that the LS fit passes through it. Moreover, the variance of the $i$th residual is then zero:

$$s^2(e_i) = \text{MSE}(1 - h_{ii}) = 0.$$

It can also be shown that there is a one-by-one correspondence between the squared Mahalanobis distance (19.2.1) for object $i$ and its $h_{ii}$:

$$h_{ii} = \frac{1}{n-1}\text{MD}_i^2 + \frac{1}{n}. \tag{19.2.3}$$

From this expression, we see that $h_{ii}$ also measures the distance of $\boldsymbol{x}_i$ to the center of the data points in the $X$-space. On the other hand, equation (19.2.3) shows that the hat-diagnostic is not robust!

### Example 1: the Telephone data set

Table 2 lists some diagnostics for the Telephone data set. We see that none of the observations has a leverage which is larger that $2p/n = 0.167$ or a squared Mahalanobis distance which exceeds $\chi^2_{1,0.05} = 3.84$. This is not surprising as the only regressor in this data set Year does not contain any outlying value. Note that the legend of this table contains a different terminology than the one used in this book: the standardized residuals $e_i^{(s)}$, defined by (13.7.1), are here denoted as 'studentized residuals' $t_i$, and the studentized residuals $e_i^*$, introduced in (19.1.2), as 'jackknifed residuals' $t(i)$. The standardized residuals of Table 1 are obtained as $e_i/s$ and can also be compared to the cut-off value 2.5.

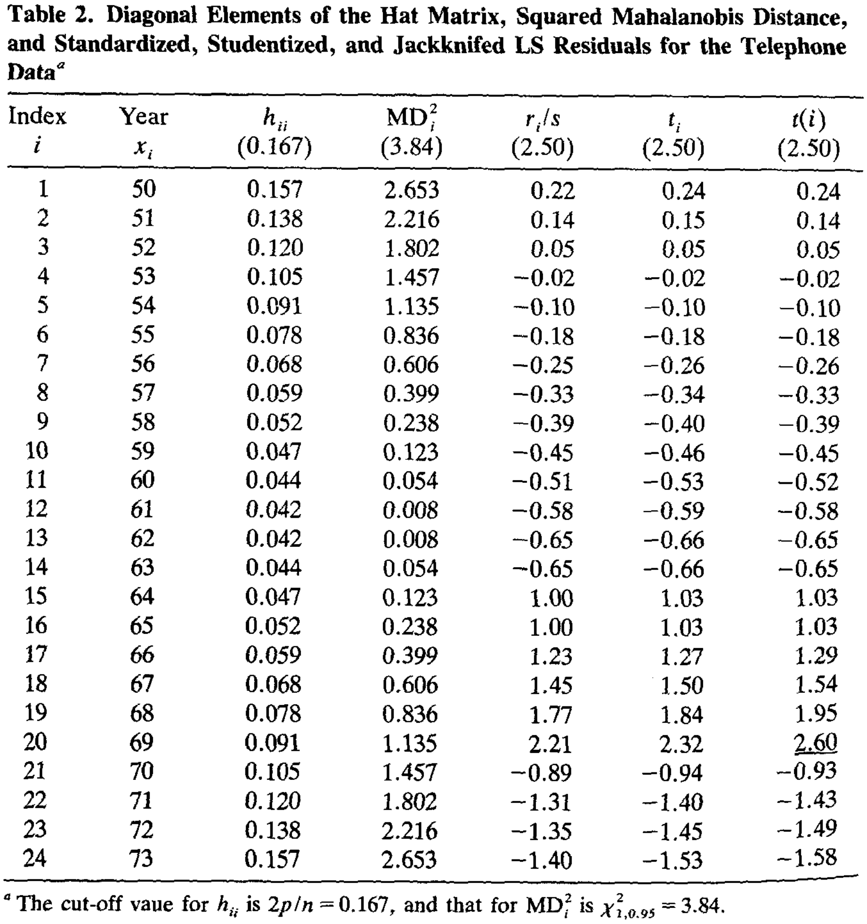


**Table 2. Diagonal Elements of the Hat Matrix, Squared Mahalanobis Distance, and Standardized, Studentized, and Jackknifed LS Residuals for the Telephone Data**[a]

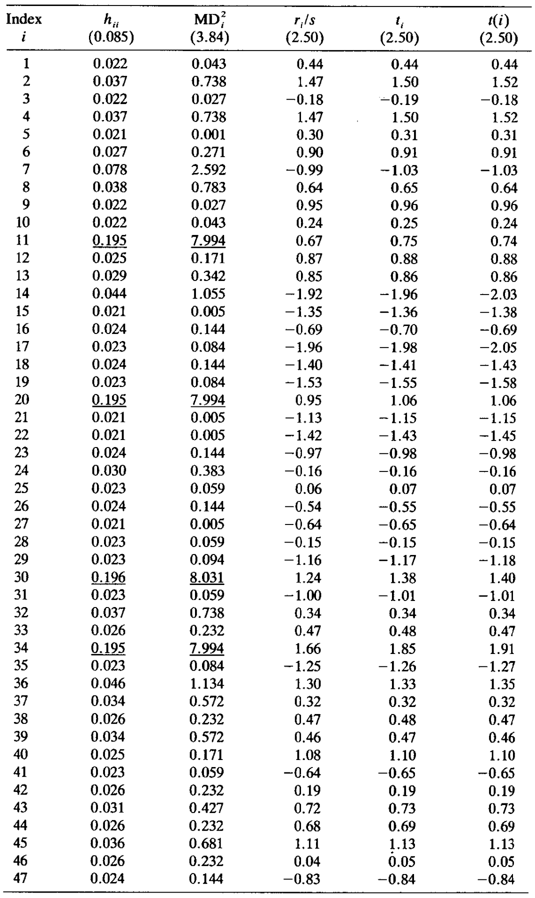
| Index $i$ | Year $x_i$ | $h_{ii}$ (0.167) | $\mathrm{MD}_i^2$ (3.84) | $r_i/s$ (2.50) | $t_i$ (2.50) | $t(i)$ (2.50) |
|---|---|---|---|---|---|---|
| 1 | 50 | 0.157 | 2.653 | 0.22 | 0.24 | 0.24 |
| 2 | 51 | 0.138 | 2.216 | 0.14 | 0.15 | 0.14 |
| 3 | 52 | 0.120 | 1.802 | 0.05 | 0.05 | 0.05 |
| 4 | 53 | 0.105 | 1.457 | −0.02 | −0.02 | −0.02 |
| 5 | 54 | 0.091 | 1.135 | −0.10 | −0.10 | −0.10 |
| 6 | 55 | 0.078 | 0.836 | −0.18 | −0.18 | −0.18 |
| 7 | 56 | 0.068 | 0.606 | −0.25 | −0.26 | −0.26 |
| 8 | 57 | 0.059 | 0.399 | −0.33 | −0.34 | −0.33 |
| 9 | 58 | 0.052 | 0.238 | −0.39 | −0.40 | −0.39 |
| 10 | 59 | 0.047 | 0.123 | −0.45 | −0.46 | −0.45 |
| 11 | 60 | 0.044 | 0.054 | −0.51 | −0.53 | −0.52 |
| 12 | 61 | 0.042 | 0.008 | −0.58 | −0.59 | −0.58 |
| 13 | 62 | 0.042 | 0.008 | −0.65 | −0.66 | −0.65 |
| 14 | 63 | 0.044 | 0.054 | −0.65 | −0.66 | −0.65 |
| 15 | 64 | 0.047 | 0.123 | 1.00 | 1.03 | 1.03 |
| 16 | 65 | 0.052 | 0.238 | 1.00 | 1.03 | 1.03 |
| 17 | 66 | 0.059 | 0.399 | 1.23 | 1.27 | 1.29 |
| 18 | 67 | 0.068 | 0.606 | 1.45 | 1.50 | 1.54 |
| 19 | 68 | 0.078 | 0.836 | 1.77 | 1.84 | 1.95 |
| 20 | 69 | 0.091 | 1.135 | 2.21 | 2.32 | $\underline{2.60}$ |
| 21 | 70 | 0.105 | 1.457 | −0.89 | −0.94 | −0.93 |
| 22 | 71 | 0.120 | 1.802 | −1.31 | −1.40 | −1.43 |
| 23 | 72 | 0.138 | 2.216 | −1.35 | −1.45 | −1.49 |
| 24 | 73 | 0.157 | 2.653 | −1.40 | −1.53 | −1.58 |

[a] The cut-off vaue for $h_{ii}$ is $2p/n = 0.167$, and that for $\mathrm{MD}_i^2$ is $\chi^2_{1,0.95} = 3.84$.

### Example 2: the Stars data set

Table 1 lists the same diagnostics for the `Stars` data set. We see that both the diagonal elements of the hat matrix and the Mahalanobis distances of the giant stars exceed their cut-off value. So in this example these diagnostics are able to identify the most extreme outliers in $\boldsymbol{x}$. But the bad leverage point 7, and the good leverage point 14 are not recognized.

**Table 1. Diagonal Elements of the Hat Matrix, Squared Mahalanobis Distance, and Standardized, Studentized, and Jackknifed LS Residuals for the Hertzsprung–Russell Diagram Data**[a]



| Index<br>$i$ | $h_{ii}$<br>(0.085) | $\text{MD}_i^2$<br>(3.84) | $r_i/s$<br>(2.50) | $t_i$<br>(2.50) | $t(i)$<br>(2.50) |
|---|---|---|---|---|---|
| 1 | 0.022 | 0.043 | 0.44 | 0.44 | 0.44 |
| 2 | 0.037 | 0.738 | 1.47 | 1.50 | 1.52 |
| 3 | 0.022 | 0.027 | −0.18 | −0.19 | −0.18 |
| 4 | 0.037 | 0.738 | 1.47 | 1.50 | 1.52 |
| 5 | 0.021 | 0.001 | 0.30 | 0.31 | 0.31 |
| 6 | 0.027 | 0.271 | 0.90 | 0.91 | 0.91 |
| 7 | 0.078 | 2.592 | −0.99 | −1.03 | −1.03 |
| 8 | 0.038 | 0.783 | 0.64 | 0.65 | 0.64 |
| 9 | 0.022 | 0.027 | 0.95 | 0.96 | 0.96 |
| 10 | 0.022 | 0.043 | 0.24 | 0.25 | 0.24 |
| 11 | 0.195 | 7.994 | 0.67 | 0.75 | 0.74 |
| 12 | 0.025 | 0.171 | 0.87 | 0.88 | 0.88 |
| 13 | 0.029 | 0.342 | 0.85 | 0.86 | 0.86 |
| 14 | 0.044 | 1.055 | −1.92 | −1.96 | −2.03 |
| 15 | 0.021 | 0.005 | −1.35 | −1.36 | −1.38 |
| 16 | 0.024 | 0.144 | −0.69 | −0.70 | −0.69 |
| 17 | 0.023 | 0.084 | −1.96 | −1.98 | −2.05 |
| 18 | 0.024 | 0.144 | −1.40 | −1.41 | −1.43 |
| 19 | 0.023 | 0.084 | −1.53 | −1.55 | −1.58 |
| 20 | 0.195 | 7.994 | 0.95 | 1.06 | 1.06 |
| 21 | 0.021 | 0.005 | −1.13 | −1.15 | −1.15 |
| 22 | 0.021 | 0.005 | −1.42 | −1.43 | −1.45 |
| 23 | 0.024 | 0.144 | −0.97 | −0.98 | −0.98 |
| 24 | 0.030 | 0.383 | −0.16 | −0.16 | −0.16 |
| 25 | 0.023 | 0.059 | 0.06 | 0.07 | 0.07 |
| 26 | 0.024 | 0.144 | −0.54 | −0.55 | −0.55 |
| 27 | 0.021 | 0.005 | −0.64 | −0.65 | −0.64 |
| 28 | 0.023 | 0.059 | −0.15 | −0.15 | −0.15 |
| 29 | 0.023 | 0.094 | −1.16 | −1.17 | −1.18 |
| 30 | 0.196 | 8.031 | 1.24 | 1.38 | 1.40 |
| 31 | 0.023 | 0.059 | −1.00 | −1.01 | −1.01 |
| 32 | 0.037 | 0.738 | 0.34 | 0.34 | 0.34 |
| 33 | 0.026 | 0.232 | 0.47 | 0.48 | 0.47 |
| 34 | 0.195 | 7.994 | 1.66 | 1.85 | 1.91 |
| 35 | 0.023 | 0.084 | −1.25 | −1.26 | −1.27 |
| 36 | 0.046 | 1.134 | 1.30 | 1.33 | 1.35 |
| 37 | 0.034 | 0.572 | 0.32 | 0.32 | 0.32 |
| 38 | 0.026 | 0.232 | 0.47 | 0.48 | 0.47 |
| 39 | 0.034 | 0.572 | 0.46 | 0.47 | 0.46 |
| 40 | 0.025 | 0.171 | 1.08 | 1.10 | 1.10 |
| 41 | 0.023 | 0.059 | −0.64 | −0.65 | −0.65 |
| 42 | 0.026 | 0.232 | 0.19 | 0.19 | 0.19 |
| 43 | 0.031 | 0.427 | 0.72 | 0.73 | 0.73 |
| 44 | 0.026 | 0.232 | 0.68 | 0.69 | 0.69 |
| 45 | 0.036 | 0.681 | 1.11 | 1.13 | 1.13 |
| 46 | 0.026 | 0.232 | 0.04 | 0.05 | 0.05 |
| 47 | 0.024 | 0.144 | −0.83 | −0.84 | −0.84 |

[a] The cut-off value for $h_{ii}$ is $2p/n = 0.85$, and that for $\text{MD}_i^2$ is $\chi^2_{1,0.95} = 3.84$.

