MSE | MASTER OF SCIENCE IN ENGINEERING

Applied Statistics and Data Analysis

Statistical Process and Quality Control

Multiple Regression

Design of Experiment

Marcel Steiner-Curtis

September 5, 2020

Prof. Dr. Marcel Steiner-Curtis
FHNW Fachhochschule Nordwestschweiz
Hochschule für Technik
Institut für Mathematik und Naturwissenschaften
Bahnhofstrasse 6
CH-5210 Windisch

marcel.steiner@fhnw.ch

Dear MSE-Students

You enrolled in the module **Applied Statistics and Data Analysis** and therefore are willing to learn something about data analysis, big data and machine learning. All of you, I hope at least, already followed in their Bachelor studies a course about *Stochastics*, others would call it *Probability Theory and Statistics*. The contents of such an introduction are of course part of the prerequisites for this module **Applied Statistics and Data Analysis**. If you feel unsure in this topic or need a bit of extra training I propose that you work through chapters 2, 7, 8, 9, 10 and 11 of my lecture notes *Wahrscheinlichkeitsrechnung und Statistik*, [37], (in German) or any other book about the fundamentals of statistics.

The module is divided into three parts:

1. **Statistical Process and Quality Control (SPC)**
2. **Multiple Regression**
3. **Design of Experiment (DoE)**

We decided to use in the exercises and examples of this course the free software environment R for statistical computing and graphics. Please download from the internet and install on your personal device the latest versions of R `http://cran.ch.r-project.org/` and RStudio `https://www.rstudio.com/` which makes R easier to use. RStudio is a free software too and includes a code editor, debugging and visualisation tools. If you have never used R, then we propose that you work through the first few chapters of the excellent R-introduction `R-intro.pdf`, which you can download from `http://cran.r-project.org/doc/manuals/R-intro.pdf`.

The exercises in these lecture notes are an integral part of the module. To properly understand the ideas and how to apply the theory to data you must solve all exercises yourselves. During classes you will be able to work on the exercises and ask individual questions. We will split the class in four exercise groups. At home you should then finish the exercises until the next exercise class. In these notes and in the distributed zip-file you will find worked solutions with R to all the exercises.

In addition, we have decided to recommend you the book of D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, [23]. Montgomery and Runger have a very pleasant writing style. The book has many examples and exercises and is for a masters course on applied statistics and data analysis very suitable.

I wrote a script for the parts **Statistical Process and Quality Control**, **Multiple Regression** and **Design of Experiment**. While creating these lecture notes, I have partly

kept to the following books [9], [22], [23], [38] and [24]. The second part on multiple regression and the third part on design of experiment was taught in German from 2008 to 2018 by Andreas Ruckstuhl, ZHAW, Winterthur. He wrote excellent German lecture notes that formed the basis for my parts on regression and design of experiment. The script does not replace the recommended book.

You are no longer the first students to work with this script. Do not judge the author too hard (and the previous readers). If you find mistakes and inconsistencies, please let me know.

September 5, 2020, Marcel Steiner-Curtis

Contents

Preface i

Contents iii

I Statistical Process and Quality Control 1

1 Introduction 3
- 1.1 What is Statistical Quality and Process Control? . . . 3
- 1.2 The Magnificent Seven . . . 4
  - 1.2.1 Check Sheet . . . 5
  - 1.2.2 Pareto Chart . . . 5
  - 1.2.3 Defect Concentration Diagram, Location Plot . . . 7
  - 1.2.4 Cause-and-Effect Diagram . . . 8
- 1.3 Exercises . . . 9

2 Control Charts 11
- 2.1 Control Charts versus Hypothesis Testing . . . 11
- 2.2 Shewhart Control Chart . . . 13
  - 2.2.1 Control Charts for $\bar{x}$ and R . . . 14
  - 2.2.2 Control Charts for $\bar{x}$ and s . . . 18
- 2.3 Individuals Control Charts . . . 20
- 2.4 Control Charts for Attributes Data – p Chart . . . 23
- 2.5 Exercises . . . 24

3 Statistical Properties of Control Charts 27
- 3.1 Interpretation of Control Charts . . . 27
- 3.2 Type I Error and Type II Error . . . 28
- 3.3 Power Function and Operating Characteristic . . . 30
- 3.4 Average Run Length . . . 34
- 3.5 Process Capability . . . 36
- 3.6 Exercises . . . 38

4 Control Charts with Memory 41
- 4.1 CUSUM . . . 42
- 4.2 EWMA . . . 45
- 4.3 Exercises . . . 48

5 Acceptance Sampling **51**
5.1 Acceptance Sampling Plans for Attributes . . . 52
5.1.1 Operating Characteristic . . . 53
5.1.2 Parameters of an Acceptance Sampling Plan . . . 56
5.2 Acceptance Sampling Plans for Variables . . . 59
5.3 Exercises . . . 59

6 R-Packages used in Statistical Process Control **61**
6.1 Package `qcc` . . . 61
6.2 Exercises . . . 61
6.3 Packages `AcceptanceSampling` . . . 62
6.4 Exercises . . . 62

II Multiple Regression **63**

7 Simple Linear Regression **65**
7.1 Simple Linear Regression Model . . . 65
7.2 Estimation of the Parameters . . . 68
7.3 Tests and Confidence Intervals . . . 73
7.3.1 Test of the Slope . . . 73
7.3.2 Test of the Intercept . . . 76
7.3.3 Confidence Interval on the Slope . . . 76
7.4 Confidence Interval on Response and Prediction Intervals . . . 77
7.4.1 Confidence Interval of the Response . . . 77
7.4.2 Prediction Interval . . . 79
7.5 Exercises . . . 82

8 Residual Analysis **85**
8.1 Introduction . . . 85
8.2 Diagnostic Tools . . . 91
8.2.1 Tukey-Anscombe Plot . . . 91
8.2.2 Scale-Location Plot . . . 94
8.2.3 Normal q-q Plot (and Histogram) . . . 95
8.3 Treatment of Model Violations . . . 97
8.3.1 Non-Constant Variance of Random Errors . . . 97
8.3.2 Outliers . . . 100
8.3.3 Heavy-Tailed Distributions . . . 101
8.4 Independence . . . 102
8.5 Exercises . . . 104

9 Multiple Linear Regression **107**
9.1 Model and Estimation . . . 108
9.1.1 Least-Squares Estimation of the Regression Coefficient . . . 108
9.1.2 Tests on the Significance of any Individual Regression Coefficient . . . 111
9.1.3 Confidence Intervals on the Regression Coefficients . . . 112
9.1.4 Confidence Interval of the Response . . . 114

9.1.5 Prediction Interval . . . 115
9.1.6 Coefficient of Determination - Multiple R-Squared . . . 115
9.2 Diversity of Modelling Possibilities . . . 116
9.2.1 Polynomial Regression . . . 116
9.2.2 Nonlinear Functions and Linear Regression . . . 117
9.2.3 Binary Explanatory Variables . . . 117
9.2.4 Factor Variables . . . 119
9.3 Comparison of Regression Models with F-Test . . . 122
9.3.1 Comparison of Two Straight Lines . . . 124
9.4 Exercises . . . 127

10 Sensitivity and Robustness **131**
10.1 Residual Analysis . . . 131
10.2 Influential Observations . . . 134
10.3 Weighted linear regression . . . 138
10.4 Robust Methods . . . 142
10.5 Exercises . . . 148

11 Variable Selection and Modelling **151**
11.1 Model Selection Methods . . . 153
11.2 Collinearity . . . 157
11.3 Modelling Strategies . . . 160
11.4 Exercises . . . 162

III Design of Experiment **165**

12 Design of Experiments **167**
12.1 Terminology and Concepts . . . 168
12.2 Experimental Design . . . 170
12.3 Planning and Conducting a Study . . . 171
12.4 Exercises . . . 173

13 Factorial Designs and Analysis of Variance **175**
13.1 One Factor Analysis of Variance . . . 175
13.2 Two Factor Analysis of Variance . . . 180
13.3 Two Factor Analysis of Variance with Interaction . . . 186
13.4 Residual Analysis . . . 188
13.5 Exercises . . . 189

14 Screening Experiments **193**
14.1 2^k Factorial Designs . . . 193
14.2 2^k Fractional Factorial Designs . . . 202
14.3 Exercises . . . 208

15 Response Surface Methodology **211**
15.1 Response Surface 211
15.2 Second-Order Response Surface with Interactions 217
15.3 R-Package `rsm` 224
15.4 Exercises 225

Appendix **226**

Tables **229**

Bibliography **245**

Index **248**

Part I

Statistical Process and Quality Control

Chapter 1

Introduction

The quality of products and services has been a much discussed topic recently. Who has not yet already undergone a Q-process, either as a student or as a lecturer? Q-management and Q-systems are on everyone's lips. In this first part of this course on statistical process and quality control it is not about fads that carry the prefix Q, but about solid statistical process and quality control. We are primarily interested in the statistical analysis of data that is generated and collected in a process. We want to examine the quality of the process on the basis of this data and control it if necessary.

1.1 What is Statistical Quality and Process Control?

In the DIN EN ISO 9000, the valid german standard for quality management, **quality**[1] is defined as follows:

> *Der Grad, in dem ein Satz inhärenter*[2] *Merkmale Anforderungen erfüllt.*

In colloquial language, the term quality is often used judgmentally, often as a counterpart to quantity. We all know the saying: *Quantity is not equal to quality.* D. C. Montgomery defines in his standard work on statistical quality control (cf. [22], p. 4) quality in two ways:

> ***Quality** means fitness for use.*

and

> ***Quality** is inversely proportional to variability.*

Statistical quality control was first introduced by Walter A. Shewhart (see Fig[3]. 1.1.i) He applied it at the *Bell Telephone Laboratories* in the twenties of the 20th century. Especially since World War II, statistical quality control has become an indispensable tool in industry. It was first and foremost the Japanese who rigorously applied statistical quality control to production, and thus built up a large advantage in keeping up with high quality standards compared to Europeans and Americans. Nowadays statistical quality control and associated with it the design of experiment (cf. third part of this course) is consistently used worldwide.

[1] lat.: *qualitas*
[2] lat.: *inhaerere*, engl.: adhere to something
[3] Source: `https://en.wikipedia.org/wiki/Walter_A._Shewhart`

Figure 1.1.i: Walter A. Shewhart, 1891-1967.

Statistical process control (SPC) is commonly understood as a way to optimise production and manufacturing processes. A nice definition by D. J. Wheeler and D. S. Chambers (cf. [41]) is:

> ***Statistical process control*** *is, first and foremost, a way of thinking which happens to have some tools attached.*

1.2 The Magnificent Seven

In this course, we are very interested in these *tools* and their applications. Often, in connection with these *tools*, we speak of **the magnificent seven**. These are the following seven statistical (graphical) methods for analysing data:

1. histogram
2. check sheet
3. Pareto chart
4. defect concentration diagram
5. cause-and-effect diagram
6. control chart
7. scatter diagram

Although these *tools* constitute an important aspect of statistical process control, these include only the technical, i.e. statistical aspect. Equally important is the attitude of the people involved in the process to continuously improve the quality and the reduction of variability.

In the following we want to introduce some of the magnificent seven. You are already familiar with the histogram and the scatter diagram from your Bachelor's degree. The control chart plays the most important role of the magnificent seven and we will deal with it in Chaps. 2, 3 and 4.

There are some examples in the R-file `Example 1.2 The Magnificent Seven.R`, which you can find in the distributed zip-file.

1.2.1 Check Sheet

The **check sheet** is a simple method of quality control. As a rule, it consists of a ready-made form to register and count possible problems in a production process.

Example 1.2.1. It has happened to us all that during a call with a mobile phone the connection was interrupted. This is very annoying and the telephone company cannot be interested in such an interruption. Therefore, an attempt is made to find the causes of the error. For this purpose, the telephone company can run a check sheet in which the possible causes are plotted against the number of complaints. So that a possible seasonality may be determined, the check sheet is broken down per quarter.

causes	interruptions per quarter				
	1.	2.	3.	4.	total
drop phone	1	0	0	2	3
battery empty	3	3	2	4	12
driving in tunnel	12	10	13	9	44
press button	1	1	0	3	5
accident	1	0	0	0	1
phone broken	1	0	2	1	4
no obvious reason	3	4	3	5	15

We find out that by far the most interruptions are due to driving through a tunnel followed by an empty battery. If the provider wants to improve the transmission security, then it should invest primarily in the tunnel infrastructure.

In Chap. 1.2.3 we will present the information of a check sheet graphically.

1.2.2 Pareto Chart

An important principle in process control is the **Pareto principle**, others say the **80-20 principle**. The two numbers 80 and 20 give the answer to the following question:

What percentage of the result is achieved with what percentage of the commitment?

The goal in quality improvement is to achieve a big impact with only a small part of the funds used. The Pareto principle is named after its inventor Vilfredo F. D. Pareto (see Fig.[4] 1.2.i).

The **Pareto chart** is a special histogram, which is used to apply the causes of a problem in function from the greatest to the least serious. It is a statistical tool for visualizing the 80-20 principle.

Example 1.2.2. We load the data of Ex. 1.2.1 with R and plot a Pareto chart.

```
#   load data as list
> data <- list("drop"     = c(1,0,0,2),
+              "battery"  = c(3,3,2,4),
+              "tunnel"   = c(12,10,13,9),
+              "button"   = c(1,1,0,3),
+              "accident" = c(1,0,0,0),
```

[4] Source: https://en.wikipedia.org/wiki/Vilfredo_Pareto

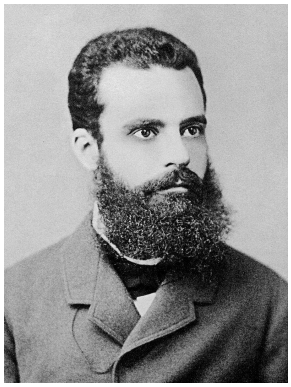


Figure 1.2.i: Vilfredo F. D. Pareto, 1891-1923.

```
+              "broken"   = c(1,0,2,1),
+              "none"     = c(3,4,3,5) )
#   save as data frame
> data <- as.data.frame(data)
> str(data)
'data.frame':   4 obs. of  7 variables:
 $ drop    : num  1 0 0 2
 $ battery : num  3 3 2 4
 $ tunnel  : num  12 10 13 9
 $ button  : num  1 1 0 3
 $ accident: num  1 0 0 0
 $ broken  : num  1 0 2 1
 $ none    : num  3 4 3 5

#   count the number of interruptions
> interrupt <- apply(data, 2, sum);   interrupt
    drop  battery   tunnel   button accident   broken     none
       3       12       44        5        1        4       15
#   sort by size
> interrupt.sort <- sort(interrupt, decreasing=TRUE);   interrupt.sort
  tunnel     none  battery   button   broken     drop accident
      44       15       12        5        4        3        1

#   Pareto chart
> barplot(interrupt.sort, main="Pareto Chart", ylim=c(0,sum(interrupt)+10))
#   cumulative sums
> lines(cumsum(interrupt.sort))
> points(cumsum(interrupt.sort), pch=20, cex=2)
```

Like this we obtain Fig. 1.2.ii, in which we can easily read off the most important reason for interruption.

If the provider wants to improve the transmission security, then it should invest primarily in the tunnel infrastructure.

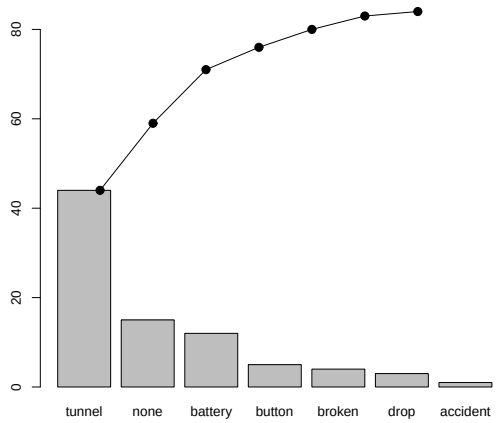


Figure 1.2.ii: Pareto chart.

1.2.3 Defect Concentration Diagram, Location Plot

The graphical counterpart to the check sheet is the **defect concentration diagram** or the **location plot**. The diagram is used to graphically visualise the locations of the various defects on a physical object.

Example 1.2.3. The defect concentration diagram[5], cf. Fig. 1.2.iii, shows the locations of the various defects in the quality control of shirts.

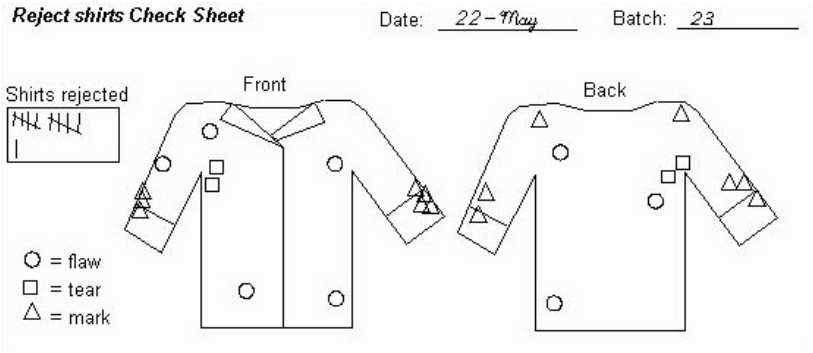


Figure 1.2.iii: Defect concentration diagram or location plot: The locations of the various defects in the quality control of shirts are marked.

We can see from the diagram that, for example, most tears form on the sleeves. With this information, we can take the necessary measures in the right place in the production chain.

[5]Source: `http://syque.com/improvement/Location%20Plot.htm`

1.2.4 Cause-and-Effect Diagram

The **cause-and-effect diagram** is a tool in the form of a fishbone, cf. Fig. 1.2.iv, for the systematic identification of causes which make problems. Here are the possible causes that trigger a specific effect, decomposed into major and minor causes, so-called **fishbone approach**. This is followed by a graphical presentation of the causes to allow a clear overall view. In this way, all causes of the problem should be identified and their dependencies displayed using the diagram.

Example 1.2.4. The increase in productivity of a company[6] represents an effect to be achieved. To reach this goal, all relevant influences must be right. For this purpose, six impact-relevant categories were selected and for each at least one cause found and added to the impact arrow.

categories	causes			
equipment	software	office		
environment	noise	heat		
people	education	motivation	fatigue	disruptions
machines	operation			
materials	quality	procurement		
methods	order	standardisation		

To increase productivity i.e. within the category *people* the causes *education*, *motivation*, *fatigue* and *disorder* have been chosen. In a further step, this list can be examined more closely, e. g. their nature of impact on the effect.

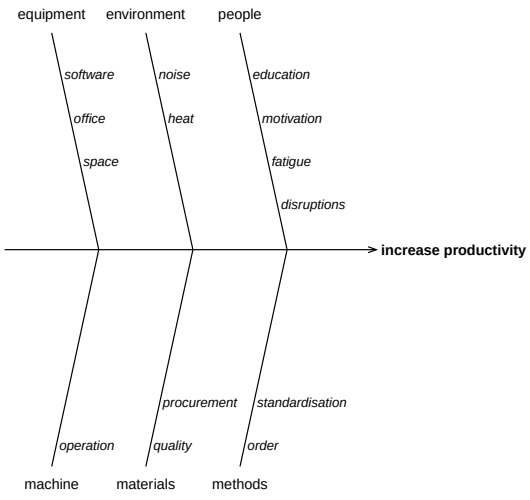


Figure 1.2.iv: Cause-and-effect diagram.

Based on this, actions can be derived for problem solving or for the goal to increase productivity. Thus to increase productivity, disruptions could be prevented, motivation increased or staff training improved.
The presented R-code uses an additional R-package `qcc`, which we introduce in Chap. 6.

[6]Source: `https://de.wikipedia.org/wiki/Ursache-Wirkungs-Diagramm`.

```
#   load package
> require(qcc)

#   cause-and-effect diagram
> cause.and.effect(
+    cause=list( equipment   = c("software","office","space"),
+                environment = c("noise","heat"),
+                people      = c("education","motivation","fatigue","disruptions"),
+                machine     = c("operation"),
+                materials   = c("quality","procurement"),
+                methods     = c("order","standardisation")),
+    effect="increase productivity")
```

1.3 Exercises

Problem 1.3.1 (**Download and Install R**). Download R and RStudio from `http://cran.ch.r-project.org/` and `https://www.rstudio.com/` and install it on your own device.

Problem 1.3.2 (**R for Beginners**). Study the excellent R-introduction `R-intro.pdf`, which you can download from `http://cran.r-project.org/doc/manuals/R-intro.pdf`.

Problem 1.3.3 (**First Steps with R**). The data set `sample.dat` contains 3 random samples with 20 entries. These are each 3 temperature measurements on 20 days. Load the data set `sample.dat` in R.

```
> setwd("C:/.../AppStat/Datasets")
> file <- "sample.dat"
> data <- read.table(file, header=TRUE)
> data
```

a. Examine the structure of the data with using the command `str(data)`.

b. With `summary(data)` you will receive a summary of the most important statistics.

c. Graph the samples with `plot`. With `help(plot)` you will obtain help.

d. Calculate the mean, median, variance and standard deviation for each day. This is elegantly done with `apply`. Add the mean value in the above graphic.

e. Create a histogram and a box plot of the mean values. Use `hist` and `boxplot`.

f. Are the variables x_1 and x_2 significantly different? Use `t.test`. Are the requirements for the t-test met? Use `qqnorm`.

Problem 1.3.4 (**Normal Distribution**). A machine produces metal plates with average thickness $\mu = 8.00\,\text{mm}$ and standard deviation $\sigma = 0.05\,\text{mm}$. The slightly fluctuating plate thickness is assumed to be normally distributed.

a. What percentage of scrap is to be expected if the thickness does not exceed 8.10 mm?

b. What percentage of scrap is to be expected if the thickness is between 7.92 mm and 8.08 mm?

c. What deviation from 8.00 mm is still allowed with at most 5% scrap?

Problem 1.3.5 (**Simulation of Random Runs**)**.** Using the command `rnorm` generate $m = 500$ random samples of size $n = 5$ from a standard normal distribution. Save the samples as an $(m \times n)$-matrix. Use `matrix`.

a. Visualise the samples as runs and be astonished about the big variation in the courses.

b. Now calculate the average $\bar{x}$ for each sample and generate a histogram of the mean values. From theory, we know that the means are normally distributed with parameters $\mu = 0$ and standard deviation $\sigma = \frac{1}{\sqrt{n}}$.

(i) Check this with a quantile-quantile plot. Use `qqnorm` and `qqline`.

(ii) Check this graphically using a histogram. Use `hist(..., freq=FALSE, ...)` and `curve(dnorm(x,mean=0,sd=1/sqrt(n)), from=-4, to=4, add=TRUE)` adds the theoretical distribution density to the histogram.

c. Repeat the task for $n = 2, 10, 100$.

Solutions

Solution 1.3.4.

a. We use the command `pnorm` and calculate

$$\begin{aligned} P(8.10\,\text{mm} \leq X) &= \texttt{pnorm(q=8.10,mean=8.00,sd=0.05,lower.tail=FALSE)} \\ &= 0.02275013. \end{aligned}$$

With `help(pnorm)` you get more information about the implementation of the normal distribution in R.

b. We calculate

$$\begin{aligned} &P\left(X \leq 7.92\,\text{mm and } 8.08\,\text{mm} \leq X\right) \\ &= \texttt{1-pnorm(q=8.08,mean=8.00,sd=0.05) + pnorm(q=7.92,mean=8.00,sd=0.05)} \\ &= 0.1095986. \end{aligned}$$

c. We use `qnorm`, then $|8.00 - \texttt{qnorm(p=0.025,mean=8.00,sd=0.05)}| = 0.0979982$.

Chapter 2

Control Charts

2.1 Control Charts versus Hypothesis Testing

The basis of a **control chart** is a statistical hypothesis test. We illustrate this with an example and compute it with R.

Example 2.1.1 (**Statistical Hypothesis Test**, cf. [38], p. 327)**.** In a turning workshop, rings of a certain type are produced. To check if the machine maintains the quality, the thickness of the individual rings is measured as a quality characteristic. Let X denote the random variable *thickness*. The target value is $\mu_0 = 10,000\,\text{mm}$.

```
#   target value
> mu0 <- 10.000
```

The deviation from the **target value** μ_0 is allowed to be between $\delta_1 = -0.012\,\text{mm}$ and $\delta_2 = 0.024\,\text{mm}$. In other words, the thickness should lie between the two **tolerance limits**

$$\text{USL} = \mu_0 + \delta_2 = 10.024\,\text{mm} \quad \text{and} \quad \text{LSL} = \mu_0 + \delta_1 = 9.988\,\text{mm}.$$

In this case, the quality requirement is met.
We denote USL the **upper specification limit** and LSL the **lower specification limit**.
The machine should therefore be set to the nominal value at the beginning of the process. Experience has shown that more or less random deviations from the target value occur during production. To check if the machine does not produce bad parts, i.e. if the thickness is within the tolerance limits, we sample from the actual production at fixed, pre-determined times, the last n produced rings and measure their thickness. This results in n realisations $x_1, \ldots, x_n$ of the random variable X.

```
#   measured values
> x <- c(10.003, 10.007, 10.009, 10.002,  9.992,
+         9.997,  9.995, 10.009,  9.995, 10.004)
```

We calculate the arithmetic mean of the sample

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

```
#   mean value
>  x.bar <- mean(x);  x.bar
[1] 10.0013
```

Now, at any point in time, we are conducting a two-sided statistical test to check the two alternative hypotheses:

$$H_0 : \mu_0 = \bar{x}, \text{ i.e. process is not disturbed.}$$
$$H_1 : \mu_0 \neq \bar{x}, \text{ i.e. process is disturbed.}$$

Let us briefly review the prerequisites of the test:

- The measured values $x_1, \ldots, x_n$ are independent and identically distributed.
- The thickness X is normally distributed with the expected value μ_0 and the known variance σ^2.
- The samples taken at different times are independent of each other.

So we know from experience the standard deviation of the process, i.e. $\sigma = 0.005\,\text{mm}$.

```
#   standard deviation
> sigma <- 0.005
```

We denote by z_q the q-quantile for the normal distribution (cf. T.2), i.e.

$$P\left(Z \leq z_q\right) = q.$$

If we assume $\alpha = 0.0027$ is the given error probability[1], then we obtain $q = 1 - \frac{\alpha}{2}$ and $z_q = 3$.

```
#   q-quantile
> alpha <- 0.0027
> z.q <- qnorm(1-alpha/2);   z.q
[1] 2.999977
```

This rejects the null hypothesis H_0 if the statistic $\bar{x}$ is outside the two **control limits**

$$\text{UCL} = \mu_0 + z_q \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \text{LCL} = \mu_0 - z_q \frac{\sigma}{\sqrt{n}}.$$

```
#   control limits
n <- length(x)
UCL <- mu0 + z.q * sigma/sqrt(n);   UCL
[1] 10.00474
LCL <- mu0 - z.q * sigma/sqrt(n);   LCL
[1] 9.995257
```

Since $9.995257 \leq \bar{x} \leq 10.00474$, the null hypothesis is accepted, and so we conclude that the process is currently under control.

[1] It is being attempted, to get three-σ limits. This explains the somewhat strange choice of significance level.

This procedure can be simplified accordingly and prepared for the periodic sampling of random samples of fixed size n from the current production of the machine. For this purpose we draw the center line μ_0 and the **lower specification limit** LCL as well as the **upper specification limit** UCL for a given value of α in a scatter plot. In this scatter plot, we then continuously record the mean value against the index of the current measurement.

The range betweeen the control limits UCL and LCL is called **control area**.

In general, process standard deviation is unfortunately unknown, and these must be determined, for example, from a preprocess that must be under statistical control. For this, the procedures presented in the next chapter, control charts, have become shown useful.

2.2 Shewhart Control Chart

In this chapter, we introduce the most commonly used Shewhart control chart for monitoring the mean and the variation of a process. It is important to remember that these charts are usually used in pairs: one chart to monitor the mean, and the other chart to monitor the variation. First and foremost, we are interested in monitoring the variation, and only in the second instance, if the variation is under control, we will consider the chart for the mean.
Shewhart originally used the **mean chart** and the **range chart**. These are still called $\bar{x}$ chart and R chart today and are still the most widely used control charts. They also serve as a prototype for understanding how all other control charts are used.
Given k samples, so-called **groups** with n values each. The total number of readings is then $N = k\,n$.
We can enter the N measured values, denoted by x_{ij} (where $i \in \{1,\ldots,k\}$, and $j \in \{1,\ldots,n\}$) into the following scheme:

random sample	sample values						mean values	standard deviations	ranges
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	$\bar{x}_1$	s_1	R_1
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2n}	$\bar{x}_2$	s_2	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{in}	$\bar{x}_i$	s_i	R_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_{k1}	x_{k2}	$\cdots$	x_{kj}	$\cdots$	x_{kn}	$\bar{x}_k$	s_k	R_k

We assume that the data comes from a normal distribution. We supplemented this scheme with the mean values and the standard deviations

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad \text{and} \quad s_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}$$

and the ranges

$$R_i = \max\{x_{ij} | j \in \{1,\ldots,n\}\} - \min\{x_{ij} | j \in \{1,\ldots,n\}\}$$

of the individual groups.

2.2.1 Control Charts for $\bar{x}$ and R

To construct an R chart, we select data from a trial run, that is, we select k samples of sample size n from the running process. In general, k is between 20 and 25 and n for example 5. The **centreline** (CL) of the R chart is denoted by $\bar{R}$ and is calculated from the arithmetic mean the ranges $R_1, \ldots, R_k$ of the k random samples, i.e.

$$\bar{R} = \frac{1}{k} \sum_{i=1}^{k} R_i.$$

Let σ_R be the standard deviation of the statistic $\bar{R}$. If we assume that the measurements of the process are approximately normally distributed then the control limits of the R chart are given by

$$\text{UCL} = D_4 \bar{R} \quad \text{and} \quad \text{LCL} = D_3 \bar{R}.$$

The constants D_3 and D_4 depend only on the sample size n and can be found in Tab. T.15.
After the centreline and control limits are determined, we add the ranges R_i against the index i in a scatter plot. In addition we draw horizontal lines for $\bar{R}$ and UCL and LCL in the graphic.
After that, the actual process control begins by looking at whether all ranges lie between the two bounds UCL and LCL. If some samples are out of bounds, it is recommended to omit these measurements and recalculate the limits.

Remark 2.2.1. We know that for an independent sample $x_1, \ldots, x_n$ from a normal distribution with parameters μ and σ the estimator (arithmetic mean)

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

satisfies

$$\text{E}(\bar{x}) = \mu \quad \text{and} \quad \text{Var}(\bar{x}) = \frac{\sigma^2}{n}.$$

That is, the arithmetic mean is an unbiased estimator with the standard deviation

$$\frac{\sigma}{\sqrt{n}}$$

(cf. Prob. 1.3.5).

Theoretically, the control limits for the $\bar{x}$ chart depend on the three-σ limits of the empirical distribution of the statistic $\bar{x}$. This defines the three-σ limits of the control chart by

$$\text{UCL} = \mu + 3\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \text{LCL} = \mu - 3\frac{\sigma}{\sqrt{n}}.$$

The abbreviation UCL stands for **upper control limit** and LCL stands for **lower control limit**. Where μ is the process mean and σ is the process variance. Of course these limits like this are unusable in practice, since μ and σ are usually unknown and must first be estimated from process data.
In order to obtain a reasonable estimate for the process variance σ, a two-step procedure is usually used. First, the process variance control chart (R chart) is brought under statistical

control to keep the process variance stable and then this process variance can serve as an estimate for σ. If we are sure that the R chart is under statistical control, then $\bar{R}$ can be used as a reasonable estimate for the mean range. The mean $\bar{R}$ is an estimate for the **process standard deviation**

$$\hat{\sigma} = \frac{\bar{R}}{d_2},$$

where the constant d_2 depends only on the sample size n and can be found in Tab. T.15. Any samples excluded for the construction of the R chart should also be disregarded for the construction of the $\bar{x}$ chart. So we have $k^\star$ valid samples from a trial run, where $k^\star \leq k$ refers to the reduced number of samples. Their mean values $\bar{x}_1, \ldots, \bar{x}_{k^\star}$ are estimators for μ, i.e.

$$\bar{\bar{x}} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} \bar{x}_i.$$

The control limits are

$$\text{UCL} = \bar{\bar{x}} + 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_2\bar{R} \quad \text{and} \quad \text{LCL} = \bar{\bar{x}} - 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_2\bar{R},$$

where we simplified

$$A_2 \approx \frac{3}{d_2\sqrt{n}}.$$

The constant A_2 depends on the sample size n and can again be found in Tab. T.15.

From these formulas, it is immediately apparent that the centreline of the R chart has a direct impact on the control limits of the $\bar{x}$ chart.

Example 2.2.1 ($\bar{x}$ **and** R **Chart**, cf. [9], p. 258, Ex. 6.1)**.** The process of making ignition keys for automobiles includes various quality critical operations on the blank. Some of the features of the product, such as the depth of the engraving, are extremely important to the correct functioning of the finished key. The permissible tolerance limits are

$$\text{USL} = 0.223\,\text{mm} \quad \text{and} \quad \text{LSL} = 0.177\,\text{mm}.$$

Due to the large production number of ignition keys, not every single one can be measured and controlled, so we are forced to take random samples: Every minute during $k = 20$ minutes, we draw a random sample of size $n = 4$ from the current production. Tab. 2.2.1 shows the depth of engraving measurements [in cm] for each of the $20 \cdot 4 = 80$ ignition keys.

i	x_1	x_2	x_3	x_4	$\bar{x}_i$	s_i	R_i
1	0.196	0.206	0.199	0.215	0.204	0.008	0.019
2	0.201	0.205	0.198	0.200	0.201	0.003	0.007
3	0.195	0.200	0.204	0.204	0.201	0.004	0.009
4	0.210	0.188	0.203	0.200	0.200	0.009	0.022
5	0.202	0.204	0.196	0.195	0.199	0.004	0.009
6	0.195	0.200	0.196	0.201	0.198	0.003	0.006
7	0.203	0.199	0.202	0.189	0.198	0.006	0.014
8	0.205	0.191	0.205	0.209	0.202	0.008	0.018
9	0.204	0.197	0.199	0.201	0.200	0.003	0.007
10	0.198	0.203	0.205	0.213	0.205	0.006	0.015
11	0.209	0.208	0.202	0.203	0.206	0.004	0.007
12	0.202	0.199	0.196	0.196	0.198	0.003	0.006
13	0.196	0.202	0.202	0.204	0.201	0.003	0.008
14	0.186	0.200	0.193	0.194	0.193	0.006	0.014
15	0.207	0.192	0.209	0.192	0.200	0.009	0.017
16	0.200	0.197	0.212	0.202	0.203	0.006	0.015
17	0.200	0.198	0.198	0.197	0.198	0.001	0.003
18	0.206	0.200	0.194	0.200	0.200	0.005	0.012
19	0.205	0.207	0.203	0.200	0.204	0.003	0.007
20	0.204	0.205	0.199	0.196	0.201	0.004	0.009

First, we prepare the data for further processing with R.

```
#   read data
> file <- "ignition-keys.dat"
> data <- read.table(file, header=TRUE)
> data
      x1    x2    x3    x4
1  0.196 0.206 0.199 0.215
2  0.201 0.205 0.198 0.200
...
20 0.204 0.205 0.199 0.196
```

Then we calculate mean, standard deviation and ranges per sample.

```
> data$mean <- apply(data[,1:4], 1, mean)
> data$sd <- apply(data[,1:4], 1, sd)
> data$R <- apply(data[,1:4], 1, function(x){ max(x) - min(x) })
> str(data)
'data.frame':   20 obs. of  7 variables:
 $ x1  : num  0.196 0.201 0.195 0.21 0.202 0.195 0.203 0.205 0.204 0.198 ...
 $ x2  : num  0.206 0.205 0.2 0.188 0.204 0.2 0.199 0.191 0.197 0.203 ...
 $ x3  : num  0.199 0.198 0.204 0.203 0.196 0.196 0.202 0.205 0.199 0.205 ...
 $ x4  : num  0.215 0.2 0.204 0.2 0.195 0.201 0.189 0.209 0.201 0.213 ...
 $ mean: num  0.204 0.201 0.201 0.200 0.199 ...
 $ sd  : num  0.00845 0.00294 0.00427 0.00918 0.00443 ...
 $ R   : num  0.019 0.007 0.009 0.022 0.009 ...
```

We graphically examine with a **q-q plot**, that is a **quantile-quantile plot** if the data is normally distributed. To do this, we need to store all the data in a vector.

```
#   save data as a vector
> data.vec <- c(data$x1, data$x2, data$x3, data$x4)
#   q-q plot
```

```
> qqnorm(data.vec, pch=20)
> qqline(data.vec)
```

We find that the points essentially scatter around the drawn line, cf. Fig. 2.2.i. That is, we can assume that the data really comes from a normal distribution. Another way to check the normal distribution is to create a **histogram**.

```
#   histogram with estimated normal distribution
> hist(data.vec, freq=FALSE, breaks=11)
> curve(dnorm(x,mean=mean(data.vec),sd=sd(data.vec)),
+       from=0.9*min(data), to=1.1*max(data), add=TRUE)
```

Once again, there is nothing against the assumption that the data comes from a normal distribution, cf. Fig. 2.2.ii.