### Example 3: the Hawkins-Bradu-Kass data set

This artifical data set contains 75 observations in four dimensions and is listed in Table 9. The first 10 observations are bad leverage points, and the next four points are good leverage points.

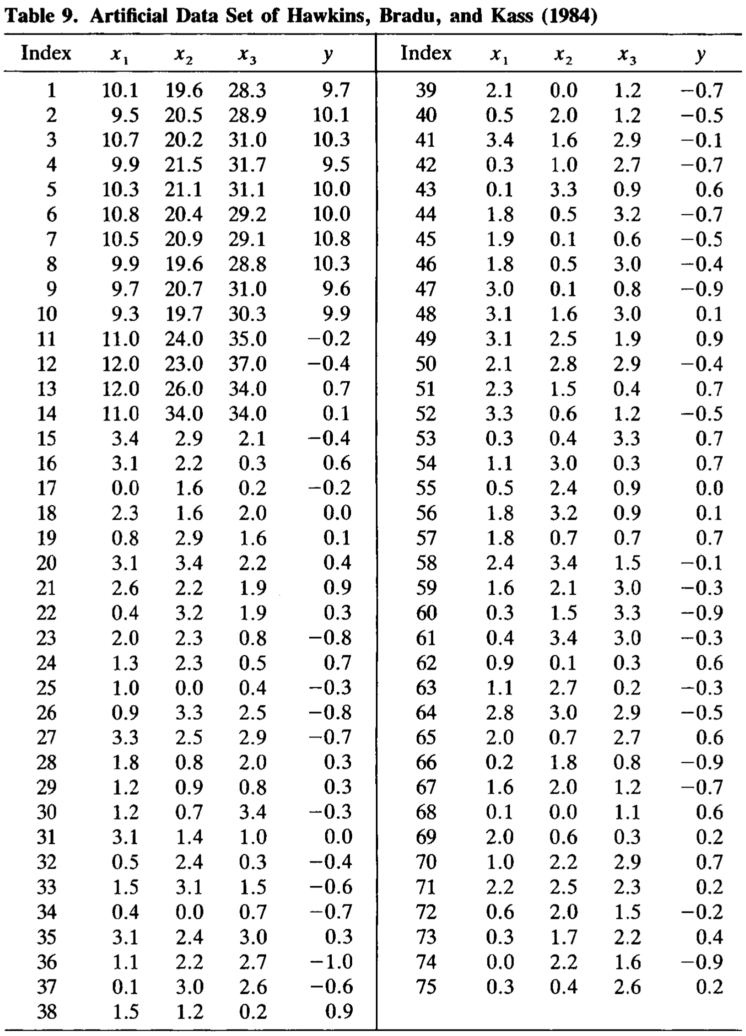


**Table 9. Artificial Data Set of Hawkins, Bradu, and Kass (1984)**

| Index | $x_1$ | $x_2$ | $x_3$ | $y$ | Index | $x_1$ | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.1 | 19.6 | 28.3 | 9.7 | 39 | 2.1 | 0.0 | 1.2 | −0.7 |
| 2 | 9.5 | 20.5 | 28.9 | 10.1 | 40 | 0.5 | 2.0 | 1.2 | −0.5 |
| 3 | 10.7 | 20.2 | 31.0 | 10.3 | 41 | 3.4 | 1.6 | 2.9 | −0.1 |
| 4 | 9.9 | 21.5 | 31.7 | 9.5 | 42 | 0.3 | 1.0 | 2.7 | −0.7 |
| 5 | 10.3 | 21.1 | 31.1 | 10.0 | 43 | 0.1 | 3.3 | 0.9 | 0.6 |
| 6 | 10.8 | 20.4 | 29.2 | 10.0 | 44 | 1.8 | 0.5 | 3.2 | −0.7 |
| 7 | 10.5 | 20.9 | 29.1 | 10.8 | 45 | 1.9 | 0.1 | 0.6 | −0.5 |
| 8 | 9.9 | 19.6 | 28.8 | 10.3 | 46 | 1.8 | 0.5 | 3.0 | −0.4 |
| 9 | 9.7 | 20.7 | 31.0 | 9.6 | 47 | 3.0 | 0.1 | 0.8 | −0.9 |
| 10 | 9.3 | 19.7 | 30.3 | 9.9 | 48 | 3.1 | 1.6 | 3.0 | 0.1 |
| 11 | 11.0 | 24.0 | 35.0 | −0.2 | 49 | 3.1 | 2.5 | 1.9 | 0.9 |
| 12 | 12.0 | 23.0 | 37.0 | −0.4 | 50 | 2.1 | 2.8 | 2.9 | −0.4 |
| 13 | 12.0 | 26.0 | 34.0 | 0.7 | 51 | 2.3 | 1.5 | 0.4 | 0.7 |
| 14 | 11.0 | 34.0 | 34.0 | 0.1 | 52 | 3.3 | 0.6 | 1.2 | −0.5 |
| 15 | 3.4 | 2.9 | 2.1 | −0.4 | 53 | 0.3 | 0.4 | 3.3 | 0.7 |
| 16 | 3.1 | 2.2 | 0.3 | 0.6 | 54 | 1.1 | 3.0 | 0.3 | 0.7 |
| 17 | 0.0 | 1.6 | 0.2 | −0.2 | 55 | 0.5 | 2.4 | 0.9 | 0.0 |
| 18 | 2.3 | 1.6 | 2.0 | 0.0 | 56 | 1.8 | 3.2 | 0.9 | 0.1 |
| 19 | 0.8 | 2.9 | 1.6 | 0.1 | 57 | 1.8 | 0.7 | 0.7 | 0.7 |
| 20 | 3.1 | 3.4 | 2.2 | 0.4 | 58 | 2.4 | 3.4 | 1.5 | −0.1 |
| 21 | 2.6 | 2.2 | 1.9 | 0.9 | 59 | 1.6 | 2.1 | 3.0 | −0.3 |
| 22 | 0.4 | 3.2 | 1.9 | 0.3 | 60 | 0.3 | 1.5 | 3.3 | −0.9 |
| 23 | 2.0 | 2.3 | 0.8 | −0.8 | 61 | 0.4 | 3.4 | 3.0 | −0.3 |
| 24 | 1.3 | 2.3 | 0.5 | 0.7 | 62 | 0.9 | 0.1 | 0.3 | 0.6 |
| 25 | 1.0 | 0.0 | 0.4 | −0.3 | 63 | 1.1 | 2.7 | 0.2 | −0.3 |
| 26 | 0.9 | 3.3 | 2.5 | −0.8 | 64 | 2.8 | 3.0 | 2.9 | −0.5 |
| 27 | 3.3 | 2.5 | 2.9 | −0.7 | 65 | 2.0 | 0.7 | 2.7 | 0.6 |
| 28 | 1.8 | 0.8 | 2.0 | 0.3 | 66 | 0.2 | 1.8 | 0.8 | −0.9 |
| 29 | 1.2 | 0.9 | 0.8 | 0.3 | 67 | 1.6 | 2.0 | 1.2 | −0.7 |
| 30 | 1.2 | 0.7 | 3.4 | −0.3 | 68 | 0.1 | 0.0 | 1.1 | 0.6 |
| 31 | 3.1 | 1.4 | 1.0 | 0.0 | 69 | 2.0 | 0.6 | 0.3 | 0.2 |
| 32 | 0.5 | 2.4 | 0.3 | −0.4 | 70 | 1.0 | 2.2 | 2.9 | 0.7 |
| 33 | 1.5 | 3.1 | 1.5 | −0.6 | 71 | 2.2 | 2.5 | 2.3 | 0.2 |
| 34 | 0.4 | 0.0 | 0.7 | −0.7 | 72 | 0.6 | 2.0 | 1.5 | −0.2 |
| 35 | 3.1 | 2.4 | 3.0 | 0.3 | 73 | 0.3 | 1.7 | 2.2 | 0.4 |
| 36 | 1.1 | 2.2 | 2.7 | −1.0 | 74 | 0.0 | 2.2 | 1.6 | −0.9 |
| 37 | 0.1 | 3.0 | 2.6 | −0.6 | 75 | 0.3 | 0.4 | 2.6 | 0.2 |
| 38 | 1.5 | 1.2 | 0.2 | 0.9 | | | | | |

If we now take a look at the $h_{ii}$ and the $\mathrm{MD}(\boldsymbol{x}_i)$ values for these 14 leverage points, we see that the classical diagnostics fail completely. The LS standardized and studentized residuals are large for the good leverage points, but not for the bad leverage points. The LS fit is thus tilted towards these bad leverage points, and renders them into regular observations. Even the diagnostics in the predictor space cannot identify the first 10 observations. The good leverage points on the other hand are converted into bad leverage points.

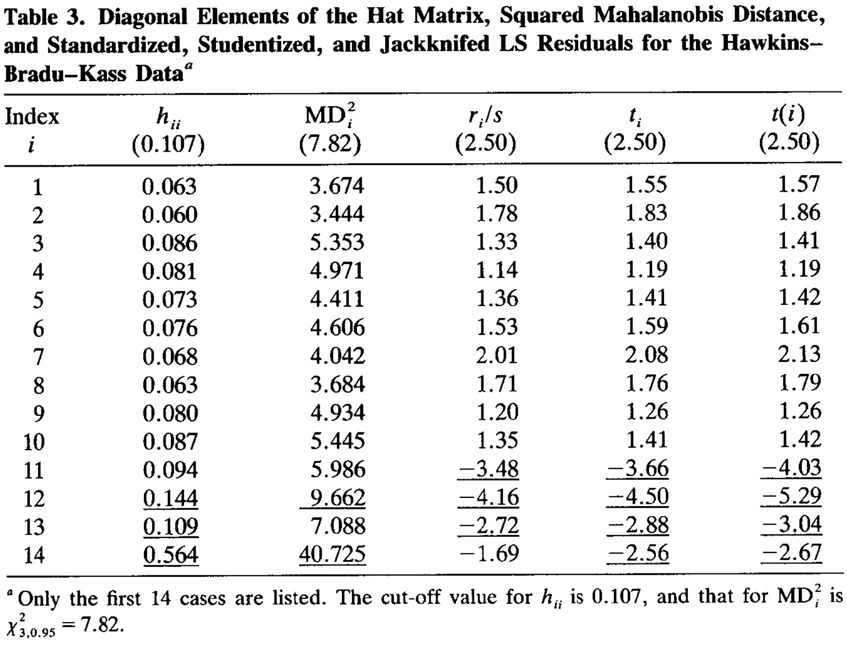


**Table 3. Diagonal Elements of the Hat Matrix, Squared Mahalanobis Distance, and Standardized, Studentized, and Jackknifed LS Residuals for the Hawkins–Bradu–Kass Data**[a]

| Index $i$ | $h_{ii}$ (0.107) | $\mathrm{MD}_i^2$ (7.82) | $r_i/s$ (2.50) | $t_i$ (2.50) | $t(i)$ (2.50) |
|---|---|---|---|---|---|
| 1 | 0.063 | 3.674 | 1.50 | 1.55 | 1.57 |
| 2 | 0.060 | 3.444 | 1.78 | 1.83 | 1.86 |
| 3 | 0.086 | 5.353 | 1.33 | 1.40 | 1.41 |
| 4 | 0.081 | 4.971 | 1.14 | 1.19 | 1.19 |
| 5 | 0.073 | 4.411 | 1.36 | 1.41 | 1.42 |
| 6 | 0.076 | 4.606 | 1.53 | 1.59 | 1.61 |
| 7 | 0.068 | 4.042 | 2.01 | 2.08 | 2.13 |
| 8 | 0.063 | 3.684 | 1.71 | 1.76 | 1.79 |
| 9 | 0.080 | 4.934 | 1.20 | 1.26 | 1.26 |
| 10 | 0.087 | 5.445 | 1.35 | 1.41 | 1.42 |
| 11 | 0.094 | 5.986 | −3.48 | −3.66 | −4.03 |
| 12 | 0.144 | 9.662 | −4.16 | −4.50 | −5.29 |
| 13 | 0.109 | 7.088 | −2.72 | −2.88 | −3.04 |
| 14 | 0.564 | 40.725 | −1.69 | −2.56 | −2.67 |

[a] Only the first 14 cases are listed. The cut-off value for $h_{ii}$ is 0.107, and that for $\mathrm{MD}_i^2$ is $\chi^2_{3,0.95} = 7.82$.