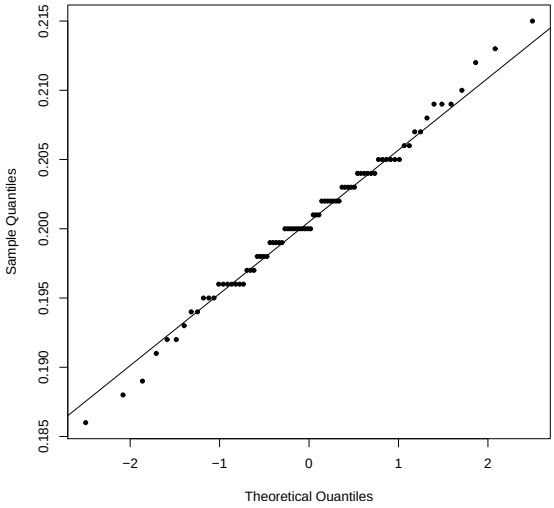


Figure 2.2.i: q-q plot to check the assumption about the normal distribution.

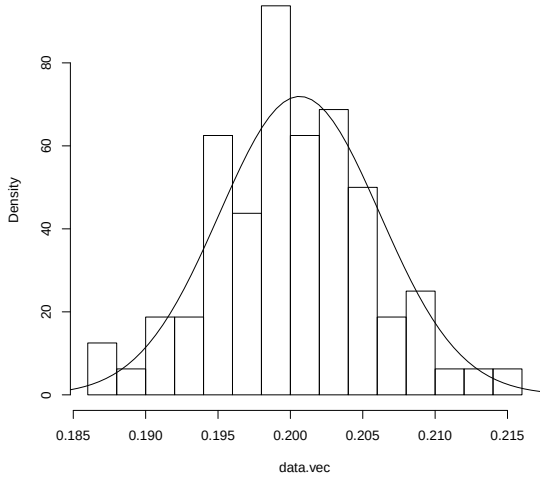


Figure 2.2.ii: Histogram with density (estimated) of the normal distribution.

Now we calculate the centreline and control limits of the R chart and add the ranges into a scatter plot. In addition, we plot the centreline and the control limits.

```
> R.bar <- mean(data$R);  R.bar
[1] 0.0112
> D3 <- 0         #   from table for n=4
> D4 <- 2.282     #   from table for n=4
> R.CL <- c(D3, 1, D4) * R.bar;  R.CL
[1] 0.0000000 0.0112000 0.0255584
> plot(data$R, pch=20, ylim=c(0,0.03), ylab="Ranges", main="R Chart")
> lines(data$R)
> abline(h=R.CL, lty=c(2,1,2))
> text(rep(1,3), R.CL, label=c("LCL", "CL", "UCL"), pos=3)
```

We find that the process is under control. In a next step, we can now construct the $\bar{x}$ chart from the R chart. For this we need an estimate for the process standard deviation,

```
> d2 <- 2.059     #   from table for n=4
> sigma.R.hat <- R.bar / d2;  sigma.R.hat
[1] 0.005439534
```

and an estimate for the process mean, i.e. for the centreline.

```
> x.barbar <- mean(as.matrix(data[,1:4]));   x.barbar
[1] 0.2006375
```

This gives us the control limits of the $\bar{x}$ chart,

```
> n <- 4
> x.bar.CL <- x.barbar + c(-3,0,3) * sigma.R.hat / sqrt(n);  x.bar.CL
[1] 0.1924782 0.2006375 0.2087968
```

and with that we can draw the $\bar{x}$ chart.

```
> plot(data$mean, pch=20, ylim=c(0.19,0.21), ylab="Mean Values",
+      main="x.bar Chart (based on R Chart)")
> lines(data$mean)
> abline(h=x.bar.CL, lty=c(2,1,2))
> text(rep(1,3), x.bar.CL, label=c("LCL", "CL", "UCL"), pos=3)
```

We note that the process is totally under control, as all mean values lie between the control limits, cf. Fig. 2.2.iii and 2.2.iv.

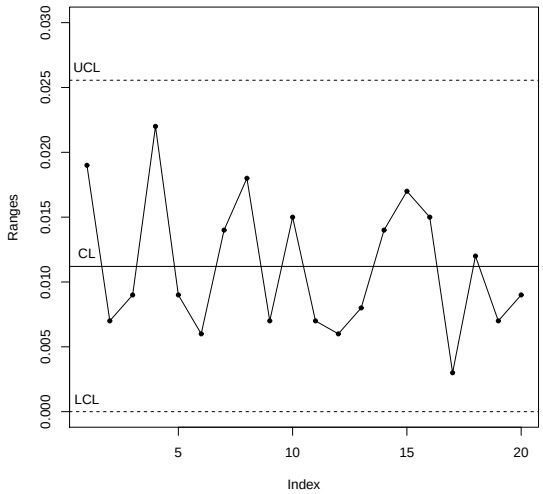


Figure 2.2.iii: R chart.

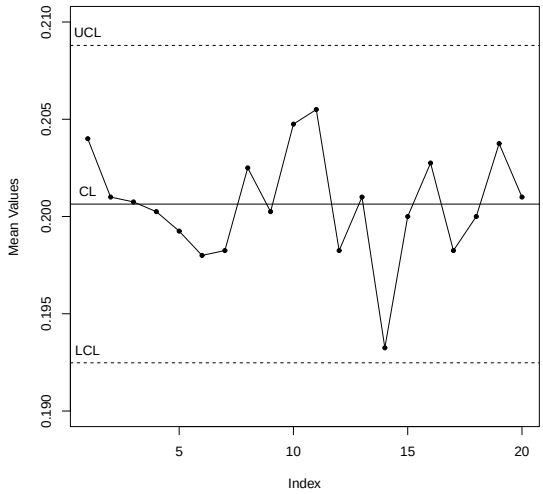


Figure 2.2.iv: $\bar{x}$ chart (based on R chart).

On the other hand, we note some suspicious runs at the beginning of the series. Whether this is an indication of a process that is not under control, Chap. 3 will tell us.

2.2.2 Control Charts for $\bar{x}$ and s

Since there are several ways to measure the variation of a process, a number of different control charts have been developed over time. One combination that is often used is the $\bar{x}$ and s charts. The construction of the $\bar{x}$ and s charts is similar to that of the $\bar{x}$ and R chart: first, the variation of the process is brought under control and then the estimation of the process variation is used to construct the $\bar{x}$ chart.
To construct an s chart, we again use data from a trial run, that is, we select k samples of sample size n from the running process.

The centreline of the s chart is denoted by $\bar{s}$ and is calculated from the arithmetic mean of the standard deviations $s_1, \ldots, s_k$ of the k random samples, i.e.

$$\bar{s} = \frac{1}{k} \sum_{i=1}^{k} s_i.$$

If we assume that the measurements of the process are approximately normally distributed, then the control limits of the s charts are given by

$$\text{UCL} = B_4 \bar{s} \quad \text{and} \quad \text{LCL} = B_3 \bar{s}.$$

The constants B_3 and B_4 only depend on the sample size n and can again be found in Tab. T.15. After the centreline and control limits are determined, we plot the standard deviations s_i against the index i in a scatter plot. Additionally, we plot horizontal lines for $\bar{s}$ and UCL and LCL into the graph.
From the s chart of a process that is under control, we can determine the process standard deviation by

$$\hat{\sigma} = \frac{\bar{s}}{c_4},$$

where the constant[2] c_4 depends only on the sample size n and can be found in Tab. T.15.
Any samples excluded for the construction of the s chart should also be disregarded for the construction of the $\bar{x}$ chart. So we have $k^\star$ valid samples from a trial run, where $k^\star \leq k$ refers to the reduced number of samples. Their mean values $\bar{x}_1, \ldots, \bar{x}_{k^\star}$ are estimators for μ, i.e.

$$\bar{\bar{x}} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} \bar{x}_i.$$

The control limits are

$$\text{UCL} = \bar{\bar{x}} + 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_3 \bar{s} \quad \text{and} \quad \text{LCL} = \bar{\bar{x}} - 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_3 \bar{s},$$

where we simplified

$$A_3 \approx \frac{3}{c_4\sqrt{n}}$$

The constant A_3 depends on the sample size n and can again be found in Tab. T.15.
From these formulas, it is immediately apparent that the centreline of the s chart has a direct impact on the control limits of the $\bar{x}$ chart.

[2] In contrast to the variance s^2 the standard deviation s of a sample is biased, i.e. $\mathrm{E}(s^2) = \sigma^2$ and $\mathrm{E}(s) \neq \sigma$. Why is that? Answer: Linear functions commute with taking the expectation and the square root is not linear. Remember that the division with $n-1$ instead of n in the calculation makes the variance unbiased but not the standard deviation.

If we assume that the sample is from a normal distribution, then there exists a correction factor

$$c_4(n) = \sqrt{n-1}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$

depending on the sample size n and where Γ is the gamma function which makes $\frac{s}{c_4}$ an unbiased estimator of the standard deviation σ. The unbiasedness of s is worse for small n since c_4 approaches 1 as n grows large, cf. Tab. T.15.

Example 2.2.2 ($\bar{x}$ **and** s **Chart**)**.** We re-examine the data from Ex. 2.2.1 and construct first an s chart and then an $\bar{x}$ chart. We calculate the centreline and the control limits of the s chart and add the standard deviations in a scatter plot. In addition, we plot the centreline and the control limits.

```
> s.bar <- mean(data$sd);  s.bar
[1] 0.00502356
> B3 <- 0         #   from table for n=4
> B4 <- 2.266     #   from table for n=4
> s.CL <- c(B3, 1, B4) * s.bar ;  s.CL
[1] 0.00000000 0.00502356 0.01138339
> plot(data$sd, pch=20, ylim=c(0,0.012), ylab="Standard Deviations",
+      main="s Chart")
> lines(data$sd)
> abline(h=s.CL, lty=c(2,1,2))
> text(rep(1,3), s.CL, label=c("LCL", "CL", "UCL"), pos=3)
```

We find that the process is under control, cf. Fig. 2.2.v. In a next step we can now construct the $\bar{x}$ chart from the s chart. For this we need an estimate for the process standard deviation.

```
> c4 <- 0.9213     #   from table for n=4
> sigma.s.hat <- s.bar / c4;    sigma.s.hat
[1] 0.005452686
```

Note that estimating the process variation using the ranges $R_1, \ldots, R_n$ returned more or less the same result. We have already calculated the estimate for the process mean in Ex. 2.2.1. Thus, we can directly calculate the control limits of the $\bar{x}$ chart and draw the $\bar{x}$ chart.

```
> n <- 4
> x.bar.CL <- x.barbar + c(-3,0,3) * sigma.s.hat / sqrt(n);  x.bar.CL
[1] 0.1924585 0.2006375 0.2088165
> plot(data$mean, pch=20, ylim=c(0.19,0.21), ylab="Mean Values",
+      main="xbar Chart (based on s Chart)")
> lines(data$mean)
> abline(h=x.bar.CL, lty=c(2,1,2))
> text(rep(1,3), x.bar.CL, label=c("LCL", "CL", "UCL"), pos=3)
```

We note that the process is totally under control, as all mean values lie between the control limits, cf. Fig. 2.2.vi.

The question of whether to use a combination of $\bar{x}$ and R chart or a combination of $\bar{x}$ and s chart is usually a matter of personal taste. The calculation of the standard deviations uses the existing information better than the calculation of the ranges and is therefore preferred by many authors. On the other hand, the calculation of the range is easier to perform than that of the standard deviation.

2.3 Individuals Control Charts

In many cases, the sample size is $n = 1$, that is, the sample consists of a single observation. We find many examples:

- In the case of an automatic check, every produced part can be controlled.

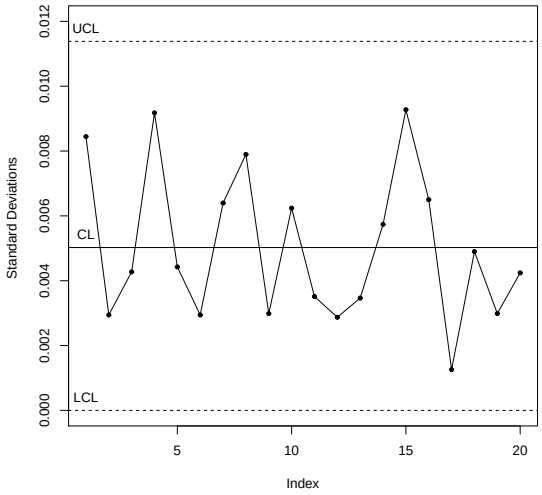


Figure 2.2.v: s chart.

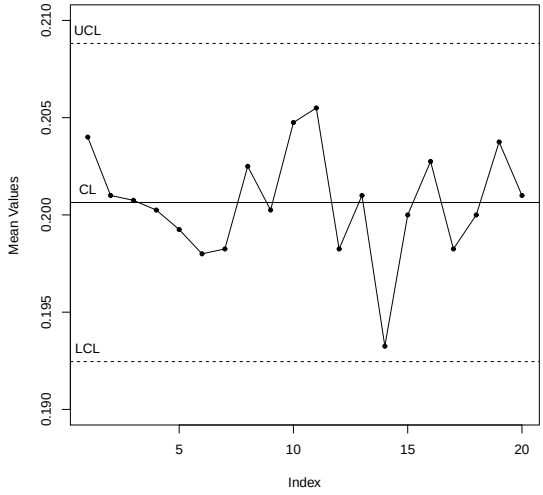


Figure 2.2.vi: $\bar{x}$ chart (based on s chart).

- If the production of a part takes a lot of time, it makes little sense to create groups larger than 1.
- If the difference of a repeated measurement cannot be attributed to the process variation, but to the measuring methods.

In such situations, it is reasonable to construct a control chart for individual measurements.

Let us take the measured values $x_1, \ldots, x_n$ without changing the order of sampling. The control chart for individual measurements use the **moving range** of two successive observations to measure the process variability. Remember, you cannot estimate variability from a single measurement.
The moving range is defined as

$$\mathrm{M}R_i = |x_{i+1} - x_i|$$

for all $i \in \{1, \ldots, n-1\}$. The arithmetic mean of all moving ranges

$$\overline{\mathrm{M}R} = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathrm{M}R_i$$

is a reasonable estimate for the process standard deviation

$$\hat{\sigma} = \frac{\overline{\mathrm{M}R}}{d_2} = \frac{\overline{\mathrm{M}R}}{1.128}.$$

Since two neighboring measurements were used to calculate the moving ranges we have

$$d_2 = 1.128.$$

Thus, we obtain the centreline of the individuals control chart by the arithmetic mean of the measured values

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the control limits

$$\text{UCL} = \bar{x} + 3\frac{\overline{MR}}{1.128} \quad \text{and} \quad \text{LCL} = \bar{x} - 3\frac{\overline{MR}}{1.128}.$$

Example 2.3.1 (**Individuals Control Charts**)**.** The following table shows 26 measured values of the diameter of different holes with the same nominal size.

i	diameter	i	diameter
1	9.99	14	9.94
2	10.12	15	9.93
3	9.81	16	10.09
4	9.73	17	9.98
5	10.14	18	10.11
6	9.96	19	9.99
7	10.06	20	10.11
8	10.11	21	9.84
9	9.95	22	9.82
10	9.92	23	10.38
11	10.09	24	9.99
12	9.85	25	10.41
13	9.92	26	10.36

We construct a control chart for individual measurements to decide if the process, i.e. the drilling of the hole, is under statistical control. To do this, we read the data with R

```
> file <- "diameter.dat"
> data <- read.table(file, header=TRUE)
> str(data)
'data.frame':   26 obs. of  1 variable:
$ diam: num   9.99 10.12  9.81  9.73 10.14 ...
```

and calculate an estimate for the process mean.

```
> x.bar <- mean(data$diam)
[1] 10.02308
```

To calculate the moving range, we use a trick and use the R-command `diff` to calculate the successive differences, of which we calculate the absolute values and then the mean value thereof.

```
> MR <- mean(abs(diff(data$diam)));   MR
[1] 0.1724
```

Now we are able to construct the control chart by calculating an estimate for the process standard deviation.

```
> d2 <- 1.128
> sigma.hat <- MR / d2;   sigma.hat
[1] 0.1528369
> CL <- x.bar + c(-3,0,3) * sigma.hat;  CL
[1] 9.564566 10.023077 10.481588
> plot(data$diam, pch=20, ylim=c(9.3,10.7), ylab="Individual Values",
+       main=main="Individuals Control Charts")
> lines(data$diam)
> abline(h=CL, lty=c(2,1,2))
> text(rep(25,3), CL, label=c("LCL", "CL", "UCL"), pos=3)
```

The process is under control, sine no measured values lie outside the control limits, cf. Fig. 2.3.i.

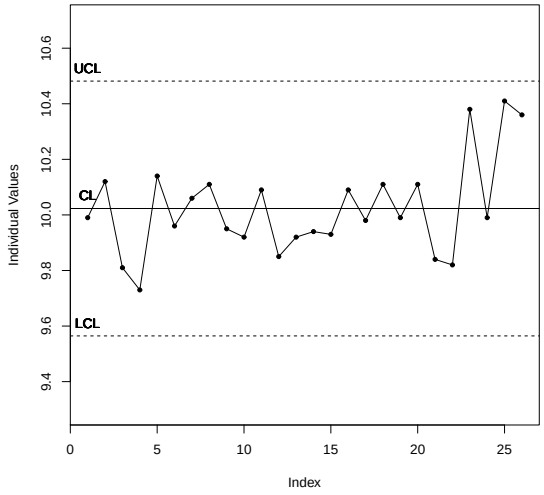


Figure 2.3.i: Individuals control charts.

2.4 Control Charts for Attributes Data – p Chart

Sometimes it is useful to mark a product as defective or useable. For example, the diameter of a bolt can be tested by whether it fits through a normalized hole or not. Such a measurement is usually much faster and cheaper than the measurements of a diameter itself. In such a case we need a **control chart for attributes data**. The proportions of nonconforming items in successive subgroups of size n_i are plotted on p **charts**. These, however, need a larger sample size than the charts based on measurements.

Remark 2.4.1. We consider a sample of size n and find D broken parts. The number of defective D under n examined parts is known to be binomial distributed with the unknown probability of success[3] p (cf. Prob. 2.5.5 and 2.5.6). The relative frequency

$$\hat{p} = \frac{D}{n}$$

is an estimator of the unknown probability of success p. In addition, the variance of the estimator $\hat{p}$ is given by

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

Given k samples with $n_1, \ldots, n_k$ values. Each of these samples has $d_1, \ldots, d_k$ defective products. That is, we get k relative frequencies

$$p_1 = \frac{d_1}{n_1}, \ldots, p_k = \frac{d_k}{n_k},$$

which we plot against the index i in a scatter plot.
The centreline and the control limits of a p chart are again determined from a stable trial run with $k^\star$ (where $k^\star \leq k$ denotes the reduced number of samples) valid samples.

[3] success here means *defective*

- If the sample sizes $n_1, \ldots, n_k$ are all equal to n, then the centreline is the arithmetic mean of the relative frequencies of the (reduced) trial run

$$\bar{p} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} p_i.$$

  In this case the control limits for a p chart are given by

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{and} \quad \text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

- On the other hand, if the sample sizes are not all the same, the control limits also depend on the index i, i.e. on the respective sample size n_i. The centreline is then given by

$$\bar{p} = \frac{d_1 + \cdots + d_{k^\star}}{n_1 + \cdots + n_{k^\star}}$$

  and the control limits by

$$\text{UCL}_i = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \quad \text{and} \quad \text{LCL}_i = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}.$$

For small sample sizes, the lower control limits may become negative. In such cases we set the lower limits to zero.

2.5 Exercises

Problem 2.5.1 (**Example 16-1 from [23]**)**.** Using R reproduce example 16-1 from [23] to create the appropriate control charts. With the following command you can read the data.

```
> file <- "vane-opening.dat"
> data <- read.table(file, header=TRUE)
```

Problem 2.5.2 (**Sheward Control Chart**)**.** A quality inspector in a beverage factory is responsible for the accuracy of the filling machine of a soft drink. She has collected 25 samples consisting of four measurements of fill level [in cm], cf. data set `soft-drinks.dat`. Read the data into R.

a. Examine if the data is from a normal distribution.

b. Use the first 20 measured values as a trial run to create an $\bar{x}$ chart from an R chart.

c. Is the trial run under statistical control?

d. Check if the next 5 measured values meet the quality requirements.

Problem 2.5.3 (**Individuals Control Chart**)**.** In the production of X-ray tubes, the leakage current is a critical quality variable. You find the data in the file `leakage-current.dat`.

a. Examine with a control chart for individual measurements whether the production process is under control.

b. Calculate an estimate for the process mean and process standard deviation.

Problem 2.5.4 (p **Chart**, cf. [38], Beispiel 21-4)**.** In order to study the production of a lathes machine, the production of the machine was checked every day for bad parts during one month ($k = 25$ random samples). Recorded were the number of bad parts and produced objects per day in the data set `rejects.dat`. Create a control chart for attributes using this data.

Problem 2.5.5 (**Binomial Distribution I**)**.** An experiment consists of five independent coin throws with an asymmetrical coin. For every roll, head is likely to top with the probability 0.53. Determine the probability for

a. exactly two times head.

b. a maximum of three times head.

c. more than one head.

Problem 2.5.6 (**Binomial Distribution II**)**.** There are two possibilities in an experiment: success with probability p and failure with probability $1 - p$. Let X be the random variable whose values $x \in \{0, 1, 2, \ldots, n\}$ represent the number of successes in n trials. Determine the probability distribution of X.

Problem 2.5.7 (**Estimation of the Process Standard Deviation**)**.** With which methods can you estimate the usually unknown process standard deviation. Find other methods in the literature, eg. in [11].

Problem 2.5.8 (**Standard Deviation is Biased**)**.** Did you know that the estimator

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

of the standard deviation σ of the sample $x_1, \ldots, x_n$ is biased? If not – check the literature and try to understand.

Solutions

Solution 2.5.5. The random variable X is binomial-distributed with parameters $n = 5$ and $p = 0.53$. Then we obtain:

a. $P(X = 2) =$ `dbinom(x=2, size=5, prob=0.53)` $= 0.2916388$

b. $P(X \leq 3) =$ `pbinom(q=3, size=5, prob=0.53)` $= 0.7727541$

c. $P(X > 1) =$ `pbinom(q=1, size=5, prob=0.53, lower.tail=FALSE)` $= 0.847754$

With `help(dbinom)` you get more information about the implementation of the binomial distribution in R.

Solution 2.5.6. For n trials, there are exactly

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

arrangements with x successes and $n - x$ failures. That is how we get the **binomial distribution**

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Chapter 3

Statistical Properties of Control Charts

3.1 Interpretation of Control Charts

In Chap. 2 we have explained the construction of a control chart from a trial run. We assumed that the trial run was under control, or at least only those measurements were used of which we were sure that they were under control.

In this section we now want to describe the test procedure using a control chart: At any given time t_i with $i \in \{1, \ldots, k\}$ we extract the last n_i parts produced by the machine and calculate the statistics of our interest, eg. for an $\bar{x}$ chart the arithmetic mean or for an R chart the range of the sample values. Then we enter the value of the statistic against the index i in the chart as a point. As long as such a point is within the range defined by the control limits UCL and LCL, the machine follows the target value. In this case the **process** is **under control**, that is, it is stable or stationary. If, on the other hand, the point lies outside the control limits, this is a signal for the interruption of the production process and for checking the settings of the machine. The **process** is **out of control**. All parts produced between this and the previous sampling must then be checked and the scrap sorted out. After such an intervention, the test procedure starts all over again.

This carrying out of a series of similar statistical tests using the control chart also provides an insight into the entire temporal production process of the machine. For this we usually connect all drawn points by a polygon. Thus, on the basis of increasing or decreasing tendencies, we can detect the occurrence of systematic errors, e.g. track the wear of the cutting tool, or the heating of a sensor, and intervene at the right time to avoid the production of scrap.
The aim of process control with control charts is to keep the process under statistical control or, if it is not initially, to put it under statistical control by improving production conditions.

For this purpose, the so-called **western electric rules** for monitoring control charts were defined, cf. Fig. 3.1.i to 3.1.iv. The process is out of control if any of the following applies:

1. Any single data point falls outside the limit defined by UCL and LCL (beyond the 3σ-limit), cf. Fig. 3.1.i.

2. Two out of three consecutive points fall beyond the limit defined by $\frac{2}{3}$UCL and $\frac{2}{3}$LCL on the same side of the centreline (beyond the 2σ-limit), cf. Fig. 3.1.ii.

3. Four out of five consecutive points fall beyond the limit defined by $\frac{1}{3}$UCL and $\frac{1}{3}$LCL on the same side of the centreline (beyond the 2σ-limit), cf. Fig. 3.1.iii.

4. Nine consecutive points fall on the same side of the centreline (so-called **run**), cf. Fig. 3.1.iv.

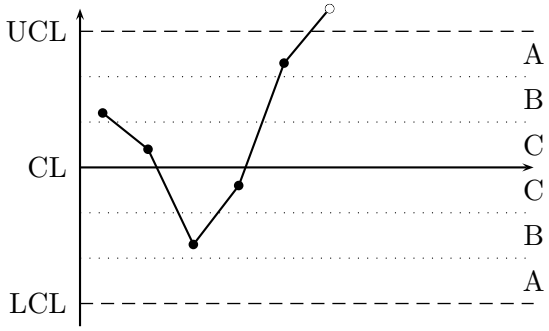


Figure 3.1.i: Rule 1: Any point beyond zone A.

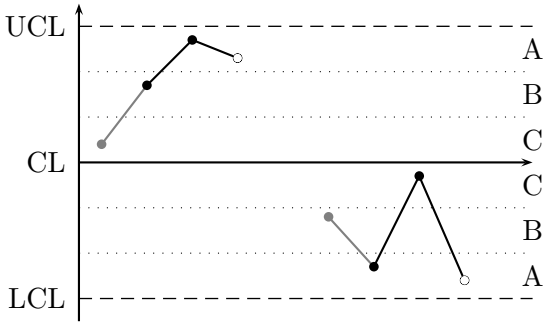


Figure 3.1.ii: Rule 2: Two out of three consecutive points fall on the same side in zone A or beyond.

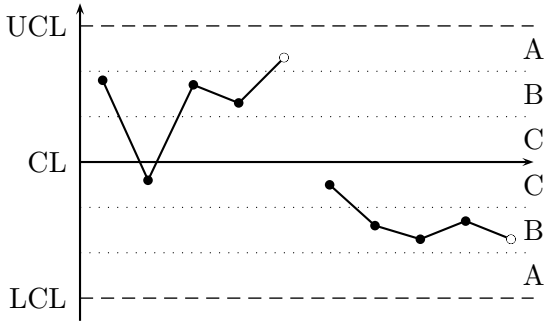


Figure 3.1.iii: Rule 3: Four out of five consecutive points fall on the same side in zone B or beyond.

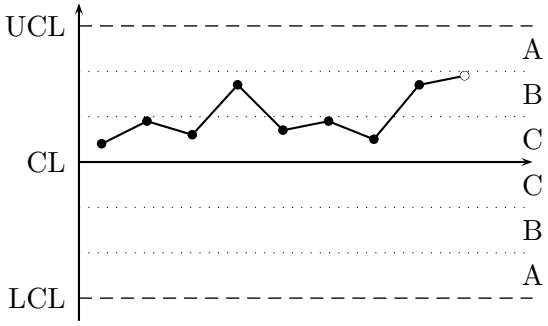


Figure 3.1.iv: Rule 4: Nine consecutive points fall on the same side of the centreline.

These rules not only look at a single point in the control chart, but try a **pattern recognition** to find specific patterns in the control chart that indicate a process that is not under control.

3.2 Type I Error and Type II Error

When monitoring a production process with a control chart, as with any statistical test, there are two wrong decisions possible.
Let μ_0 be the target value of a process and the two alternative hypotheses are:

$$H_0 : \mu_0 = \mu, \text{ i.e. process is not disturbed.}$$
$$H_1 : \mu_0 \neq \mu, \text{ i.e. process is disturbed, } \mu_1 \text{ is true.}$$

Let us denote by μ the considered statistic calculated from the measured values $x_1, \ldots, x_n$, eg. $\mu = \bar{x}$ when using the chart for $\bar{x}$ or $\mu = R$ when using the chart for R, and by μ_1 the real value of the considered statistic.
We make a statistical conclusion in favor of the null hypothesis H_0 or the alternative hypothesis H_1. In both cases, certain inferences are drawn from one sample to the corresponding population. We must remember that there are absolutely no sure conclusions. In a test decision there is always a certain probability that the decision made is wrong. There are two types of errors:

Definition 3.2.1.

(a) If a *true* null hypothesis H_0 is *rejected* we make a **type I error**. An intervention in the process is necessary, because the control limits are exceeded, although the process is not disturbed. This is called a **false alarm**.

(b) If a *false* null hypothesis H_0 is *accepted* we make a **type II error**. There is no intervention, since the control limits are not exceeded, although the process is disturbed. This is called an **omitted alarm**.

The probability of a type I error corresponds to the error probability for which we have specified the significance level α. The probability of a type II error is denoted by β and depends on the actual mean μ_1 of the measured feature.

We now explain the possible cases by means of a one-sided test, in which the null hypothesis H_0: $\mu = \mu_0$ is tested against the alternative hypothesis H_1: $\mu > \mu_0$. Let μ denote the unknown parameter of the distribution to be tested.

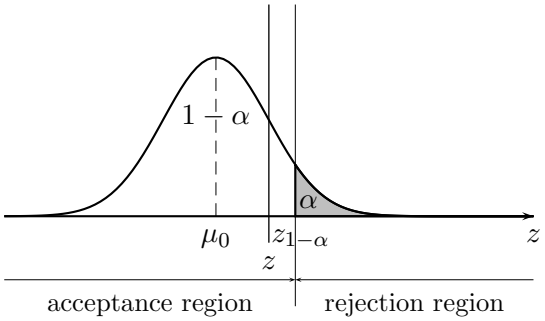


Figure 3.2.i: Let H_0 be true: Since $z < z_{1-\alpha}$ the null hypothesis is accepted. This is the right decision, which is made with probability $1 - \alpha$.

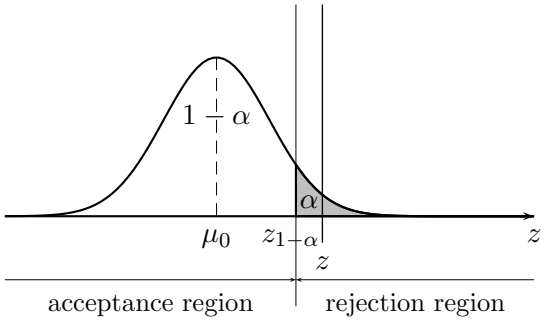


Figure 3.2.ii: Let H_0 be true: Since $z \geq z_{1-\alpha}$ the null hypothesis is rejected. This is the wrong decision (**type I error**), which is made with probability α.

The power of a hypothesis test is the probability $1 - \beta$ that the test correctly rejects the null hypothesis when the alternative hypothesis is true, cf. Fig. 3.2.iv, i.e.

$$\text{power} = P(\text{reject } H_0 | \, H_1 \text{ is true}) = 1 - \beta.$$

In practice, we try to keep the type I and II errors simultaneously small (i.e. α and β at the same time). Observe Figs. 3.2.iv and 3.2.iii and notice that a reduction of α (shift the critical

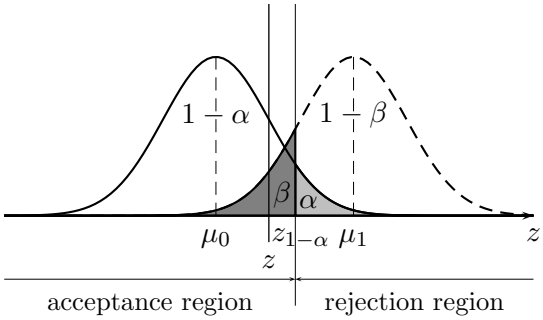


Figure 3.2.iii: Let H_0 be false, H_1 true, i.e. the dashed density is true: Since $z < z_{1-\alpha}$ the null hypothesis is accepted. This is the wrong decision (**type II error**), which is made with probability β.

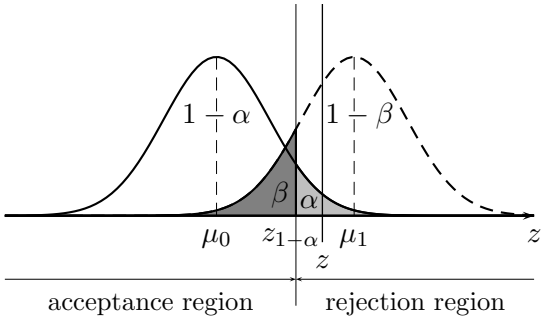


Figure 3.2.iv: Let H_0 be false, H_1 true, i.e. the dashed density is true: Since $z \geq z_{1-\alpha}$ the null hypothesis is rejected. This is the correct decision, which is made with probability $1 - \beta$ (**power**).

size $z_{1-\alpha}$ to the right) automatically increases β and vice versa. If we decide for a small α and thus for a small risk to reject a correct null hypothesis, then at the same time we accept a significantly increased risk for a type II error. So we have to decide case by case which of the two errors has the bigger consequences.
However, if the risk for a type II error β should be reduced, without at the same time having to increase the probability α for a type I error we only can increase the sample size[1] (improving the power of the test). In addition, we see in Fig. 3.2.iii that the power depends on the alternative hypothesis H_1, i.e. on the location of μ_1.
This functional relationship between the power of a test and μ_1 is denoted by **operating characteristic**.

3.3 Power Function and Operating Characteristic

The **power function** is the probability to reject the null hypothesis H_0 if H_1, i.e. μ_1 is true. Of course, this probability depends on the location of μ_1. We are usually not interested in μ_1 alone, but always in relation to the target value μ_0 and the process standard deviation σ. Therefore, we introduce the new normalised variable

$$\delta = \frac{\mu_1 - \mu_0}{\sigma},$$

or $\mu_1 = \delta\sigma + \mu_0$. The variable δ is a measure for the deviation of the disturbed from the undisturbed process in units of σ. In statistical process control the **power function** is denoted by

$$g(\mu_1) = g(\delta\sigma + \mu_0) = \tilde{g}(\delta).$$

It is a measure for the probability of an intervention in the process.

- For an undisturbed process, i.e. $\mu = \mu_0$, we have

$$g(\mu_0) = \tilde{g}(0) = \alpha, \tag{3.3.a}$$

[1] The distributions become thinner and thus α and β simultaneously smaller.

where α is the error probability. In this script, we have consistently set $\alpha = 0.0027$ so that the control limits UCL and LCL become three-σ limits.

- For a disturbed process, $g(\mu_1)$ should be as large as possible and thus the probability for an omitted alarm $1 - g(\mu_1)$ as small as possible.

In the following we want to calculate the power function for a simple $\bar{x}$ chart. The error probability α is given. We assume that the measured values $x_1, \ldots, x_n$ come from a normal distribution with known parameters μ_0 and σ^2. Thus, the mean $\bar{x}$ is also normally distributed with parameters μ_0 and $\frac{\sigma^2}{n}$. Here we make the unrealistic assumption that the process standard deviation σ is known. Of course, in reality the process standard deviation must be estimated from a stable trial run. In the real case, the following calculations are only approximately true. For exact formulas see [29].
Denote by μ the considered statistic calculated from the measured values $x_1, \ldots, x_n$, i.e. $\mu = \bar{x}$ with the chart for $\bar{x}$. So we get for the power function, cf. Fig. 3.3.i,

$$g(\mu_1) = P(\mu \le \mathrm{LCL}) + P(\mathrm{UCL} \le \mu) \tag{3.3.b}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mathrm{LCL}} \mathrm{e}^{-\frac{(z-\mu_1)^2}{2\sigma^2} n} \mathrm{d}z + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathrm{UCL}}^{\infty} \mathrm{e}^{-\frac{(z-\mu_1)^2}{2\sigma^2} n} \mathrm{d}z, \tag{3.3.c}$$

with the two control limits

$$\mathrm{UCL} = \mu_0 + z_q \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mathrm{LCL} = \mu_0 - z_q \frac{\sigma}{\sqrt{n}},$$

where z_q with $q = 1 - \frac{\alpha}{2}$ is the two-sided q-quantile of the normal distribution.

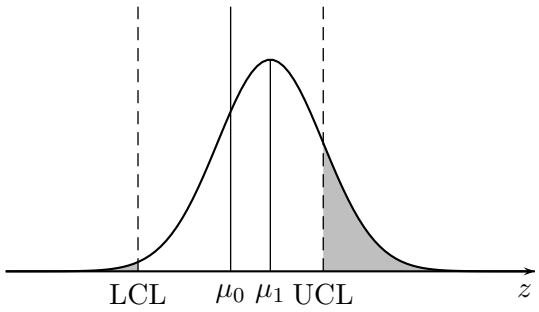


Figure 3.3.i: Definition of the power function using the actual distribution with parameters μ_1 and $\frac{\sigma^2}{n}$ drawn on the control chart.

In the following, $\Phi(x, \mu, \sigma^2)$ denotes the probability distribution of the **normal distribution** with the parameters μ and σ^2, i.e.

$$\Phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}} \mathrm{d}z.$$

We calculate

$$\begin{aligned} g(\mu_1) &= 1 - \Phi\left(\mathrm{UCL}, \mu_1, \frac{\sigma^2}{n}\right) + \Phi\left(\mathrm{LCL}, \mu_1, \frac{\sigma^2}{n}\right) \\ &= 1 - \Phi\left(\frac{\mathrm{UCL} - \mu_1}{\sigma}\sqrt{n}, 0, 1\right) + \Phi\left(\frac{\mathrm{LCL} - \mu_1}{\sigma}\sqrt{n}, 0, 1\right) \\ &= 1 - \Phi\left(\frac{\mu_0 + z_q\frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma}\sqrt{n}, 0, 1\right) + \Phi\left(\frac{\mu_0 - z_q\frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma}\sqrt{n}, 0, 1\right) \\ &= 1 - \Phi\left(-\delta\sqrt{n} + z_q, 0, 1\right) + \Phi\left(-\delta\sqrt{n} - z_q, 0, 1\right). \end{aligned}$$

Using the symmetry $\Phi(x, 0, 1) = 1 - \Phi(-x, 0, 1)$ we obtain the power function

$$\tilde{g}(\delta) = \Phi\left(\delta\sqrt{n} - z_q, 0, 1\right) + \Phi\left(-\delta\sqrt{n} - z_q, 0, 1\right). \tag{3.3.d}$$

From Eqn. (3.3.d) we can immediately conclude that the operating characteristic is an even function, i.e. that $\tilde{g}(-\delta) = \tilde{g}(\delta)$. Thus, in the literature, often only one side, i.e. $\delta \geq 0$ is considered.
In statistics the **operating characteristic** (OC curve) is often used instead of the power function. This is the functional relation between the type II error and the position of μ_1, i.e.

$$\mathrm{OC}(\mu_1) = 1 - g(\mu_1).$$

Example 3.3.1. We want to calculate the power function of the statistical hypothesis test described in Ex. 2.1.1. The error probability is $\alpha = 0.0027$. There we assumed that the measured values $x_1, \ldots, x_n$ came from a normal distribution with the known parameters $\mu_0 = 10.000\,\mathrm{mm}$ and $\sigma = 0.005\,\mathrm{mm}$. In our example we want to draw the power functions for different sample sizes with R.

```
#   target value
> mu0 <- 10.000
#   known standard deviation
> sigma <- 0.005
#   q-quantile
> alpha <- 0.0027
z.q <- qnorm(1-alpha/2)
#   define power function
> func.g.tilde <- function(delta,n)
+                  { pnorm(delta*sqrt(n)-z.q) + pnorm(-delta*sqrt(n)-z.q) }
#   graphic
> curve(func.g.tilde(delta=x,n= 2), from=-4, to=4, lty=2,
+       xlab="delta", ylab="Power Function", main="Power Function")
> curve(func.g.tilde(delta=x,n= 5), from=-4, to=4, lty=3, add=TRUE)
> curve(func.g.tilde(delta=x,n=10), from=-4, to=4, lty=4, add=TRUE)
> curve(func.g.tilde(delta=x,n=20), from=-4, to=4, lty=5, add=TRUE)
> abline(v=0, h=0)
#   add legend
> legend(-4,0.2, lty=2:5, legend=c("n=2","n=5","n=10", "n=20"), bty="n")
```

In Fig. 3.3.ii four power function curves for different sample sizes $n = 2$, $n = 5$, $n = 10$ and $n = 20$ are plotted.

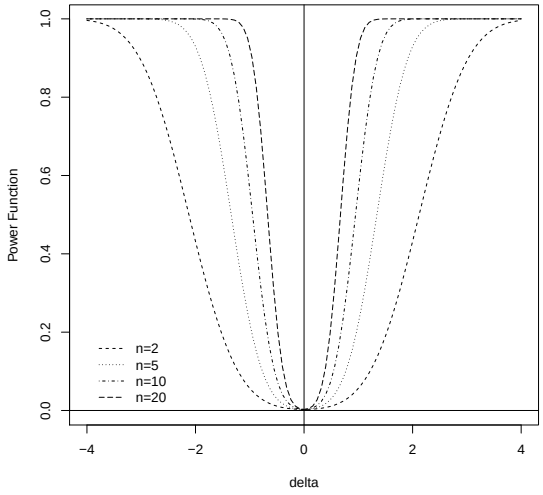


Figure 3.3.ii: Four power function curves for different sample sizes $n = 2$, 5, 10 and 20.

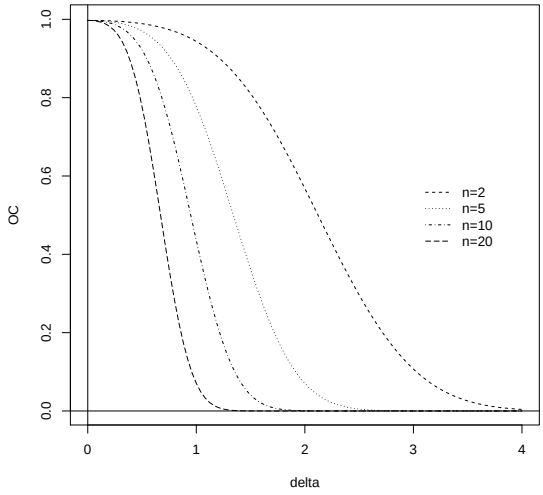


Figure 3.3.iii: Four operating characteristic curves for different sample sizes $n = 2$, 5, 10 and 20.

For a given δ (deviation of the actual value of the process to the target value measured in units of $\sigma = 0.005\,\text{mm}$) we can read off the plot the probability of a type II error, i.e. an omitted alarm
In Fig. 3.3.iii four operating characteristic curves for different sample sizes $n = 2$, $n = 5$, $n = 10$ and $n = 20$ are plotted.

```
#   define OC curve
func.OC <- function(delta,n) { 1 - func.g.tilde(delta, n) }
#   Grafik
> curve(func.OC(delta=x,n= 2), from=0, to=4, lty=2,
+       xlab="delta", ylab="beta", main="operating characteristic")
> curve(func.OC(delta=x,n= 5), from=0, to=4, lty=3, add=TRUE)
> curve(func.OC(delta=x,n=10), from=0, to=4, lty=4, add=TRUE)
> curve(func.OC(delta=x,n=20), from=0, to=4, lty=5, add=TRUE)
> abline(v=0, h=0)
#   add legend
> legend(3,0.6, lty=2:5, legend=c("n=2","n=5","n=10", "n=20"), bty="n")
```

For instance for $n = 5$ and $\delta = 1$, that is for $\mu_1 = \delta\sigma + \mu_0 = 10.005\,\text{mm}$ the probability of an omitted alarm is $\text{OC}(10.005\,\text{mm}) = 1 - g(10.005\,\text{mm}) = 0.7775392$, cf. the following R-code.

```
#   probability of an omitted alarm
> delta <- 1
> func.OC(delta,n=5)
[1] 0.7775392
```

We observe that with increasing n the slope of the power function increases. On the other hand, for fixed n, $\tilde{g}(\delta)$ tends to 1, if δ goes to $\pm\infty$. The probability to intervene into process tends to 1 with increasing deviation of the actual process value μ_1 from the target value μ_0.

Remark 3.3.1. The formulas for the power function and the operating characteristic are only approximations, if the target value μ_0 and the process standard deviation σ must be estimated from a trial run.

3.4 Average Run Length

The statements in Chap. 3.3 about the probabilities for an omitted resp. false alarm refer to the performance of a single test and thus to a fixed point in time. At any time t_i with $i \in \{1, 2, \ldots, k\}$ a statistical test is made when running a control chart based on a random sample. We are interested in the time to a first intervention, the so-called **run length**.
We assume that n random samples are taken at fixed intervals and independently from the actual production process.
If the i_{RL}-th sample is the first to result in an intervention, i.e. this sample is beyond the control limits, then i_{RL} is called the **run length** of the control chart.
The run length I_{RL} is a discrete random variable[2], which takes values $k \in \{1, 2, 3, \ldots\}$ with the probabilities

$$p_k = P(I_{\mathrm{RL}} = k). \tag{3.4.a}$$

The expected value

$$\mathrm{ARL} = \mathrm{E}(I_{\mathrm{RL}}) = \sum_{k=1}^{\infty} k\, p_k$$

of the random variable I_{RL} is called **average run length** (ARL). It describes the mean time until the first intervention in the current process.
The average run length depends on the actual mean μ_1 of the process as well as the control limits. A chart would be *ideal* if, for an undisturbed process, i.e. $\mu = \mu_0$, no intervention and no false alarm occurred; in other words, it would be $\mathrm{ARL} = \mathrm{E}(I_{\mathrm{RL}}) = \infty$. On the other hand, with an ideal chart, an immediate alarm would be given if the process is so disturbed that one intervention is more economical than keep up production; in other words, it would be $\mathrm{ARL} = \mathrm{E}(I_{\mathrm{RL}}) = 1$. However, such a control chart cannot be expected with random sampling.
To calculate the average run length we need the probabilities $p_k = P(I_{\mathrm{RL}} = k)$. The event $I_{\mathrm{RL}} = k$ occurs if the kth sample gives an alarm and the previous $k-1$ samples did not. Since the samples and thus the decisions are independent of each other, the random variable I_{RL} has a **geometric distribution** (cf. Prob. 3.6.3 and 3.6.4) with the probabilities

$$p_k = p(1-p)^{k-1}$$

and the expected value (cf. Prob. 3.6.5)

$$\mathrm{E}(I_{\mathrm{RL}}) = \frac{1}{p}, \tag{3.4.b}$$

where the parameter p with $0 \le p \le 1$ denotes the intervention probability, i.e. the likelihood of exceeding the control limits when performing a test. This probability p depends on the actual process mean μ_1 and on the control limits UCL and LCL. That is

$$p = p(\mu_1) = P(\mu \le \mathrm{LCL}) + P(\mathrm{UCL} \le \mu).$$

[2]Random variables are generally given capital Latin characters $X, Y, Z, \ldots, I_{\mathrm{RL}}$, and their realisations small Latin characters $x, y, z, \ldots, k$.

It is nothing else than the power function of the statistical test (cf. Eqn. (3.3.b)), i.e.

$$p = p(\mu_1) = g(\mu_1) = \tilde{g}(\delta) \quad \text{with} \quad \delta = \frac{\mu_1 - \mu_0}{\sigma}.$$

Therefore we immediately find the average run length

$$\mathrm{ARL}(\delta) = \frac{1}{p(\mu)} = \frac{1}{g(\mu_1)} = \frac{1}{\tilde{g}(\delta)}.$$

Because of this relationship, the average run length is also used instead of the power function to evaluate and compare various control charts.
If we had an undisturbed process with $\mu_1 = \mu_0$, i.e. with $\delta = 0$, then it would follow from Eq. (3.3.a) that

$$\mathrm{ARL}(0) = \frac{1}{\alpha}.$$

Example 3.4.1. With the error probability $\alpha = 0.0027$ used in this script it follows

$$\mathrm{ARL}(0) = \frac{1}{0.0027} = 370.3704.$$

That is, for an undisturbed process on average only after 370 samples a false alarm occurs.

To determine the average run length of a disturbed process with $\mu = \mu_1$, i.e. $\delta = \frac{\mu_1 - \mu_0}{\sigma}$ we use the power function $\tilde{g}(\delta)$ from Eq. (3.3.d).

Example 3.4.2. Again, let us consider Ex. 2.1.1. The error probability is $\alpha = 0.0027$. There we assumed that the measured values $x_1, \ldots, x_n$ came from a normal distribution with the known parameters $\mu_0 = 10.000\,\mathrm{mm}$ and $\sigma = 0.005\,\mathrm{mm}$. For our example we want to plot the average run length $\mathrm{ARL}(\delta)$ for different sample sizes with R, cf. Fig. 3.4.i.

```
#   target value
> mu0 <- 10.000
#   known standard deviation
> sigma <- 0.005
#   q-quantile
> alpha <- 0.0027
> z.q <- qnorm(1-alpha/2)
#   average run length of the undisturbed process
> ARL.0 <- 1/alpha;   ARL.0
[1] 370.3704
#   define power function
> func.g.tilde <- function(delta,n) {
+                      pnorm(delta*sqrt(n)-z.q) + pnorm(-delta*sqrt(n)-z.q)
+                 }
#   graphic
> curve(1/func.g.tilde(delta=x,n= 2), from=0, to=4, lty=2,
+       ylim=c(0, 400), xlab="delta", ylab="ARL", main="Average Run Length")
> curve(1/func.g.tilde(delta=x,n= 5), from=0, to=4, lty=3, add=TRUE)
> curve(1/func.g.tilde(delta=x,n=10), from=0, to=4, lty=4, add=TRUE)
> curve(1/func.g.tilde(delta=x,n=20), from=0, to=4, lty=5, add=TRUE)
> abline(v=0, h=0)
#   add legend
> legend(3,100, lty=2:5, legend=c("n=2","n=5","n=10", "n=20"), bty="n")
> points(0, ARL.0, pch="-", cex=2)
> text(0, ARL.0, label=floor(ARL.0), pos=4)
```

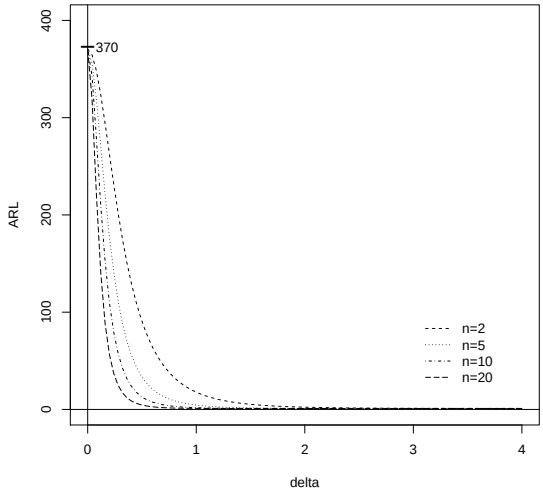


Figure 3.4.i: Average run lengths curves for different sample sizes n and with error probability $\alpha = 0.0027$.

For a given δ (deviation of the actual value of the process to the target value measured in units of $\sigma = 0.005\,\text{mm}$) we can read off the plot how many samples are taken on average until the control charts signals an intervention.
For instance for $n = 5$ and $\delta = 1$, that is for $\mu_1 = \delta\sigma + \mu_0 = 10.005\,\text{mm}$ the average run length is $\text{ARL}(1) = \frac{1}{\tilde{g}(1)} \approx 4$, cf. the following R-code.

```
#   average run length for a given deviation
> delta <- 1
> 1/func.g.tilde(delta,n=5)
[1] 4.495174
```

At fixed sample size n: the larger δ, the smaller the average run length. For big δ the average run length ARL ≈ 1; so on average, there is an intervention every sample.
Also, in the Fig. 3.4.i we observe that keeping δ constant and if the sample size n increases, then ARL gets smaller, i.e. the average time until the first intervention gets shorter.

Remark 3.4.1. The formulas for the power function, the operating characteristic and the average run length are only approximations, if the target value μ_0 and the process standard deviation σ must be estimated from a trial run.

3.5 Process Capability

To derive the control limits of a control chart, we assumed that the random variable X is normally distributed with parameters μ_0 and σ^2. In general, the parameters are unknown and must be estimated from a trial run.
Thus, choosing the error probability $\alpha = 0.0027$, there are 99.73% of all measured values, i.e. almost all, within the three-σ zone $[\mu_0 - 3\sigma, \mu_0 + 3\sigma]$. The width of the three-σ zone is therefore 6σ.

Note that the **control limits** $\mathrm{UCL} = \mu_0 + 3\frac{\sigma}{\sqrt{n}}$ and $\mathrm{LCL} = \mu_0 - 3\frac{\sigma}{\sqrt{n}}$ define a zone that is typically narrower than the six-σ zone, unless the sample size is $n = 1$.
On the other hand, in industry technical tolerances for quality characteristics are defined in the form of two **tolerance limits**: **upper specification limit** (USL) and **lower specification limit** (LSL), which have to be met in production. If the measured value x of the random variable X is beyond the tolerance limits, then the part is rejected. A control chart is only suitable for monitoring the process if as little bad parts as possible are produced.
The **specification limit** (SL) is defined by

$$\mathrm{SL} = \frac{\mathrm{USL} + \mathrm{LSL}}{2}.$$

This performance is measured with so-called **capability process ratios** (PCR). The simplest process capability index is

$$C_p = \frac{\mathrm{USL} - \mathrm{LSL}}{6\sigma}.$$

The capability process ratio C_p expresses the ratio of the width of the tolerance range to the width of the process range. Numerous other process capability ratios can be found in the literature, cf. e. g. [22] and [33].
If we assume that the process is under control, i.e. if $\mu_1 = \mu_0$ and $\mathrm{SL} = \mu_0$, then

- $C_p = 1$ implies a reject rate of $\alpha \cdot 100\% = 0.27\%$.
- $C_p < 1$ implies a reject rate of more than $\alpha \cdot 100\% = 0.27\%$, i.e. the process capability is not guaranteed.
- $C_p > 1$ implies a reject rate of less than $\alpha \cdot 100\% = 0.27\%$, i.e. the process capability is guaranteed.

In practice the target value μ_0 and the process standard error σ are unknown and are estimated from a trial run. In this case we only get an estimate for the process capability ratio. Note that this estimate is very sensitive if the process is not under control or if the measured values are not normally distributed.

Example 3.5.1. Again, we consider Ex. 2.1.1. The error probability is $\alpha = 0.0027$. We assumed that the measured values $x_1, \ldots, x_n$ came from a normal distribution with known parameters $\mu_0 = 10.000\,\mathrm{mm}$ and $\sigma = 0.005\,\mathrm{mm}$. The allowed deviation from the target value μ_0 are $\delta_1 = -0.012\,\mathrm{mm}$ and $\delta_2 = 0.024\,\mathrm{mm}$. In other words, the thickness should be between the tolerance limits

$$\mathrm{USL} = \mu_0 + \delta_2 = 10.024\,\mathrm{mm} \quad \text{and} \quad \mathrm{LSL} = \mu_0 + \delta_1 = 9.988\,\mathrm{mm}.$$

In this case, the quality requirements are met.
We calculate

$$C_p = \frac{10.024\,\mathrm{mm} - 9.988\,\mathrm{mm}}{6 \cdot 0.005\,\mathrm{mm}} = 1.2.$$

This puts the reject rate below 0.27% and the process capability is guaranteed.

3.6 Exercises

Problem 3.6.1 (**Western Electric Rules**)**.** Examine if the process of Ex. 2.2.1 is under control according to the Western Electric Rules. Just look at the chart for $\bar{x}$ which is based on the chart for R.

Problem 3.6.2 (**OC Curve**)**.** Determine an approximate operating characteristic curve of the chart for $\bar{x}$ (based on the s chart) of Ex. 2.2.1. The target value μ_0 and the process standard deviation σ are unknown and can be estimated from the data. Calculate the probability of an omitted alarm if $\Delta\mu = \mu_1 - \hat{\mu_0} = \delta\hat{\sigma} = 0.015\,\text{cm}$.

Problem 3.6.3 (**Geometric Distribution I**)**.** For each throw the probability to hit a basketball in the opponent's basket is $p = 0.3$. The ball is thrown on the basket until it strucks for the first time. Determine the probability for

a. for exactly two throws.

b. at most three throws.

c. more than one throw.

Problem 3.6.4 (**Geometric Distribution II**)**.** Workpieces are produced from blanks. The probability of making a usable piece from a blank is p. Determine the probability distribution for the number X of the blanks used.

Problem 3.6.5 (**Expected Value of the Geometric Distribution**)**.** Suppose that X is geometrically distributed with parameter p with $0 \leq p \leq 1$, i.e. $P(X = k) = p(1-p)^{k-1}$. Show that $\mathrm{E}(X) = \frac{1}{p}$, i.e. prove Eq. 3.4.b.

Problem 3.6.6 (**ARL**)**.** Determine the average run length of the chart for $\bar{x}$ (based on the s chart) of Ex. 2.2.1. The target value μ_0 and the process standard deviation σ are unknown and can be estimated from the data. Calculate the average run length if $\Delta\mu = \mu_1 - \hat{\mu_0} = \delta\hat{\sigma} = 0.015\,\text{cm}$.

Problem 3.6.7 (**PCR**)**.** Estimate the capability process ratio of Ex. 2.2.1 with the tolerance limits USL $= 0.223\,\text{mm}$ and LSL $= 0.177\,\text{mm}$. The target value μ_0 and the process standard deviation σ are unknown and can be estimated from the data.

Solutions

Solution 3.6.3. The random variable X counts the number of throws. Thus we obtain

a. $P(X = 2) =$ `dgeom(x=2-1, prob=0.3)` $= 0.21$.

b. $P(X \leq 3) =$ `pgeom(q=3-1, prob=0.3)` $= 0.657$.

c. $P(X > 1) =$ `pgeom(q=1-1, prob=0.3, lower.tail=FALSE)` $= 0.7$.

With `help(dgeom)` you can get more information about the implementation of the geometric distribution in R. Note that `x` and `q` in the functions `dgeom` and `pgeom` are the number of failed attempts until the first success.

Solution 3.6.4. Let us denote $X = k$ the number of blanks used.

k	1	2	3	4	...	k	...
$P(X=k)$	p	$p(1-p)$	$p(1-p)^2$	$p(1-p)^3$	$\cdots$	$p(1-p)^{k-1}$	$\cdots$

A distribution has to be normalised, therefore we control our result

$$\sum_{k=1}^{\infty} P(X=k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p\frac{1}{1-(1-p)} = 1.$$

In the second-to-last equation, we have applied the sum formula of a geometric series

$$\sum_{k=1}^{\infty} q^{k-1} = \frac{1}{1-q}$$

with $|q| = |1-p| < 1$.

Solution 3.6.5. We calculate

$$\begin{aligned}\sum_{k=1}^{\infty} kP(X=k) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p\sum_{k=0}^{\infty} k(1-p)^{k-1} \\ &= -p\sum_{k=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}p}(1-p)^k = -p\frac{\mathrm{d}}{\mathrm{d}p}\frac{1}{1-(1-p)} = -p\frac{\mathrm{d}}{\mathrm{d}p}\frac{1}{p} = p\frac{1}{p^2} \\ &= \frac{1}{p}.\end{aligned}$$

In the fourth-last equation we again have again applied the sum formula of a geometric series

$$\sum_{k=1}^{\infty} q^{k-1} = \frac{1}{1-q}$$

with $|q| = |1-p| < 1$. Is that not a very nice proof?

Chapter 4

Control Charts with Memory

Using a classic Shewhart control chart, the decision to interfere with the production process is based only on the results of the current sample. The development of the manufacturing process in the past is not considered in the decision. On the other hand, control charts developed in recent years are also based on random samples taken in the past. These charts are called **control charts with memory**. Because these charts use more information about the process, they can detect a disturbance faster than the classical charts. We will look at the two most common charts with memory.
In a manufacturing process, random samples of size $n_i \geq 1$ are collected. Therefore, we have the following data set:

sample	measured values						mean	standard deviation
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1n_1}	$\bar{x}_1$	s_1
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2n_2}	$\bar{x}_2$	s_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{in_i}	$\bar{x}_i$	s_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
k	x_{k1}	x_{k2}	$\cdots$	x_{kj}	$\cdots$	x_{kn_k}	$\bar{x}_k$	s_k

The mean values and standard deviations of the individual samples were added to this data set. We assume that the data comes from a normal distribution.
Control charts with memory are based on a linear combination of the mean values of some or all random samples taken up to the time with index i, i.e.

$$y_i = \alpha_i + \sum_{j=1}^{i} \beta_j \bar{x}_j, \tag{4.0.a}$$

where α_i and the weights $\beta_1, \ldots, \beta_i$ can be arbitrary real numbers with

$$\sum_{j=1}^{i} \beta_j = 1.$$

Depending on how the weights are chosen, we get another type of control chart.

For example, we get the Shewhart chart for $\bar{x}$ introduced in Chap. 2.2 if we set $\alpha_i = 0$ and $\beta_1 = \cdots = \beta_{i-1} = 0$ and $\beta_i = 1$, i.e.

$$y_i = \bar{x}_i.$$

This chart has no memory.

4.1 CUSUM

In this chapter we present the **cumulative sum control chart**, short **CUSUM** chart, to control the mean. The CUSUM chart plots the cumulative sums of deviations of measurement values from the target value. CUSUMs can be constructed for single observations as well as for sample means.
So let $\bar{x}_i$ be the mean of the i-th random sample taken from the production process. If the process is under control, then we assume that $\bar{x}_i$ comes from a normal distribution with expected value μ_0 and standard deviation $\frac{\sigma}{\sqrt{n_i}}$. We assume that the standard deviation is either known or can be estimated.
The value μ_0 is the **target value** for the quality characteristic X. If the process drifts away from this target, the CUSUM chart will give us fast a signal, and we will have to readjust the process.
Using two[1] statistics C^+, resp. C^- the CUSUM chart sums up deviations above, resp. below the target value. These statistics are calculated recursively as follows

$$\begin{aligned} C_i^+ &= \max\{0, \bar{x}_i - (\mu_0 + K) + C_{i-1}^+\}, \\ C_i^- &= \max\{0, (\mu_0 - K) - \bar{x}_i + C_{i-1}^-\}. \end{aligned}$$

The starting values of the recursions are $C_0^+ = 0$ and $C_0^- = 0$.
The constant K is called **reference value**. If we want to detect a shift of, say, Δ, then K is in general set to

$$K = \frac{\Delta}{2}.$$

Note that the two statistics C^+ and C^- sum up only those absolute deviations from the target value which are greater than the reference value K.
If the process is under control, i.e. if for all i

$$\mu_0 - K \leq \mathrm{E}(\bar{x}_i) \leq \mu_0 + K,$$

then the expected values of the statistics are both zero

$$\mathrm{E}(C_i^+) = 0 \quad \text{and} \quad \mathrm{E}(C_i^-) = 0.$$

In this case the two random variables C^+ and C^- fluctuate around the line at zero.

[1] In the literature (cf. [38]) there is also a simpler version with only one statistic

$$y_i = -i\mu_0 + \sum_{j=1}^{i} \bar{x}_j = \sum_{j=1}^{i} (\bar{x}_j - \mu_0).$$

It can be derived directly from Eq. (4.0.a) by setting $\alpha_i = -i\mu_0$ and $\beta_1 = \cdots = \beta_i = 1$. The statistic y_i is therefore a cumulative sum of the deviations of all previous sample mean values from the target value μ_0.

If, on the other hand, the process is out of control beyond a certain index i_0, say

$$\mathrm{E}(\bar{x}_i) = \mu \geq \mu_0 + K \quad \text{for all } i \geq i_0,$$

then the statistic C^+ sums up the deviations from $\mu_0 + K$, i.e.

$$\begin{aligned} \mathrm{E}(C_i^+) &\geq (\mu - (\mu_0 + K)) \cdot (i - i_0) = (\Delta\mu - K) \cdot (i - i_0) \text{ und} \\ \mathrm{E}(C_i^-) &= 0. \end{aligned}$$

From the moment i_0 at which the disturbance occurs, the points (i, C_i^+) lie on average above a straight line with slope $\mu - (\mu_0 + K)$, cf. Fig. 4.1.i. The same thing happens, of course, with a different sign, if the process is not under control in the other direction.
The process is out of control if either C^+ or C^- exceeds a certain constant H. The constant H is called **decision interval**. If the CUSUM chart indicates that the process is out of control, then we should stop the process and look for the cause. After the problem has been resolved, the CUSUM chart must be restarted, that is, the recursion starts again at $C_0^+ = 0$ and $C_0^- = 0$.
There are numerous studies on how to choose the reference value K and the decision interval H. In principle, the two constants are chosen so that the CUSUM chart has a good average run length ARL, cf. Kap. 3.4. As a rule of thumb, the following values can be used

$$K = \frac{\hat{\sigma}}{2} \quad \text{and} \quad H = 5\,\hat{\sigma},$$

where $\hat{\sigma}$ is an estimate for the process standard deviation. That is, we want to detect approximately a shift of one standard deviation. For more details we refer to the literature, eg. [22], Chap. 8-1.3.

Example 4.1.1 (**CUSUM**, cf. [23], Example 15-2, data slightly modified)**.** In the chemical industry, concentrations are continuously measured during a production process. We get the following 20 measurement values:

i	x_i	i	x_i
1	102.0	11	101.3
2	94.8	12	98.7
3	98.3	13	101.1
4	98.4	14	101.4
5	102.0	15	101.0
6	98.5	16	101.7
7	99.0	17	102.3
8	97.7	18	102.4
9	100.0	19	101.2
10	98.1	20	101.0

In this example, $n_1 = \cdots = n_{20} = 1$ and thus $\bar{x}_i = x_i$ for all $i \in \{1, \ldots, 20\}$. The target value of the process is $\mu_0 = 99$. First, we read the data with R.

```
#   read data
> file <- "concentration.dat"
> data <- read.table(file, header=TRUE)
> str(data)
'data.frame':   20 obs. of  1 variable:
 $ x: num  102 94.8 98.3 98.4 102 98.5 99 97.7 100 98.1 ...
#   target value
> mu0 <- 99
```

Next, we estimate the process standard deviation using the standard deviation of the data[2],

```
#   process standard deviation
> sigma.hat <- sd(data$x);  sigma.hat
[1] 2.006955
```

and calculate from that the reference value and the decision interval

```
#   reference value
> K <- 0.5 * sigma.hat;   K
[1] 1.003478
#   decision interval
> H <- 5 * sigma.hat; H
[1] 10.03478
```

In our case we have the following recursions for the CUSUMs

$$\begin{aligned} C_i^+ &= \max\{0, x_i - 100 + C_{i-1}^+\} \quad \text{with } C_0^+ = 0, \\ C_i^- &= \max\{0, 98 - x_i + C_{i-1}^-\} \quad \text{with } C_0^- = 0. \end{aligned}$$

This yields the following cumulative sums for $i = 1$

$$\begin{aligned} C_1^+ &= \max\{0, 102.0 - 100 + 0\} = 2.0, \\ C_1^- &= \max\{0, 98 - 102.0 + 0\} = 0.0 \end{aligned}$$

and for $i = 2$

$$\begin{aligned} C_2^+ &= \max\{0, 94.8 - 100 + 2.0\} = 0.0, \\ C_2^- &= \max\{0, 98 - 94.8 + 0.0\} = 3.2 \end{aligned}$$

and so on. The following R code with a `for` loop[3] is used to recursively calculate the two statistics.

```
#   C^+ and C^- CUSUMs
> C.plus <- NULL
> C.plus[1] <- 0
> C.mius <- NULL
> C.mius[1] <- 0
> for(i in 1:20) {
+     C.plus[i+1] <- max(0, data$x[i] - (mu0+K) + C.plus[i])
+     C.mius[i+1] <- max(0, (mu0-K) - data$x[i] + C.mius[i])
+ }
> C.plus
 [1]  0.000000  1.996522  0.000000  0.000000  0.000000  1.996522  0.493045
 [8]  0.000000  0.000000  0.000000  0.000000  1.296522  0.000000  1.096522
[15]  2.493045  3.489567  5.186090  7.482612  9.879135 11.075657 12.072180
> C.mius
 [1] 0.0000000 0.0000000 3.1965225 2.8930450 2.4895675 0.0000000 0.0000000
 [8] 0.0000000 0.2965225 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
[15] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
```

[2] For $n = 1$ we have no correction factor c_4.

[3] Since R works consistently with matrices, `for` loops are rarely used. It is in general possible to replace `for` loops with more sophisticated techniques, eg. the command `apply`. However, the readability of the code often suffers clarity.

Now we can draw the CUSUM chart.

```
#   CUSUM-Karte
> plot(0:20, C.plus, col="black", pch=20, ylim=c(-12,12),
+      xlab="Index", ylab="Cumulative Sums", main="CUSUM")
> lines(0:20, C.plus, col="black")
> points(0:20, -C.mius, pch=20, col="gray")
> lines(0:20, -C.mius, col="gray")
> abline(h=c(-1,0,1)*H, lty=c(2,1,2))
> text(rep(2,3), c(-1,1)*H, label=c("Lower CUSUM", "Upper CUSUM"), pos=3)
```

We note that the upper CUSUMs in this example exceed the decision interval, cf. Fig. 4.1.i, that is, the process is not under control. It drifts upwards away from the target value.

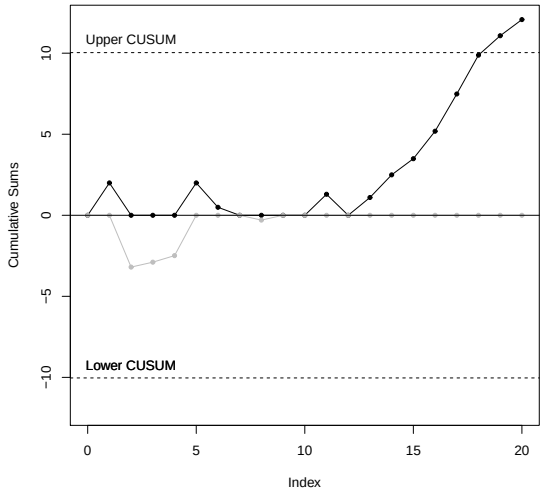


Figure 4.1.i: CUSUM. Starting from measurement index $i = 19$ the process is out of control.

A comparison of CUSUM charts with classic Shewhart charts shows that, given a fixed sample size n, the average run length ARL of the CUSUM chart is significantly shorter than with a $\bar{x}$ chart, especially for small

$$|\delta| = \frac{|\mu_1 - \mu_0|}{\sigma}$$

and not too small error probability α. CUSUM charts are especially recommended for processes where small disturbances are critical and should be detected fast.