## 19.3 Single-case diagnostics

After the outlying observations in $X$- or $Y$-space are identified, classical diagnostics proceed in asserting the influence of these outlying cases on the regression fit. There exist several single-case diagnostics that are based on the omission of a single case to measure its influence. Since all these diagnostics are characterized by some function of the LS residuals, or the diagonal elements of the hat matrix, they will however not be able to identify the true influential data points when the outliers in the data set are not isolated.

### 19.3.1 DFFITS

A measure of the influence that case $i$ has on the fitted value $\hat{y}_i$ is

$$\boxed{\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{s^2_{(i)} h_{ii}}}} \tag{19.3.1}$$

Since $\hat{y}_i = \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}$, the variance of the fitted value equals

$$\text{Var}(\hat{y}_i) = \boldsymbol{x}_i^t \sigma^2 (X^t X)^{-1} \boldsymbol{x}_i = \sigma^2 h_{ii}$$

and is usually estimated by

$$s^2(\hat{y}_i) = s^2 h_{ii}.$$

In (19.3.1) the denominator is the estimated standard deviation of $\hat{y}_i$ but the unknown error variance is now estimated by the MSE obtained by omitting the $i$th case from the data set. The DFFITS value thus measures the (standardized) effect on the prediction when an observation is deleted.

Like other single-case diagnostics, e.g. the deleted residual (17.2.3), the DFFITS values can be computed by the results from fitting the entire data set:

$$\text{DFFITS}_i = e_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

The DFFITS$_i$ value thus depends on the size of the studentized residual $e_i^*$ and the leverage value $h_{ii}$, and will be large if either $e_i^*$ is large, or $h_{ii}$ is large or they are both large. A case is considered to be influential if

$$|\text{DFFITS}_i| > 2\sqrt{p/n}.$$

### 19.3.2 Cook's distance

Cook's distance $D_i$ measures the influence of the $i$th case on all $n$ fitted values. Let $\hat{y}_{j(i)} = \boldsymbol{x}_j^t \hat{\boldsymbol{\beta}}_{(i)}$ then it is defined as

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2} \tag{19.3.2}$$

$$= \frac{e_i^2}{ps^2} \frac{h_{ii}}{(1 - h_{ii})^2} \tag{19.3.3}$$

$$= (e_i^{(s)})^2 \frac{h_{ii}}{1 - h_{ii}} \frac{1}{p} \tag{19.3.4}$$

and thus $D_i$ is essentially the same as the square of the $\text{DFFITS}_i$.

In matrix notation, it can be written as

$$D_i = (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})^t (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})/(ps^2)$$

so $D_i$ is the squared distance between $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_{(i)}$, divided by $ps^2$. Since $(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}) = X(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$ it is also equivalent to

$$D_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^t (X^t X)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})/(ps^2).$$

Hence $D_i$ also measures the influence of the $i$th case on the regression coefficients.

There is no formal test to decide when $D_i$ is large. We can simply compare the sizes of the large $D_i$ with the base level indicated by the majority of the distances. Some authors suggest to declare a data point influential if

$$D_i > 1.$$

### 19.3.3 DFBETAS

The DFBETAS measure computes for each case $i$ its influence of each regression coefficient $\hat{\beta}_j$:

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s_{(i)}^2 (X^t X)_{jj}^{-1}}} \tag{19.3.5}$$

for $j = 1, \ldots, p-1$. The denominator is an estimate of the standard error of $\hat{\beta}_j$, because from (12.2.15) it follows that

$$s(\hat{\beta}_j) = \sqrt{\sigma^2 (X^t X)_{jj}^{-1}}.$$

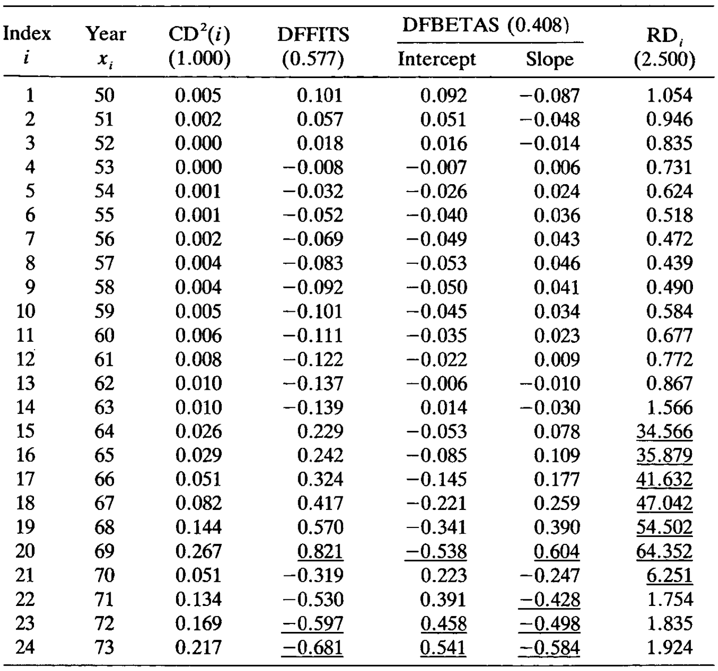
A guideline to identify influential cases is

$$|\mathrm{DFBETAS}_{ij}| > 2/\sqrt{n}.$$

### 19.3.4 Examples

Let us look at the values of the single-case diagnostics DFFITS$_i$, $D_i$ and DFBETAS$_i$ for the three data sets considered before. Table 5 lists the outliers diagnostics for the `Telephone` data. None of the diagnostics identifies the vertical outliers 15-19, and only the DFFITS and DFBETAS values for observation 20 exceed the corresponding cut-off values. Moreover, also two regular data points (23 and 24) are declared as influential!

**Table 5. Outlier Diagnostics (Including the Resistant Diagnostic RD$_i$) for the Telephone Data**



| Index | Year | CD$^2(i)$ | DFFITS | DFBETAS (0.408) | | RD$_i$ |
|---|---|---|---|---|---|---|
| $i$ | $x_i$ | (1.000) | (0.577) | Intercept | Slope | (2.500) |
| 1 | 50 | 0.005 | 0.101 | 0.092 | −0.087 | 1.054 |
| 2 | 51 | 0.002 | 0.057 | 0.051 | −0.048 | 0.946 |
| 3 | 52 | 0.000 | 0.018 | 0.016 | −0.014 | 0.835 |
| 4 | 53 | 0.000 | −0.008 | −0.007 | 0.006 | 0.731 |
| 5 | 54 | 0.001 | −0.032 | −0.026 | 0.024 | 0.624 |
| 6 | 55 | 0.001 | −0.052 | −0.040 | 0.036 | 0.518 |
| 7 | 56 | 0.002 | −0.069 | −0.049 | 0.043 | 0.472 |
| 8 | 57 | 0.004 | −0.083 | −0.053 | 0.046 | 0.439 |
| 9 | 58 | 0.004 | −0.092 | −0.050 | 0.041 | 0.490 |
| 10 | 59 | 0.005 | −0.101 | −0.045 | 0.034 | 0.584 |
| 11 | 60 | 0.006 | −0.111 | −0.035 | 0.023 | 0.677 |
| 12 | 61 | 0.008 | −0.122 | −0.022 | 0.009 | 0.772 |
| 13 | 62 | 0.010 | −0.137 | −0.006 | −0.010 | 0.867 |
| 14 | 63 | 0.010 | −0.139 | 0.014 | −0.030 | 1.566 |
| 15 | 64 | 0.026 | 0.229 | −0.053 | 0.078 | 34.566 |
| 16 | 65 | 0.029 | 0.242 | −0.085 | 0.109 | 35.879 |
| 17 | 66 | 0.051 | 0.324 | −0.145 | 0.177 | 41.632 |
| 18 | 67 | 0.082 | 0.417 | −0.221 | 0.259 | 47.042 |
| 19 | 68 | 0.144 | 0.570 | −0.341 | 0.390 | 54.502 |
| 20 | 69 | 0.267 | 0.821 | −0.538 | 0.604 | 64.352 |
| 21 | 70 | 0.051 | −0.319 | 0.223 | −0.247 | 6.251 |
| 22 | 71 | 0.134 | −0.530 | 0.391 | −0.428 | 1.754 |
| 23 | 72 | 0.169 | −0.597 | 0.458 | −0.498 | 1.835 |
| 24 | 73 | 0.217 | −0.681 | 0.541 | −0.584 | 1.924 |

Note that these diagnostics can be computed in R with the following commands:

```r
e <- residuals(phones.lm)
phones.dfbetas <- dfbetas(phones.lm)
phones.dffits <- h^0.5*e/(si*(1-h))
p <- phones.lm$rank
phones.stres <- stdres(phones.lm)
phones.cd <- (1/p * phones.stres^2 * h)/(1 - h)
```

Next, we consider the outlier diagnostics for the `Stars` data set, listed in Table 4. We see that the giant stars 11, 20, 30 and 34 are not noticed by Cook's distance $D_i$, but DFFITS and DFBETAS are more powerful.

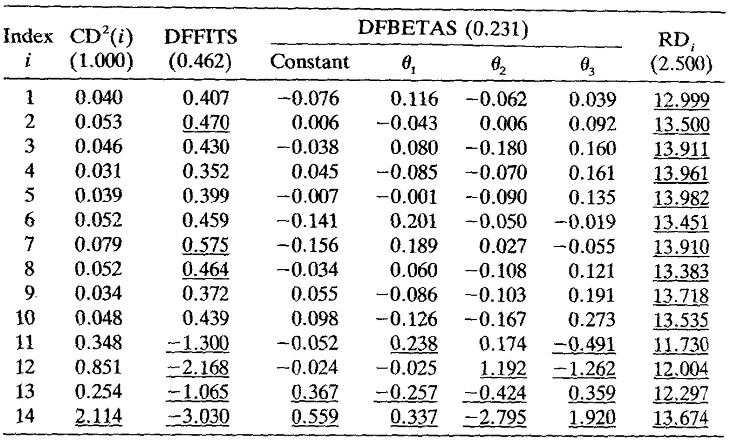
**Table 4. Outlier Diagnostics (Including the Resistant Diagnostic $\mathrm{RD}_i$, which will be explained in Section 6) for the Hertzsprung–Russell Diagram Data[a]**