4.2 EWMA

Again, let $\bar{x}_i$ be the mean of the i-th sample of a process. We assume that all samples have the same sample size n, i.e. $n_1 = \cdots = n_k = n$. Let μ_0 be the **target value** for the quality characteristic X and λ a constant with $0 \leq \lambda \leq 1$.
If we put the weights in Eq. (4.0.a) as follows

$$\begin{aligned} \alpha_i &= (1-\lambda)^i \mu_0, \\ \beta_j &= \lambda(1-\lambda)^{i-j} \quad \text{with} \quad j \in \{1, 2, \ldots, i\}, \end{aligned}$$

then we get

$$y_i = (1-\lambda)^i \mu_0 + \lambda \sum_{j=1}^{i} (1-\lambda)^{i-j} \bar{x}_j. \tag{4.2.a}$$

This is the so-called **exponentially weighted moving average control chart**, short **EWMA** chart to control means. The name follows from the fact that the weights β_j decay exponentially. The **smoothing parameter** λ determines the influence of the previous sample mean $\bar{x}_j$ on the statistic y_i.
The smaller λ, the more values $\bar{x}_j$ are used for the decision. The EWMA chart thus has an non-constant, unlimited memory. For $\lambda = 1$ we get the well-known Shewhart $\bar{x}$ chart (cf. Chap. 2.2). As with the CUSUM chart, we can calculate the statistic y_i recursively

$$y_i = (1-\lambda) y_{i-1} + \lambda \bar{x}_i \tag{4.2.b}$$

starting with $y_0 = \mu_0$. In the following, we want to show how control limits can be calculated for the EWMA chart.
If the process is under control, then we assume that $\bar{x}_i$ comes from a normal distribution with the expected value μ_0 and the standard deviation $\frac{\sigma}{\sqrt{n}}$. We assume that the standard deviation is either known or can be estimated from data. The statistic y_i is then also normally distributed with

$$\mathrm{E}(y_i) = \mu_0 \quad \text{and} \quad \mathrm{Var}(y_i) = \frac{\lambda}{2-\lambda} \left(1 - (1-\lambda)^{2i}\right) \frac{\sigma^2}{n}.$$

This results in three-σ control limits

$$\mathrm{UCL}_i = \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}\left(1-(1-\lambda)^{2i}\right)}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mathrm{LCL}_i = \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}\left(1-(1-\lambda)^{2i}\right)}\frac{\sigma}{\sqrt{n}}.$$

Note that the limits depend on the measurement index i. Already for rather small i we get the asymptotic control limits

$$\mathrm{UCL} = \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mathrm{LCL} = \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}},$$

which still depend on the target value μ_0, the sample size n, the process standard deviation σ and the smoothing parameter λ.
The process standard deviation can be estimated with $\hat{\sigma} = \frac{\bar{s}}{c_4}$, where $\bar{s}$ is the mean of the standard deviations of the individual samples, and the constant c_4 depends only on the sample size n and can be found in Tab. T.15.

Example 4.2.1. We examine again the data from the Ex. 4.1.1. For the sake of completeness, we re-read the data and redefine the target value.

```
#   read data
> file <- "concentration.dat"
> data <- read.table(file, header=TRUE)
> str(data)
'data.frame':   20 obs. of  1 variable:
 $ x: num  102 94.8 98.3 98.4 102 98.5 99 97.7 100 98.1 ...
#   sample size
```

```
> n <- ncol(data);  n
[1] 1
#   target value
> mu0 <- 99.0
```

Next we estimate the process standard deviation using the standard deviation[4].

```
#   process standard deviation
> sigma.hat <- sd(data$x);  sigma.hat
[1] 2.006955
```

The EWMA should have a long memory, so we set the smoothing parameter $\lambda = 0.1$.

```
#   smoothing parameter
> lambda <- 0.1
```

In this case we have the following recursion for the EWMA

$$\begin{aligned} y_0 &= 99.0, \\ y_1 &= (1-0.1)\cdot 99.0 + 0.1 \cdot 102.0 = 99.30, \\ y_2 &= (1-0.1)\cdot 99.3 + 0.1 \cdot 94.8 = 98.85, \\ y_3 &= \cdots \end{aligned}$$

The following code in R with a `for` loop will do the job.

```
#   recursion
> y <- NULL
> y[1] <- mu0
> for(i in 1:20) {
+     y[i+1] <- (1-lambda) * y[i] + lambda * data$x[i]
+ }
> y
 [1]  99.00000  99.30000  98.85000  98.79500  98.75550  99.07995  99.02196
 [8]  99.01976  98.88778  98.99901  98.90910  99.14819  99.10337  99.30304
[15]  99.51273  99.66146  99.86531 100.10878 100.33790 100.42411 100.48170
```

Now we also have to calculate the control limits.

```
#   control limits
> sigma.pro <- sigma.hat/sqrt(n) *
+              sqrt(lambda/(2-lambda) * (1-(1-lambda)^(2*(0:20))))
> UCL <- mu0 + 3*sigma.pro;  UCL
 [1]  99.00000  99.60209  99.81002  99.94551 100.04242 100.11476 100.17008
 [8] 100.21304 100.24675 100.27341 100.29460 100.31151 100.32505 100.33591
[15] 100.34465 100.35169 100.35736 100.36194 100.36563 100.36862 100.37103
> LCL <- mu0 - 3*sigma.pro;  LCL
 [1] 99.00000 98.39791 98.18998 98.05449 97.95758 97.88524 97.82992 97.78696
 [9] 97.75325 97.72659 97.70540 97.68849 97.67495 97.66409 97.65535 97.64831
[17] 97.64264 97.63806 97.63437 97.63138 97.62897
```

[4]For $n = 1$ we have no correction factor c_4. Therefore, we simply set the process standard deviation equal to the standard deviation of the data.

Now we can plot the EWMA chart.

```
#   EWMA chart
> plot(0:20, y, pch=20, ylim=c(-1,1)*1.6+mu0,
+      xlab="Index", ylab="EWMA", main="EWMA")
> lines(0:20, y)
#   add control limits
> abline(h=mu0)
> lines(0:20, UCL, lty=2, type="S")
> lines(0:20, LCL, lty=2, type="S")
> text(rep(20,2), c(LCL[21],UCL[21]), label=c("LCL", "UCL"), pos=1)
```

We observe that the EWMA exceeds the control limits at the very end, i.e. the process is not under control, cf. Fig. 4.2.i and 4.2.ii.

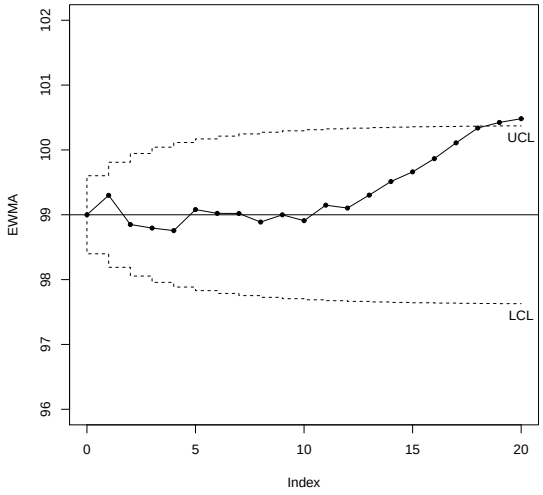


Figure 4.2.i: EWMA with smoothing parameter $\lambda = 0.1$ (long memory). The process gets out of control.

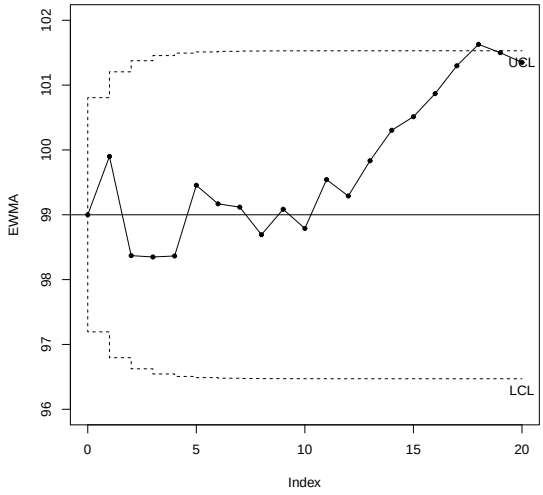


Figure 4.2.ii: EWMA with smoothing parameter $\lambda = 0.3$ (short memory). The process gets out of control.

A comparison of EWMA charts with classic Shewhart charts shows that, given a fixed sample size n, the average run length ARL of the EWMA chart is significantly shorter than with a $\bar{x}$ chart, especially for small

$$|\delta| = \frac{|\mu_1 - \mu_0|}{\sigma}$$

and not too small error probability α. EWMA charts are especially recommended for processes where small disturbances are critical and should be detected fast.

4.3 Exercises

Problem 4.3.1 (**CUSUM of Individual Measurements**)**.** The diameter of different holes with the same nominal size is regularly measured automatically. You can find the data in `diameter.dat` (cf. Prob. 2.3.1). Examine if the process is under control with a CUSUM chart. The target value is $\mu_0 = 10.0$.

Problem 4.3.2 (**CUSUM of Random Samples**)**.** A quality inspector in a beverage factory is responsible for the accuracy of the bottling machine of a soft drink. She collected 25 random samples consisting of four measurements of fill level [in cm], see data set `soft-drinks.dat`. Examine if the process is under control with a CUSUM chart. The target value is $\mu_0 = 16.0$. You can estimate the process standard deviation with $\hat{\sigma} = \frac{\bar{s}}{c_4}$, where $\bar{s}$ is the mean of the standard deviations of the individual samples, and the constant c_4 depends only on the sample size n and can be found in Tab. T.15.

Problem 4.3.3 (**Monitor Simulated Process with CUSUM**)**.** Simulate $k = 50$ random samples of size $n = 8$ each from a normal distribution with target value $\mu_0 = 10.0$ and standard deviation $\sigma = 2.0$. Use the command `rnorm`. Create a data set consisting of k rows and n columns with `data.frame`.

a. Create a CUSUM chart of the simulated data. You can estimate the process standard deviation from the simulated data.

b. Now add a synthetic process drift of $\Delta\mu = 1.4$ from sample $i = 21$ to the data and look at the resulting CUSUM chart. Use `data.mod <- rbind(data[1:20,1:n], data[21:50,1:n]+1.4)` to modify the simulated data.

c. From time $i = 20$ plot the expected line on which the process drifts away.

d. Play with the parameters of your simulation, that is, vary $\Delta\mu$.

Problem 4.3.4 (**EWMA of Random Samples**)**.** A quality inspector in a beverage factory is responsible for the accuracy of the bottling machine of a soft drink. She collected 25 random samples consisting of four measurements of fill level [in cm], see data set `soft-drinks.dat`. Examine if the process is under control with a EWMA chart. The target value is $\mu_0 = 16.0$. You can estimate the process standard deviation with $\hat{\sigma} = \frac{\bar{s}}{c_4}$, where $\bar{s}$ is the mean of the standard deviations of the individual samples, and the constant c_4 depends only on the sample size n and can be found in Tab. T.15. For λ, choose a long memory.

Problem 4.3.5 (**Monitor Simulated Process with EWMA**)**.** Simulate $k = 50$ random samples of size $n = 8$ each from a normal distribution with target value $\mu_0 = 10.0$ and standard deviation $\sigma = 2.0$. Use the command `rnorm`. Create a data set consisting of k rows and n columns with `data.frame`.

a. Create an EWMA chart of the simulated data. You can estimate the process standard deviation from the simulated data.

b. Now add a synthetic process drift of $\Delta\mu = 1.4$ from sample $i = 21$ to the data and look at the resulting EWMA chart. Use `data.mod <- rbind(data[1:20,1:n], data[21:50,1:n]+1.4)` to modify the simulated data.

c. Play with the parameters of your simulation, that is, vary the length of the memory λ and $\Delta\mu$.

d. Compare with the CUSUM simulation in Prob. 4.3.3.

Problem 4.3.6 (**Theory of EWMA**)**.** Prove that the recursive definition in Eq. (4.2.b) leads to the statistic in Eq. (4.2.a).

Solution

Solution 4.3.6. We start the recursion with $y_0 = \mu_0$ and simply put it in the recursive equation

$$y_i = (1 - \lambda)y_{i-1} + \lambda\bar{x}_i$$

and get the explicit form

$$y_i = (1 - \lambda)^i\mu_0 + \lambda\sum_{j=1}^{i}(1 - \lambda)^{i-j}\bar{x}_j.$$

Just be careful when sorting and multiplying the terms.

Chapter 5

Acceptance Sampling

Acceptance sampling is an important topic in statistical quality control. It was applied for the first time during World War II by the US military in quality control of bullets. The problem was that not all bullets can be tested in advance, otherwise there will be no more left[1]. On the other hand, if no bullet is tested, it can cause a malfunction during the battle.
In order to meet both needs, H. Dodge and H. Romig had the idea to sample some bullets, test them and make a decision on the quality of all the bullets based on the information of the sample.
It is therefore a compromise between no control and 100% control. In general, the sample can either be accepted (*go*) or rejected (*no-go*). Such a procedure will be called **acceptance sampling**.
Acceptance sampling is most likely to be useful when

- testing is destructive,
- the cost of a 100% control would be too large, or
- a 100% control would take too long.

Acceptance control is based on **acceptance sampling plans**, which contain instructions based on which the acceptance or return of a lot is decided. We should remember that the actual point of acceptance control is not to value quality, but to decide whether or not delivery will likely pass quality control. So with acceptance sampling, we cannot intervene directly in the production process to improve quality, but we can only decide whether to accept the delivery or not.
According to DIN 55350, Teil 31, a **lot** is defined as the bulk of a product that has been made under conditions that are considered to be consistent.
In analogy to process control we also distinguish in acceptance sampling classification by variables and attributes. Variables are quality characteristics that are measured on a numerical scale. Attributes are quality characteristics that are expressed on a *go*, *no-go* basis.
In the following we discuss lot-by-lot acceptance sampling plans for attributes.

[1] Actually a pity that this is not the case.

5.1 Acceptance Sampling Plans for Attributes

We have a lot of N units. The number of defective items m with $0 \leq m \leq N$ is observed. Thus, the **percent defective** is

$$p = \frac{m}{N}.$$

In reality, however, m and thus the percent defective p are unknown.
The aim of an **acceptance sampling plan** is to estimate the percent defective p and to decide whether we accept the lot or not with the help of a random sample of size $n \leq N$. Denote x the number of defective units in the sample with size n. An acceptance sampling plan consists of

- the **sample size** n,
- the **acceptance number** c, where $0 \leq c < n$, and
- a **rule**: If $x \leq c$, then the lot is accepted. If $x > c$, then the lot is rejected and may be returned to the producer.

The sample size n and the acceptance number c are the parameters of the acceptance sampling plan. These are quality requirements that are put on the lot. The quality requirements are agreed between the consumer and the producer. The consumer and the producer, however, have opposing interests.

- The **producer** is interested in accepting a lot with a small percent defective p based on a sample taken from the lot.
- On the other hand, the **consumer** wants to make sure that lots with a large percent defective p are rejected during the acceptance sampling.

The rule of the acceptance sampling plan corresponds to alternative hypotheses of a statistical test.

$$H_0 : x \leq c, \text{ i.e. the lot is accepted.}$$
$$H_1 : x > c, \text{ i.e. the lot is rejected.}$$

As with any statistical test, two false decisions are possible:

(a) If the consumer returns a good lot with a small percent defective p on the basis of a random sample a **type I error** occurs.
The probability for this false decision is called **producer risk** and is denoted by α.

(b) If the consumer accepts a bad lot with a large percent defective p on the basis of a random sample a **type II error** occurs.
The probability for this false decision is called **consumer risk** and is denoted by β.

In practice, we typically have a probability of acceptance p_α defined by the producer and a probability of rejection p_β defined by the consumer. There must be $p_\alpha < p_\beta$. If possible, the producer does not want to take back lots with $p < p_\alpha$. On the other hand, the consumer wants to reject lots with $p_\beta < p$ whenever possible.
An effective acceptance sampling plan must be designed so that the producer risk α and the consumer risk β are as small as possible. This means that no false decisions are made.
In practice, the two probabilities α and β are specified in acceptance sampling plans. They are chosen so that the consequences of further processing defective parts is justifiable.

5.1.1 Operating Characteristic

The **operating characteristic** of an acceptance sampling plan relates the four variables α, β and p_α, p_β, which define the quality agreements between producer and consumer with the two parameters of an acceptance sampling plan, that is, the sample size n and the acceptance number c.
The **operating characteristic**, short OC, is the probability of accepting a lot as a function of the percent defective p. It is called **acceptance probability** and is a function

$$\mathrm{OC} : [0,1] \to [0,1].$$

With a good acceptance sampling plan, the larger the percent defective p, the smaller OC should be. If the lot has no defective parts, i.e. $p = 0$, then the lot should be accepted, thus $\mathrm{OC}(0) = 1$. On the other hand, if the lot consists only of defective parts, i.e. $p = 1$, then the lot should be returned, thus $\mathrm{OC}(1) = 0$.
Therefore **OC curves** have the form shown in Figs. 5.1.ii and 5.1.iii.

- Upon control of all units of the lot, i.e. $n = N$, the percent defective p can be determined exactly. This makes it possible to specify the form of the OC curve. If $p^\star$ is the allowable percent defective, all lots are accepted if $p \le p^\star$ and all are returned if $p^\star < p$. Therefore we obtain the following OC curve

$$\mathrm{OC}(p) = \begin{cases} 1 & \text{if } p \le p^\star, \\ 0 & \text{if } p^\star < p. \end{cases}$$

  This **ideal** OC **curve** is shown in Fig. 5.1.i.

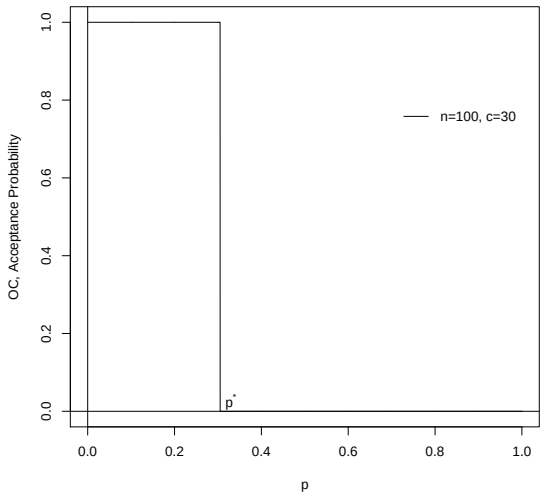


Figure 5.1.i: OC curve of an ideal acceptance sampling plan.

- Normally, however, a complete control is not possible and so we need to draw a sample of sample size n from the lot of size N. If the number x of defective parts is at most equal to the acceptance number c, then the lot is accepted.

The number of defective units X is a hypergeometric random variable (cf. Prob. 5.3.1 and 5.3.2). There are N parts in total, among which m defective and $N-m$ are good parts. So a sample of size n is drawn at once. The probability that among these n parts are at most c parts defective is given by

$$\mathrm{OC}(p) = P(X \le c) = \sum_{k=0}^{c} P(X=k) = \sum_{k=0}^{c} \frac{\binom{pN}{k}\binom{N-pN}{n-k}}{\binom{N}{n}}. \tag{5.1.a}$$

Note that $m = pN$ denotes the number of defective parts and $N - pN$ the number of good parts.

For a given lot size N, the parameters n and c of the acceptance sampling plan determine the form of the OC curve, cf. Fig. 5.1.ii and 5.1.iii. An acceptance sampling plan is more effective the steeper the OC curve, i.e. the more the OC curve equals an ideal OC curve.

Example 5.1.1. We want to draw some OC curves with R. We choose a rather small lot of size $N = 100$.

```
#   lot size
> N <- 100
```

Then we define the operating characteristic as a function.

```
#   define operating characteristic
#       p: percent defective
#       N: lot size
#       n: sample size
#       c: acceptance number
> func.OC <- function(p,N,n,c) { phyper(q=c,m=p*N,n=(1-p)*N,k=n) }
```

In a first step, we plot the ideal OC curve for a 100% control, i.e. $n = N = 100$. The acceptance number $c = 30$ is arbitrary selected, i.e. $p^\star = \frac{c}{N} = 0.3$.

```
#   ideal operating characteristic
> curve(func.OC(p=x,N,n=100,c=30), from=0, to=1, n=1001,
+       xlab="p", ylab="OC, Acceptance Probability",
+       main="Ideal Operating Characteristic Curve")
> abline(h=0, v=0)
#   add legend
> legend(0.7,0.8, legend=c("n=100, c=30"), lty=1, bty="n")
> text(0.3,0, pos=1, offset=0.1, labels=expression(paste("p"^"*")))
```

Then we draw OC curves for a constant sample size $n = 40$ and varying acceptance number c, cf. Fig. 5.1.ii.

```
#   OC curves: n=40 constant, c variable
> curve(func.OC(p=x,N,n=40,c= 2), from=0, to=1, n=10001, lty=2,
+       xlab="p", ylab="OC, Acceptance Probability",
+       main="Operating Characteristic Curves")
> curve(func.OC(p=x,N,n=40,c= 5), from=0, to=1, n=10001, lty=3, add=TRUE)
> curve(func.OC(p=x,N,n=40,c=10), from=0, to=1, n=10001, lty=4, add=TRUE)
> curve(func.OC(p=x,N,n=40,c=20), from=0, to=1, n=10001, lty=5, add=TRUE)
> abline(h=0, v=0)
```

```
#   add legend
> legend(0.7,0.8, legend=c("n=40, c= 2",
+                          "n=40, c= 5",
+                          "n=40, c=10",
+                          "n=40, c=20"),
+        lty=2:5, bty="n")
```

Then we plot OC curves for a varying sample size n and constant acceptance number $c = 10$, cf. Fig. 5.1.iii.

```
#   OC curves: n variable, c=10 constant
> curve(func.OC(p=x,N,n=20,c=10), from=0, to=1, n=10001, lty=2,
+       xlab="p", ylab="OC, Acceptance Probability",
+       main="Operating Characteristic Curves")
> curve(func.OC(p=x,N,n=30,c=10), from=0, to=1, n=10001, lty=3, add=TRUE)
> curve(func.OC(p=x,N,n=50,c=10), from=0, to=1, n=10001, lty=4, add=TRUE)
> curve(func.OC(p=x,N,n=80,c=10), from=0, to=1, n=10001, lty=5, add=TRUE)
> abline(h=0, v=0)
#   add legend
> legend(0.7,0.8, legend=c("n=20, c=10",
+                          "n=30, c=10",
+                          "n=50, c=10",
+                          "n=80, c=10"),
+       lty=2:5, bty="n")
```

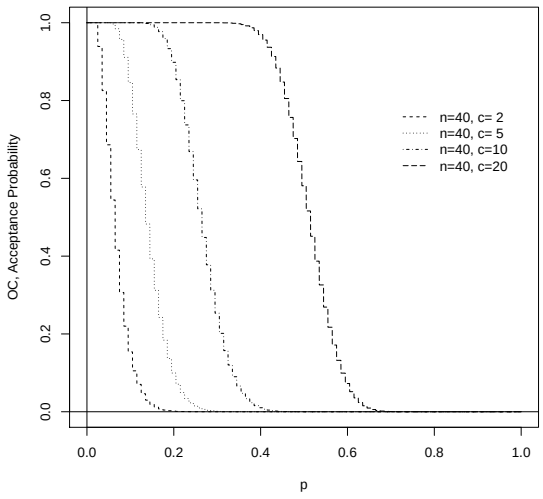


Figure 5.1.ii: Four different OC curves for constant sample size $n = 40$ and varying acceptance number $c = 2$, 5, 10 and 20.

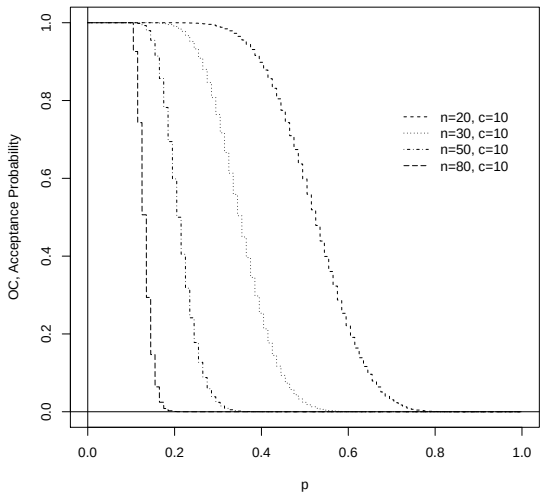


Figure 5.1.iii: Four different OC curves for constant acceptance number $c = 10$ and varying sample size $n = 20$, 30, 50 and 80.

There are two relations between the possible false decisions, type I and II errors, and the operating characteristic.
First, since α is the probability to reject a lot with the probability of acceptance p_α and hence $1 - \alpha$ is the probability of accepting the lot, we have

$$\mathrm{OC}(p_\alpha) = 1 - \alpha \tag{5.1.b}$$

and second, we have

$$\mathrm{OC}(p_\beta) = \beta, \tag{5.1.c}$$

where $p_\alpha < p_\beta$ is assumed, cf. Fig. 5.1.iv.

5.1.2 Parameters of an Acceptance Sampling Plan

In the following we assume that the producer and the consumer agree on the probability of acceptance p_α and the producer risk α on the one hand, and on the probability of rejection p_β and the consumer risk β on the other . The pair $(p_\alpha, 1-\alpha)$ is called **producer risk point** and the pair (p_β, β) is called **consumer risk point**. These two points ar given on an OC curve with the unknown parameters n and c.
Our goal now is to determine the parameters n and c of the acceptance sampling plan. The sample size n and the acceptance number c must both be integers with $0 \leq c < n \leq N$. However, this usually means that Eqs. (5.1.b) and (5.1.c) cannot be fulfilled perfectly. We set up slightly weaker conditions:

- the smallest possible sample size n,
- the producer risk is at most equal to α, i.e. $\mathrm{OC}(p_\alpha) \geq 1-\alpha$, and
- the consumer risk is at most equal to β, i.e. $\mathrm{OC}(p_\beta) \leq \beta$.

Unfortunately, the solution of the two inequalities with the two unknowns n and c is not explicitly possible. In the following example we give a numerical brute-force solution.

Example 5.1.2. We consider the acceptance sampling of bolts. The quality feature is met if the bolts fit through a standardised hole, otherwise they are rejected. The lot size is $N = 100$. We want to find an acceptance sampling plan with the following specifications agreed on by the producer and the consumer:

- The probability of acceptance is $p_\alpha = 0.05$ and the producer risk is $\alpha = 0.2$.
- The probability of rejection is $p_\beta = 0.1$ and the consumer risk is $\beta = 0.1$.

Enter in R.

```
#   producer risk
> alpha <- 0.2
#   consumer risk
> beta <- 0.1
#   probability of acceptance
> p.alpha <- 0.05
#   probability of rejection
> p.beta <- 0.1
```

Thus we have to solve the two inequalities

$$\mathrm{OC}(0.05) \geq 1 - 0.2 = 0.8, \tag{5.1.d}$$

$$\mathrm{OC}(0.1) \leq 0.1 \tag{5.1.e}$$

numerically, under the condition that n is minimal. First, we define the operating characteristic as a function.

```
#   define operating characteristic
#       p: percent defective
#       N: lot size
#       n: sample size
#       c: acceptance number
> func.OC <- function(p,N,n,c) { phyper(q=c,m=p*N,n=(1-p)*N,k=n) }
```

I cannot think of anything more clever than to test the two inequalities (5.1.d) and (5.1.e) for all useful combinations $n \in \{1, \ldots, 100\}$ and $c \in \{1, \ldots, n-1\}$ starting with $n = 1$ and then to choose the first pair $(n_{\text{good}}, c_{\text{good}})$ which fulfills both inequalities.

```
#   brute-force solution: test all useful combinations
> n.good <- NULL
> c.good <- NULL
> for(ni in 1:N) {
+     for(ci in 1:(ni-1)) {
+         a.flag <- func.OC(p=p.alpha,N,n=ni,c=ci) >= 1-alpha
+         b.flag <- func.OC(p=p.beta, N,n=ni,c=ci) <= beta
+         if(a.flag & b.flag){
+             n.good <- c(n.good, ni)
+             c.good <- c(c.good, ci)
+         }
+     }
+ }
#   choose smallest sample size
> n <- n.good[1]; n
[1] 64
#   choose acceptance number
> c <- c.good[1]; c
[1] 4
```

Of course, I admit that this is only feasible for relatively small lot sizes, say $N < 20\,000$.

```
#   graphics
> curve(func.OC(p=x,N,n=n,c=c), from=0, to=0.2, n=10001,
+       xlab="p", ylab="OC, Acceptance Probability",
+       main="Operating Characteristic")
> abline(h=0, v=0)
> lines(c(0,p.alpha,p.alpha), c(1-alpha,1-alpha,0), lty=2)
> lines(c(0,p.beta,p.beta), c(beta,beta,0), lty=2)
> points(c(p.alpha, p.beta), c(1-alpha,beta), pch=20, cex=2)
#   add legend
> text(c(0,0,p.alpha-0.005,p.beta-0.005), c(1-alpha+0.02,beta+0.02,0,0),
+      pos=c(4,4,3,3), labels=c(expression(1-alpha),expression(beta),
+                               expression(p[alpha]),expression(p[beta])))
```

We want to control how good our solution really is.

```
#   real producer risk
1-func.OC(p=p.alpha,N,n=n,c=c)
[1] 0.1012719
```

The real producer risk for our acceptance sampling plan with $n = 64$ and $c = 4$ is thus significantly smaller than the targeted $\alpha = 0.2$.

```
#   real consumer risk
> func.OC(p=p.beta,N,n=n,c=c)
[1] 0.09547271
```

The real consumer risk for our acceptance sampling plan with $n = 64$ and $c = 4$ is thus slightly smaller than the targeted $\beta = 0.1$. All in all – a good, usable acceptance sampling plan.
It works, Fig 5.1.iv shows it impressively!

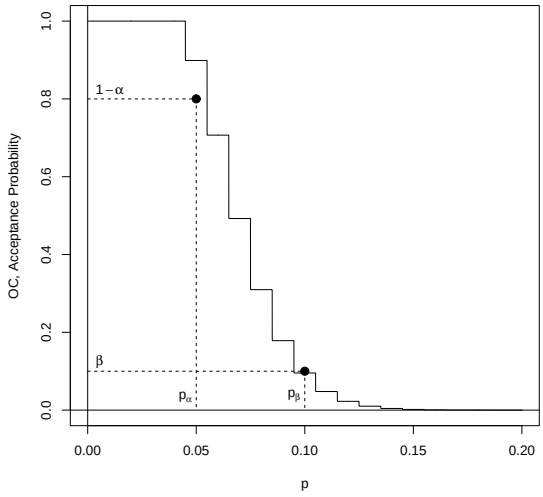


Figure 5.1.iv: OC curve of a real acceptance sampling plan.

Of course, it immediately raises the question of whether finding the parameters is not also easier. Everything is simple with R: there is an `AcceptanceSampling` package. We are anticipating from Chap. 6.3. After having installed the package `AcceptanceSampling`, we load it in R.

```
#   load package
> require(AcceptanceSampling)
```

The calculation of the parameters is now easy.

```
#   producer risk
> alpha <- 0.2
#   consumer risk
> beta <- 0.1
#   probability of acceptance
> p.alpha <- 0.05
#   probability of rejection
> p.beta <- 0.1

#   Utility function for finding sampling plans
find.plan(PRP=c(p.alpha, 1-alpha), CRP=c(p.beta, beta), type="hypergeom", N=100)
$n
[1] 64
$c
```

```
[1] 4
$r
[1] 5
```

This even gives the same result! Not bad? Only problem: What exactly does the command `find.plan` do? Find it out!

In practice lots of acceptance sampling plans have been published. Eg. for acceptance sampling plans for attributes consult DIN ISO 2859, *Annahmestichprobenprüfung anhand der Anzahl fehlerhaften Einheiten oder Fehler (Attributprüfung).*

5.2 Acceptance Sampling Plans for Variables

Of course, we can test the quality of a lot by measuring a specific quality characteristic X. For the characteristic X there are two tolerance limits USL and LSL.
Then we have an analogue theory as in Chap. 5.1. We would assume that the characteristic X is eg. normally distributed. But we do not want to do this here, because it is always possible to reduce an acceptance sampling plan for variables to an acceptance sampling plan for attributes by saying:

- If $\text{LSL} \leq x \leq \text{USL}$, then the part is fit for use.
- If $x < \text{LSL}$ oder $\text{USL} < x$, then the part is rejected.

By counting the number of rejected parts, we again have an acceptance sampling plan for attributes and are back in Chap. 5.1.
For acceptance sampling plans for variables I refer to the literature, eg. [23], [38] or consult DIN ISO 3951, *Verfahren für die Stichprobenprüfung anhand quantitativer Merkmale (Variablenprüfung).*

5.3 Exercises

Problem 5.3.1 (**Hypergeometric Distribution I**)**.** Four out of ten lots win. Determine the probability that out of five randomly selected lots there will be

a. exactly two winning lots.

b. a maximum of three winning lots.

c. more than one winning lot.

Problem 5.3.2 (**Hypergeometric Distribution II**)**.** There are $N = m+n$ balls in an urn, of which m are red and n balls are black. We draw k balls at once. Determine the probability that there are exactly x red balls among the k balls.

Problem 5.3.3 (**Acceptance Sampling Plan**, cf. [38], Beispiel 22-2)**.** In acceptance sampling lots with $N = 11\,000$ bolts are tested. Lots with less than 1% of bad bolts are rejected in 10% of all cases, though they would have to be accepted as fit for use. On the other hand, lots with more than 5% of bad bolts are accepted at most in 5% of all cases, though their quality does not meet the terms of delivery.

a. What is the producer risk α and the consumer risk β?

b. What is the probability of acceptance p_α and a probability of rejection p_β?

c. Find the parameters of the acceptance sampling plan, i.e. the sample size n and the acceptance number c.

Solutions

Solution 5.3.1. The random variable X counts the number of winning lots amongst $k = 5$ drawn lots. Therefore we obtain

a. $P(X = 2) =$ `dhyper(x=2, m=4, n=10-4, k=5)` $= 0.4761905$.

b. $P(X \leq 3) =$ `phyper(q=3, m=4, n=10-4, k=5)` $= 0.9761905$.

c. $P(X > 1) =$ `phyper(q=1, m=4, n=10-4, k=5, lower.tail=FALSE)` $= 0.7380952$.

With `help(dhyper)` you get more information about implementation of the hypergeometric distribution in R.

Solution 5.3.2. There are $\binom{m+n}{k}$ possibilities in total to arrange k balls among $m+n$ balls, and exactly $\binom{m}{x}\binom{n}{k-x}$ possibilities to arrange x red balls among k balls. The factor $\binom{m}{x}$ counts the number of arrangements with x red balls among m red balls, and the factor $\binom{n}{k-x}$ counts the number of arrangements with $k-x$ black balls among n black balls. Therefore we obtain

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}.$$

Chapter 6

R-Packages used in Statistical Process Control

It is clear that there are many computer programs for statistical quality control. In this course we use the free software environment R for statistical computing and graphics, cf. `http://cran.ch.r-project.org/`. In addition to the built-in functions, we can install more packages, eg. the package for statistical quality control `qcc` von L. Scrucca, cf. [33] and [34] or the package for acceptance sampling `AcceptanceSampling` from A. Kirweiler, cf. [13].

6.1 Package `qcc`

In the following I refer to the paper [33], which describes the package perfectly. Also note the manual [34]. You can find some worked examples in `qcc.R` packed in the distributed zip-file.

6.2 Exercises

Problem 6.2.1 (**Installation**)**.** Install the package `qcc` in R, cf. [34]. Load the package with `require(qcc)`.

Problem 6.2.2 (**Shewhart Control Charts with** `qcc`)**.** A quality inspector in a beverage factory is responsible for the accuracy of the filling machine of a soft drink. She has collected 25 samples consisting of four measurements of fill level [in cm], cf. data set `soft-drinks.dat`. Use the first 20 measured values as a trial run to make the following charts and check if subsequent measurements are under control.

a. Chart for $\bar{x}$.

b. Chart for R.

c. Chart for s.

d. Compare to the charts in Prob. 2.5.2.

Problem 6.2.3 (**Group Data, Shewhart Control Chart with** `qcc`, cf. [23], Ex. 15-91)**.** The diameter of fuse pins used in an aircraft engine application is an important quality characteristic. Twenty-five samples of three pins each were measured, cf. the data set `fuse-pins.dat`. The target value is $\mu_0 = 64.000\,\text{mm}$.

a. Display the data graphically, i.e. measurements, resp. means of measurements versus sample index.

b. Group the data so that it can be further processed with R and the package `qcc`.

c. Create charts for $\bar{x}$ and s.

Problem 6.2.4 (**CUSUM and EWMA with** `qcc`, cf. [23], Ex. 15-91)**.** The diameter of fuse pins used in an aircraft engine application is an important quality characteristic. Twenty-five samples of three pins each were measured, cf. the data set `fuse-pins.dat`. The target value is $\mu_0 = 64.000\,\text{mm}$.

a. Create a CUSUM chart.

b. Create a EWMA chart with $\lambda = 0.3$.

c. The trial run consists of the 25 measurements. Is the new sample 63.997, 63.991 and 64.004 under control?

Problem 6.2.5 (**OC with** `qcc`)**.** A quality inspector in a beverage factory is responsible for the accuracy of the filling machine of a soft drink. She has collected 25 samples consisting of four measurements of fill level [in cm], cf. data set `soft-drinks.dat`. Create operating characteristic curves for different sample sizes n from the stable trial run.

Problem 6.2.6 (**PCR with** `qcc`)**.** A quality inspector in a beverage factory is responsible for the accuracy of the filling machine of a soft drink. She has collected 25 samples consisting of four measurements of fill level [in cm], cf. data set `soft-drinks.dat`. Estimate the process capability index C_P from the stable trial run for tolerance limits USL = 18.1 cm and LSL = 15.6 cm. Study the graphics and try to understand the R output. The information in `help(process.capability)` might help.

6.3 Packages `AcceptanceSampling`

In the following I refer to the paper [13], which describes the package perfectly. Also note the manual [14]. You can find some worked examples in `AcceptanceSampling.R` packed in the distributed zip-file.

6.4 Exercises

Problem 6.4.1 (**Installation**)**.** Install the package `AcceptanceSampling` in R, cf. [14]. Load the package with `require(AcceptanceSampling)`.

Problem 6.4.2 (**Sampling Plan with** `AcceptanceSampling`, cf. [38], Beispiel 22-2)**.** In acceptance sampling lots with $N = 11\,000$ bolts are tested. Lots with less than 1% of bad bolts are rejected in 10% of all cases, though they would have to be accepted as fit for use. On the other hand, lots with more than 5% of bad bolts are accepted at most in 5% of all cases, though their quality does not meet the terms of delivery. Find the parameters of the acceptance sampling plan, i.e. the sample size n and the acceptance number c.

Part II

Multiple Regression

Chapter 7

Simple Linear Regression

One of the most important and widely used statistical technique is regression analysis. This is a statistical technique for investigating and modelling relationships between variables.
We find applications in almost every scientific field which deals with data and is based on quantitative statements, i.e. engineering, physical and chemical sciences, economics, management, life and biological sciences and social sciences.

7.1 Simple Linear Regression Model

A simple linear regression model is a model with a single predictor variable that has a relationship with a response that is a straight line. Let us look at a simple example, cf. [24].

Example 7.1.1. An industrial engineer has set himself the task to analyse the product delivery and the service operations for beverage vending machines in more detail.
He suspects that the time it takes a service person to load and service the vending machines primarily depends on the number of cases of product delivered. Therefore he measures the delivery time and the volume of product delivered at 25 randomly selected locations with beverage vending machines. Tab. 7.1.1 shows the delivery time and the volume of product for each of the 25 vending machines.

i	Time [min]	Volume [ ]	i	Time [min]	Volume [ ]
1	16.68	7	14	19.75	6
2	11.50	3	15	24.00	9
3	12.03	3	16	29.00	10
4	14.88	4	17	15.35	6
5	13.75	6	18	19.00	7
6	18.11	7	19	9.50	3
7	8.00	2	20	35.10	17
8	17.83	7	21	17.90	10
9	79.24	30	22	52.32	26
10	21.50	5	23	18.75	9
11	40.33	16	24	19.83	8
12	21.00	10	25	10.75	4
13	13.50	4			

First, we read the full data set in R, second we plot it.

```
#   read data
> file <- "vending-machines.dat"
> data <- read.table(file, header=TRUE)
> str(data)
'data.frame':   25 obs. of  4 variables:
 $ Time    : num  16.7 11.5 12 14.9 13.8 ...
 $ Volume  : int  7 3 3 4 6 7 2 7 30 5 ...
 $ Distance: int  560 220 340 80 150 330 110 210 1460 605 ...
 $ Town    : Factor w/ 4 levels "Austin","Boston",..: 4 4 4 4 4 4 4 2 2 2 ...

#   Scatter diagram: Time versus Volume
> plot(Time ~ Volume, data,
+      pch=20, xlim=c(0,30), ylim=c(0,80),
+      main="Delivery Time versus Delivery Volume")
> grid()
```

Notice that, tilde, i.e. the symbol ~, is used in R to separate the left- and right-hand sides in a model formula.
The delivery time and the delivery volume are shown in Fig. 7.1.i in a scatter diagram.

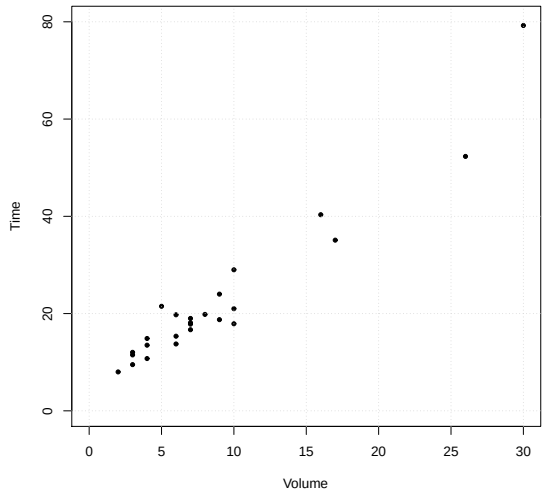


Figure 7.1.i: Delivery time versus delivery volume.

The graphic clearly indicates that there is a relationship between the delivery time and the delivery volume. The data points are not exactly on a straight line, but they scatter very close to one.

If in the above example we denote y the delivery time and x the delivery volume, then the equation of a straight line relating the two variables is

$$y = \beta_0 + \beta_1 x,$$

where β_0 is the intercept and β_1 the slope. Since the data points do not exactly lie on a line we have to modify the above model to account for this error. Let the difference between the observed value of y and the straight line $\beta_0 + \beta_1 x$ be an **error** ε. The error ε is a random

variable that accounts for the failure of the model to fit the data exactly. Therefore, the more plausible model is

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

The errors are assumed to have mean zero

$$\mathrm{E}(\varepsilon) = 0$$

and unknown variance

$$\mathrm{Var}(\varepsilon) = \sigma^2.$$

Furthermore we assume that the errors are uncorrelated.
The variable x is often called the independent variable and the variable y the dependent variable. Unfortunately this is misleading and causes confusion with the concept of statistical independence. Therefore we refer to x as the **explanatory** or **predictor variable** and y as the **response variable**. We call this model a **(simple) linear regression model**[1].

The regression equation is in real applications only an approximation to the true functional relationship between the variables of interest.[2] The functional relationships are often based on physical, chemical or other engineering or scientific theory, i.e. they are derived from first principles. We call these **mechanistic models**.
In many cases on the other hand, especially in economics, management, life and biological sciences and social sciences, these functional relationships are unknown and so we have to find **empirical models**.

Regression models are used for the following purposes:

1. **Description.** Scientists use equations to summarise or describe a set of data. For instance in Ex. 7.1.1 it is certainly possible to summarise efficiently the collected data with a straight line.

2. **Prediction, forecast, interpolation, extrapolation.** Many applications of regression involve prediction of the response variable. Often not only the average expected value is of interest, but also an upper and lower limit beyond which the true value might be only with a small known probability. Extrapolating response variables beyond the interval of predictor variables should be done with extreme care since the regression model is often only local approximation.

3. **Parameter estimation.** With mechanistic models the parameters in the model often have a physical meaning. For instance, Ohm's law states that the current I through a conductor between two points is directly proportional to the voltage U across the two points. The resistance R can be estimated from data as the slope of the simple regression model
$$U = R\,I + \varepsilon$$
with no intercept, where I denotes the predictor and U the response variable.

[1] It is a simple model because it only contains one predictor variable.
[2] It is always possible to approximate a pretty smooth process with a straight line, i.e. Taylor approximation of the first order.

4. **Determination of significant variables.** Fitting models with several predictor variables to data often raises the question of which variables have the largest influence on the response variable and which ones can be omitted, i.e. which variables are significant?

5. **Optimisation.** The search for optimal production conditions is often done with regression models. We will discuss response surfaces modelling and design of experiment in the third part of this course.

In order to apply regression models and answer questions in depth we must make sure that the selected model describes the data well enough. Checking the model adequacy is therefore the central topic in statistical modelling.

7.2 Estimation of the Parameters

First we want to think a bit deeper about parameter estimation:

1. We live and collect data in the **real world**. In the real world we calculate statistical quantities from a sample data set like the mean $\bar{x}$ or the standard deviation s, most often denoted with small Latin letters.

2. In contrast there is the **model world**. Our thoughts, models and hypotheses live in the model world. In the model world, for example, we believe that a process follows a normal distribution with the parameters μ and σ, most often denoted by Greek letters.

We do not know the model world or only incompletely. We can only approximate it with a finite number of measurements. Intuitively it is clear that the more measurements we make of it, the better we understand the model world. In a thought experiment we can move from the real world to the model world by increasing the number of measurements n towards infinity. Doing this we reduce the random error and become more and more secure.

Estimation is matching the **real world** with the **model world** as well as possible.

Let us assume that we have a data set of n points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Now we want to estimate the unknown parameters β_0, β_1 and ε of the model

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

using the data.

We use the **method of least squares** to estimate β_0, β_1 and σ^2. The idea is to minimise the sum of squares of the difference between the observations y_i and the straight line, cf. Fig. 7.2.i.
Thus, the least-squares criterion is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.2.a}$$

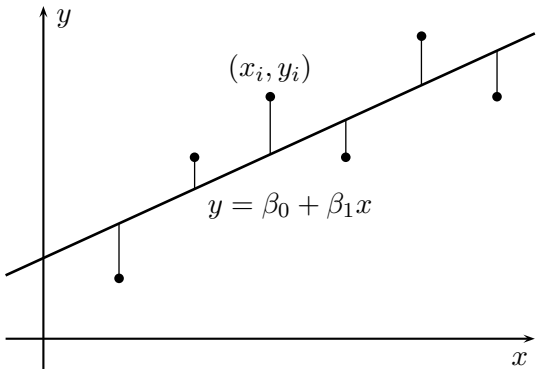


Figure 7.2.i: Method of least squares.

We call $\rho(r) = r^2$ the least squares **loss function**, cf. Fig. 10.4.i.a. The least-squares estimators[3] $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 must therefore satisfy

$$\begin{aligned} \frac{\partial S}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = 0. \end{aligned}$$

Simplifying these two equations yields to a linear system of equations

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{7.2.b}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{7.2.c}$$

Eqns. 7.2.b and 7.2.c are called the **least-squares normal equations**. From Eqn. 7.2.b we immediately conclude that

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \tag{7.2.d}$$

i.e. the centre of gravity

$$(\bar{x}, \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i \right)$$

of the point cloud lies on the straight line.
The least-squares normal equations are linear in $\hat{\beta}_0$ and $\hat{\beta}_1$ and therefore easy to solve. We find

$$\hat{\beta}_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}. \tag{7.2.e}$$

[3] We denote estimators with a hat.

In the literature the solution is often written in the form

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{with} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{7.2.f}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{7.2.g}$$

In our model there is also the error term ε which is a random variable with

$$\mathrm{E}(\varepsilon) = 0$$

and unknown variance

$$\mathrm{Var}(\varepsilon) = \sigma^2.$$

The variance σ^2 is not used to find the best straight line, but it will be used later to investigate model adequacy and to detect departures from underlying assumptions.
The difference between the ith observed value y_i and its **fitted value** $\hat{y}_i$ is the **residual**

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad \text{all} \ \ i \in \{1, \ldots, n\}.$$

A straight forward calculation using Eqn. 7.2.d shows that

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0.$$

An unbiased estimator $\hat{\sigma}^2$ is obtained from the **error sum of squares**

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2. \tag{7.2.h}$$

Notice that we have to divide by the **degrees of freedom** $n - 2$ to make the estimator unbiased.

Example 7.2.1. We want to find the best straight line in the sense of least squares for the vending machine data, cf. Ex. 7.1.1.

First we apply Eqn. 7.2.f and 7.2.h directly.

```
#   estimation of the parameters (explicit formulae)
> Volume.bar <- mean(data$Volume)
> Time.bar <- mean(data$Time)
> Sxx <- sum((data$Volume-Volume.bar)^2)
> Sxy <- sum((data$Volume-Volume.bar)*(data$Time-Time.bar))
#   slope
> beta1 <- Sxy/Sxx;   beta1
[1] 2.176167
#   intercept
> beta0 <- Time.bar - beta1*Volume.bar;   beta0
[1] 3.32078

#   fitted values
```

```
> Time.hat <- beta0 + beta1*data$Volume
#   residuals
> res <- data$Time - Time.hat
> mean(res)
[1] -9.239484e-16
#   error sum of squares
> sigma2.hat <- sum(res^2)/(nrow(data)-2)
#   residual standard error
> sqrt(sigma2.hat)
[1] 4.181397
```

Second, we want to use the much better `lm`-command which is one of the most used commands in R. Models for `lm` are specified symbolically. A typical model has the form

```
response ~ terms
```

where `response` is the (numeric) response vector and `terms` is a series of terms which specifies a linear predictor for `response`. In our example the model is

```
Time ~ Volume
```

and we also have to specify the data.

```
#   estimation of the parameters (lm)
> mod <- lm(Time ~ Volume, data);   mod
Call:
lm(formula = Time ~ Volume, data = data)

Coefficients:
(Intercept)       Volume
      3.321        2.176
```

The `lm`-command creates a big object of which we can display the internal structure using `str(mod)`. We can easily access all the necessary values.

```
#   coefficients
> mod$coefficients
(Intercept)      Volume
   3.320780    2.176167

#   residuals
> mod$residuals
         1          2          3          4          5 ...         24         25
-1.8739466  1.6507201  2.1807201  2.8545534 -2.6277800 ... -0.9001133 -1.2754466

#   fitted values
> mod$fitted.values
        1         2         3         4         5 ...        24        25
18.553947  9.849280  9.849280 12.025447 16.377780 ... 20.730113 12.025447

#   deviance
> dev <- deviance(mod);   dev
[1] 402.1338
```

```
#   degrees of freedom
> dfree <- df.residual(mod);   dfree
[1] 23

#   residual standard error
> sigma.hat <- sqrt(dev/dfree);   sigma.hat
[1] 4.181397
```

R is also very efficient to add the best model to an existing scatter diagram.

```
#   Scatter diagram: Time versus Volume
> plot(Time ~ Volume, data,
+      pch=20, xlim=c(0,30), ylim=c(0,80),
+      main="Delivery Time versus Delivery Volume with best model")
#   add best model
> abline(mod)
```

It is always wise to control the best model graphically.

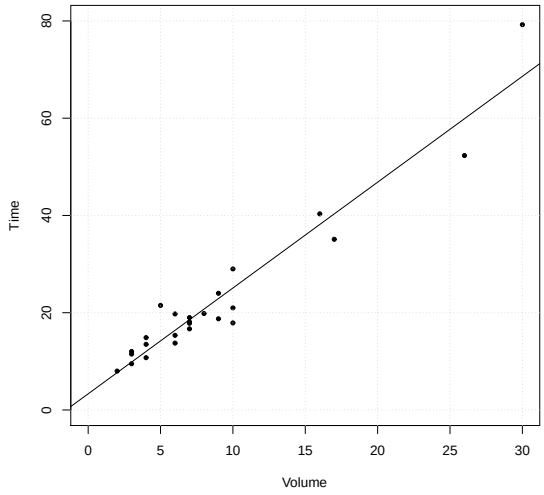


Figure 7.2.ii: Delivery time versus delivery volume with the best model.

It should be obvious that we will use mainly the `lm`-command instead of the explicit formulae. We will soon see that the `lm`-command is even much more powerful.

Having the least-squares fit ready we are not yet finished with our analysis. We need to ask several important questions:

1. How well does the model fit the data?
2. Can the model be used as a reliable predictor?
3. Are the assumptions of constant variance and uncorrelated errors met?

To be able to answer these questions we need to know the **distributions of the estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$. To do that we make the additional assumptions that all the errors ε_i are normally distributed with expected value 0 and variance σ^2. Using probability theory it is possible to

find the distributions directly from the distribution of the errors (cf. [24], Chap. 2.2.2). The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are therefore also normally distributed

$$\hat{\beta}_0 \sim \mathcal{N}\Big(\beta_0, \sigma^2(\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}})\Big) \tag{7.2.i}$$

and

$$\hat{\beta}_1 \sim \mathcal{N}\Big(\beta_1, \tfrac{\sigma^2}{S_{xx}}\Big) \tag{7.2.j}$$

where S_{xx} is given in Eqn. 7.2.g.

We directly observe that the size of S_{xx}, the sample size n and the variance of the errors directly affect the precision of the estimators.

7.3 Tests and Confidence Intervals

In many cases we are not only interested in estimating the model parameters but also in testing hypothesis and constructing confidence intervals.

7.3.1 Test of the Slope

In Ex. 7.2.1 we estimated the slope $\hat{\beta}_1 = 2.176$. Since this is an estimate of a random variable with a certain variance, it might well be that the slope is 2. How do we know if a slope of 2 does not match the data? This question will be answered with a statistical test.
Let us test the hypothesis that the slope equals the constant $\beta_{1,0}$. The two alternative hypotheses are

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0}$$

The null hypothesis states that the observations follow the model of simple linear regression with $\beta_1 = \beta_{1,0}$ and arbitrary β_0 and σ.
The difference between the estimated slope $\hat{\beta}_1$ and the assumed value $\beta_{1,0}$ is a measure for the discrepancy of the data and the model. To be able to judge which difference is big we need the distribution of the slope assuming the null hypothesis to be true, Eqn. 7.2.j. Since the variance σ^2 of the errors is unknown we will estimate it with the error sum of squares, Eqn. 7.2.h, thus we find that estimated **standard error** is

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}. \tag{7.3.a}$$

Therefore the statistic to test the hypothesis about the slope is

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{se}(\hat{\beta}_1)}. \tag{7.3.b}$$

Under the null hypothesis the test statistic T follows a Student's t-distribution with $n - 2$ degrees of freedom, cf. T.3. The degrees of freedom equals the denominator of $\hat{\sigma}$.

In many applications we test the two alternative hypotheses

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

These hypothesis relate the **significance of regression**. If we fail to reject $H_0 : \beta_1 = 0$ then there is no linear relationship between x and y. Or the other way round: if $H_0 : \beta_1 = 0$ is rejected then the variable x is capable of explaining some variability in y.

Example 7.3.1. We test the slope for significance of regression in the vending machine regression model of Ex. 7.1.1 and Ex. 7.2.1. The two alternative hypotheses are

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

First, we will do the test by calculating each step using some of the `lm`-Output:

```
#   null hypothesis
> beta10 <- 0
#   estimator of slope
> beta1.hat <- as.numeric(mod$coefficients["Volume"]);   beta1.hat
[1] 2.176167
#   standard error (Sxx from Ex. 7.2.1)
> se.beta1 <- sqrt(sigma.hat^2/Sxx);   se.beta1
[1] 0.1240296
#   test statistic
> Test.beta1 <- (beta1.hat-beta10)/se.beta1;   Test.beta1
[1] 17.54555
#   critical value (two-sided)
> alpha <- 0.05
#   the degrees of freedom df were defined in Ex. 7.2.1
> t.crit <- qt(p=1-alpha/2, df=dfree);   t.crit
[1] 2.068658
```

For $t_{0.975,23} = 2.068658$ see also T.3. Since $|T| = 17.54555 > t_{0.975,23} = 2.068658$ we reject the null hypothesis and conclude that the delivery volume depends significantly on the delivery time.

If we want to test a different null hypothesis, eg. $H_0 : \beta_1 = 2$, then we need to adapt the above code slightly and get

```
#   new null hypothesis
> beta10 <- 2
#   test statistic
> (beta1.hat-beta10)/se.beta1
[1] 1.42036
```

Since $|T| = 1.42036 \leq t_{0.975,23} = 2.068658$ we accept the null hypothesis and conclude that the data agrees also with a model with slope 2.

In R there is a much simpler way to do these tests. We create a very useful summary of the whole estimation process.

```
> summary(mod)
Call:
lm(formula = Time ~ Volume, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371   2.422   0.0237 *
Volume         2.176      0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

In the section `Coefficients` we find the standard `Error` for the slope. To test significance of regression we only have to calculate the test statistic $T = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$ and compare it with the `t value`. Note that the `t value` is calculated on the 5%-level for a two-sided test.

On the other hand, there is a simpler way to test significance of regression looking at the so-called P-**value**. The P-value is defined as the probability, under the null hypothesis, of obtaining a result equal to or more extreme than what was actually estimated. In Fig. 7.3.i we can see the distribution of $\hat{\beta}_1$ under the null hypothesis $H_0 : \beta_1 = \beta_{1,0}$. The P-value for the two-sided alternative is

$$P = P(\beta_1 \leq -|\hat{\beta}_1| \,|H_0 \text{ is true}) + P(|\hat{\beta}_1| \leq \beta_1 | H_0 \text{ is true}).$$

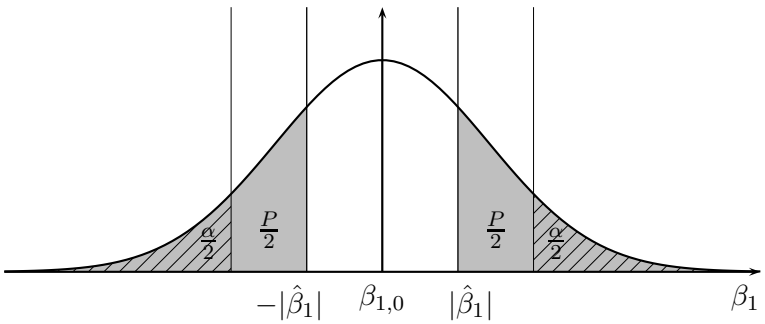


Figure 7.3.i: Two-sided P-value (gray) for test of significance of regression $H_0 : \beta_1 = \beta_{1,0}$ with significance level α (hatched).

The convention of the statistical test procedure with a given significance level α (often $\alpha = 5\%$) is the following:

- Reject H_0 and believe that the alternative is true if the P-value is smaller than the significance level α. The data significantly differs from the null hypothesis.

- Accept H_0 if the P-value is bigger or equal than the significance level α. The data agrees with the null hypothesis.

Example 7.3.2. We test the slope for significance of regression in the vending machine regression model of Ex. 7.1.1, 7.2.1 and Ex. 7.3.1 calculating the P-value.

```
#   the degrees of freedom were defined in Ex. 7.2.1
> pt(-abs(Test.beta1), df=dfree) + (1-pt(abs(Test.beta1), df=dfree)))
[1] 8.217921e-15
```

Since the P-value is much smaller than the significance level α we reject the null hypothesis.

We can also directly read off the P-value `Pr(>|t|)` from the output of `summary(mod)`. In addition this output provides us with a very quick overview of which variables are significant looking at the significance code behind the reported P-values. In our example find three stars which means that the P-value is smaller than 0.001.

The idea of a statistical test is to find out if a particular model can be true or not. The test is a rule that decides in which cases we can answer this question with *yes* and when with *no*, or even better with *plausible* and *implausible*.
To establish this rule we use a kind of a proof by contradiction: We assume that the observed random variables correspond to the model of the null hypothesis. If the observed value lies in the range of implausible values, then we consider this as a contradiction, i.e. as statistical proof that the null hypothesis does not hold. As a consequence, we use the actual research hypothesis in the alternative, if possible.

We are in the process of connecting models and data, i.e. using the data to say something about values of the parameters of the model that seem plausible. We can ask the following questions:

1. Which value is the most plausible for the parameter under consideration? The answer is given by the estimator.

2. Is a certain value plausible? The decision is based on a statistical hypothesis test.

3. Which values are plausible overall? The answer is a whole lot of plausible values which form the **confidence interval**.

The confidence interval contains all parameter values that are not rejected due to a statistical test. Each confidence interval therefore corresponds to a specific test rule.

7.3.2 Test of the Intercept

Very similar procedures can be used to test hypothesis about the intercept. We do not give the details here since R provides us with all the necessary quantities.

7.3.3 Confidence Interval on the Slope

The **confidence interval** on the **slope** in a simple regression model can be found by looking at the test statistic T, cf. Eqn. 7.3.b. In a two-sided test the region of acceptance for T is

$$t_{\frac{\alpha}{2},n-2} \leq T \leq t_{1-\frac{\alpha}{2},n-2},$$

where $t_{p,m}$ denotes the critical value of Student's t-distribution to the probability p and with m degrees of freedom, cf. T.3. Now we solve Eqn. 7.3.b for $\beta_{1,0}$, i.e. the value in the null hypothesis H_0. Like this we obtain all possible slopes which are not rejected under the null hypothesis, that is, we find that the slope β_1 can be in the interval

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{\beta}_1). \tag{7.3.c}$$

The width of this $100(1-\alpha)$ percent confidence interval is a measure of the overall quality of the regression line.

The usual **frequentist interpretation** of such a confidence interval is as follows: If we were to take repeated samples of the same sample size at the same levels and construct, eg. 95% confidence intervals on the slope for each sample, then 95% of those intervals will contain the true value of β_1.

Example 7.3.3. We construct the 95% confidence interval on the slope of the simple linear regression in the vending machine data of Ex. 7.1.1.

```
#   95% confidence interval on slope (explicit formulae)
> beta1.hat + c(-1,1)*t.crit*se.beta1
[1] 1.919592 2.432741

#   95% confidence interval on slope (lm)
> confint(mod, level=0.95)
                2.5 %   97.5 %
(Intercept) 0.4844979 6.157062
Volume      1.9195920 2.432741
```

The confidence interval on the slope contains $\beta_1 = 2$ and therefore a slope of 2 is compatible with the data.

7.4 Confidence Interval on Response and Prediction Intervals

In the following we are going to construct two different intervals: First the **confidence interval on the expected response** and second the **prediction interval** for a new observation. It will be clear that the prediction interval will be wider than the confidence interval on the response, since the variance of the future observation has to be considered too.

7.4.1 Confidence Interval of the Response

A major use of a regression model is to estimate the response y for a particular value of the explanatory variable x. Applying the fitted model we immediately get the estimated response

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Now we want to find a **confidence interval on the response**. Such a confidence interval corresponds to the null hypothesis

$$H_0 : \hat{y}_0 = \mu_0.$$

It is possible to show (cf. [24], Chap. 2.4.2) that the estimator $\hat{y}_0$ is normally distributed, unbiased and has variance

$$\mathrm{Var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \tag{7.4.a}$$

As usual σ^2 is in general unknown and has to be estimated from the data, cf. 7.2.h. The test statistic about the response is again of the same type as in Eqn. 7.3.b, that is

$$T = \frac{\hat{y}_0 - \mu_0}{\mathrm{se}(\hat{y}_0)} \tag{7.4.b}$$

with

$$\mathrm{se}(\hat{y}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Under the null hypothesis the test statistic T follows a Student's t-distribution with $n-2$ degrees of freedom, cf. T.3.

Therefore a $100(1-\alpha)$ percent **confidence interval on the response** at the point x_0 is

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{y}_0) \leq \hat{\beta}_0 + \hat{\beta}_1 x_0 \leq \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{y}_0). \tag{7.4.c}$$

Note that the width of the confidence interval depends on x_0. The interval width is minimal for $x_0 = \bar{x}$ and gets bigger as $|x_0 - \bar{x}|$ increases. This makes completely sense, since we would expect that our best estimates of y to be made at values x near the centre of the data and that the precision of the estimates gets worse the further we move to the boundaries.

If we calculate for every x_0 a confidence interval on the response we obtain two symmetric curves around the fitted straight line, cf. Fig. 7.4.i.

Example 7.4.1 (**Confidence Intervals on the Response**)**.** We construct 95% confidence intervals on the response for Ex. 7.1.1. Now we switch from using explicit formulae to beautiful commands in R.

```
#   define a new data.frame of equally spaced x-values
#   covering at least the range of Volume-data
> data.new <- data.frame(Volume=seq(-10, 40, length=101))
#   predict using interval="confidence"
> Time.Conf <- predict(mod, newdata=data.new, interval="confidence", level=0.95)
> head(Time.Conf)
        fit       lwr       upr
1 -18.44089 -23.55568 -13.32610
2 -17.35280 -22.34705 -12.35855
3 -16.26472 -21.13883 -11.39061
4 -15.17664 -19.93103 -10.42224
5 -14.08855 -18.72369  -9.45342
6 -13.00047 -17.51684  -8.48410

#   scatter diagram: Time versus Volume with confidence intervals on the response
> plot(Time ~ Volume, data,
+      pch=20, xlim=c(0,30), ylim=c(0,80),
+      main="Time ~ Volume, best model, CIs on the response")
```

```
> abline(v=mean(data$Volume), lty=3)
> grid()
#   add best model
> lines(data.new$Volume, Time.Conf[,"fit"], lty=1)
#   add confidence intervals on the response
> lines(data.new$Volume, Time.Conf[,"lwr"], lty=2)
> lines(data.new$Volume, Time.Conf[,"upr"], lty=2)
#   add legend
> legend("topleft",
+        legend=c("best model",
+                 "confidence intervals on the response"),
+        lty=c(1,2),
+        bty="n")
```

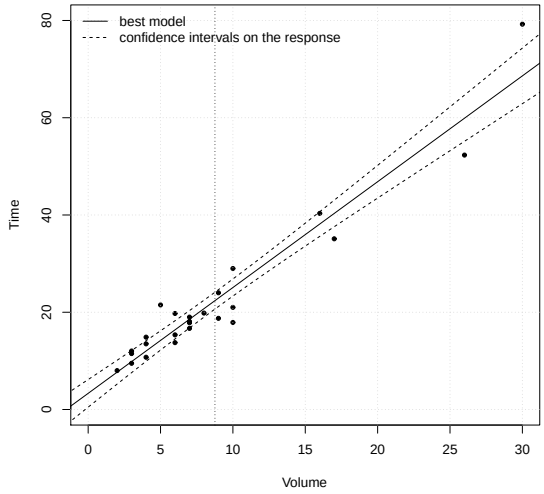


Figure 7.4.i: 95% Confidence intervals of the response. Vertical dotted line is at the narrowest position `mean(Volume)` = 8.76.

7.4.2 Prediction Interval

Prediction of a new observations y corresponding to a specified level of the explanatory variable x is a very important application of the regression model. Let us assume that x_0 is the value of the explanatory variable of interest, then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimate of the new value of the response y_0.

Now our aim is to find a **prediction interval** for a future observation. It is very important to understand that in this situation the randomness of the future observation y_0 has to be considered too. Therefore we now consider the random variable

$$y_0 - \hat{y}_0.$$

Both random variables[4] y_0 and $\hat{y}_0$ are normally distributed. The expected value of $y_0 - \hat{y}_0$ is zero. They are independent because y_0 only depends on the future observation and $\hat{y}_0$ on the past observations $y_1, \ldots, y_n$. To calculate its variance we use the formulae for the difference of two uncorrelated random variables X and Y (cf. [23], Chap. 5-4)

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Now using Eqn. 7.4.a we obtain

$$\begin{aligned} \mathrm{Var}(y_0 - \hat{y}_0) &= \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \end{aligned}$$

The prediction interval is wider than the confidence interval on the response, since the variance of the future observation has to be considered too. As usual σ^2 is in general unknown and has to be estimated from the data, cf. 7.2.h.

Therefore a $100(1-\alpha)$ percent **prediction interval** at the point x_0 is

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(y_0 - \hat{y}_0) \le \hat{y}_0 \le \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(y_0 - \hat{y}_0) \tag{7.4.d}$$

with

$$\mathrm{se}(y_0 - \hat{y}_0) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

where $t_{p,m}$ denotes the critical value of Student's t-distribution to the probability p and with m degrees of freedom, cf. T.3.

Example 7.4.2 (**Prediction Intervals**)**.** We want to construct 95% prediction intervals for Ex. 7.1.1.

```
#   predict using interval="prediction"
> Time.Pred <- predict(mod, newdata=data.new, interval="prediction", level=0.95)
> head(Time.Pred)
        fit       lwr       upr
1 -18.44089 -28.48984 -8.391934
2 -17.35280 -27.34094 -7.364664
3 -16.26472 -26.19333 -6.336109
4 -15.17664 -25.04703 -5.306245
5 -14.08855 -23.90206 -4.275050
6 -13.00047 -22.75844 -3.242499

#   scatter diagram: Time versus Volume with confidence intervals on the response
#   and prediction intervals
> plot(Time ~ Volume, data,
+      pch=20, xlim=c(0,30), ylim=c(0,80),
+      main="Time ~ Volume, best model, CIs on the response, PIs")
```

[4] Here we are bit sloppy with our notation, since random variables should be written in capital letters. But we don't want to further complicate things.