| Index | $\mathrm{CD}^2(i)$ | DFFITS | DFBETAS (0.292) | | $\mathrm{RD}_i$ |
|---|---|---|---|---|---|
| $i$ | (1.000) | (0.413) | Intercept | Slope | (2.500) |
| 1 | 0.002 | 0.066 | −0.009 | 0.014 | 0.841 |
| 2 | 0.044 | 0.300 | −0.181 | 0.197 | 1.111 |
| 3 | 0.000 | −0.027 | −0.006 | 0.005 | 1.194 |
| 4 | 0.044 | 0.300 | −0.181 | 0.197 | 1.111 |
| 5 | 0.001 | 0.045 | 0.005 | −0.002 | 1.189 |
| 6 | 0.012 | 0.151 | −0.061 | 0.071 | 0.772 |
| 7 | 0.044 | −0.298 | −0.264 | 0.254 | 3.505 |
| 8 | 0.008 | 0.128 | −0.079 | 0.086 | 0.882 |
| 9 | 0.010 | 0.144 | 0.033 | −0.024 | 1.885 |
| 10 | 0.001 | 0.037 | −0.005 | 0.008 | 0.700 |
| 11 | 0.068 | 0.366 | 0.353 | −0.345 | 6.891 |
| 12 | 0.010 | 0.141 | −0.046 | 0.054 | 0.833 |
| 13 | 0.011 | 0.147 | −0.066 | 0.075 | 0.769 |
| 14 | 0.089 | −0.437 | −0.334 | 0.315 | 2.290 |
| 15 | 0.020 | −0.203 | −0.027 | 0.014 | 1.422 |
| 16 | 0.006 | −0.109 | 0.032 | −0.039 | 0.977 |
| 17 | 0.046 | −0.315 | −0.109 | 0.089 | 1.819 |
| 18 | 0.025 | −0.226 | 0.067 | −0.081 | 1.421 |
| 19 | 0.028 | −0.242 | −0.083 | 0.068 | 1.553 |
| 20 | 0.137 | 0.524 | 0.505 | −0.495 | 7.163 |
| 21 | 0.014 | −0.170 | −0.023 | 0.012 | 1.289 |
| 22 | 0.022 | −0.215 | −0.029 | 0.014 | 1.467 |
| 23 | 0.012 | −0.155 | 0.046 | −0.056 | 1.154 |
| 24 | 0.000 | −0.028 | 0.013 | −0.015 | 0.845 |
| 25 | 0.000 | 0.010 | −0.002 | 0.002 | 0.599 |
| 26 | 0.004 | −0.087 | 0.026 | −0.031 | 0.888 |
| 27 | 0.005 | −0.095 | −0.013 | 0.006 | 0.979 |
| 28 | 0.000 | −0.023 | 0.004 | −0.005 | 0.650 |
| 29 | 0.016 | −0.182 | −0.065 | 0.054 | 1.320 |
| 30 | 0.233 | 0.689 | 0.664 | −0.651 | 7.364 |
| 31 | 0.012 | −0.154 | 0.026 | −0.037 | 1.182 |
| 32 | 0.002 | 0.067 | −0.041 | 0.044 | 1.000 |
| 33 | 0.003 | 0.078 | −0.029 | 0.034 | 0.585 |
| 34 | 0.416 | 0.939 | 0.905 | −0.886 | 7.641 |
| 35 | 0.019 | −0.196 | −0.067 | 0.055 | 1.376 |
| 36 | 0.043 | 0.295 | −0.202 | 0.216 | 1.108 |
| 37 | 0.002 | 0.059 | −0.033 | 0.036 | 0.806 |
| 38 | 0.003 | 0.078 | −0.029 | 0.034 | 0.585 |
| 39 | 0.004 | 0.086 | −0.048 | 0.052 | 0.725 |
| 40 | 0.015 | 0.176 | −0.057 | 0.068 | 0.978 |
| 41 | 0.005 | −0.098 | 0.017 | −0.023 | 0.960 |
| 42 | 0.000 | 0.031 | −0.012 | 0.014 | 0.516 |
| 43 | 0.008 | 0.129 | −0.064 | 0.071 | 0.734 |
| 44 | 0.006 | 0.113 | −0.043 | 0.050 | 0.682 |
| 45 | 0.024 | 0.219 | −0.129 | 0.141 | 0.947 |
| 46 | 0.000 | 0.007 | −0.003 | 0.003 | 0.522 |
| 47 | 0.009 | −0.132 | 0.039 | −0.047 | 1.066 |

[a] The cut-off value for DFFITS is $2(p/n)^{1/2} = 0.413$, and for DFBETAS it is $2/\sqrt{n} = 0.292$.

Finally, we look at the diagnostics for the `Hawkins-Bradu-Kass` data. Again we see that none of the classical diagnostics succeed in separating the ‘bad’ points from the good ones. The robust diagnostic $RD_i$ on the other hand (do not confuse the last column in the tables with the robust distance (19.2.2)!) which is based on robust residuals finds all the outlying data points. Its definition will not be given here, since the good and the bad leverage points can be detected by means of the diagnostic plot.

**Table 6. Outlier Diagnostics (Including the Resistant Diagnostic $\mathbf{RD}_i$) for the First 14 Cases of the Hawkins–Bradu–Kass Data**



| Index | $CD^2(i)$ | DFFITS | DFBETAS (0.231) | | | | $RD_i$ |
|---|---|---|---|---|---|---|---|
| $i$ | (1.000) | (0.462) | Constant | $\theta_1$ | $\theta_2$ | $\theta_3$ | (2.500) |
| 1 | 0.040 | 0.407 | −0.076 | 0.116 | −0.062 | 0.039 | 12.999 |
| 2 | 0.053 | 0.470 | 0.006 | −0.043 | 0.006 | 0.092 | 13.500 |
| 3 | 0.046 | 0.430 | −0.038 | 0.080 | −0.180 | 0.160 | 13.911 |
| 4 | 0.031 | 0.352 | 0.045 | −0.085 | −0.070 | 0.161 | 13.961 |
| 5 | 0.039 | 0.399 | −0.007 | −0.001 | −0.090 | 0.135 | 13.982 |
| 6 | 0.052 | 0.459 | −0.141 | 0.201 | −0.050 | −0.019 | 13.451 |
| 7 | 0.079 | 0.575 | −0.156 | 0.189 | 0.027 | −0.055 | 13.910 |
| 8 | 0.052 | 0.464 | −0.034 | 0.060 | −0.108 | 0.121 | 13.383 |
| 9 | 0.034 | 0.372 | 0.055 | −0.086 | −0.103 | 0.191 | 13.718 |
| 10 | 0.048 | 0.439 | 0.098 | −0.126 | −0.167 | 0.273 | 13.535 |
| 11 | 0.348 | −1.300 | −0.052 | 0.238 | 0.174 | −0.491 | 11.730 |
| 12 | 0.851 | −2.168 | −0.024 | −0.025 | 1.192 | −1.262 | 12.004 |
| 13 | 0.254 | −1.065 | 0.367 | −0.257 | −0.424 | 0.359 | 12.297 |
| 14 | 2.114 | −3.030 | 0.559 | 0.337 | −2.795 | 1.920 | 13.674 |

## 19.4 The LTS estimator

### 19.4.1 Parameter estimates

The **Least Trimmed Squares** estimator is a highly robust regression estimator. It is defined as

$$\boxed{\hat{\boldsymbol{\beta}}_{\text{LTS}} = \underset{\hat{\boldsymbol{\beta}}}{\text{argmin}} \sum_{i=1}^{h} (e^2(\hat{\boldsymbol{\beta}}))_{i:n}} \tag{19.4.1}$$

with $h$ an integer between $[n+p+1]/2$ and $n$, and $e^2_{i:n}$ the $i$th smallest squared residual. For any candidate $\hat{\boldsymbol{\beta}}$ we thus rank the squared residuals from smallest to largest $(e^2(\hat{\boldsymbol{\beta}}))_{1:n} \leqslant (e^2(\hat{\boldsymbol{\beta}}))_{2:n} \leqslant \ldots \leqslant (e^2(\hat{\boldsymbol{\beta}}))_{n:n}$ and compute the sum of the $h$ smallest squared residuals. The LTS fit then corresponds to that $\hat{\boldsymbol{\beta}}$ which yields the smallest sum. The LTS estimator does not try to make all the residuals as small as possible, but only the 'majority', where the 'majority' is defined as $h/n$.

The robustness of an estimator can be measured by its **breakdown value** which says how many of the $n$ observations need to be replaced before the estimate is carried away. Formally, the finite-sample breakdown value of any regression estimator $T(Z) = T(X, \boldsymbol{y})$ is given by

$$\boxed{\epsilon^*_n = \epsilon^*_n(T, Z) = \ \min \{\frac{m}{n}; \sup_{Z'} \|T(Z')\| = \infty\}}$$

where $Z' = (X', \boldsymbol{y}')$ ranges over all data sets obtained by replacing any $m$ observations of $Z = (X, \boldsymbol{y})$ by arbitrary points.

The breakdown value of the LTS estimator satisfies:

$$\boxed{\epsilon^*_n = (n - h + 1)/n}$$

and is maximal for $h = [(n+p+1)/2]$. Roughly speaking, the maximal breakdown value of 50% is obtained for $h \approx n/2$. If we choose $h = 0.75n$, the breakdown value is approximately 25% etc. If $h$ gets closer to $n$, the LTS estimator approaches the LS estimator. The larger we choose $h$ the better the finite-sample efficiency of the LTS estimator will be, but the lower its resistance towards outliers!

With the parameter estimates $\hat{\boldsymbol{\beta}}_{\text{LTS}}$ we can associate an estimator of the error scale $\sigma$:

$$s_{\text{LTS}} = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h}(e^2(\hat{\boldsymbol{\beta}}_{\text{LTS}}))_{i:n}}.$$

The constant $d_{h,n}$ is chosen to make the scale estimator consistent at the gaussian model, which gives

$$\begin{aligned} c_{h,n} &= 1/\Phi^{-1}(\frac{h+n}{2n}) \\ d_{h,n} &= 1/\sqrt{1-\frac{2n}{hc_{h,n}}\phi(1/c_{h,n})}. \end{aligned}$$

### 19.4.2 Computation

Contrary to the LS estimator, the objective function of the LTS estimator

$$\sum_{i=1}^{h}(e^2(\hat{\boldsymbol{\beta}}))_{i:n} \tag{19.4.2}$$

is not convex and has many local minima. Therefore, one has to rely on approximate algorithms to compute the LTS estimator. Several approaches exist which differ in speed and/or accuracy.

The **$p$-subset algorithm**, such as PROGRESS, starts by drawing a random subset of $p$ observations out of the whole data set $n$. Then, the hyperplane $\hat{\boldsymbol{\beta}}$ through these $p$ data points is computed and the objective function (19.4.2) is evaluated in $\hat{\boldsymbol{\beta}}$. By drawing many random $p$-subsets (500-3000, depending on $n$ and $p$) we obtain many candidate fits, from which the $\hat{\boldsymbol{\beta}}_{\text{LTS}}$ with the smallest objective function can be selected. The FAST-LTS algorithm also starts with random subsets but it then uses more advanced steps that decrease the objective function.

### 19.4.3 Reweighted LTS

Although the LTS estimator is asymptotically normal, its asymptotic and finite-sample efficieny is not very high. This implies that its variance at uncontaminated data is much larger than the variance of $\hat{\boldsymbol{\beta}}_{\text{LS}}$. To improve the efficiency of LTS, we can apply a reweighted procedure. Another advantage of the reweighted LTS is that it yields inferential information such as standard errors of the estimates, $t$ and $p$-values and so on, which can be used for model improvement.

First, the standardized residuals are computed

$$e_i(\hat{\boldsymbol{\beta}}_{\text{LTS}})/s_{\text{LTS}}$$

and a weight function is applied to them. An example of a weight function is

$$w_i = \begin{cases} 1 & \text{if } |e_i(\hat{\boldsymbol{\beta}}_{\text{LTS}})/s_{\text{LTS}}| < 2.5 \\ 0 & \text{otherwise.} \end{cases}$$

This is called *hard rejection* and produces a clear distinction between accepted and rejected points. Next, a weighted LS fit is computed, which is equivalent to apply OLS on the transformed observations $(\sqrt{w_i}\boldsymbol{x}_i, \sqrt{w_i}y_i)$ as discussed in Chapter 16, Section 16.4.3. If we denote the resulting parameter estimates as $\hat{\boldsymbol{\beta}}_{\text{RLTS}}$, we can again compute the corresponding residuals $e_i(\hat{\boldsymbol{\beta}}_{\text{RLTS}})$ and the scale estimate

$$s_{\text{RLTS}} = \sqrt{\frac{\sum_i w_i e_i^2(\hat{\boldsymbol{\beta}}_{\text{RLTS}})}{\sum_i w_i - p}}.$$

In the R package `robustbase` the default output is actually the reweighted LTS.

## 19.5 The MCD estimator

### 19.5.1 Parameter estimates

The **Minimum Covariance Determinant** estimator is a robust estimator for the center $\boldsymbol{\mu}$ and the shape $\Sigma$ of a multivariate data set. In the regression context, it will be applied to the set of predictor variables $X_1, \dots, X_{p-1}$ to detect good and bad leverage points. In this section we therefore assume that our data points are $(p-1)$-dimensional.

The MCD estimator is defined by:

- find the $h$ observations out of $n$ whose classical covariance matrix has the lowest determinant
- then, $\hat{\boldsymbol{\mu}}_0$ is the average of those $h$ observations, and $\hat{\Sigma}_0$ is the covariance matrix of those $h$ observations (multiplied by a consistency factor)

with $[n+p]/2 \leqslant h \leqslant n$. This definition is inspired by the following relation: let $\bar{\boldsymbol{x}}_h$ denote the mean and $S_h$ the covariance matrix of $h$ observations. Consider now the tolerance ellipsoid

$$\{\boldsymbol{x}; (\boldsymbol{x} - \bar{\boldsymbol{x}}_h)^t S_h^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_h) \leqslant c^2\}$$

for some constant $c$. Then it can be shown that the volume of this ellipsoid is proportional to the square root of the determinant of $S_h$.

The breakdown value of the MCD estimator is $\epsilon_n^* = (n-h+1)/n$. For any scatter matrix $\hat{\Sigma}$, breakdown means that the largest eigenvalue becomes arbitrary large or that the smallest eigenvalue becomes zero:

$$\epsilon_n^*(\hat{\Sigma}, X) = \min\{\frac{m}{n} \sup_{X'} \frac{\lambda_{\max(X')}}{\lambda_{\min(X')}} = \infty\}$$

with $X'$ obtained by replacing $m$ points out of $X$. This implies that either the tolerance ellipsoid explodes (i.e. becomes unbounded) or that it implodes (i.e. is flattened to a lower dimension and deflated to a zero volume). Remember that the determinant of a square matrix is equal to the product of its eigenvalues.