```
> grid()
#   add best model
> lines(data.new$Volume, Time.Pred[,"fit"], lty=1)
#   add confidence intervals on the response
> lines(data.new$Volume, Time.Conf[,"lwr"], lty=2)
> lines(data.new$Volume, Time.Conf[,"upr"], lty=2)
#   add prediction intervals
> lines(data.new$Volume, Time.Pred[,"lwr"], lty=3)
> lines(data.new$Volume, Time.Pred[,"upr"], lty=3)
#   add legend
> legend("topleft",
+        legend=c("best model",
+                 "confidence intervals on the response",
+                 "prediction intervals"),
+        lty=c(1,2,3),
+        bty="n")
```

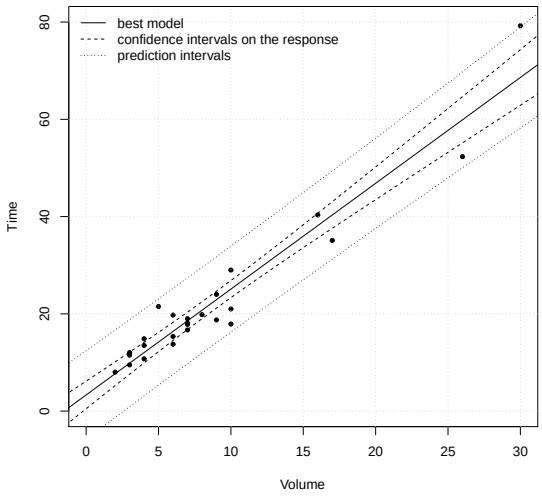


Figure 7.4.ii: 95% Confidence intervals of the response and prediction intervals.

Remark 1. It is obvious that regression models can also be used to **extrapolate** beyond the range of the original data. Fig. 7.4.i clearly illustrates the dangers inherent in extrapolation: the quality of extrapolation deteriorates rapidly the further we move away from the original region of x space. A nice example is the weather forecast which always gets less and less reliable the further we look in the future[5].

Remark 2. We have to be a bit careful interpreting prediction intervals. The probability that one single new observation lies in the prediction interval is by construction exactly equal to $1 - \alpha$.
If we want to make statements about more than one new observation, then we have to consider that the number of observations in the prediction intervals is not binomially distributed. The

[5]Niels Bohr, 1885-1962: *Prediction is very difficult, especially about the future.*

events that individual future observations lie in the prediction intervals are not independent. They are dependent due to the random limits of the prediction intervals.

7.5 Exercises

Problem 7.5.1 (**Venice Sea Level**, cf. [8], Example 5.1)**.** The annual maximum sea levels [in cm] in Venice, 1931-1981 have been recorded by P. A. Pirazzoli. The data set `venice.dat` contains the annual maximum tides at Venice for the 51 years.

a. Represent the data in a scatter diagram sea level versus year and describe the functional context in words.

b. Fit a straight line to the data points. Give the estimated parameter values.

c. Add the model in the scatter diagram. Comment on the solution.

d. Does the data support the hypothesis that Venice sinks?

Problem 7.5.2 (**Windmill**, cf. [24], Example 5.2)**.** A research engineer is investigating the use of a windmill to generate electricity. She has collected data on DC output [in A] from her windmill and the corresponding wind velocity [in m/s]. The data set `windmills.dat` contains the wind velocities and the DC outputs.

a. Plot a scatter diagram of DC output versus wind velocity and a scatter diagram of DC output versus the inverse wind velocity. What are your observations?

b. Fit the model

$$\texttt{DCOutput}(\texttt{WindVelocity}) = \beta_0 + \frac{\beta_1}{\texttt{WindVelocity}}$$

to the data. Give the estimated parameter values and the standard errors.

c. Calculate the 99% confidence interval on β_1.

d. Add the best model to the scatter diagram DC output versus the inverse wind velocity. Give a physical interpretation of the parameters β_0 and β_1.

e. Calculate the estimated DC output, the 95% confidence interval of the response and the 95% prediction interval with a wind velocity of one and ten metres per second. Comment your results.

Problem 7.5.3 (**Atmospheric Pressure versus Boiling Point of Water**, cf. [43], Chap. 1.1)**.** In 1857 the Scottish physicist James D. Forbes discussed an experiment concerning the relationship between atmospheric pressure and the boiling point of water. At the time it was known that altitude could be determined from atmospheric pressure. In the middle of the nineteenth century, barometers were fragile instruments, and Forbes had the idea to replace barometers by a simpler measurement of the boiling point of water.
He measured in the Alps and in Scotland at each location the pressure [in inHg] with a barometer and the boiling point [in °F] using a thermometer. The data set `forbes.dat` contains the pressure and the boiling point.

a. Plot the logarithm (base 10) of the boiling point versus the pressure. What are your observations?

b. Fit and plot a straight line to the transformed data points. Give the estimated parameter values. What are your observations?

c. Fit and plot a straight line to the transformed data points without the 12th observation. Use the argument `subset` in `lm`. Compare the estimated parameter values, the standard errors and the residual standard error with the fit of the full data. What are your observations?

Solve the next questions without the 12th observation.

d. Test the slope for significance of regression on the 5% level.

e. Calculate the 95% confidence level for the slope.

f. Find an estimation of the boiling point at the pressure 26 inHg using the model fit to the transformed data. Calculate the 95% confidence interval of the response of this estimator and add the complete confidence interval band to the scatter diagram.

g. Calculate the 95% prediction interval and add the complete prediction interval band to the scatter diagram.

Problem 7.5.4 (**Simulation of the Distribution of Regression Parameters**, cf. [32], Reg1, Problem 4)**.** We want to simulate the distribution of the regression parameters β_0 and β_1 keeping the explanatory variable x constant only varying the measurement errors ε randomly. Let us choose the simple linear model $y = 5+3x$ and $n = 20$ uniformly distributed x-values in the interval from $x_{\min} = 0$ to $x_{\max} = 50$. The measurement errors $\varepsilon_1, \ldots, \varepsilon_n$ are supposed to be normally distributed with mean zero and standard error $\sigma = 5$.

The following code should help you to start the programming. Have fun!

```
> set.seed(1)
> x <- runif(n=20, min=0, max=50)
> y <- 5+3*x
> for(i in 1:100){
+     Y <- y + rnorm(n=20, mean=0, sd=5)
+     plot(x, Y, main="Data Set with Simulated Errors")
+     abline(a=5, b=3)
+     abline(lm(Y ~ x), col="red")
+     dev.flush()
+     Sys.sleep(0.2)
+ }
```

We use the command `set.seed()` so that we can always reproduce exactly the same simulation.

a. Using a `for`-loop plot successively 100 data sets with simulated errors. For each simulated data set add the best straight line. Observe the scattering of the points and its corresponding regression line.

b. For each data set with simulated errors record the estimated parameters β_0 and β_1 in two separate vectors. Plot the two empirical distributions of these estimated parameters using histograms with the argument `freq=F`. For each histogram add the density of the theoretical distribution. Compare the means and the standard deviations of the simulated parameter vectors with the theoretical parameters.

c. Repeat the above with 10000 simulations. What do you observe?

Chapter 8

Residual Analysis

8.1 Introduction

In our study of regression analysis in Chap. 7 we made several assumptions. The discussed estimation of parameters and statistical tests are based on assumptions about the model:

1. The relationship between response and the regressors is linear.
2. The error ε has mean zero.
3. The error ε has constant variance[1] σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

If the first assumption about linearity is violated then the errors do not have mean zero anymore.
We usually cannot detect departures from underlying assumptions by examination of standard summary statistics, such as t or R^2. These are global model properties and thus do not ensure model adequacy.
In other words, if some assumptions are violated we should be able to see them in the errors ε_i. On the other hand, the errors are unknown to us, so we have to deal with the residuals

$$e_i = y_i - \hat{y}_i \quad \text{for all} \quad i \in \{1, \ldots, n\}$$

instead. The residuals are estimators of the random errors.
If the errors are normally distributed, so are the residuals of a least-squares estimate. Since the variance of the residuals depends on the explanatory variables it is not equal to the variance of the errors. Therefore we use **scaled residuals**, cf. [24], Chap. 4.2.2,

$$\tilde{e}_i = \frac{e_i}{\sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}} \quad \text{for all} \quad i \in \{1, \ldots, n\}. \tag{8.1.a}$$

The scaling depends on $(x_i - \bar{x})^2$, i.e. the further the observation is away from the center of gravity, the smaller the scaling factor.

[1] Constant variance is often called **homoscedasticity** and the contrary **heteroscedasticity**.

Often it is more convenient to use **standardised residuals**

$$\tilde{e}_{\mathrm{std},i} = \frac{e_i}{\hat{\sigma}\sqrt{1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}} \quad \text{for all } \; i \in \{1,\ldots,n\}, \tag{8.1.b}$$

where $\hat{\sigma}$ is estimated from the data, cf. 7.2.h. In R standardised residuals can be obtained from the `lm`-output using the command `stdres()`.

Checking these assumptions is usually essential in regression analysis. It is not primarily about a justification, but about the chance to develop a better model from possible deviations.
Having the chance to improve a model is the idea of exploratory data analysis. In this area, it is less about precise mathematical statements, optimality of statistical methods or statistical significance, but about methods that are useful for creatively designing appropriate models for existing data.

Residual analysis is based on some graphical methods and also some formal statistical tests. These methods can give indications that a model does not exactly describe the data. With these indications it is sometimes possible to find the reason. If several aspects of the model are wrong and therefore corresponding effects superimpose such an analysis can be cumbersome. The development of a suitable model then needs intuition, experience and often a lot of creativity. Beside these soft skills we have to rely on powerful tools to detect violations of the assumptions on the errors.
To be able to illustrate the concepts in this chapter we introduce a new example from chemistry (cf. [28], Problem 14.9.37).

Example 8.1.1 (**Chemical Reaction, Dissociation Pressure**)**.** In 1970 R. H. Orcutt recorded the dissociation pressure P [in torr] for a reaction involving barium nitride as a function of the absolute temperature T [in K]. The second law of thermodynamics gives the approximate relationship, i.e. a mechanistic model

$$P = A\exp\left(\frac{B}{T}\right), \tag{8.1.c}$$

where A and B are parameters depending on the material. This functional relationship is clearly visible in the data, cf. Fig. 8.1.i.
On the other hand, it is possible to transform Eqn. 8.1.c

$$\ln(P) = \ln(A) + \frac{B}{T}, \tag{8.1.d}$$

which clearly shows that the natural logarithm[2] of the pressure is linearly dependent on the inverse of the temperature, cf. Fig. 8.1.ii.
Eqn. 8.1.c and 8.1.d mathematically represent exactly the same model. If we define new parameters $\beta_0 = \ln(A)$ and $\beta_1 = B$ then the advantage of Eqn. 8.1.d is that it is a linear model in the transformed variables $y = \ln(P)$ and $x = \frac{1}{T}$.

```
#   read data
> file <- "dissociation-pressure.dat"
> data <- read.table(file, header=TRUE)
```

[2]Here we use the natural logarithm because Eqn. 8.1.c is formulated with base e.

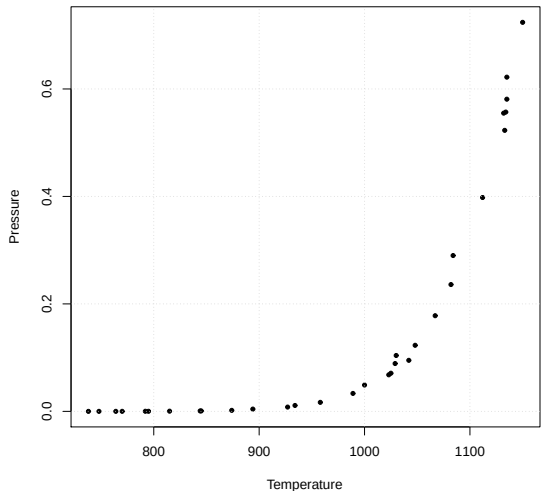


Figure 8.1.i: Scatter diagram of pressure versus temperature.

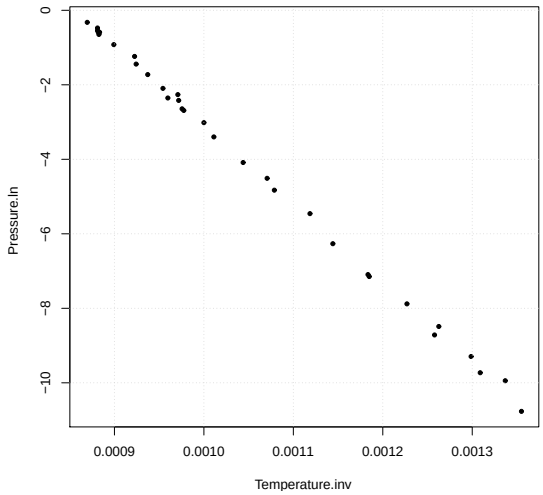


Figure 8.1.ii: Scatter diagram of logarithm of pressure versus inverse temperature.

```
> str(data)
data.frame':   32 obs. of  2 variables:
$ Temperature: int  738 748 764 770 792 795 815 844 845 874 ...
$ Pressure   : num  2.11e-05 4.80e-05 5.95e-05 9.20e-05 ...

#   transform variables
> data$Temperature.inv <- 1/data$Temperature
#   in R log denotes the natural logarithm
> data$Pressure.ln <- log(data$Pressure)

#   scatter diagram: Pressure versus Temperature
> plot(Pressure ~ Temperature, data,
+      pch=20, main="Pressure versus Temperature")
> grid()
#   scatter diagram: ln(Pressure) versus 1/Temperature with best model
> plot(Pressure.ln ~ Temperature.inv, data,
+      pch=20, main="ln(Pressure) versus 1/Temperature with best model")
> grid()
#   linear model
> mod <- lm(Pressure.ln ~ Temperature.inv, data)
> abline(mod)
```

If the model described the data perfectly the response values would perfectly match the fitted values. Therefore we look at the scatter diagram of the fitted values versus the response values. In the perfect case they would lie on the diagonal, Fig. 8.1.iv.

```
> plot(Pressure.ln ~ fitted(mod), data,
+      xlim=c(-12,0), ylim=c(-12,0), pch=20,
+      main="ln(Pressure) versus fitted values")
> grid()
> abline(a=0, b=1, lty=3)
```

Our first tool the check the model assumptions is simply a visual qualitativ inspection of the best model cf. Fig. 8.1.iii. We can easily see that the model is a very good one.

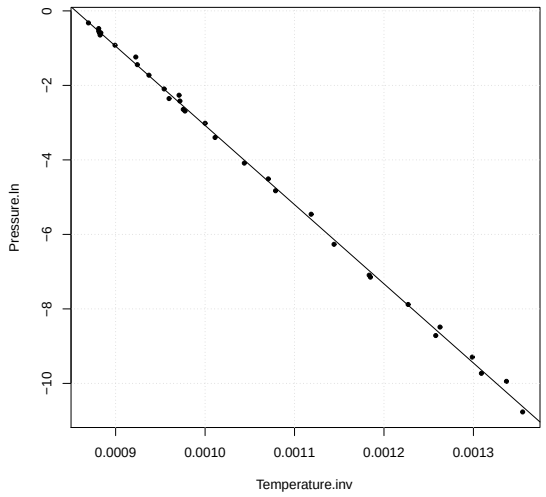


Figure 8.1.iii: Scatter diagram of pressure versus temperature with best linear model.

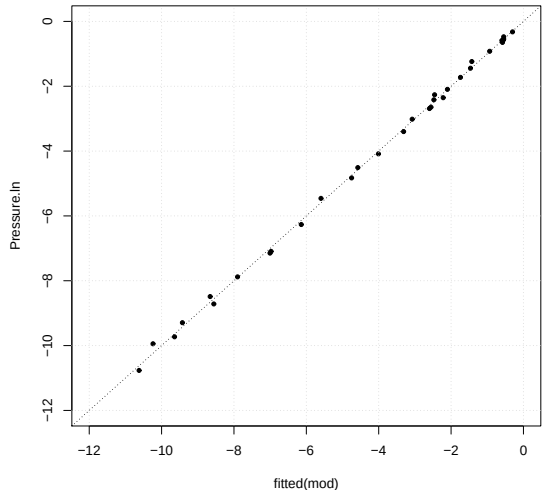


Figure 8.1.iv: Scatter diagram of logarithm of pressure versus fitted values with diagonal line (dashed).

But is this visual qualitativ assessment enough? We can do better!

In the following we want to quantify the goodness of fit[3] with the **coefficient of determination**. It is the proportion of the variance in the dependent variable that is predictable from the independent variable.

- In **simple linear regression** with only one explanatory variable the coefficient of determination is defined by

$$R^2 = \frac{\mathrm{SS}_{\mathrm{fit}}}{S_{yy}},$$

where

$$\begin{aligned} \mathrm{SS}_{\mathrm{fit}} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^{n} ((\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^{n} (\bar{x} - x_i)^2 = \hat{\beta}_1^2 S_{xx}, \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2. \end{aligned}$$

Therefore the coefficient of determination is

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}. \tag{8.1.e}$$

[3]The arguments are only valid if there is an intercept β_0 in the model.

Combining Eqn. 7.2.f with Eqn. 8.1.e we immediately obtain an other useful formula

$$R^2 = \frac{S_{xy}^2}{S_{xx}\,S_{yy}} = \mathrm{Cor}(x,y)^2 = \frac{\left(\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}.$$

This is the squared correlation between the explanatory x and the response variable y.

- In **multiple regression** and **simple linear regression** with at least one explanatory variable the coefficient of determination is identical to the squared correlation between the response variable y and the fitted values $\hat{y}$

$$R^2 = \mathrm{Cor}(y,\hat{y})^2 = \frac{S_{y\hat{y}}^2}{S_{yy}\,S_{\hat{y}\hat{y}}}.$$

The coefficient of determination is a measure of the linear relationship between the response variable and the fit, cf. Fig. 8.1.iv.

An R^2 of 1 indicates that the fitted model explains all variability in y, while a vanishing R^2 indicates no linear relationship. A value such as $R^2 = 0.8$ may be interpreted as follows: Eighty percent of the variance in the response variable can be explained by the explanatory variables. The remaining twenty percent can be attributed to unknown variables or inherent variability.
What is a good R^2? The answer depends a lot on the field. In technical and natural sciences values bigger than 0.9 are common, while people in social sciences are often already happy with values around 0.6.

Example 8.1.2 (**Chemical Reaction, Dissociation Pressure**)**.** We will now look at the coefficient of determination in Ex. 8.1.1. The simplest way is to check the `Multiple R-squared` in the R-summary-output.

```
> summary(mod)
Call:
lm(formula = Pressure.ln ~ Temperature.inv, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.158596 -0.087968  0.004495  0.061145  0.293036

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.818e+01  1.419e-01   128.2   <2e-16 ***
Temperature.inv -2.126e+04  1.335e+02  -159.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1164 on 30 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9988
F-statistic: 2.538e+04 on 1 and 30 DF,  p-value: < 2.2e-16

> summary(mod)$r.squared
[1] 0.9988193
```

Of course, we can also calculate the coefficient of determination by the above explicit formula.

```
> cor(data$Pressure.ln, data$Temperature.inv)^2
[1] 0.9988193
> cor(data$Pressure.ln, fitted(mod))^2
[1] 0.9988193
```

We conclude that in this example circa 0.12% of the variance remains unexplained. This is extremely good even for a mechanistic model.

The coefficient of determination is a global measure for the goodness of fit and it says nothing about the suitability of the regression model.

Example 8.1.3 (**Anscombe Quartet**)**.** In 1973 F. J. Anscombe published, cf. [1], four different data sets each containing an x- and a y-variable. All four artificial data sets can be found in `anscombe.dat`.

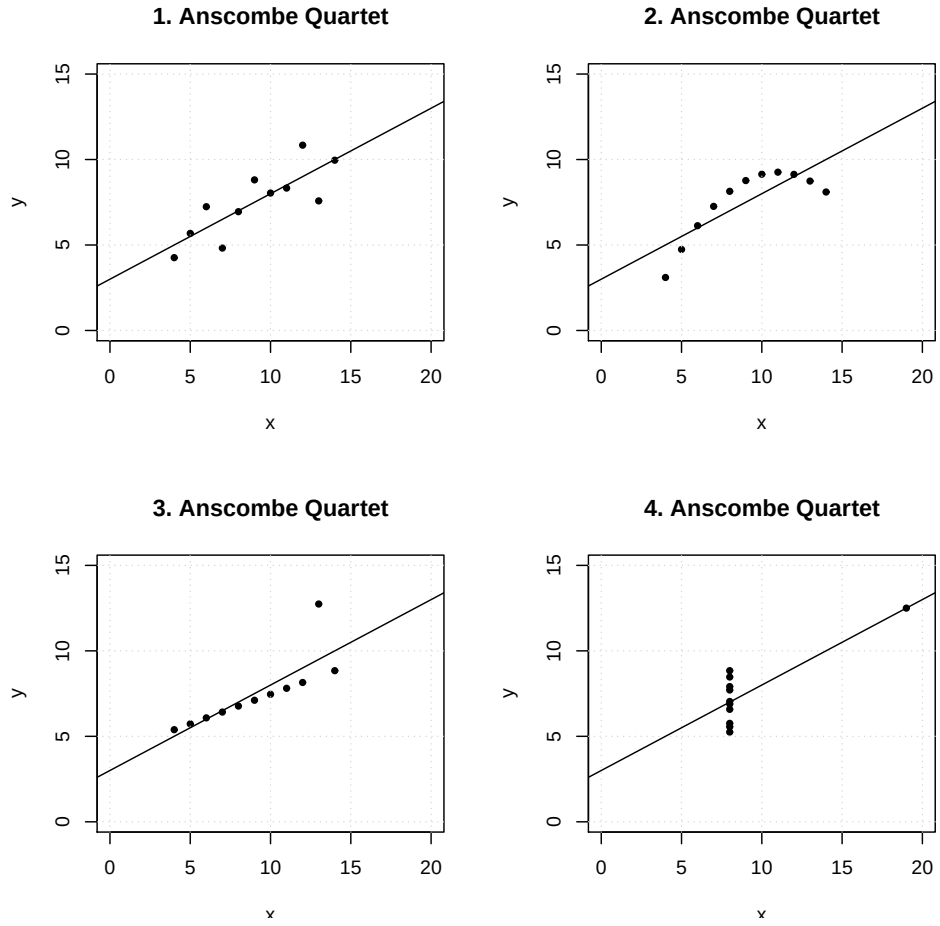


Figure 8.1.v: Anscombe Quartet. The coefficient of determination is the same for all four data sets: $R^2 = 0.6667$. But only the first Anscombe quartet makes sense in our framework.

In Fig. 8.1.v we observe that the estimated best straight line is identical for all four data sets. Even the standard errors and the coefficient of determination for each model are the same.

If we investigate these four data sets we immediately realise that the coefficient of determination tells us nothing about the suitability of the regression model. Therefore, the coefficient of determination should only be taken into account as a valid measure of the goodness of fit if the residual analysis described below does not indicate any violation of the model assumptions.

Let us now return to Fig. 8.1.iii to investigate the suitability of the model. What does this scatter diagram with the best straight line tell us about the individual assumptions?
We realised already earlier that the model fits nicely. But if we look a bit closer we can observe that the points scatter around the straight line more on the right than on the left. This implies that in this example the assumption of constant variance of the errors might be violated.
The assumption of a zero mean of the errors seems to be in order. Whether the errors are normally distributed, we cannot decide with this diagram.
It is clear to us that with the scatter diagram and the best model it is difficult to verify the model assumptions. That is why we need more effective diagnostic tools.

8.2 Diagnostic Tools

We will present three graphical diagnostic tools to check model assumptions 1, 2, 3 and 5. The aim is to be sure that there are no dangerous discrepancies from the assumptions in the data. All three tools are based on the residuals which are representations of the unknown errors. To obtain useful conclusions we will use bootstrapping, which is a very powerful tool in modern statistics, cf. [11].
There exist also formal statistical test, so-called **goodness of fit tests**. In this course we will not discuss these procedures. The author (and many other professional statistician) believe that graphical diagnostic tools are superior in the sense that the eye might detect unexpected patterns which can later help to improve the model.

8.2.1 Tukey-Anscombe Plot

The first diagnostic tool is the **Tukey-Anscombe plot** which is an indispensable component of any residual analysis.
The investigations of Fig. 8.1.iii (response versus explanatory variable) and Fig. 8.1.iv (fitted versus explanatory variable) can be made more accurate if we rotate the plot slightly.
Instead of the observed values y_i we plot the residuals e_i in the vertical direction and the fitted values $\hat{y}_i$ in the horizontal direction. Thus, the line of reference now lies horizontally in the diagram and the deviations (errors) from the straight line are correctly perceived by the eye. This modification also helps to see deviations more clearly if the points scatter only little around the straight line, i.e. if the coefficient of determination R^2 is high and the residuals therefore small relative to the range of the response variables. In the **Tukey-Anscombe plot** the points should spread uniformly around the horizontal line.

Example 8.2.1 (**Atmospheric Pressure versus Boiling Point of Water**)**.** We consider Prob. 7.5.3 again. The following analysis has been carried out without the outlier which was

detected in Prob. 7.5.3. The aim is to find out which of the two linear regression models

$$\texttt{Boiling} = \beta_0 + \beta_1 \cdot \texttt{Pressure}$$

or the model with the log-transformed pressure

$$\texttt{Boiling} = \beta_0 + \beta_1 \cdot \log(\texttt{Pressure})$$

fits the data better.

Therefore we investigate Fig. 8.2.i and 8.2.ii.

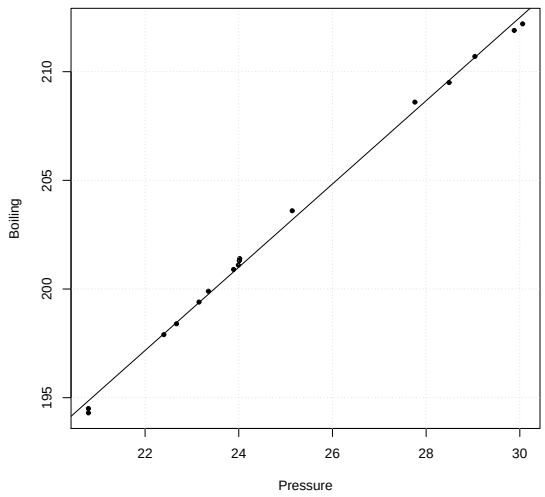


Figure 8.2.i: Boiling point versus pressure with best linear model.

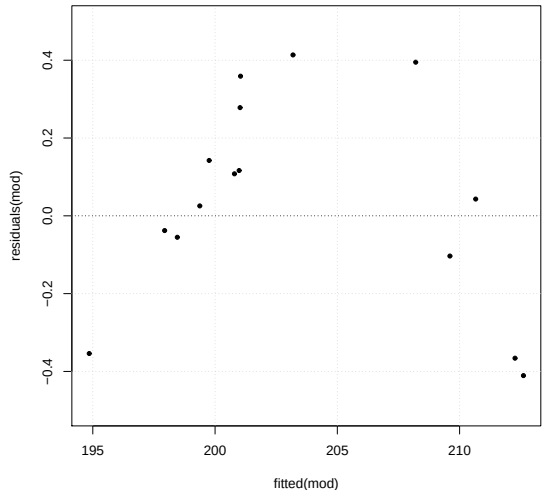


Figure 8.2.ii: Tukey-Anscombe plot of the model with the original pressure variable.

We can clearly observe that there is an unexplained structure in the Tukey-Anscombe plot of Fig. 8.2.ii which cannot be random. This tells us that the first model is not good.

On the other hand, this non random structure of the residuals disappears if we take the second model with the log-transformed pressure variable, cf. Fig. 8.2.iv. Therefore the second model describes the data better.

Model assumption number two states that the errors should have mean zero. This implies that the residuals in any interval of the Tukey-Anscombe plot should vary randomly around the horizontal line at zero. To check this assumption it is best to smooth the data in the Tukey-Anscombe plot first. A very useful and robust smoothing procedure is `loess`. This smoother fits a polynomial curve determined by one or more numerical predictors, using local fitting, cf. [35].
It is clear that this smoother will somehow vary around the horizontal line at zero, cf. Fig. 8.2.v. But we are still left with the question: Is such a curved curve due to chance possible? Or is it a real deviation that we might make disappear by improving the model?

An informal method to find answers is based on **bootstrap simulations**, cf. [11]:

0. Step The best model $y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$ with $\mathrm{Var}(\varepsilon) = \hat{\sigma}^2$ is estimated and the smooth curve of the residuals in the Tukey-Anscombe plot is calculated.

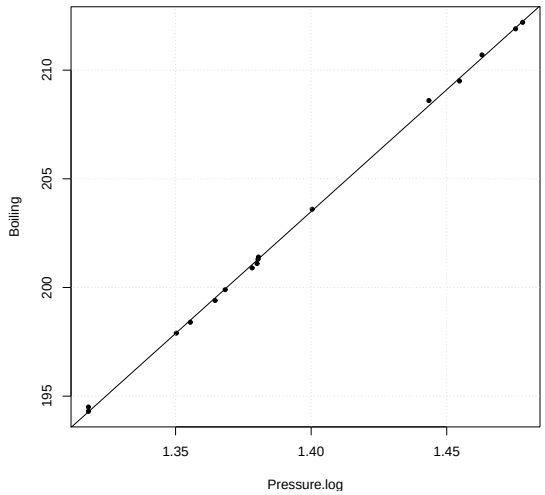


Figure 8.2.iii: Boiling point versus logarithm of pressure with best linear model.

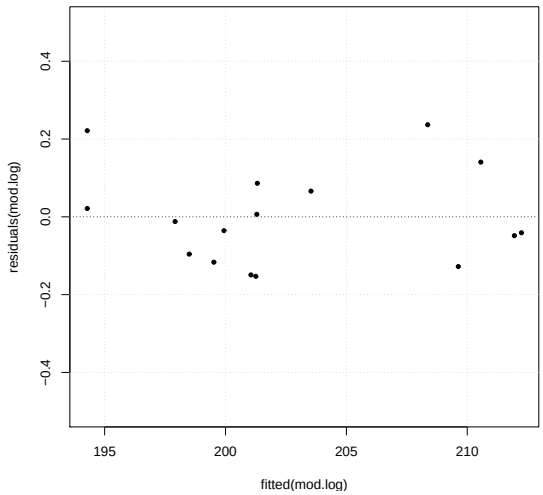


Figure 8.2.iv: Tukey-Anscombe plot with log-transformed pressure.

1. Step Simulate new observations y_i^* which correspond to the fitted model in the 0th step: Draw n random numbers e_i^* from a standard normal distribution and create the simulated observations

$$y_i^* = \hat{y}_i + \hat{\sigma} e_i^*,$$

where $\hat{y}_i$ is the ith fitted value and $\hat{\sigma}$ the estimated error sum of squares.

2. Step Fit the same model to the simulated data consisting of the original explanatory variable x and the simulated response variable y^*. Calculate the smoothed residuals of the Tukey-Anscombe plot and add it to the diagram.

3. Step Repeat the 2nd step $m_{\text{sim}} = 19$ times[4].

The generated curves are due to random fluctuations in the error term. The model values y_i^* follow exactly the linear regression model, which was estimated from the data. The generated curves thus show the inherent stochastic fluctuation in the smooth curve. Now the eye's ability to recognise patterns is used to informally judge whether the curve in the original Tukey-Anscombe plot looks more extreme than the other 19 simulated curves. We should not only pay attention to whether the original smooth curve remains within the bandwidth of the simulated curves, also an untypical shape might give us a hint.

Example 8.2.2 (**Chemical Reaction, Dissociation Pressure**)**.** We want to come back to Ex. 8.1.1 and want to check the model assumptions. We look at the Tukey-Anscombe plot in Fig. 8.2.v.
The Tukey-Anscombe plot with the original model is easily created with R and the generic plot-command.

```
#   Tukey-Anscombe plot with R-command
> plot(mod, which=1, pch=20)
```

[4] A statistical test on the $5\% = \frac{1}{20}$ level is rejected if 1 out of $1 + 19 = 20$ observations is extreme cf. [7].

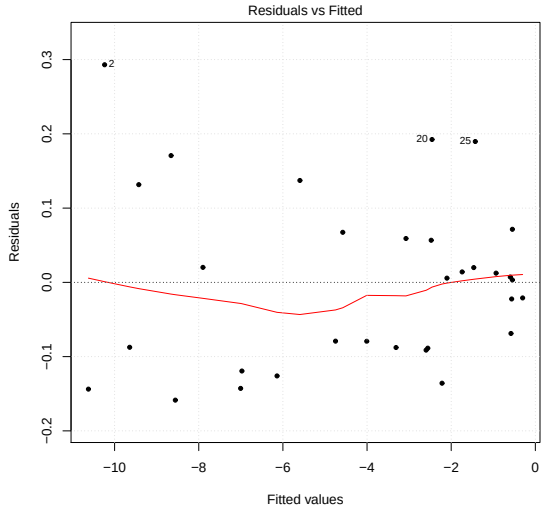


Figure 8.2.v: Tukey-Anscombe plot with smoothed residuals (red).

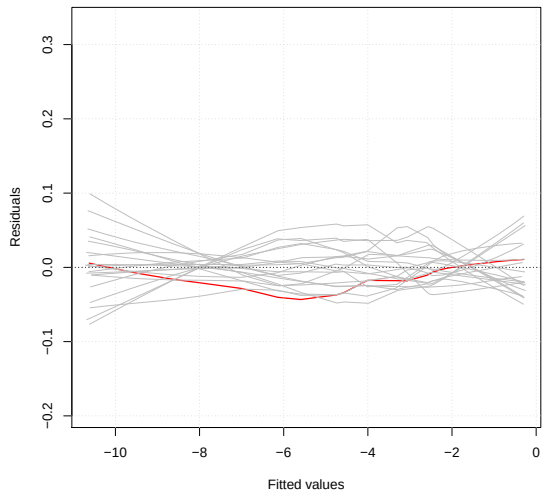


Figure 8.2.vi: Tukey-Anscombe plot with 19 simulated smoothed residuals (gray).

```
> grid()
> abline(h=0, lty=3)
```

To simulate 19 other curves we use the function `plot.lmSim.R` written by A. Ruckstuhl.

```
#   load R-function to do residual analysis (written by A. Ruckstuhl)
> source("RFn_Plot-lmSim.R")
#   Tukey-Anscombe plot with simulated data
> plot.lmSim(mod, which=1, SEED=1)
> grid()
> abline(h=0, lty=3)
```

To make the Fig. 8.2.vi easier to read we omitted the individual residuals. We can observe that although the smoothed original residuals show a clear curvature it is not an extreme curve among all 20 curves. Therefore we conclude that this curvature is not critical and the model fits the data well.

8.2.2 Scale-Location Plot

The second diagnostic tool is the **scale-location plot**. Model assumption number three states that the errors should have equal variance.
In analogy to the Tukey-Anscombe plot in Chap. 8.2.1 we now investigate a scatter diagram with the square-root of the absolute standardised residuals $\tilde{e}_{\text{std},i}$ versus the fitted values $\hat{y}_i$. If the 3rd model assumption is true, then the smoothed curve of the square-root of the standardised residuals should be approximately horizontal.

Example 8.2.3 (**Chemical Reaction, Dissociation Pressure**)**.** We want to continue with Ex. 8.1.1 and want to check the 3rd model assumptions about the constant variance of the errors. We look at the scale-location plot in Fig. 8.2.vii.
The scale-location plot with the original model is easily created with R and the generic plot-command.

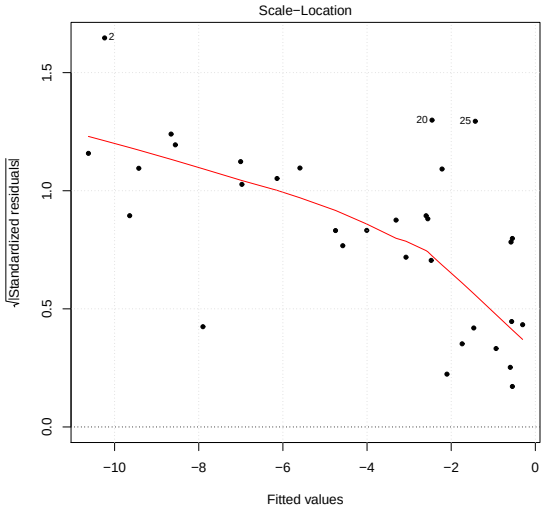


Figure 8.2.vii: Scale-location plot with smoothed residuals (red).

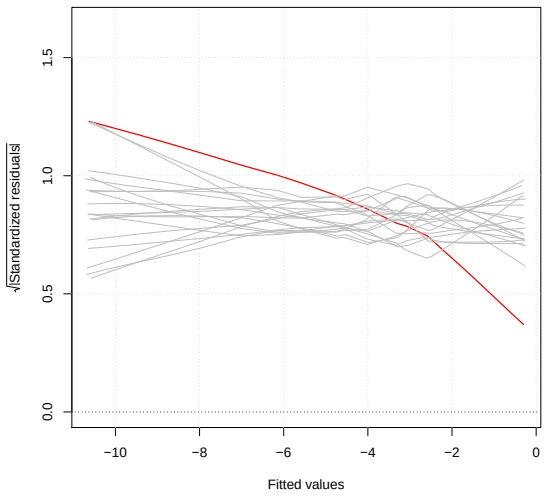


Figure 8.2.viii: Scale-location plot with 19 simulated smoothed residuals (gray).

```
#   scale-location plot with R-command
> plot(mod, which=3, pch=20)
> grid()
> abline(h=0, lty=3)
```

To simulate 19 other curves we use the function `plot.lmSim.R` written by A. Ruckstuhl.