### 19.5.2 Computation

The computation of the MCD estimator is very difficult. Exhaustive search over all $h$-subsets is usually too time-consuming. Again a $p$-subset approach can be followed.

It starts by drawing random $p$ points out of $n$, and computing their mean $m_0$ and covariance matrix $C_0$. Since the observations are $(p-1)$-dimensional, this covariance matrix will be non-singular unless the $p$ observations lie on a hyperplane of $\mathbb{R}^{p-1}$. We then compute the $h$ observations with smallest robust distance (with respect to $m_0$ and $C_0$). Finally the mean $m_1$ and covariance matrix $C_1$ of these $h$ data points is computed and the determinant of $C_1$ is evaluated.

Improvements to this elemental approach are incorporated in the FAST-MCD algorithm, which can be used in R with the `covMcd` function in package `robustbase`.

### 19.5.3 Reweighted MCD-estimator

The **one-step reweighted** MCD-estimator is defined analogously to the reweighted LTS estimator. Based on $\hat{\boldsymbol{\mu}}_0$ and $\hat{\Sigma}_0$ we compute the robust distances as in (19.2.2):

$$\mathrm{RD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)^t \hat{\Sigma}_0^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)}$$

and assign a weight to each observation:

$$w_i = \begin{cases} 1 & \text{if } \mathrm{RD}(\boldsymbol{x}_i) \leqslant \sqrt{\chi^2_{p-1,0.025}} \\ 0 & \text{otherwise} \end{cases}$$

Finally the weighted mean and weighted covariance matrix are obtained:

$$\hat{\boldsymbol{\mu}}_1 = \left(\sum_{i=1}^n w_i \boldsymbol{x}_i\right) \Big/ \left(\sum_{i=1}^n w_i\right)$$

$$\hat{\Sigma}_1 = \left(\sum_{i=1}^n w_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)^t\right) \Big/ \left(\sum_{i=1}^n w_i - 1\right)$$

These reweighted estimates attain a higher finite-sample efficiency than the raw MCD estimates, but retain the same breakdown value and are the default output of the `covMcd` function in R package `robustbase`.

## 19.6 A robust R-squared

The classical coefficient of determination (12.3.2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

can also be written as

$$R^2_{\text{ML}} = 1 - \frac{\hat{\sigma}^2_{\text{ML}}(X, \boldsymbol{y})}{\hat{\sigma}^2_{\text{ML}}(\mathbf{1}, \boldsymbol{y})}$$

with $\hat{\sigma}^2_{\text{ML}}$ the maximum likelihood estimator for $\sigma^2$ (under normal errors), given by $\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum e_i^2$.

Remember that the mean of the $y_i$ corresponds with the univariate least squares estimator:

$$\bar{y} = \underset{\hat{\mu}}{\text{argmin}} \sum_{i=1}^{n}(y_i - \hat{\mu})^2.$$

Therefore a robust $R^2$ can be defined analogously as

$$\boxed{R^2_{LTS} = 1 - \frac{s^2_{LTS}(X, \boldsymbol{y})}{s^2_{LTS}(\mathbf{1}, \boldsymbol{y})}.}$$

The denominator is equal to the squared univariate LTS or MCD scale estimator. It is defined by the variance of the $h$-subset with smallest variance, and can be computed by an explicit algorithm of $\text{O}(n \log n)$. For this we first have to sort the (univariate) observations and compute the variance of $h$ successive points.

## 19.7 Model selection

All the variable selection methods discussed in Chapter 17 are based on LS estimates and hence they are very sensitive to outliers! Simple robust alternatives are e.g. based on the robust $R^2$ value.

One should however be very cautious when outliers are detected. They are found not to satisfy the linear model that is followed by the majority of the data points. It is very important to investigate the reason why they differ: it can be due to the fact that they indeed belong to another population and hence satisfy another relation. But they can also point us to a model-misspecification. The inclusion of a quadratic term or a transformation of a variable can e.g. accommodate the outliers. Knowledge about the problem at hand is thus indispensable for model building and outlier detection!!

# Chapter 20

# Logistic regression

## 20.1 Introduction

In the previous chapters, we have always assumed that the response variable $Y$ is quantitative or continuous. The logistic regression model deals with a binary response variable.

Applications of logistic regression are often encountered in medical sciences, to evaluate the extent to which an exposure, like smoking, is associated with a certain disease. The response variable is then the disease outcome, which is usually represented with 0 for not diseased, and 1 for diseased. In those medical applications, one can distinguish the exposure variables (smoking) from control variables, such as age, race, sex, ... which are not of primary interest.

In general, we assume that as before, we want to model the relation between $p-$ 1 explanatory variables $X_1, \ldots, X_{p-1}$ and $Y$. Note that, as in linear regression, the $X_j$ may be quantitative, qualitative, interaction effects, or transformed observed variables.

## 20.2 The logistic regression model

We can apply regression techniques on this type of data by considering regression as a conditional mean. For each $\boldsymbol{x}_i$, let $\pi_i$ denote the conditional mean of

$Y$:

$$\pi_i = E(Y|\, \boldsymbol{x}_i) \tag{20.2.1}$$

Because $Y$ is binary,

$$\begin{aligned} E(Y|\, \boldsymbol{x}_i) &= 1P(Y = 1|\, \boldsymbol{x}_i) + 0P(Y = 0|\, \boldsymbol{x}_i) \\ &= P(Y = 1|\, \boldsymbol{x}_i) \end{aligned}$$

hence $\pi_i$ represents the proportion of 1's among those persons with $\boldsymbol{x} = \boldsymbol{x}_i$. If the $X_j$ are discrete, we could compute these proportions at each of their observed values. If all or some of the $X_j$ are continuous, we could apply local regression estimates. This is illustrated in Figure 15.1, where the broken line represents a lowess estimate.

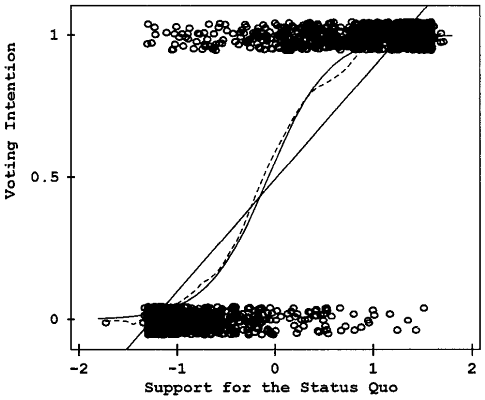


Figure 15.1. Scatterplot of voting intention (1 represents "yes," 0 represents "no") by a scale of support for the status quo, for a sample of Chilean voters surveyed prior to the 1988 plebiscite. The points are jittered vertically to minimize overplotting. The solid straight line shows the linear least-squares fit; the solid curved line shows the fit of the logistic-regression model (described in the next section); the broken line represents a lowess nonparametric regression.

However, nonparametric regression does not yield an expression for the response function, so we prefer to model explicitly the relation between $Y$ and the predictor variables.

We could try the *linear probability model* as in (12.1.1):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \tag{20.2.2}$$

with the $\epsilon_i \sim N(0, \sigma^2)$, which, using (20.2.1), means that

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}$$

Model (20.2.2) has several drawbacks:

1. the $y_i$ are 0/1, hence the errors can also only take 2 values (at each $\boldsymbol{x}_i$). If $y_i = 1$ (which happens with probability $\pi_i$), then $\epsilon_i = y_i - E(y_i) = 1 - \pi_i$, whereas if $y_i = 0$ then $\epsilon_i = 0 - \pi_i = -\pi_i$.

2. the variance of the errors can not be constant:

$$\begin{aligned}\mathrm{Var}(\epsilon_i) &= (1-\pi_i)^2\pi_i + (-\pi_i)^2(1-\pi_i)\\ &= \pi_i(1-\pi_i)\\ &= \boldsymbol{x}_i^t\boldsymbol{\beta}(1-\boldsymbol{x}_i^t\boldsymbol{\beta}).\end{aligned}$$

   The error variance thus depends on $\boldsymbol{x}_i$ and the unknown parameters $\boldsymbol{\beta}$, which makes it difficult to apply weighted least squares.

3. For a fixed $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^t$, the evaluation of the linear function $\boldsymbol{x}^t\boldsymbol{\beta}$ will not be restricted to values between 0 and 1 (see also Figure 15.1) although it represents a probability. This makes interpretation of the model meaningless.

If there is only one regressor, one could apply the *constrained linear probability model*, which constraints $\pi$ to the unit interval, while retaining the linear relation between $\pi$ and $\boldsymbol{x}$ within this interval:

$$\pi = \begin{cases} 0 & \text{for } 0 > \beta_0 + \beta_1 X\\ \beta_0 + \beta_1 X & \text{for } 0 \le \beta_0 + \beta_1 X \le 1\\ 1 & \text{for } \beta_0 + \beta_1 X > 1\end{cases}$$

This model, illustrated in Figure 15.2, also involves many difficulties. The abrupt changes of the slopes are unreasonable. And at which $x_i$ values should the slopes be changes? And how can this model be generalized if there are several regressors?

To make sure that the response function stays in [0,1], we can apply a transformation to $\boldsymbol{x}^t\boldsymbol{\beta}$, using a positive monotone function whose image is [0,1]. Any cumulative probability distribution function (cdf) can be selected for this purpose. We thus specify the model as

$$\boxed{\pi_i = F(\boldsymbol{x}_i^t\boldsymbol{\beta}) = F(\eta_i)} \qquad (20.2.3)$$

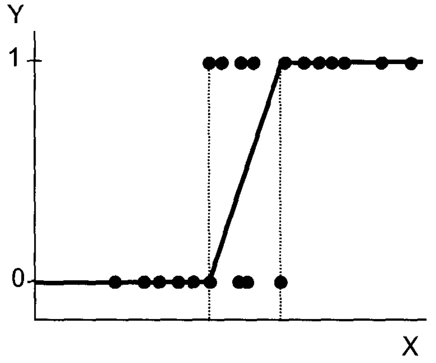


Figure 15.2. The constrained linear probability model. The estimate of the line $\pi = \alpha + \beta X$ is determined by the leftmost 1 and the rightmost 0, as shown by the vertical lines.

with

$$\eta_i = \boldsymbol{x}_i^t \boldsymbol{\beta}.$$

If we choose a smooth, symmetric and strictly increasing cdf, we can rewrite model (20.2.3) as

$$F^{-1}(\pi_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} = \eta_i \tag{20.2.4}$$

It is thus a nonlinear transformation of $\pi$ that yields a linear model in the regression parameters $\boldsymbol{\beta}$.

Typical choices for $F$ are the gaussian distribution function $\Phi(z)$ or the logistic distribution:

$$F_L(z) = \frac{1}{1 + e^{-z}}$$

leading respectively to the linear **probit model**:

$$\begin{aligned} \pi_i &= \Phi(\boldsymbol{x}_i^t \boldsymbol{\beta}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{x}_i^t \boldsymbol{\beta}} e^{-z^2/2} dz \end{aligned}$$

and the linear **logistic/logit** regression model:

$$\boxed{\pi_i = \frac{1}{1 + e^{-\boldsymbol{x}_i^t \boldsymbol{\beta}}}} \tag{20.2.5}$$

Both response functions are S-shaped, and are nearly linear in the middle of their range, as can be seen in Figure 15.3. This S-shape is very appealing to epidemiologist if the variable $z = \boldsymbol{x}^t \boldsymbol{\beta}$ is viewed as representing an index that combines contributions of several risk factors, and $F_L(z)$ represents the risk for a given value of $z$. The effect of $z$ on an individual's risk is minimal for low $z$'s until some threshold is reached. The risk then rises rapidly over a certain range of intermediate $z$ values, and then remains extremely high around 1 once $z$ is large enough.

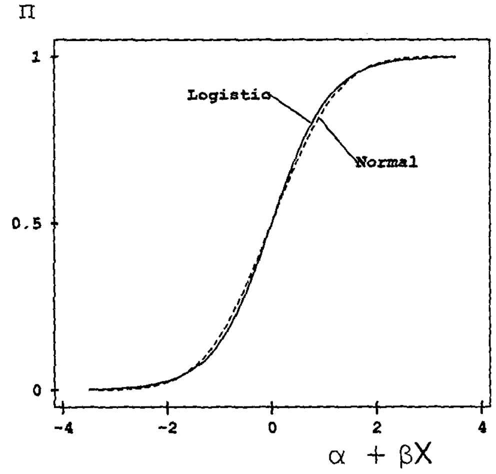


**Figure 15.3.** Once their variances are equated, the cumulative logistic and cumulative normal distributions—used here to transform $\alpha + \beta X$ to the unit interval—are virtually indistinguishable.

Although the logit and the probit model are very similar, the logit model is easier to work with because it does not involve an unevaluated integral, and because it allows a direct interpretation of the parameters.

## 20.3 Interpretation of the regression parameters

The **odds** of an event is defined as the ratio of the expected number of times that an event will occur to the expected number of times it will not occur. If $p$ is the probability of an event, then the odds of the event is defined by

$$\text{odds(event)} = \frac{p}{1-p} = \frac{P(\text{event})}{P(\text{no event})}.$$

Odds are always positive, but have no upper bound. Odds smaller than 1 correspond with $p < 0.5$. Taking the (natural) logarithm of an odd gives the log-odd or the **logit** of an event, which attains both negative and positive values. Below we show the connection between probability, odds and logit.

```
      prob   odds logit
 [1,] 0.01  0.010 -4.60
 [2,] 0.05  0.053 -2.94
 [3,] 0.10  0.111 -2.20
 [4,] 0.30  0.429 -0.85
 [5,] 0.50  1.000  0.00
 [6,] 0.70  2.333  0.85
 [7,] 0.90  9.000  2.20
 [8,] 0.95 19.000  2.94
 [9,] 0.99 99.000  4.60
```

From (20.2.5) we deduce:

$$\begin{aligned}
\log\Big(\frac{\pi_i}{1-\pi_i}\Big) &= \log\Big(\frac{1/(1+e^{-\eta_i})}{1-1/(1+e^{-\eta_i})}\Big) \\
&= \ldots = \log\Big(\frac{1}{e^{-\eta_i}}\Big) \\
&= \eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta}. \qquad (20.3.1)
\end{aligned}$$

The logistic model thus states that the log of the odds that $Y$ is 1 rather than 0 for given $\boldsymbol{x}_i$ is linear in $\boldsymbol{\beta}$ (and in $\boldsymbol{x}_i$). Consequently, if we hold all variables constant and increase $X_j$ by 1, the difference in the log odds is

$$\begin{aligned}
\text{logit}(x_{ij}+1) - \text{logit}(x_{ij}) &= \beta_0 + \ldots + \beta_j(x_{ij}+1) + \ldots \beta_{p-1}x_{i,p-1} \\
&\quad - \beta_0 - \ldots - \beta_j x_j - \ldots - \beta_{p-1}x_{i,p-1} \\
&= \beta_j. \qquad (20.3.2)
\end{aligned}$$

The **odds ratio** is defined as the ratio of the odds at $x_{ij}+1$ versus the odds at $x_{ij}$. From (20.3.2) we find that

$$\log(\text{odds}(x_{ij}+1)) - \log(\text{odds}(x_{ij})) = \log\Big(\frac{\text{odds}(x_{ij}+1)}{\text{odds}(x_{ij})}\Big) = \log(\text{OR}) = \beta_j$$

or

$$\text{OR} = e^{\beta_j}$$

and

$$\text{odds}(x_{ij}+1) = e^{\beta_j}\,\text{odds}(x_{ij}).$$

The estimated odds are thus multiplied by $e^{\beta_j}$ for any unit increase in $X_j$.

**Example: odds ratio.**

Consider a table of race of dependant by death sentence for 147 penalty-trial cases.

| | blacks | non-blacks | total |
|---|---|---|---|
| death | 28 | 22 | 50 |
| life | 45 | 52 | 97 |
| total | 73 | 74 | 147 |

For blacks, the odds of a death sentence is $28/45 = 0.62$. For non-blacks, the odds of a death sentence is $22/52 = 0.42$. The odds ratio is:

$$\frac{\text{odds of blacks}}{\text{odds of non-blacks}} = \frac{0.62}{0.42} = \frac{28*52}{22*45} = 1.47$$

The odds of a death sentence for blacks is thus 47% higher than for non-blacks.

**Example: disease outbreak.**

A health study investigates the epidemic outbreak of a disease that is spread by mosquitoes. Individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease (then, $y_i = 1$). The predictor variables are

1. age ($X_1$, continuous)
2. socioeconomic status of household (categorical with three levels, coded with two binary variables $X_2$, $X_3$). The upper class has $X_2 = 0$ and $X_3 = 0$, the middle class takes $X_2 = 1$ and $X_3 = 0$ whereas the lower class is coded with $X_2 = 0$ and $X_3 = 1$.

3. sector within the city ($X_4$, categorical with two levels). The second sector has $X_4 = 1$.

The data for 196 individuals were collected. The first 98 cases were selected as the training data set, so they are used to fit the model.
A first-order logistic regression model was considered:

$$\pi_i = [1 + \exp(-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}))]^{-1}.$$

In R we obtain the coefficient estimates

```
outbreakdisease[,"Sector"] = as.factor(outbreakdisease[,"Sector"])
outbreakdisease[,"Socio"] = as.factor(outbreakdisease[,"Socio"])
outbreakdisease[,"Disease"] = as.factor(outbreakdisease[,"Disease"])
trainingdata <- outbreakdisease[1:98,]
train.glm <- glm(Disease~Age+Socio+Sector, data=trainingdata, family=binomial)
coefficients(train.glm)

(Intercept)         Age      Socio2      Socio3     Sector2
-2.31293482  0.02975009  0.40879024 -0.30525456  1.57474923
```

The estimated logistic response function thus is:

$$\hat{\pi} = [1 + \exp(-(-2.313 + 0.03X_1 + 0.409X_2 - -0.305X_3 + 1.575X_4))]^{-1}$$

from which the odds ratios for the predictor variables follow:

- the odds ratio for age is $e^{0.03} = 1.03$: for a given socioeconomic status and sector location, the odds of a person having contracted the disease is increased by about 3% with each additional year of age
- the odds ratio for sector is $e^{1.575} = 4.83$: the odds of a person in sector 2 having contracted the disease is almost 5 times as high as for a person in sector 1, given age and socioeconomic status.

## 20.4 Computation

The logistic regression parameters are estimated via the maximum likelihood approach. For a Bernoulli variable $Y_i$, the density can be written as:

$$f(y_i) = P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

where $y_i = 0/1$. The joint probability of $n$ independent observations $y_1, \ldots, y_n$ is then given by

$$\begin{aligned} f(y_1, \ldots, y_n) &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \Big(\frac{\pi_i}{1 - \pi_i}\Big)^{y_i}(1 - \pi_i) \end{aligned}$$

Using (20.3.1) and (20.2.5) we obtain the likelihood function

$$L(\boldsymbol{\beta}) = f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \exp(\boldsymbol{x}_i^t\boldsymbol{\beta})^{y_i}[1 + \exp(\boldsymbol{x}_i^t\boldsymbol{\beta})]^{-1}$$

and the log-likelihood function

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i\boldsymbol{x}_i^t\boldsymbol{\beta} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x}_i^t\boldsymbol{\beta})]. \tag{20.4.1}$$

The partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned} \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} y_i\boldsymbol{x}_i - \sum_{i=1}^{n} \Big(\frac{\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}\Big)\boldsymbol{x}_i & (20.4.2) \\ &= \sum_{i=1}^{n} y_i\boldsymbol{x}_i - \sum_{i=1}^{n} \Big(\frac{1}{1 + \exp(-\boldsymbol{x}_i^t\boldsymbol{\beta})}\Big)\boldsymbol{x}_i \\ &= \sum_{i=1}^{n} y_i\boldsymbol{x}_i - \sum_{i=1}^{n} \pi_i\boldsymbol{x}_i & (20.4.3) \end{aligned}$$

The maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ are obtained by setting the partial derivatives to 0:

$$\sum_{i=1}^{n} \hat{\pi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}})\boldsymbol{x}_i = \sum_{i=1}^{n} y_i\boldsymbol{x}_i \tag{20.4.4}$$

where the fitted values $\hat{\pi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$ are given by

$$\hat{\pi}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \frac{1}{1 + \exp(-\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{\mathrm{ML}})}.$$

In matrix notation, we thus obtain estimating equations

$$X^t \boldsymbol{p} = X^t \boldsymbol{y}$$

where $\boldsymbol{p}$ is the vector of fitted values. Note that this equation is similar to the least-squares normal equations $(X^t X)\hat{\boldsymbol{\beta}} = X^t \boldsymbol{y}$ which also satisfy $X^t \hat{\mathbf{y}} = X^t \boldsymbol{y}$.

Because the estimating equations (20.4.4) are nonlinear in $\hat{\boldsymbol{\beta}}$, they have to be solved numerically and iteratively.

The **Newton-Raphson** approach uses a Taylor expansion of the log-likelihood function:

$$\log L(\boldsymbol{\beta}) \approx \log L(\hat{\boldsymbol{\beta}}^{(0)}) + \Big[\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big]^t_{\hat{\boldsymbol{\beta}}^{(0)}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)})^t H(\hat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)})$$

with $H$ the Hessian matrix

$$H(\boldsymbol{\beta}) = \Big[\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t}\Big].$$

Taking derivatives yields

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}^{(0)}} + H(\hat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)})$$

and setting them to zero:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}^{(0)}} + H(\hat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(0)}) = 0$$

from which we find the iteration step:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(1)} &= \hat{\boldsymbol{\beta}}^{(0)} - H(\hat{\boldsymbol{\beta}}^{(0)})^{-1} \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}^{(0)}} \\ &= \hat{\boldsymbol{\beta}}^{(0)} - H(\hat{\boldsymbol{\beta}}^{(0)})^{-1} X^t(\boldsymbol{y} - \hat{\pi}(\hat{\boldsymbol{\beta}}^{(0)})) \end{aligned} \tag{20.4.5}$$

where the last equality follows from (20.4.3). To compute the Hessian matrix, we differentiate (20.4.2):

$$\begin{aligned} H(\boldsymbol{\beta}) &= -\sum_{i=1}^n \frac{\exp(-\boldsymbol{x}_i^t \boldsymbol{\beta})}{(1 + \exp(-\boldsymbol{x}_i^t \boldsymbol{\beta}))^2} \boldsymbol{x}_i \boldsymbol{x}_i^t \\ &= -\sum_{i=1}^n \pi_i (1 - \pi_i) \boldsymbol{x}_i \boldsymbol{x}_i^t \\ &= -X^t V X \end{aligned} \tag{20.4.6}$$

with $V(\boldsymbol{\beta}) = \operatorname{diag}(\pi_i(1-\pi_i))$.
When convergence takes place, (20.4.5) says that

$$(X^tVX)^{-1}X^t(\boldsymbol{y} - \hat{\pi}(\hat{\boldsymbol{\beta}})) \approx 0$$

hence the estimating equations (20.4.4) are approximately satisfied.

Assume that the Newton-Raphson procedure has attained convergence. From (20.4.5) we have that

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \hat{\boldsymbol{\beta}}_{\text{ML}} + (X^tVX)^{-1}X^t(\boldsymbol{y} - \hat{\pi}(\hat{\boldsymbol{\beta}}_{\text{ML}}))$$

or

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = (X^tVX)^{-1}X^tV\boldsymbol{y}^*$$

with

$$\boldsymbol{y}^* = X\hat{\boldsymbol{\beta}}_{\text{ML}} + V^{-1}(\boldsymbol{y} - \hat{\pi}(\hat{\boldsymbol{\beta}}_{\text{ML}})).$$

These formulas suggest an algorithm based on **iteratively reweighted least squares:**

1. starting with $\hat{\boldsymbol{\beta}}^{(0)}$, compute fitted values $\hat{\pi}(\hat{\boldsymbol{\beta}}^{(0)})$, the matrix $V_0 = V(\hat{\boldsymbol{\beta}}^{(0)})$ and pseudo-response variables $\boldsymbol{y}_0^* = X\hat{\boldsymbol{\beta}}^{(0)} + V_0^{-1}(\boldsymbol{y} - \hat{\pi}(\hat{\boldsymbol{\beta}}^{(0)}))$
2. compute updated (weighted least squares) estimates:

$$\hat{\boldsymbol{\beta}}^{(1)} = (X^t V_0 X)^{-1} X^t V_0 \boldsymbol{y}_0^*$$

**Example: outbreak disease data.**

```
summary(train.glm)

Call:
glm(formula = Disease ~ Age + Socio + Sector, family = binomial,
    data = trainingdata)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6552  -0.7529  -0.4788   0.8558   2.0977


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
Age          0.02975    0.01350   2.203 0.027577 *
Socio2       0.40879    0.59900   0.682 0.494954
Socio3      -0.30525    0.60413  -0.505 0.613362
Sector2      1.57475    0.50162   3.139 0.001693 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05
```

```

Number of Fisher Scoring iterations: 4
```

## 20.5 Inference

### 20.5.1 Inference for a single parameter

The large-sample theory of maximum likelihood estimation learns that the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ can be estimated by

$$\begin{aligned}\hat{\Sigma}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) &= [-H(\hat{\boldsymbol{\beta}}_{\mathrm{ML}})]^{-1} \\ &= [X^t V(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) X]^{-1}\end{aligned} \tag{20.5.1}$$

using (20.4.6). This estimate is thus a by-product of the numerical procedure. Its diagonal elements give estimates of the standard errors of $\hat{\beta}_j$, whereas inferences are based on the approximation

$$\frac{\hat{\beta}_j - \beta_j}{s(\beta_j)} \approx N(0,1)$$

Hence, approximate confidence limits for $\beta_j$ are given by

$$\hat{\beta}_j \pm z_{\alpha/2} s(\beta_j)$$

and confidence limits for the odds ratio $e^{\beta_j}$ are

$$\exp[\hat{\beta}_j \pm z_{\alpha/2} s(\beta_j)]$$

The corresponding test statistic for $\beta_j = \beta_{j0}$ is called the Wald test.