```
#   scale-location plot with simulated data
> plot.lmSim(mod, which=3, SEED=1)
> grid()
> abline(h=0, lty=3)
```

In Fig. 8.2.vii we observe a clearly downward trend in residual scattering. The simulation in Fig. 8.2.viii confirms the extraordinary behavior. Therefore the variance of the errors is not constant and the 3rd model assumption violated.

The smoothed curve of the square root of the standardised residuals does not tend to the standard deviation for an infinite number of normally distributed observations. Since we are not interested in the standard deviation itself, but only in a possible change in the variance over different ranges of fitted values, we can nevertheless work with this diagram.

8.2.3 Normal q-q Plot (and Histogram)

The third diagnostic tool is the **normal q-q Plot**. We will check model assumption number five about the normality of the errors graphically.
The first idea is to look at the empirical histogram of the residuals e_i which are the realisations of the errors and compare it with the normal density function with parameters 0 and $\hat{\sigma}^2$. There are some disadvantages of this procedure, cf. Fig. 8.2.ix:

- It is often difficult to compare a histogram with a bell shaped curve.

- The histogram is very sensitive to the number of histogram cells and the breakpoints between cells.

The second idea is to use the **normal q-q plot**. In the q-q plot the quantiles of the empirical distribution of the of the standardised residuals $\tilde{e}_{\text{std},i}$ are plotted versus the quantiles of the normal distribution. If the data is from a normal distribution, the points in the q-q plot scatter around a straight line.

Example 8.2.4 (**Chemical Reaction, Dissociation Pressure**)**.** We want to continue with Ex. 8.1.1 and want to check the 5th model assumptions about the normality of the errors. First we look at the histogram in Fig. 8.2.ix.

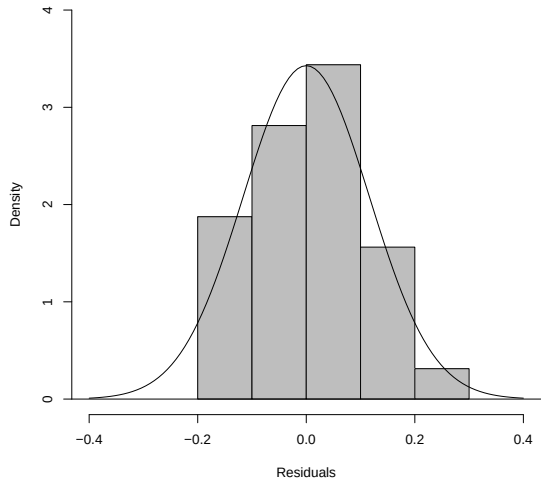


Figure 8.2.ix: Histogram of residuals with normal density function with mean zero and estimated standard deviation $\hat{\sigma}$.

It seems that the histogram does at least not contradict the assumption about the normality of the errors. To get a more precise statement using a histogram we would need to increase the number of bins, which is in this example not possible due to the relativ small sample size of 32 observations.

```
#   histogram
> hist(residuals(mod), freq=F, breaks=5,
+      col="gray", xlim=0.4*c(-1,1), ylim=c(0,4), xlab="Residuals")
> curve(dnorm(x,mean=0,sd=summary(mod)$sigma), add=T)
> abline(h=0)
```

Second we look at the q-q plot in Fig. 8.2.x.
The q-q plot with the original model is easily created with R and the generic plot-command.

```
#   q-q plot with R-command
> plot(mod, which=2, pch=20)
> grid()
```

To simulate 19 q-q plots of standardised residuals we use the function `plot.lmSim.R` written by A. Ruckstuhl.

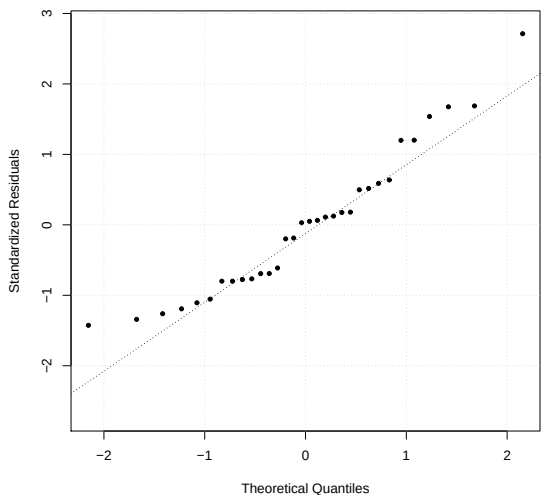


Figure 8.2.x: q-q plot with standardised residuals.

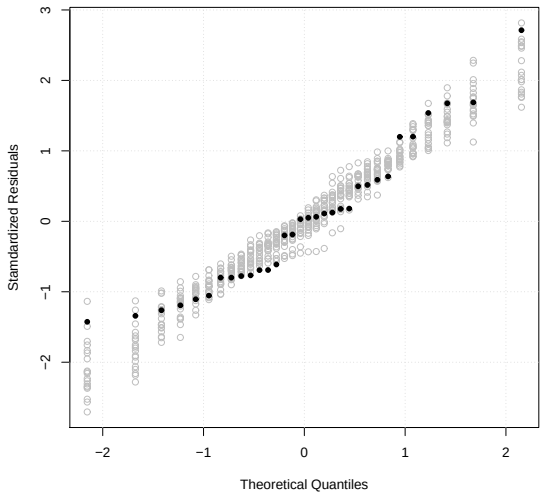


Figure 8.2.xi: q-q plot with 19 simulated standardised residuals (gray).

```
#   q-q plot with simulated data
> plot.lmSim(mod, which=2, SEED=1)
> grid()
```

In the normal q-q plot, cf. Fig. 8.2.x, we can observe a discrepancy to normality: on the left we have a short-tailed distribution (points lie above reference line) and on the right a slightly long-tailed distribution (points lie above reference line).
The simulation in Fig. 8.2.xi shows us that this discrepancy might be due to random fluctuations. Therefore the normality of the errors can be assumed and the 5th model assumption is satisfied.

Remark. Note that a normal q-q plot is completely meaningless for the response y, since all y_i have different expected values.

8.3 Treatment of Model Violations

With these diagnostic tools, we can characterise inadequacies, investigate their causes and, if possible, eliminate them. However, it is easier to assume that there are frequent violations of the model assumptions and characterise the effects in our diagnostic instruments.

8.3.1 Non-Constant Variance of Random Errors

Often the assumption is violated that the variance of random errors is constant. A dependency of the variance on the values of the response variable can be seen in our diagnostic tools, cf. Fig. 8.3.i: The model violation is visible in the scale-location plot as clear increase of the variance and in the q-q plot, where we can observe a right tailed distribution of the errors.

This model violation can be changed by a transformation of the response variable. If the standard deviation of the residuals is more or less proportional to the response variable, then

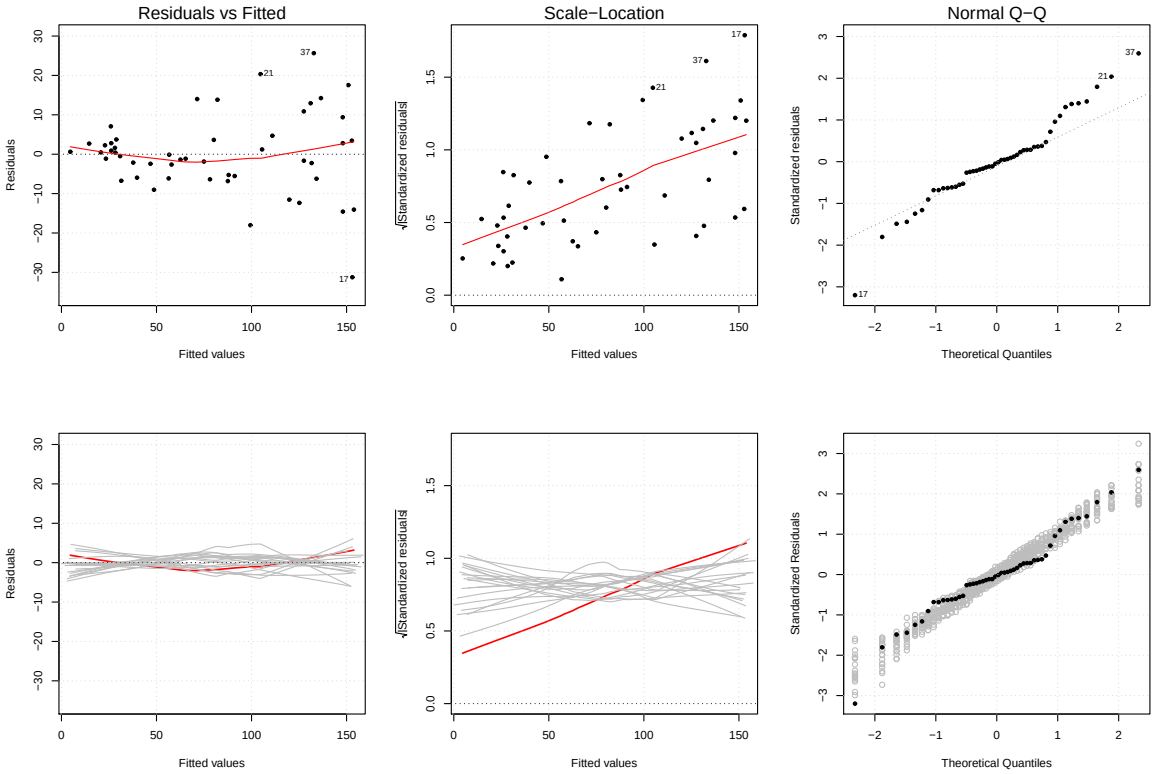


Figure 8.3.i: Residual analysis: Artificial data with error proportional to the response variable. Model violation is clearly visible in the scale-location plot and the q-q plot.

a logarithmic transformation of the response variable might help. If this is a too strong transformation then a square-root transformation might be more adequate.

J. W. Tukey gave the name to the following very useful **first aid transformations**.

- **Logarithmic transformation** for continuous positive variables:

$$z \mapsto \log(z)$$

- **Square-root transformation** for continuous and discrete positive variables:

$$z \mapsto \sqrt{z}$$

- **Arcsine transformation** for proportions:

$$z \mapsto \arcsin(\sqrt{z})$$

- **Logit-transformation** for proportions:

$$z \mapsto \log\left(\frac{z+\epsilon_1}{1+\epsilon_2-z}\right)$$

  where ϵ_1 and ϵ_2 are some small nonnegative numbers so that the logarithm is defined.

They should always be applied if something is wrong with the model assumptions.

Remark 8.3.1 (**Which Logarithm?**). For the logarithm we simply wrote log, not specifying which base we use. In general the base can be exchanged by the formula

$$\log_b(x) = \frac{1}{\log_a(b)} \log_a(x). \tag{8.3.a}$$

We observe that changing the base from b to a is a multiplication with the constant $\frac{1}{\log_a(b)}$. In estimating coefficients this only affects the coefficient by a factor of $\log_a(b)$. For instance, if only the response y is transformed logarithmically, then the model

$$\log(y) = \frac{1}{\ln(10)} \ln(y) = \beta_0 + \beta_1 x + \varepsilon$$

is the same as

$$\ln(y) = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\varepsilon}$$

with $\tilde{\beta}_0 = \ln(10)\beta_0$ and $\tilde{\beta}_1 = \ln(10)\beta_1$. Similar but different formulae hold if the explanatory variable x or the response y and the explanatory variable x are transformed logarithmically.

Therefore we need not to specify the base of the logarithm. On the other hand, with physical problems, eg. acoustics, we will definitely use the logarithm with base 10. In these lecture notes we use ln for the natural logarithm and log for the logarithm to the base 10.
Notice, on the other hand, that in R and in many other softwares `log` denotes the natural logarithm and `log10` the logarithm to the base 10.

Remark 8.3.2 (**Additive and Multiplicative Error Terms**). It is important to realise that a transformation of the response variable also changes the form of the error. Under the logarithmic transformation of the response $\tilde{y} = \log(y)$ and the explanatory variable $\tilde{x} = \log(x)$ the simple linear regression model with additive error term

$$\begin{aligned} \tilde{y} &= \beta_0 + \beta_1 \tilde{x} + \varepsilon \\ \log(y) &= \beta_0 + \beta_1 \log(x) + \varepsilon \end{aligned}$$

is transformed to a multiplicative model

$$y = 10^{\beta_0 + \beta_1 \log(x) + \varepsilon} = 10^{\beta_0} 10^{\beta_1 \log(x)} 10^{\varepsilon} = 10^{\beta_0} x^{\beta_1} 10^{\varepsilon}$$

with multiplicative random error 10^{ε}.

If we work with a transformed response variable we will also get predicted values in this transformed space. To get predicted values in the original space, we need to transform the predictors back. However, eg. for the logarithmic transformation, not the expected value but the median is predicted. If the expected value is needed, then work with the actual distribution of the actual response values, which then often leads to nonlinear models.

In the Tukey-Anscombe plot we often see clear systematic deviations of the expected value. They can often be made to disappear by a transformation of the explanatory variable, eg. Ex. 8.2.1.

If no transformation of the explanatory variable is successful we can add a second additional term x^2 in the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

This is a simple quadratic linear regression[5].

8.3.2 Outliers

We have talked about outliers a couple of times already. However, the term outlier is not clearly defined. It is an observation that fits badly with a model that is suitable for most data. In the case of an univariate sample, an **outlier** is an observation that is, compared to the scattering of the data, far away from the median.

Least-squares estimate are extremely sensitive to outliers, cf. Fig. 8.3.ii and 8.3.iii.

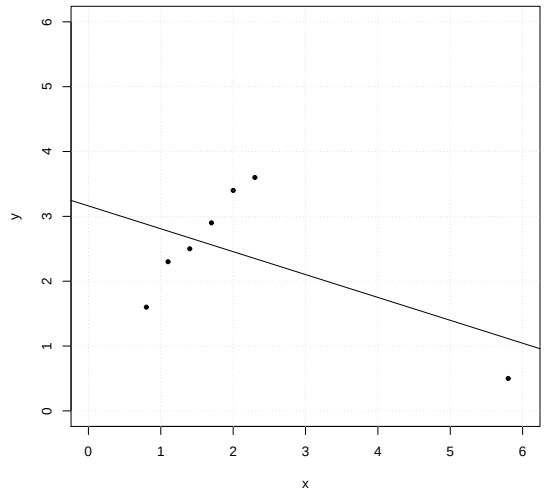


Figure 8.3.ii: Outlier in the explanatory variable. The best straight line is almost orthogonal to line through the data without outlier.

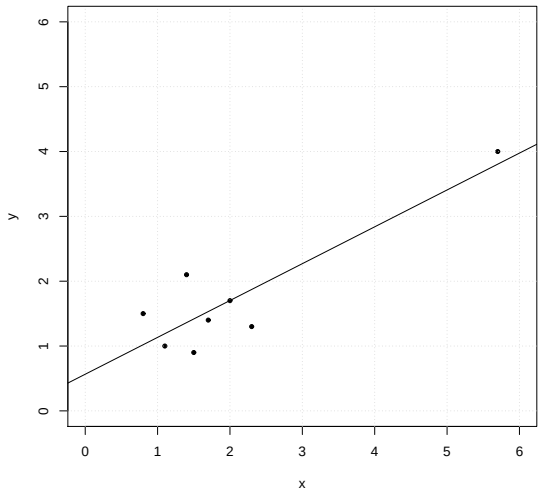


Figure 8.3.iii: Outlier in the explanatory and response variable. Without the outlier there is no relationship between x and y, but with the outlier there seems to be one.

Outliers are often clearly visible in the Tukey-Anscombe and the normal q-q plot, cf. Fig. 8.3.iv.

A single outlier can affect the smooth curve in the Tukey-Anscombe plot in a way that there is a clear trend. In Fig. 8.3.iv this is not the case. R will always mark the three most extreme points. This does not imply that these points are all suspicious. In our artificial example only observation with index $i = 37$ was manipulated.

The data should be checked for correctness in the case of outliers. If no good explanations for doubtful observations can be found, then try to transform the response and/or explanatory variables.

[5]It is simple because there is only one explanatory variable involved. It is linear in the parameters β_0, β_1 and β_2

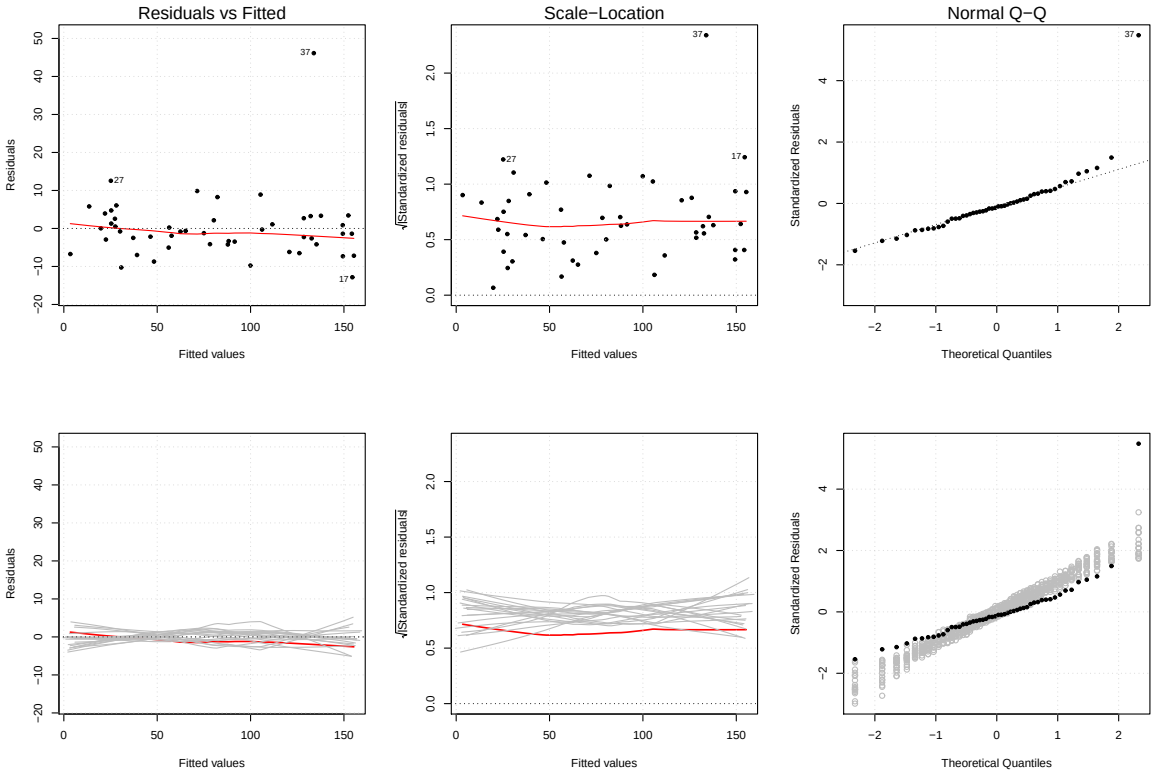


Figure 8.3.iv: Residual analysis: Artificial data with an outlier in the response variable at position $i = 37$. Model violation is clearly visible in all three diagnostic plots.

It is also possible that unexpected observations are completely random. For instance if the underlying random process is not normal but follows a long-tailed Student's t distribution with a small number of degrees of freedom.

Do not forget: Often in science the unexpected observations are the most interesting and might show some new discovery.

8.3.3 Heavy-Tailed Distributions

If the q-q plot shows a symmetric distribution but with heavy tails, cf. Fig. 8.3.v, then transformations are useless.
One possibility is to omit the most extreme observations such that the heavy tails disappear, (not more than 5% of the data). In statistics this is often called **trimming**. Results obtained with a trimmed data set should be treated with care: the error probability of statistical tests will be wrong.
In any case it is wise and also best practice to report omitted observations in a report.

Least-squares methods are not optimal with errors that follow distributions with heavy tails. In this case **robust methods** are much better. We will discuss these methods in Chap. 10.4.

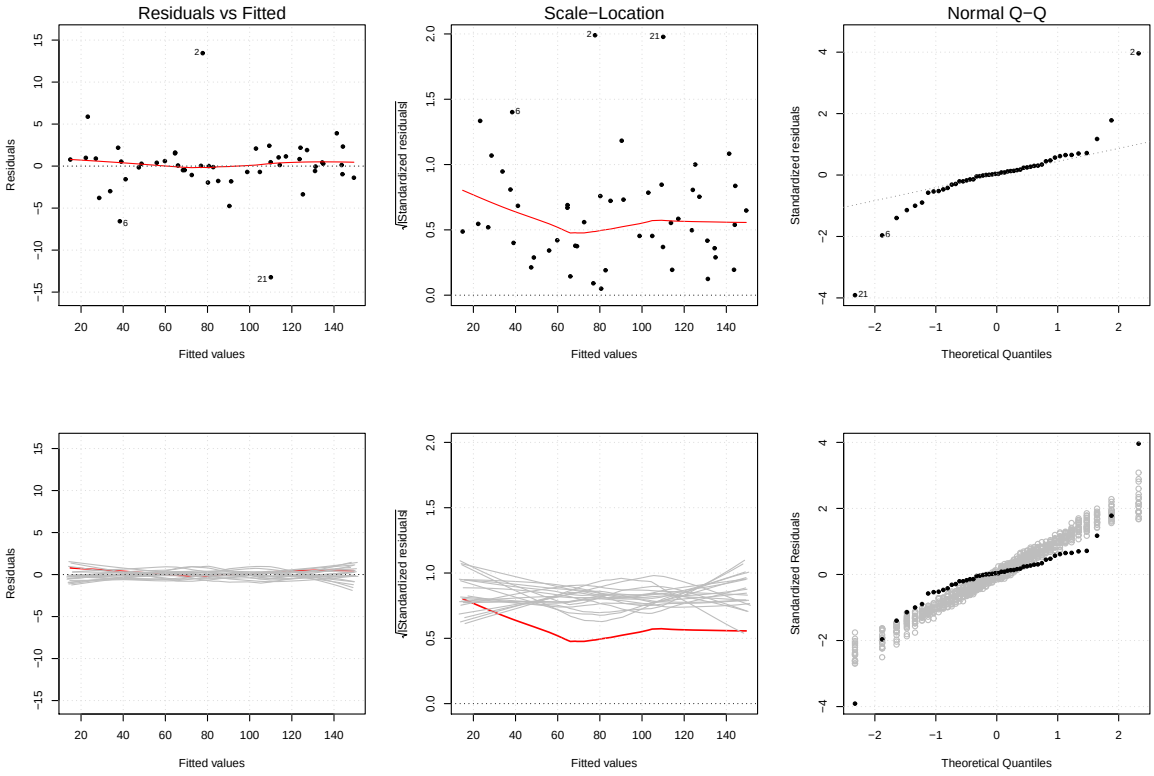


Figure 8.3.v: Residual analysis: Artificial data with errors that follow a distribution with a long tail. Model violation is clearly visible in the q-q plot.

8.4 Independence

Model assumption four about stochastic independence of the errors remains to be checked. The concept of stochastic independent random variables is important to develop theoretical probability theorems like the distribution of the regression parameters. In applied statistics it is very difficult to test. On the other hand the related but weaker assumption of uncorrelated errors is easier to check. It is important to note that independence implies uncorrelatedness but not conversely.

If the observations are ordered chronologically, then it can happen that errors are correlated, i.e. neighbor residuals e_i are more similar than residuals far apart. In such a situation we say that the errors are **autocorrelated** and the model assumption are violated. If the errors are correlated then we can see a certain pattern if we plot the residuals versus the index, i.e. time, cf. Fig. 8.4.i, upper row.

- **Negative correlation.** After a positive (negative) residual the chance of observing a negative (positive) residual is high, i.e. very strong alternation of the sign of the residuals.

- **No correlation.** The residuals are completely random, no obvious pattern present.

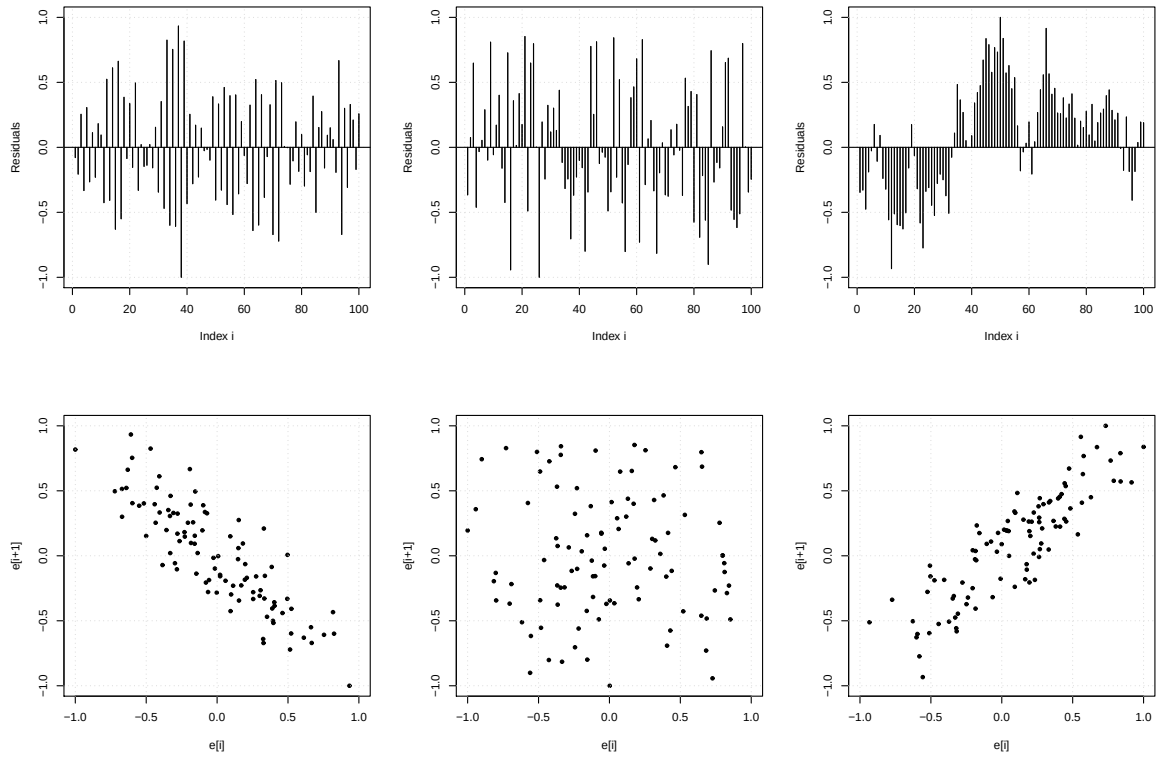


Figure 8.4.i: Detect simple correlations of lag one: Residuals versus time (upper row), e_{i+1} versus e_i, negative correlation (lower row, left), no correlation (lower row, middle), positive correlation (lower row, right).

- **Positive correlation.** After a positive (negative) residual the chance of observing a positive (negative) residual is high, i.e. the residuals show the same sign over certain periods.

Alternatively we can also plot the pairs (e_i, e_{i+1}) with $i \in \{1, \ldots, n-1\}$ in a scatter diagram, cf. Fig. 8.4.i, lower row.

- **Negative correlation.** The points lie in an ellipse with negative slope of the first principal axis.
- **No correlation.** The points show no pattern
- **Positive correlation.** The points lie in an ellipse with positive slope of the first principal axis.

The smaller is the ellipse the bigger the correlation over time.

In many cases the order in the original data set might be lost and then we cannot check for independence anymore.

If the errors are correlated then the reported P-values are grossly erroneous. Furthermore the confidence intervals are not at all reliable. Statistical methods which deal with correlations

exist and can be found under the name **generalised least squares** . In R with `gls` from package `nlme`.

In same simple cases, eg. if saisonal effects are present, a factor in the multivariate regression model might help.

Remark. Applying residual analysis to real data in real life does not always lead to a unique conclusion. Often we have to deal with bigger or smaller deviations from the model assumptions and it can be difficult to make the right decisions. It might be helpful to remember us of the following:

- The specified precision of the scientific question can help making the correct decisions.
- Every statistical test has the risk of a type I error and type II error.
- The best statistical method cannot compensate for bad data.
- Do not work alone! Ask the other non-statistical scientists involved in the project. They often know more about the background of the data and the circumstances in which the data was collected.

In the end this is all the fun of statistics and leaves us a lot of leeway to interpret and argue!

8.5 Exercises

Problem 8.5.1 (**Vending Machines**)**.** Let us look at the data `vending-machines.dat` of Ex. 7.1.1 from the softdrink vending machines.

a. Fit the simple linear model

$$\texttt{Time} = \beta_0 + \beta_1 \cdot \texttt{Volume} + \varepsilon \tag{8.5.a}$$

to the data.

b. Perform a complete residual analysis assuming that the errors are independent, normally distributed with expected value zero and constant variance. Report all your findings and conclusions.

c. Apply the following Tukey's first aid transformation:

$$\log(\texttt{Time}) = \beta_0 + \beta_1 \cdot \log(\texttt{Volume}) + \varepsilon \tag{8.5.b}$$

$$\log(\texttt{Time}) = \beta_0 + \beta_1 \cdot \sqrt{\texttt{Volume}} + \varepsilon \tag{8.5.c}$$

Perform a complete residual analysis for both variants. Report all your findings and conclusions, compare with the model in Eqn. 8.5.a.

Problem 8.5.2 (**Anscombe Quartet**)**.** In 1973 F. J. Anscombe published, cf. [1], four different data sets each containing an x- and a y-variable, cf. Ex. 8.1.3. All four artificial data sets can be found in `anscombe.dat`.

a. Fit for each of the four data sets a simple linear model. Report in a table all intercepts, slopes, corresponding standard errors, residual standard errors and coefficients of determination.

b. Reproduce Fig. 8.1.v and report all your findings and conclusions.

Problem 8.5.3 (**Understand q-q Plot with Simulation** cf. [32], Reg2, Problem 3)**.** With the following simulation you will become familiar mit the q-q plot.

a. **Normal distribution.** Draw $n = 10$, 20, 50 and 100 random numbers from a standard normal distribution and check normality with a q-q plot. Repeat this several times to get an idea of the randomness of the deviations in the q-q plot.

b. **Heavy-tailed distribution.** Draw $n = 10$ and 100 random numbers from a Student's t-distribution with 20, 7 and 3 degrees of freedom and check normality with a q-q plot. Repeat this several times to get an idea of the randomness of the deviations in the q-q plot.

c. **Skewed Distribution.** Draw $n = 10$ and 100 random numbers from a χ^2-distribution with 20 and 1 degrees of freedom and check normality with a q-q plot. Repeat this several times to get an idea of the randomness of the deviations in the q-q plot.

Problem 8.5.4 (**Windmill**)**.** We come back to Prob. 7.5.2. Investigate the following regression models:

$$\texttt{DCOutput} = \beta_0 + \beta_1 \cdot \texttt{WindVelocity} + \varepsilon \tag{8.5.d}$$

$$\log(\texttt{DCOutput}) = \beta_0 + \beta_1 \cdot \log(\texttt{WindVelocity}) + \varepsilon \tag{8.5.e}$$

$$\texttt{DCOutput} = \beta_0 + \frac{\beta_1}{\texttt{WindVelocity}} + \varepsilon \tag{8.5.f}$$

Perform a complete residual analysis for the variants. Report all your findings and conclusions. Which model is the best?

Problem 8.5.5 (**Atmospheric Pressure versus Boiling Point of Water**)**.** We come back to Prob. 7.5.3. Investigate the following regression model

$$\texttt{Boiling} = \beta_0 + \beta_1 \cdot \log(\texttt{Pressure}) + \varepsilon \tag{8.5.g}$$

in connection with the complete data set. Perform a complete residual analysis. Report all your findings and conclusions.

Problem 8.5.6 (**Study R-Function**)**.** Study the code in `RFn_Plot-lmSim.R` written by A. Ruckstuhl.

Chapter 9

Multiple Linear Regression

A **multiple regression model** is a model which involves more than one regressor variable.

To illustrate the concepts in this chapter we introduce a new example from vibration control (source of the example: [32], Beispiel b, Chap. 1.2 and source of the data: Ingenieurbüro Basler und Hoffmann, Zürich).

Example 9.0.1 (**Blasting**)**.** Building a road tunnel near a village is difficult. Blasting is one option. But it is important that the vibrations in the houses do not exceed a certain critical limit. Therefore blasting close to civilisation has to be done with lots of care causing higher costs. It would be desirable to have rules which define the charge for several situations.
In Fig. 9.0.i we can see vibrations versus the distance of 48 recorded blastings. The charge is indicated with little thermometers: the thermometers are filled proportionally to the charge, see the legend in the figure.

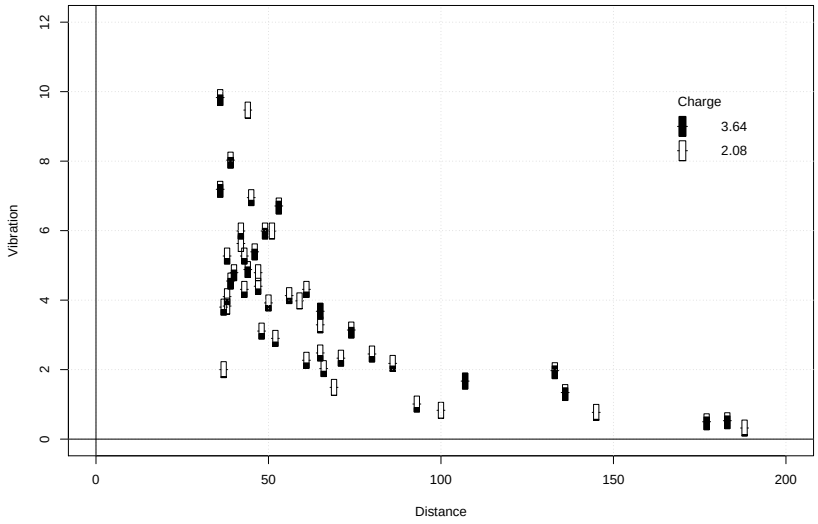


Figure 9.0.i: In blasting vibrations depend on the distance and the charge.

Vibrations which are measured in the basement of houses depend on the charge, the distance and the underground between the house and the blasting position, the blasting position itself

and several further variables. The physical and geological situations are so difficult and hardly quantifiable that a mechanistic model is not available.
We can observe that the vibrations for similar distances scatter quite a bit. There is hope that this variance can be explained with other variables than the distance. The charge might be a good candidate.

9.1 Model and Estimation

In this chapter we generalise the concept of a simple linear regression model with only one explanatory variable to a **multiple linear regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{9.1.a}$$

with m explanatory variables $x_1, x_2, \ldots, x_m$ and $m+1$ unknown coefficients β_0 and $\beta_1, \beta_2, \ldots, \beta_m$. The assumptions about the errors are the same as for the simple regression model.
The coefficient β_0 is the intercept and β_k for $k \in \{1, 2, \ldots, m\}$ is the partial derivative, i.e. the slope, in the x_k direction.

9.1.1 Least-Squares Estimation of the Regression Coefficient

We will again estimate the coefficients β_0 and $\beta_1, \beta_2, \ldots, \beta_m$ with the method of least-squares. To do this we will use matrix notation. This allows a very compact display of the model, data and results. Let us write the model for each measurement $(x_{1i}, \ldots, x_{mi}, y_i)$ with $i \in \{1, 2, \ldots, n\}$ in the following form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \varepsilon_i = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ik} + \varepsilon_i. \tag{9.1.b}$$

The following formulas only hold for a model with an intercept[1]. Now we define[2] the $n \times 1$ vector $\mathbf{y}$ of the observations and the $n \times (m+1)$ **design matrix** $\mathbf{X}$ of the levels of the explanatory variables

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}. \tag{9.1.c}$$

In addition we define the $n \times 1$ vector $\boldsymbol{\varepsilon}$ of the errors and the $(m+1) \times 1$ vector $\boldsymbol{\beta}$ of the unknown parameters

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}.$$

[1] If the model has no intercept, then $\mathbf{X}$ is only a $n \times m$ matrix without the first column and $\boldsymbol{\beta}$ an $m \times 1$ vector without the first entry.

[2] Vectors and matrices are denoted with bold letters.

The model in Eqn. 9.1.b in matrix notation is now as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Now we want to find the vector of least-squares estimators $\boldsymbol{\beta}$, which minimise

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The sum is over all measurements i and t denotes matrix transposition. With a bit of matrix algebra we can write

$$S(\boldsymbol{\beta}) = \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}.$$

The least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ must now satisfy

$$\frac{\partial S}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = -2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = 0,$$

which simplifies to the so-called **least-squares normal equations**

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}}.$$

The solution of the normal equations is the least-squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{9.1.d}$$

in matrix notation. If the regressors are linearly independent, i.e. no column in $\mathbf{X}$ is a linear combination of the other columns, then the inverse of $\mathbf{X}^t\mathbf{X}$ exists.

It is now straight forward to calculate the fitted values

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{9.1.e}$$

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is usually called the **hat matrix**. It maps the vector of observed values $\mathbf{y}$ into a vector of fitted values $\hat{\mathbf{y}}$. It puts $\mathbf{y}$ the hat on!

The $n \times 1$ vector of residuals $\mathbf{e}$ can be written as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Example 9.1.1 (**Blasting**)**.** In Ex. 9.0.1 the vibrations not only depend on the distance but also on the charge. Observing Fig. 9.0.i it is obvious that a good model cannot rely on the original variables. We try logarithmically transformed variables. A multiple linear regression model for this situation might be

$$\log(\texttt{Vibration}) = \beta_0 + \beta_1 \cdot \log(\texttt{Distance}) + \beta_2 \cdot \log(\texttt{Charge}) + \varepsilon. \tag{9.1.f}$$

First we read and log-transform (to the base 10) the data.