**Example: outbreak disease data.**

```
train.coef <- summary(train.glm)$coef
print(train.coef,digits=2)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.31      0.643   -3.60  0.00032
Age             0.03      0.014    2.20  0.02758
Socio2          0.41      0.599    0.68  0.49495
Socio3         -0.31      0.604   -0.51  0.61336
Sector2         1.57      0.502    3.14  0.00169

train.ci <- cbind(train.coef[,1]-qnorm(0.975)*train.coef[,2],
                  train.coef[,1]+qnorm(0.975)*train.coef[,2])
print(cbind(coef(train.glm),train.ci),digits=2)
```

```
             [,1]    [,2]   [,3]
(Intercept) -2.31 -3.5724 -1.053
Age          0.03  0.0033  0.056
Socio2       0.41 -0.7652  1.583
Socio3      -0.31 -1.4893  0.879
Sector2      1.57  0.5916  2.558
```

```
train.or <- exp(coef(train.glm)[2:5])
train.cior <- exp(train.ci[2:5,])
print(cbind(train.or,train.cior),digits=3)
```

```
        train.or
Age        1.030 1.003  1.06
Socio2     1.505 0.465  4.87
Socio3     0.737 0.226  2.41
Sector2    4.830 1.807 12.91
```

For the hypothesis $H_0 : \beta_1 = 0$ the Wald test statistic becomes

$$z = \frac{\hat{\beta}_1}{s(\beta_1)} = \frac{0.03}{0.014} = 2.203$$

with corresponding p-value: $2P(Z \geq z) = 2\Phi(z) = 0.028$. Hence, the null hypothesis is rejected at the 5% significance level. Note that we could already see this from the 95% confidence interval for $\beta_1$ which does not include 0.

### 20.5.2 Test for several parameters

If we want to test whether

$$\begin{aligned} &H_0 : \beta_{p-q} = \beta_{p-q+1} = \ldots = \beta_{p-1} = 0 \\ &H_1 : \text{not all } \beta_j \text{ equal zero } (j = p-q, \ldots, p-1) \end{aligned}$$

we can follow two (equivalent) approaches.

The **likelihood ratio** test is based on the test statistic

$$\begin{aligned} G_0^2 &= 2 \log L(H_1) - 2 \log L(H_0) \\ &= -2 \log\Big(\frac{L(H_0)}{L(H_1)}\Big) \qquad \approx_{H_0} \chi_q^2 \end{aligned}$$

(if $n$ is sufficiently large). Here, $L(H_0)$ is the value of the maximum likelihood function evaluated at $\hat{\boldsymbol{\beta}}(H_0)$, which is smaller than the likelihood under the full model $L(H_1)$. Hence, the likelihood ratio $LR = L(H_0)/L(H_1) \leq 1$, and $-2\log LR$ is positive. If the $H_0$ hypothesis is correct, $G_0^2 \approx 0$. If $X_{p-q}, \ldots, X_{p-1}$ are significant, $L(H_0)/L(H_1) \ll 1$ and $G_0^2$ becomes large.

Alternatively, we can use the **model deviance**. The deviance of a fitted model compares the log-likelihood of the fitted model to the log-likelihood of a model with $n$ parameters that fits the $n$ observations perfectly. This is called a *saturated model.* From (20.4.1) we derive that the log-likelihood at $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is given by

$$\begin{aligned}\log L(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} y_i \log\Big(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\Big) - \sum_{i=1}^{n} \log[1+\exp(\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}})] \\ &= \sum_{i=1}^{n}[y_i \log\hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]\end{aligned}$$

If the fitted model can perfectly predict the $y$ values ($\hat{\pi}_i = 1$ when $y_i = 1$ and $\hat{\pi}_i = 0$ when $y_i = 0$) we see that $\log L(\hat{\boldsymbol{\beta}}_S) = 0$. So if the predictions are not perfect, $\log L(\hat{\boldsymbol{\beta}}) < 0$ (and $0 < L < 1$). The deviance is defined as

$$\mathrm{DEV}(X_0, \ldots, X_{p-1}) = 2\log L(\hat{\boldsymbol{\beta}}_S) - 2\log L(\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$$

which here reduces to

$$\boxed{\mathrm{DEV}(X_0, \ldots, X_{p-1}) = -2\textstyle\sum_{i=1}^{n}[y_i \log\hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]} \qquad (20.5.2)$$

Note that in the normal error linear regression model, the definition of the deviance is slightly modified (multiplied by $\sigma^2$), and then turns out to be the error sum of squares

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

We can compute the deviance for each fitted model. The difference between the deviances for two fitted models is called a *partial deviance* and is used to test the significance of a set of predictor variables in a model:

$$\begin{aligned}\mathrm{DEV}(X_{p-q}, \ldots, X_{p-1}|X_0, \ldots, X_{p-q-1}) = \mathrm{DEV}&(X_0, \ldots, X_{p-q-1}) \\ -\mathrm{DEV}&(X_0, \ldots, X_{p-1}).\end{aligned}$$

Note the analogy with the extra sums of squares in linear regression. Because $\log L(\hat{\boldsymbol{\beta}}_S) = 0$, the deviance is the same as $-2\log L(\hat{\boldsymbol{\beta}})$, and the partial deviance is exactly the same as $G_0^2$.

**Example: outbreak disease data.**

We can now also test whether $\beta_1$ for age is significant using the likelihood ratio test:

```
train.noage <- update(train.glm, .~.-Age)
anova(train.noage,train.glm)

Analysis of Deviance Table


Model 1: Disease ~ Socio + Sector
Model 2: Disease ~ Age + Socio + Sector
  Resid. Df Resid. Dev Df Deviance
1        94     106.20
2        93     101.05  1   5.1495
```

which yields the p-value: $P(U \geq 5.15) = 0.023$ because $U \sim \chi_1^2$. Hence, the null hypothesis is again rejected at the 5% significance level.

We can also test whether interactions terms are required in the model. The larger model contains 5 extra variables: $X_1{*}X_2, X_1{*}X_3, X_1{*}X_4, X_2{*}X_4, X_3{*}X_4$. Comparing the two models yields

```
train.interact <- update(train.glm, .~.^2)

Analysis of Deviance Table


Model 1: Disease ~ Age + Socio + Sector
Model 2: Disease ~ Age + Socio + Sector + Age:Socio + Age:Sector + Socio:Sector
  Resid. Df Resid. Dev Df Deviance
1        93    101.054
2        88     93.996  5   7.0583
```

with p-value: $P(U \geq 7.058) = 0.216$ because $U \sim \chi_5^2$. Hence, we conclude that the interactions are not significant.

### 20.5.3 Goodness of fit

The model deviance can be used as a goodness of fit criterion. The larger the model deviance, the poorer is the fit. It can also be used to check the model assumption that the logistic response function is the correct one.

$$H_0 : \pi = [1 + \exp(-\boldsymbol{x}^t\boldsymbol{\beta})]^{-1}$$
$$H_1 : \pi \neq [1 + \exp(-\boldsymbol{x}^t\boldsymbol{\beta})]^{-1}$$

Under $H_0$,

$$\boxed{\mathrm{DEV}(X_0, \ldots, X_{p-1}) \approx_{H_0} \chi^2_{n-p}}$$

hence if $\mathrm{DEV}(X_0, \ldots, X_{p-1}) \leq \chi^2_{n-p,\alpha}$ we accept $H_0$, otherwise we accept $H_1$.

**Example: outbreak disease data.**
The model deviance is

```
train.glm$deviance

[1] 101.0542
```

with $n - p = 98 - 5 = 93$ degrees of freedom. Since $\chi^2_{93,0.05} = 116.5$ we accept the logistic regression model at the 5% significance level.

### 20.5.4 Inference about mean response

At a certain fixed value of the predictor variables $\boldsymbol{x}_0$, the fitted value estimates the mean response $E(y_i|\boldsymbol{x}_0)$:

$$\hat{\pi}_0 = [1 + \exp(-\boldsymbol{x}_0^t\hat{\boldsymbol{\beta}})]^{-1}.$$

To obtain confidence limits for $\pi_0$, we first derive confidence limits for the logit mean response:

$$\eta_0 = \log\left(\frac{\pi_0}{1 - \pi_0}\right) = \boldsymbol{x}_0^t\boldsymbol{\beta}.$$

Since, $\hat{\eta}_0 = \boldsymbol{x}_0^t\hat{\boldsymbol{\beta}}$,

$$s^2(\hat{\eta}_0) = \boldsymbol{x}_0^t\hat{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0, \tag{20.5.3}$$

from which lower and upper confidence limits are derived:

$$L = \hat{\eta}_0 - z_{\alpha/2}s(\hat{\eta}_0)$$
$$U = \hat{\eta}_0 + z_{\alpha/2}s(\hat{\eta}_0).$$

Because the relation between $\pi_0$ and $\eta_0$ is monotone:

$$\pi_0 = [1 + \exp(-\eta_0)]^{-1}$$

we can convert $L$ and $U$ to confidence limits for $\pi_0$:

$$L^* = [1 + \exp(-L)]^{-1}$$
$$U^* = [1 + \exp(-U)]^{-1}.$$

**Example: outbreak disease data.**

We compute confidence intervals for the probability that persons of 10 years old, who are of lower socioeconomic status and live in sector 1, have contracted the disease.

```
x0 = data.frame(Age=10, Socio=factor("3",levels=levels(trainingdata$Socio)),
                Sector=factor("1",levels=levels(trainingdata$Sector)))
predx0 <- predict(train.glm,x0,se.fit=T)
predx0

$fit
        1
-2.320688

$se.fit
[1] 0.5426989

$residual.scale
[1] 1

loweta <- predx0$fit - qnorm(0.975)*predx0$se.fit
uppeta <- predx0$fit + qnorm(0.975)*predx0$se.fit
print(c(loweta,uppeta),digits=3)

    1     1
-3.38 -1.26
```

For the probability $\pi_0$ we find

```
predx0.prob <- predict(train.glm, x0, type="response")
predx0.prob

         1
0.08942399

lowpi <- 1/(1+exp(-loweta))
upppi <- 1/(1+exp(-uppeta))
print(c(lowpi,upppi),digits=3)

     1      1
0.0328 0.2215
```

The approximate 95% confidence interval for the mean response $\pi_0$ is thus

$$0.03 \leq \pi_0 \leq 0.22$$

Note that this interval is not symmetric around the point estimate $\hat{\pi}_0 = 0.089$ because $\pi_0$ is not a linear function of $\eta_0$.

## 20.6 Model diagnostics

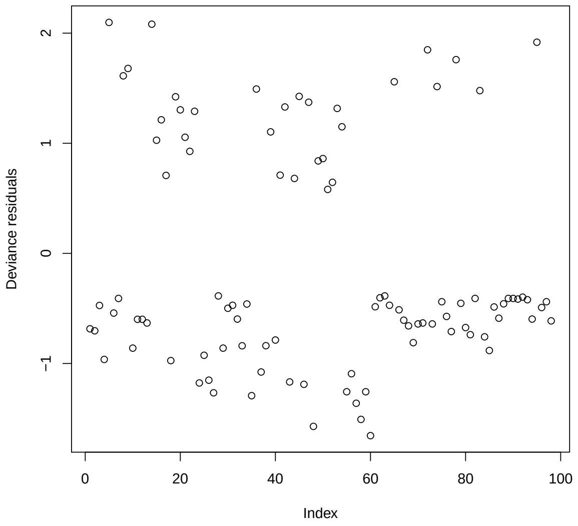
Residual analysis for logistic regression is more difficult than for linear regression models because the residuals are not normally distributed (their distribution is unknown). Plots of residuals against fitted values or predictor variables are usually uninformative.

**Deviance residuals** are defined as

$$\text{dev}_i = \text{sign}(y_i - \hat{\pi}_i)\Big(-2(y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i))\Big)^{1/2}.$$

An important property of the deviance residuals is that the sum of the squared deviance residuals equals the model deviance (20.5.2). The deviance residuals can be plotted to identify outlying residuals.

```r
plot(residuals(train.glm),ylab="Deviance residuals")
```



## 20.7 Prediction of a new observation

Logistic regression provides estimates of the probability of success, but does not forecast a binary outcome for a new observation $\boldsymbol{x}_0$. An outcome 1 can be predicted if the estimated response $\hat{\pi}_0$ is large, and the outcome 0 if $\hat{\pi}_0$ is small. We can just simply use 0.5 as cutoff value, which gives the prediction rule:

$$\text{if } \hat{\pi}_0 > 0.5, \text{ predict 1; otherwise predict 0.}$$

One can also try to find the best cutoff value for a validation set. This approach involves evaluating different cutoffs. For each trial, the rule is employed on the $m$ cases in the validation set and the proportion of incorrect predictions is computed. If no validation set is available, a cross-validation approach using the training set is to be preferred.

**Example: outbreak disease data.**

In the training set of 98 observations, 31 had contracted the disease, hence $31/98 = 0.316$ can be used as a starting point in the search for the best cutoff in the prediction rule.

Let us apply this to the training set.

```
library(descr)
summary(trainingdata$Disease)

 0  1
67 31

estimtrain <- ifelse(fitted(train.glm) > 0.316,1,0)
CrossTable(trainingdata[,"Disease"],estimtrain,prop.chisq=F)

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


==================================================
                             estimtrain
trainingdata[, "Disease"]        0       1   Total
--------------------------------------------------
0                               49      18      67
                             0.731   0.269   0.684
                             0.860   0.439
                             0.500   0.184
--------------------------------------------------
1                                8      23      31
                             0.258   0.742   0.316
                             0.140   0.561
                             0.082   0.235
--------------------------------------------------
Total                           57      41      98
                             0.582   0.418
==================================================
```

We thus find 8 persons with the disease who are predicted not to contract the disease, whereas 18 persons without the disease would be incorrectly predicted to have contracted the disease. The total prediction error is thus (18+8)/98 = 26.5%.

We now apply it to the validation set.

```
validationdata <- outbreakdisease[99:196,]
summary(validationdata[,"Disease"])

 0  1
```

```
72 26

val.pred <- predict(train.glm,validationdata,type="response")
estimvalid <- ifelse(val.pred >= 0.316,1,0)
CrossTable(validationdata[,"Disease"],estimvalid,prop.chisq=F)

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


====================================================
                               estimvalid
validationdata[, "Disease"]        0       1   Total
----------------------------------------------------
0                                 44      28      72
                               0.611   0.389   0.735
                               0.786   0.667
                               0.449   0.286
----------------------------------------------------
1                                 12      14      26
                               0.462   0.538   0.265
                               0.214   0.333
                               0.122   0.143
----------------------------------------------------
Total                             56      42      98
                               0.571   0.429
====================================================
```

We now obtain a misclassification error of $(12 + 28)/98 = 40.8\%$.

# Chapter 21

# Appendix

## 21.1 Maximum likelihood (ML) schatters voor $\boldsymbol{\mu}$ en $\boldsymbol{\Sigma}$

**Stelling.** Indien $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dan is de ML schatter $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ gelijk aan

$$\begin{cases} \hat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} \\ \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\tau} = \frac{n-1}{n}\boldsymbol{S} \end{cases}$$

Hiervoor hebben we enkele matrix-eigenschappen nodig:

**Lemma 11.** *Neem $\boldsymbol{A}$ $(k \times k)$ symmetrisch, $\boldsymbol{B}$ $(m \times k)$ en $\boldsymbol{C}$ $(k \times m)$. Zij $\boldsymbol{x}$ $(k \times 1)$ een vector.*

1. $tr(\boldsymbol{BC}) = tr(\boldsymbol{CB})$

2. $\boldsymbol{x}^{\tau}\boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}^{\tau}\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^{\tau})$

3. $tr(\boldsymbol{A}) = \sum_{j=1}^{k}\lambda_j$ *waarbij* $\lambda_j$ *de eigenwaarden van* $\boldsymbol{A}$ *zijn*

*Proof.* 1. $tr(\boldsymbol{BC}) = \sum_{j=1}^{m}(\boldsymbol{BC})_{jj} = \sum_{j=1}^{m}(\sum_{i=1}^{k} b_{ji}c_{ij})$

$tr(\boldsymbol{CB}) = \sum_{i=1}^{k}(\boldsymbol{CB})_{ii} = \sum_{i=1}^{k}(\sum_{j=1}^{m} c_{ij}b_{ji})$

2. $\underbrace{\boldsymbol{x}^{\tau}\boldsymbol{A}\boldsymbol{x}}_{\in\mathbb{R}} = tr(\underbrace{\boldsymbol{x}^{\tau}}_{\boldsymbol{B}}\underbrace{\boldsymbol{A}\boldsymbol{x}}_{\boldsymbol{C}}) = tr(\underbrace{\boldsymbol{A}\boldsymbol{x}}_{\boldsymbol{C}}\underbrace{\boldsymbol{x}^{\tau}}_{\boldsymbol{B}})$

3. spectraalontbinding: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\tau$ met $\boldsymbol{U}\boldsymbol{U}^\tau = \boldsymbol{I}_k$ en

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_k \end{pmatrix} \quad \text{met } \lambda_j \in \mathbb{R}$$

$$\Rightarrow tr(\boldsymbol{A}) = tr(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\tau) = tr(\boldsymbol{\Lambda}\underbrace{\boldsymbol{U}\boldsymbol{U}^\tau}_{\boldsymbol{I}_k}) = tr(\boldsymbol{\Lambda}) = \lambda_1 + \ldots + \lambda_k$$

□

We leiden nu eerst de ML schatter voor $\boldsymbol{\mu}$ af.

**Result.** Voor elke $\boldsymbol{\Sigma} \in$ PD(p) geldt:

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} \text{ is de ML schatter van } \boldsymbol{\mu}$$

en de formule van $\hat{\boldsymbol{\mu}}$ hangt niet af van $\boldsymbol{\Sigma}$.

*Proof.* Met behulp van het vorige lemma, kunnen we de exponent van $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ anders schrijven:

$$\begin{aligned} (\boldsymbol{x}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) &= tr\left[(\boldsymbol{x}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right] \\ &= tr\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\tau\right] \end{aligned}$$

Dus,

$$\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = tr\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\tau\right)\right]$$

Daar bovenop,

$$\begin{aligned}
\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\tau &= \sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\boldsymbol{x}_i-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau \\
&= \sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^\tau+\sum_{i=1}^{n}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau \\
&+ \underbrace{\left(\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})\right)(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau}_{=0}+\underbrace{\sum_{i=1}^{n}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^\tau}_{=0} \\
&= \underbrace{\boldsymbol{W}}_{\text{SSCP matrix}}+n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau
\end{aligned}$$

Bijgevolg,

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \frac{1}{(2\pi)^{\frac{np}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}}e^{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{W}+n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau)\right]}$$

Voor elke matrix $\boldsymbol{\Sigma} \in$ PD(p) geldt dus

$L(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ is maximaal
$\Leftrightarrow tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}+\boldsymbol{\Sigma}^{-1}n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau)\right]$ is minimaal
$\Leftrightarrow tr\left[\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau\right]$ is minimaal
$\Leftrightarrow (\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\tau\underbrace{\boldsymbol{\Sigma}^{-1}}_{PD}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \geqslant 0$ is minimaal
$\Leftrightarrow (\overline{\boldsymbol{x}}-\boldsymbol{\mu})=\boldsymbol{0}$

□

Vervolgens zoeken we de ML schatter van $\boldsymbol{\Sigma}$:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \operatorname*{argmax}_{\boldsymbol{\Sigma}} L(\overline{\boldsymbol{x}},\boldsymbol{\Sigma};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \\
&= \operatorname*{argmax}_{\boldsymbol{\Sigma}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}}e^{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right]}
\end{aligned}$$

Deze wordt gegeven door volgend resultaat:

**Result.** Gegeven de matrix $\boldsymbol{W} \in$ PD(p),

$$\frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right]} \leqslant \frac{1}{|\frac{1}{n}\boldsymbol{W}|^{\frac{n}{2}}} e^{-\frac{np}{2}}$$

voor elke $\boldsymbol{\Sigma} \in$ PD(p).

Dit wordt een gelijkheid $\Leftrightarrow \boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{W}$.

*Proof.* Laat $\boldsymbol{W}^{\frac{1}{2}}$ de unieke symmetrische vierkantswortel zijn van $\boldsymbol{W}$.

$$tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) = tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}}) = tr(\underbrace{\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{\frac{1}{2}}}_{\in PD(p)})$$

Stel $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$ de eigenwaarden van $\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{\frac{1}{2}}$

$$\Rightarrow \begin{cases} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) & = \sum_{j=1}^{p} \lambda_j \\ \underbrace{|\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{\frac{1}{2}}|}_{=|\boldsymbol{W}^{\frac{1}{2}}||\boldsymbol{\Sigma}^{-1}||\boldsymbol{W}^{\frac{1}{2}}|=\frac{|\boldsymbol{W}|}{|\boldsymbol{\Sigma}|}} & = \prod_{j=1}^{p} \lambda_j \quad \Rightarrow \frac{1}{|\boldsymbol{\Sigma}|} = \frac{1}{|\boldsymbol{W}|}\prod_{j=1}^{p} \lambda_j \end{cases}$$

Dus,

$$\begin{aligned} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right]} &= \frac{1}{|\boldsymbol{W}|^{\frac{n}{2}}}\left(\prod_{j=1}^{p} \lambda_j^{\frac{n}{2}}\right) e^{-\frac{1}{2}\sum_{j=1}^{p}\lambda_j} \\ &= \frac{1}{|\boldsymbol{W}|^{\frac{n}{2}}} \prod_{j=1}^{p} \left(\lambda_j^{\frac{n}{2}} e^{-\frac{\lambda_j}{2}}\right) \\ &\overset{\bigstar}{\leqslant} \frac{1}{|\boldsymbol{W}|^{\frac{n}{2}}} n^{\frac{np}{2}} e^{-\frac{np}{2}} = \frac{1}{|\frac{1}{n}\boldsymbol{W}|^{\frac{n}{2}}} e^{-\frac{np}{2}} \end{aligned}$$

$\bigstar$ geldt omdat (calculus) het maximum van de afbeelding

$$h : \mathbb{R}^+ \to \mathbb{R}^+ : \lambda \to \lambda^{\frac{n}{2}} e^{-\frac{\lambda}{2}}$$

gelijk is aan $n^{\frac{n}{2}} e^{-\frac{n}{2}}$ (het unieke maximum wordt enkel bereikt in $\lambda = n$).

De ongelijkheid $\bigstar$ wordt een gelijkheid

$\Leftrightarrow$ alle $\lambda_j = n$

$\Leftrightarrow \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{\frac{1}{2}} = n\boldsymbol{I}_p$

$\Leftrightarrow \boldsymbol{\Sigma}^{-1} = n\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{I}_p\boldsymbol{W}^{-\frac{1}{2}} = n\boldsymbol{W}^{-1}$

$\Leftrightarrow \boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{W}$

□

**Opmerking:**

$$\begin{aligned}
\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\
&= \boxed{\frac{1}{(2\pi)^{\frac{np}{2}}} \frac{1}{|\hat{\boldsymbol{\Sigma}}|^{\frac{n}{2}}} e^{-\frac{np}{2}}} \\
&= \frac{c}{|\boldsymbol{S}|^{\frac{n}{2}}} \qquad\qquad (\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{S})
\end{aligned}$$

# Bibliography

Draper, N., Smith, H. (1998), *Applied Regression Analysis,* John Wiley, New York.

Fox, J. (1997), *Applied Regression Analysis, Linear Models, and Related Methods,* Sage Publications, Thousand Oaks.

Gunst, R.F., Mason, R.L. (1980), *Regression Analysis and its Application. A Data-oriented Approach.* Marcel Dekker, New York.

Johnson, R.A., Wichern, D.W. (2002), Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Upper Saddle River.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996), *Applied Linear Statistical Models, 4th edition,* McGraw Hill, New York.

Ramsay, F., Schafer, D. (2013), *The Statistical Sleuth,* 3rd Edition, Brooks/Cole Cengage Learning.

Kaufman L., Rousseeuw P.J. (1990), *Finding Groups in Data, an Introduction to Cluster Analysis,* John Wiley, New York.

Rousseeuw, P.J., Leroy, A. (1987), *Robust Regression and Outlier Detection,* John Wiley, New York.

Ryan, T. (1997), *Modern Regression Methods,* John Wiley, New York.

Sen, A., Srivastava (1990), *Regression Analysis: Theory, Methods and Applications,* Springer, New York.

KU Leuven
Department of Mathematics
Celestijnenlaan 200B, 3001 Leuven
Tel. +32 16 37 23 83
Stefan.VanAelst@kuleuven.be