```
#   read data
file <- "blasting.dat"
data <- read.table(file, header=TRUE)
str(data)
'data.frame':   48 obs. of  4 variables:
 $ Position : int  1 1 1 2 2 2 2 2 2 2 ...
 $ Charge   : num  2.18 3.33 3.33 3.33 3.12 3.12 3.12 2.6 2.08 2.08 ...
 $ Distance : int  188 183 177 53 49 46 44 42 42 44 ...
 $ Vibration: num  0.32 0.53 0.5 6.71 5.99 5.39 4.88 5.99 5.63 9.47 ...

#   define new log-transformed variables
data$Charge.log    <- log10(data$Charge)
data$Distance.log  <- log10(data$Distance)
data$Vibration.log <- log10(data$Vibration)
```

Then we want to estimate the parameters using the explicit formulae in Eqn. 9.1.d.

```
#   estimation of the parameters (explicit formulae)
> n <- nrow(data);   n
[1] 48
#   vector of the observations
> y <- as.vector(data$Vibration.log)
> length(y)
[1] 48
#   matrix of the levels of the regressor variables
> X <- as.matrix(cbind(Intercept=rep(1,n),
+                      data[,c("Distance.log", "Charge.log")]))
> dim(X)
[1] 48  3
> X
      Intercept Distance.log Charge.log
 [1,]         1     2.274158  0.3384565
 [2,]         1     2.262451  0.5224442
 [3,]         1     2.247973  0.5224442
 ...
[48,]         1     1.579784  0.4149733

#   least-squares estimator
> beta.hat <- solve(t(X)%*%X) %*% t(X) %*% y;   beta.hat
                   [,1]
Intercept     2.8322552
Distance.log -1.5107128
Charge.log    0.8083438
```

It is of course much more comfortable to rely on the R-command `lm`. We need to learn the notation of multiple models in R. In this example we write the model in Eqn. 9.1.a as follows:

```
Vibration.log ~ Distance.log + Charge.log
```

Note that the intercept and the error term are omitted and the plus sign adds additional explanatory variables to the model.

```
#   estimation of the parameters (lm)
```

```
> mod <- lm(Vibration.log ~ Distance.log + Charge.log, data)
> summary(mod)
Call:
lm(formula = Vibration.log ~ Distance.log + Charge.log, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43571 -0.11580  0.00227  0.09715  0.36978

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.8323     0.2229  12.707   <2e-16 ***
Distance.log  -1.5107     0.1111 -13.592   <2e-16 ***
Charge.log     0.8083     0.3042   2.658   0.0109 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3521 on 45 degrees of freedom
Multiple R-squared:  0.8048,    Adjusted R-squared:  0.7962
F-statistic: 92.79 on 2 and 45 DF,  p-value: < 2.2e-16
```

9.1.2 Tests on the Significance of any Individual Regression Coefficient

It is possible to show (cf. [24], Chap. 3.2.3) that $\hat{\boldsymbol{\beta}}$ follows an $(m+1)$-dimensional multivariate normal distribution with expected value

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

and $(m+1) \times (m+1)$ covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}.$$

Using the above we can derive a test procedure and corresponding confidence intervals for each coefficient β_0 and β_k for $k \in \{1, 2, \ldots, m\}$. The hypothesis for testing the significance of any individual regression coefficient, such as β_k, are

$$H_0 : \beta_k = \beta_{k,0}$$
$$H_1 : \beta_k \neq \beta_{k,0}.$$

The test statistic for this hypothesis is

$$T_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\mathrm{se}(\hat{\beta}_k)} \quad \text{with} \quad \mathrm{se}(\hat{\beta}_k) = \hat{\sigma} \sqrt{((\mathbf{X}^t \mathbf{X})^{-1})_{kk}}, \tag{9.1.g}$$

where $((\mathbf{X}^t \mathbf{X})^{-1})_{kk}$ is the kth diagonal element of the matrix $(\mathbf{X}^t \mathbf{X})^{-1}$. Under the null hypothesis the test statistic T_k follows a Student's t-distribution with $n - (m+1)$ degrees of freedom, cf. T.3. The degrees of freedom equals the denominator of

$$\hat{\sigma}^2 = \frac{1}{n-(m+1)} \sum_{i=1}^{n} e_i^2 = \frac{\mathbf{e}^t \mathbf{e}}{n-(m+1)},$$

where $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is the $n \times 1$ vector of residuals.

If $H_0 : \beta_k = 0$ is not rejected, then this indicates that the regressor x_k can be omitted from the model.

Example 9.1.2 (**Blasting**). We come back the Ex. 9.1.1 and want to test the hypothesis for the significance of any individual regression coefficient.

We want to estimate the test statistics T_0, T_1 and T_2 using the explicit formulae in Eqn. 9.1.g. Of course we could also use the summary R-output of the `lm`-object. We will definitely do that in future, but now we want to learn about multiple regression and want to understand the formulae.

```
#   hat matrix
> H <- X %*% solve(t(X)%*%X) %*% t(X)
> dim(H)
[1] 48 48
#   fitted values
> Vibration.log.hat <- H %*% y
#   residuals
> res <- data$Vibration.log - Vibration.log.hat
#   error sum of squares
> sigma2.hat <- sum(res^2)/(nrow(data)-3)
#   residual standard error
> sqrt(sigma2.hat)
[1] 0.1529102
#   three standard errors
> se.beta <- sqrt(sigma2.hat) * sqrt(diag(solve(t(X)%*%X)));   se.beta
Intercept Distance.log   Charge.log
0.2228918    0.1111472    0.3041724

#   null hypothesis
> beta0 <- rep(0,3)
#   three test statistics
> Test.beta <- (beta.hat-beta0)/se.beta;   Test.beta
                   [,1]
Intercept     12.706862
Distance.log -13.592000
Charge.log     2.657519

#   three P-values
> 2*(1-pt(abs(Test.beta), df=n-3))
                     [,1]
Intercept    2.220446e-16
Distance.log 0.000000e+00
Charge.log   1.085725e-02
```

We conclude that both variables `Distance.log` and `Charge.log` have a significant influence on the response variable `Vibration.log`. In addition the intercept cannot be omitted in the model.

9.1.3 Confidence Intervals on the Regression Coefficients

The corresponding $100(1-\alpha)$ percent **confidence interval** on β_k is

$$\hat{\beta}_k - t_{\frac{\alpha}{2},n-(m+1)} \cdot \mathrm{se}(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{1-\frac{\alpha}{2},n-(m+1)} \cdot \mathrm{se}(\hat{\beta}_k), \tag{9.1.h}$$

where $\mathrm{se}(\hat{\beta}_k)$ can be found in Eqn. 9.1.g and $t_{p,n-(m+1)}$ denotes the critical value of Student's t-distribution to the probability p and with $n-(m+1)$ degrees of freedom, cf. T.3.

Example 9.1.3 (**Blasting**). We look at the model

$$\log(\texttt{Vibration}) = \beta_0 + \beta_1 \cdot \log(\texttt{Distance}) + \beta_2 \cdot \log(\texttt{Charge}) + \varepsilon,$$

cf. Eqn. 9.1.f, which was fitted to the data of the blasting measurements.

Mathematically this model is equivalent to

$$\begin{aligned} \texttt{Vibration} &= 10^{\beta_0+\beta_1\cdot\log(\texttt{Distance})+\beta_2\cdot\log(\texttt{Charge})+\varepsilon} \\ &= 10^{\beta_0}10^{\beta_1\cdot\log(\texttt{Distance})}10^{\beta_2\cdot\log(\texttt{Charge})}10^{\varepsilon} \\ &= 10^{\beta_0}\texttt{Distance}^{\beta_1}\texttt{Charge}^{\beta_2}10^{\varepsilon}. \end{aligned}$$

First principles from physics suggest that the spreading of a blasting wave follows a law with

$$E \sim r^{-2},$$

where E is an energy and r a radius. The exponent -2 corresponds to the unimpeded propagation of the energy in three dimensions. The energy is then inversely proportional to the spherical surface and thus to the squared radius.

Therefore we want to test the Hypothesis

$$\begin{aligned} H_0 &: \beta_1 = -2 \\ H_1 &: \beta_1 \neq -2. \end{aligned}$$

We check if the value -2 is in the 95% confidence interval on β_1. We calculate the three confidence intervals of the regression coefficients using the explicit formulae

```
#   significance level
alpha <- 0.05
#   three 95% confidence intervals (explicit formulae)
> t.crit <- qt(p=1-alpha/2, df=nrow(data)-3);   t.crit
> CI <- cbind(beta.hat - t.crit*se.beta, beta.hat + t.crit*se.beta)
> colnames(CI) <- c("lower CI", "upper CI");   CI
               lower CI  upper CI
Intercept     2.3833281  3.281182
Distance.log -1.7345748 -1.286851
Charge.log    0.1957093  1.420978
```

and the summary R-output of the `lm`-object.

```
#   three 95% confidence intervals (lm)
> confint(mod, level=1-alpha)
                  2.5 %    97.5 %
(Intercept)   2.3833281  3.281182
Distance.log -1.7345748 -1.286851
Charge.log    0.1957093  1.420978
```

We realise that $-2 \notin [-1.7345748, -1.286851]$ which contradicts the naive physical concept. If the energy is reflected at certain layers in the underground, then a less pronounced decrease with distance is plausible.

9.1.4 Confidence Interval of the Response

We want to construct a confidence interval on the response at a particular point

$$(x_{01}, x_{02}, \ldots, x_{0m}).$$

Define the $(m+1) \times 1$ vector

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0m} \end{pmatrix} \tag{9.1.i}$$

and calculate the estimated value

$$\begin{aligned} \hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_m x_{0m} \\ &= \hat{\beta}_0 \cdot 1 + \sum_{k=1}^{m} \hat{\beta}_k x_{0k} \\ &= \mathbf{x}_0^t \hat{\boldsymbol{\beta}}. \end{aligned}$$

This is an unbiased, normally distributed estimator with variance

$$\mathrm{Var}(\hat{y}_0) = \sigma^2 \mathbf{x}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_0,$$

cf. [24], Chap. 3.4.2.

Therefore a $100(1-\alpha)$ percent **confidence interval on the response** at the point

$$(x_{01}, x_{02}, \ldots, x_{0m})$$

is

$$\hat{y}_0 - t_{\frac{\alpha}{2}, n-(m+1)} \cdot \mathrm{se}(\hat{y}_0) \leq \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \leq \hat{y}_0 + t_{1-\frac{\alpha}{2}, n-(m+1)} \cdot \mathrm{se}(\hat{y}_0) \tag{9.1.j}$$

with

$$\mathrm{se}(\hat{y}_0) = \hat{\sigma} \sqrt{\mathbf{x}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_0} \tag{9.1.k}$$

and where $t_{p,n-(m+1)}$ denotes the critical value of Student's t-distribution to the probability p and with $n-(m+1)$ degrees of freedom, cf. T.3.

Example 9.1.4 (**Blasting**). We are interested in blasting at a `Distance` = 50 with a `Charge` = 3.0. The estimated value of vibration with a 95% confidence interval on the response can easily be calculated with R.

```
#   define new point
> data.new <- data.frame(Distance.log=log10(50), Charge.log=log10(3.0))
#   predict using interval="confidence"
> Vibration.log.Conf <- predict(mod, newdata=data.new,
+                               interval="confidence", level=0.95)
> 10^Vibration.log.Conf
      fit      lwr     upr
4.479994 3.910365 5.132602
```

To obtain a vibration in the original scale we had to undo the logarithm.

9.1.5 Prediction Interval

The regression model can be used to predict future observations on y corresponding to particular values of regressor variables, for example, $(x_{01}, x_{02}, \ldots, x_{0m})$. A point estimate of the future observation is

$$\hat{y}_0 = \mathbf{x}_0^t \hat{\boldsymbol{\beta}},$$

where $\mathbf{x}_0$ was defined in Eqn. 9.1.i.

A $100(1-\alpha)$ percent **prediction interval** for the future observation

$$(x_{01}, x_{02}, \ldots, x_{0m})$$

is

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-(m+1)} \cdot \sqrt{\hat{\sigma}^2 + \mathrm{se}(\hat{y}_0)^2} \leq \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \leq \hat{y}_0 + t_{1-\frac{\alpha}{2},n-(m+1)} \cdot \sqrt{\hat{\sigma}^2 + \mathrm{se}(\hat{y}_0)^2}, \qquad (9.1.l)$$

where $\mathrm{se}(\hat{y}_0)$ was defined in Eqn. 9.1.k and $t_{p,n-(m+1)}$ denotes the critical value of Student's t-distribution to the probability p and with $n-(m+1)$ degrees of freedom, cf. T.3.

Example 9.1.5 (**Blasting**)**.** The limit of vibration in a house is 7. We are interested if blasting at a `Distance` = 50 with a `Charge` = 3.0 will be within this limit. Therefore we calculate the 95% prediction interval (one-sided) and are only interested in the upper bound of the 90% prediction interval (two-sided).

```
#   predict using interval="prediction" (one-sided, therefore level=0.9)
> Vibration.log.Pred <- predict(mod, newdata=data.new,
+                               interval="prediction", level=0.9)
> 10^Vibration.log.Pred
     fit     (lwr)      upr
4.479994 2.453559 8.180095
```

Since $8.180095 > 7$ we conclude that highly likely blasting with these parameters will exceed the limit.

9.1.6 Coefficient of Determination - Multiple R-Squared

In multiple regression the **coefficient of determination**, often called **multiple** R^2, is identical to the squared correlation between the response variable y and the fitted values $\hat{y}$

$$R^2 = \mathrm{Cor}(y, \hat{y})^2.$$

The coefficient of determination is a measure of the linear relationship between the response variable and the fit.

Example 9.1.6 (**Blasting**)**.** We will check goodness of fit calculating the multiple R^2 of our model.

```
#   multiple R-squared (explicit formulae)
> cor(data$Vibration.log, fitted(mod))^2
[1] 0.8048385
#   multiple R-squared (lm)
> summary(mod)$r.squared
[1] 0.8048385
```

Therefore our model is capable of explaining 80.48% of the variance in the response variable and for the remaining 19.52% we have at the moment no explanation.

9.2 Diversity of Modelling Possibilities

In the model of multiple regression, no assumptions are made about the explanatory variables. In all our examples, they were always continuous variables, but that is not necessary. Therefore, they do not have to be of a particular data type, let alone follow a particular distribution, because they are not considered as random variables. Nothing is assumed about their dependence on each other. In principle, it is therefore also permitted that explanatory variables are strongly (linearly) correlated. However, this leads to difficulties in interpreting the results of the regression analysis and should be avoided if possible.

9.2.1 Polynomial Regression

The polynomial function is an important function for formulating curved relationships. If we want to use a polynomial of degree two to describe a relationship, then the model has the following form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

Because only one explanatory variable is involved we call this model a simple **quadratic regression**.

On the other hand, if we define the new variables $\texttt{lin} = x$ and $\texttt{quad} = x^2$, then we obtain mathematically the same model

$$y = \beta_0 + \beta_1 \cdot \texttt{lin} + \beta_2 \cdot \texttt{quad} + \varepsilon,$$

which is now a multiple linear regression model. In the same way, higher powers can be introduced, resulting in a **polynomial regression**.

Example 9.2.1 (**Stopping Distance**)**.** The kinetic energy of a moving mass is proportional to its velocity squared. Therefore a reasonable model to describe the relationship between the stopping distance d of a car and its final velocity v is

$$d = \beta_2 v^2 + \varepsilon.$$

This is a quadratic regression model with missing intercept and missing linear term. The corresponding R-syntax of such a model is:

```
Distance ~ -1 + I(Velocity^2)
```

Notice that `-1` removes the intercept term and the function`I()` inhibits the interpretation of the power operator as formula operator, so it is used as an arithmetical operator[3].

Remark. The word *linear* in the expression multiple linear regression means that the unknown coefficients β_0 and $\beta_1, \beta_2, \ldots, \beta_m$ are linear in the model equation. It does not relate to the relationship between response and explanatory variables.

[3]When arithmetic expressions involve operators which are also used symbolically in model formulae, there can be confusion between arithmetic and symbolic operator use. To avoid this confusion, the function `I()` can be used to bracket those portions of a model formula where the operators are used in their arithmetic sense.

9.2.2 Nonlinear Functions and Linear Regression

Experience shows that in most real-world regression problems, the variables must be transformed to get the best results. In Chap. 8.3 we presented Tukey's useful first aid transformations.

Especially mechanistic models are most often nonlinear functions in the explanatory variable x. In some cases these models can be linearised to obtain linear regression models. We will give four examples:

- Inverting the nonlinear equations

$$y = \frac{1}{a + b\,\mathrm{e}^{-x}} \qquad \text{gives} \qquad \frac{1}{y} = a + b\,\mathrm{e}^{-x}$$
$$y = \frac{ax}{b + x} \qquad \text{gives} \qquad \frac{1}{y} = \frac{1}{a} + \frac{b}{a}\frac{1}{x}.$$

  In the second example we will also substitute the coefficients with $\tilde{a} = \frac{1}{a}$ and $\tilde{b} = \frac{b}{a}$. After the estimation is done we will substitute the estimators $\hat{\tilde{a}}$ and $\hat{\tilde{b}}$ back to obtain estimates of the original coefficients $\hat{a} = \frac{1}{\hat{\tilde{a}}}$ and $\hat{b} = \hat{a}\hat{\tilde{b}} = \frac{\hat{\tilde{b}}}{\hat{\tilde{a}}}$.

- Taking logarithms of the nonlinear equations

$$y = a\,x^b \qquad \text{gives} \qquad \ln(y) = \ln(a) + b\ln(x)$$
$$y = a\,\mathrm{e}^{b\,g(x)} \qquad \text{gives} \qquad \ln(y) = \ln(a) + b\,g(x),$$

  where g is any function. In both examples we will substitute the coefficient $\tilde{a} = \ln(a)$. After the estimation is done we will substitute the estimator $\hat{\tilde{a}}$ back to obtain the estimate of the original coefficients $\hat{a} = \mathrm{e}^{\hat{\tilde{a}}}$.

It is important to realise that also such transformations change the form of the error term. If we are not allowed to alter the original additive error term, then we need to use **nonlinear least-squares regression**. Like any other nonlinear problem this is much more difficult, cf. [2] and in R with `nls`.

9.2.3 Binary Explanatory Variables

In all our examples, the explanatory variables have been continuous so far. This is not necessary. It is for instance also possible that a variable is binary with values 0 and 1, eg. the gender.
If the binary variable x_b is the only variable in the model $y = \beta_0 + \beta_1 x_\mathrm{b} + \varepsilon$, then

$$y = \beta_0 + \varepsilon \qquad \text{if } \; x_\mathrm{b} = 0,$$
$$y = \beta_0 + \beta_1 + \varepsilon \qquad \text{if } \; x_\mathrm{b} = 1.$$

This regression model is equivalent to the model of two independent random samples of which we are interested in the difference of their means. This is the unpaired **Student's t test**.

Example 9.2.2 (**Student's Sleep Data**)**.** Data which show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.

```
#   classical example: sleep data
help(sleep)
sleep
##       extra group ID
##    1    0.7     1  1
##    2   -1.6     1  2
##    3   -0.2     1  3
##    ...
##    20   3.4     2 10

#   graphic
plot(extra ~ group, data=sleep)
```

First, we perform a regression with the binary variable `group` of the data set `sleep`:

```
#   regression
> mod <- lm(extra ~ group, data=sleep)
> summary(mod)
Call:
lm(formula = extra ~ group, data = sleep)

Residuals:
   Min     1Q Median     3Q    Max
-2.430 -1.305 -0.580  1.455  3.170

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7500     0.6004   1.249   0.2276
group2        1.5800     0.8491   1.861   0.0792 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.899 on 18 degrees of freedom
Multiple R-squared:  0.1613,    Adjusted R-squared:  0.1147
F-statistic: 3.463 on 1 and 18 DF,  p-value: 0.07919
```

In other words, the binary variable `group` has no significant impact on the increase in hours of sleep on the 5% level.

Second, we perform an unpaired Student's t-test with equal variance:

```
> t.test(extra ~ group, data=sleep, paired=FALSE, var.equal=TRUE)

        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

We observe that

- the `(Intercept)` in the regression is equal to `mean in group 1` in the Student t-test,
- `(Intercept)+group2` in regression is equal to `mean in group 2` in the Student t-test,
- `t value` of `group2` in regression is equal to `abs(t)` in the Student t-test and
- `Pr(>|t|)` of `group2` in regression is equal to `p-value` in the Student t-test.

9.2.4 Factor Variables

If a discrete variable has more than two levels it is called a **factor**.

Let $\{l_1, \ldots, l_s\}$ be the s levels in the factor x_k containing n observations. To be able to use such a factor in a regression model we have to introduce for every level $j \in \{2, \ldots, s\}$ a so-called **dummy-variable**

$$x_k^{(j)} = (x_{k,1}^{(j)}, \ldots, x_{k,i}^{(j)}, \ldots, x_{k,n}^{(j)}).$$

For each $i \in \{1, \ldots, n\}$ we define its component

$$x_{k,i}^{(j)} = \begin{cases} 1 & \text{if } i\text{th observation } x_{k,i} \text{ is equal to } j\text{th level } l_j \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } x_{k,i} = l_j \\ 0 & \text{if } x_{k,i} \neq l_j \end{cases}$$

Now we can add these $s-1$ dummy variables to the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + (\beta_{k,2} x_k^{(2)} + \cdots + \beta_{k,s} x_k^{(s)}) + \cdots + \beta_m x_m + \varepsilon. \tag{9.2.a}$$

If you read carefully, then you realise that for a factor with s levels we only added $s-1$ dummy-variables. This is done to make the estimation unique, cf. Ex. 9.1.6, by setting artificially $\beta_{k,1} = 0$.

Example 9.2.3 (**Blasting**). We come back to Ex. 9.1.1. It is possible that the position of the measurement has an effect on the vibrations.

```
#   read data
> file <- "blasting.dat"
> data <- read.table(file, header=TRUE)
> str(data)
data.frame':   48 obs. of  4 variables:
$ Position : int  1 1 1 2 2 2 2 2 2 2 ...
$ Charge   : num  2.18 3.33 3.33 3.33 3.12 3.12 3.12 2.6 2.08 2.08 ...
$ Distance : int  188 183 177 53 49 46 44 42 42 44 ...
$ Vibration: num  0.32 0.53 0.5 6.71 5.99 5.39 4.88 5.99 5.63 9.47 ...

#   define new log10-transformed variables
data$Charge.log    <- log10(data$Charge)
data$Distance.log  <- log10(data$Distance)
data$Vibration.log <- log10(data$Vibration)
```

The position was also recorded and is a discrete variable with integer entries 1 for the first, 2 for the second, 3 for the third and 4 for the fourth position.
Now we will define 4 dummy-variables (later we will only use the last three).

```
#   define 4 dummy-variables
> data$Position1.dum <- as.numeric(data$Position==1)
> data$Position2.dum <- as.numeric(data$Position==2)
> data$Position3.dum <- as.numeric(data$Position==3)
> data$Position4.dum <- as.numeric(data$Position==4)

> str(data, vec.len=ncol(data))
'data.frame':   48 obs. of  11 variables:
 $ Position     : int  1 1 1 2 2 2 2 2 2 2 2 2 2 1 1 ... 1 1 3 3 3 4 ... 4 3 ... 3
 ...
 $ Position1.dum: num  1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 ... 1 1 0 0 0 0 ... 0 0 ... 0
 $ Position2.dum: num  0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 0 0 0 ... 0 0 ... 0
 $ Position3.dum: num  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 1 1 1 0 ... 0 1 ... 1
 $ Position4.dum: num  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 1 ... 1 0 ... 0
```

Observe the relation of the dummy-variables to the original variable `Position`. Now we will estimate the coefficients with all four dummy-variables.

```
#   estimation of the parameters
> mod.dum.full <- lm(Vibration.log ~ Distance.log + Charge.log
+                                    + Position1.dum
+                                    + Position2.dum
+                                    + Position3.dum
+                                    + Position4.dum, data)
> summary(mod.dum.full)
Call:
lm(formula = Vibration.log ~ Distance.log + Charge.log + Position1.dum +
    Position2.dum + Position3.dum + Position4.dum, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36732 -0.09607  0.01385  0.08883  0.28017

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.62124    0.24687  10.618 1.82e-13 ***
Distance.log  -1.33779    0.14073  -9.506 4.97e-12 ***
Charge.log     0.69179    0.29666   2.332   0.0246 *
Position1.dum -0.11080    0.07477  -1.482   0.1459
Position2.dum  0.05351    0.06564   0.815   0.4196
Position3.dum -0.08910    0.06205  -1.436   0.1585
Position4.dum       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3379 on 42 degrees of freedom
Multiple R-squared:  0.8322,    Adjusted R-squared:  0.8122
F-statistic: 41.66 on 5 and 42 DF,  p-value: 3.194e-15
```

We realise that R was unable to estimate the coefficient of the last dummy-variable. This is due to the fact that we could add a constant to all the coefficients of the four dummy-variables and subtract it again from the intercept without changing anything. Therefore the

least-squares estimators are not uniquely defined. The constraint $\beta_{3,1} = 0$ will make it unique. Other more symmetrical constraints are of course possible. Therefore, to make the estimation unique, we omit `Position1.dum`.

```
#   estimation of the parameters
> mod.dum <- lm(Vibration.log ~ Distance.log + Charge.log
+                               + Position2.dum
+                               + Position3.dum
+                               + Position4.dum, data)
> summary(mod.dum)
Call:
lm(formula = Vibration.log ~ Distance.log + Charge.log + Position2.dum +
    Position3.dum + Position4.dum, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36732 -0.09607  0.01385  0.08883  0.28017

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.51044    0.28215   8.898 3.25e-11 ***
Distance.log  -1.33779    0.14073  -9.506 4.97e-12 ***
Charge.log     0.69179    0.29666   2.332   0.0246 *
Position2.dum  0.16430    0.07494   2.192   0.0340 *
Position3.dum  0.02170    0.06366   0.341   0.7349
Position4.dum  0.11080    0.07477   1.482   0.1459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3379 on 42 degrees of freedom
Multiple R-squared:  0.8322,    Adjusted R-squared:  0.8122
F-statistic: 41.66 on 5 and 42 DF,  p-value: 3.194e-15
```

If you have the impression that this is complicated then you might be right. In R, on the other hand, adding factors to a model can be done much easier. Internally R will define corresponding dummy-variables.

```
#   define Position as a factor
> data$Position.fac  <- as.factor(data$Position)
> str(data$Position.fac)
Factor w/ 4 levels "1","2","3","4": 1 1 1 2 2 2 2 2 2 2 ...

#   estimation of the parameters
> mod.fac <- lm(Vibration.log ~ Distance.log + Charge.log + Position.fac, data)
> summary(mod.fac)
Call:
lm(formula = Vibration.log ~ Distance.log + Charge.log + Position.fac, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36732 -0.09607  0.01385  0.08883  0.28017
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.51044    0.28215   8.898 3.25e-11 ***
Distance.log  -1.33779    0.14073  -9.506 4.97e-12 ***
Charge.log     0.69179    0.29666   2.332   0.0246 *
Position.fac2  0.16430    0.07494   2.192   0.0340 *
Position.fac3  0.02170    0.06366   0.341   0.7349
Position.fac4  0.11080    0.07477   1.482   0.1459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3379 on 42 degrees of freedom
Multiple R-squared:  0.8322,    Adjusted R-squared:  0.8122
F-statistic: 41.66 on 5 and 42 DF,  p-value: 3.194e-15
```

We can observe the following things:

- We need to understand how to interpret the estimated coefficients of the factor levels: For every position we get a different estimated model. These four models have the slope $\hat{\beta}_1 = -1.33779$ (coefficient of `Distance.log`) and the slope $\hat{\beta}_2 = 0.69179$ (coefficient of `Charge.log`) in common.

  But they all have different intercepts. The model at the first position has the intercept $\hat{\beta}_0 = 2.51044$ (reference model due to our constraint), the model at the second position has the intercept $\hat{\beta}_0 + \hat{\beta}_{3,2} = 2.51044 + 0.16430 = 2.67474$, the model at the third position has the intercept $\hat{\beta}_0 + \hat{\beta}_{3,3} = 2.51044 + 0.02170 = 2.53214$ and the model at the forth position has the intercept $\hat{\beta}_0 + \hat{\beta}_{3,4} = 2.51044 + 0.11080 = 2.62124$.

- The t-statistic and the P-value of the factor levels have little meaning. With our constraint $\beta_{3,1} = 0$ they simply show if the difference between the corresponding positions and the first (reference) position is significant.

- Adding a further factor increases the multiple R^2 by a few percent. This is no surprise because the new model is also more flexible. We will discuss that later.

The above models differ only by an additive constant. It would also be possible, that the relationship between vibration, distance and charge differs in another way. A straight forward model would include also different slopes for the four positions, so-called **interactions**.
In the model notation of R interactions between the variables `A` and `B` are denoted by `A*B`.

9.3 Comparison of Regression Models with F-Test

The idea of factors in a model is extremely fruitful. Often additional factors improve the goodness of fit tremendously.
What is missing at this point is a test which allows us to ask the question if a certain factor has a significant influence. We cannot simply look at the t-statistics and P-values of individual levels in the R-output. We need to be able to test not only one coefficient but a whole set of coefficients simultaneously.

In general we want to compare two multiple linear regression models. To do that we fit a big model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

with $p = m + 1$ coefficients[4] and a smaller model with q coefficients $\beta_{j_1}, \ldots, \beta_{j_q}$ put to zero, i.e. with only $p - q$ coefficients.

The question is, if the smaller regression model can fit the data as well as the big one? The alternative hypothesis to that question are

$$H_0 : \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_q} = 0$$
$$H_1 : \beta_{j_k} \neq 0 \text{ for at least one } k \in \{1, \ldots, q\}.$$

The **F-test to compare two models** tests whether any of the explanatory variables in a multiple linear regression model are significant.
To answer this question we calculate the F-statistic

$$F = \frac{n-p}{q} \frac{\mathrm{SS}_E^* - \mathrm{SS}_E}{\mathrm{SS}_E}, \tag{9.3.a}$$

where SS_E^* is the sum of squares of the errors of the small model and SS_E is the sum of squares of the errors of the big model.
Under the null hypothesis the F-statistic follows an F-distribution with $(q, n-p)$ degrees of freedom. Critical values of the F-distribution can be found in T.6–T.14.

Example 9.3.1 (**Blasting**). We come back to Ex. 9.1.4 and want to ask the question whether the factor `Position.fac` has a significant influence on the vibrations. We first fit a model with the factor `Position.fac` and one without it.

```
#   big model with Position.fac (already fitted in Ex. 9.2.3)
> mod.fac <- lm(Vibration.log ~ Distance.log + Charge.log + Position.fac, data)
#   fit small model without Position.fac (already fitted in Ex. 9.1.1)
> mod <- lm(Vibration.log ~ Distance.log + Charge.log, data)
```

Then we test the hypothesis that the factor `Position.fac` can be omitted by comparing the two models with the above F-test.

```
#   F-test to compare two models (explicit formulae)
> SSE      <- sum(residuals(mod.fac)^2)
> SSE.star <- sum(residuals(mod)^2)
> n <- nrow(data);   n
[1] 48
> p <- 1+2+3   #  intercept, Distance.log, Charge.log, 3 times Position.fac
> q <- 3       #  3 times Position.fac
> f <- (n-p)/q * (SSE.star - SSE) / SSE;   f
[1] 2.282667
#   P-value
> 1-pf(f, df1=q, df2=n-p)
[1] 0.09296647
```

In R this can be done with the command `anova`.

[4] $p = m$ coefficients in the case of a model with no intercept

```
#   ANOVA for linear model fits
> anova(mod.fac, mod)
Analysis of Variance Table

Model 1: Vibration.log ~ Distance.log + Charge.log + Position.fac
Model 2: Vibration.log ~ Distance.log + Charge.log
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     42 0.90467
2     45 1.05217 -3   -0.1475 2.2827 0.09297 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the R-output we find $F = 2.2827$ with a P-value of 0.09297. Therefore the influence of the factor `Position.fac` is not significant on the 5% level.

An other, more powerful, R-command to perform an F-test to compare two models is `drop1`, which drops all possible single terms to a model, fits those models and computes a table of the changes in fit. At the moment we are only interested in the last column of this table.

```
> drop1(mod.fac, test="F")
Single term deletions

Model:
Vibration.log ~ Distance.log + Charge.log + Position.fac
             Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                    0.90467 -178.63
Distance.log  1   1.94659 2.85125 -125.53 90.3722 4.973e-12 ***
Charge.log    1   0.11713 1.02180 -174.78  5.4379   0.02457 *
Position.fac  3   0.14750 1.05217 -177.38  2.2827   0.09297 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the last column we find the P-values for the variables `Distance.log` and `Charge.log` which were already calculated in Ex. 9.1.6. These two variables are significant on the 5% level. We find also the P-value of 0.09297 for the factor `Position.fac`. Therefore its influence is not significant on the 5% level.

9.3.1 Comparison of Two Straight Lines

We ask the question if two straight lines are equal? Two different straight lines can have different intercepts, different slopes or both.
To be able to answer this question we define a factor variable which indicates the group

$$g_i = \begin{cases} 0 & \text{observation } i \text{ belongs to the first line} \\ 1 & \text{observation } i \text{ belongs to the second line} \end{cases}$$

It is in fact a binary variable.
Now we use the multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to answer the above question. We define new variables $x_1 = x$, $x_2 = g$ and $x_3 = x\,g$ and simplify the notation of the coefficients $\beta_0 = \alpha$, $\beta_1 = \beta$, $\beta_2 = \Delta\alpha$ and $\beta_3 = \Delta\beta$. Then we obtain the model

$$y = \alpha + \beta x + \Delta\alpha \cdot g + \Delta\beta \cdot x\,g + \varepsilon. \tag{9.3.b}$$

We notice that for the first straight line with $g = 0$ we obtain the equation

$$y = \alpha + \beta x$$

and for the second straight line with $g = 1$ we obtain

$$y = (\alpha + \Delta\alpha) + (\beta + \Delta\beta)x.$$

The above questions can now be answered with two different alternative hypothesis on the difference of the intercept $\Delta\alpha$ and on the difference of the slope $\Delta\beta$:

$$H_{\alpha,0} : \Delta\alpha = 0 \qquad H_{\beta,0} : \Delta\beta = 0 \tag{9.3.c}$$

$$H_{\alpha,1} : \Delta\alpha \neq 0 \qquad H_{\beta,1} : \Delta\beta \neq 0 \tag{9.3.d}$$

The following test statistics will answer the questions.

- Testing equal intercepts we use the t-statistic and P-value of the coefficient $\Delta\alpha$.
- Testing equal slopes we use the t-statistic and P-value of the coefficient $\Delta\beta$.
- Testing equal lines, i.e. both hypothesis simultaneously, we need the F-statistic to compare two models. The big model is given in Eqn. 9.3.b and the small model, which only fits one single straight line, is $y = \alpha + \beta x + \varepsilon$.

Example 9.3.2 (**Two Straight Lines**). Let us look at the data set `two-lines.dat`.

```
#   read data
> file <- "two-lines.dat"
> data <- read.table(file, header=TRUE)
> str(data)
##    'data.frame':   63 obs. of  3 variables:
##     $ x    : num  1 1.3 1.4 1.6 1.7 1.9 2 2.1 2.4 2.6 ...
##     $ y    : num  7 6.9 6.7 6.5 6.3 5.4 5.4 5.3 5.2 5.1 ...
##     $ group: int  0 0 0 0 0 0 0 0 0 0 ...

#   define group as a factor
> data$group.fac  <- as.factor(data$group)
> str(data$group.fac)
##  Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

#   graphic
> plot(y ~ x, data, pch=as.character(group), main="Two Straight Lines")
> grid()
```

Now we fit the model in Eqn. 9.3.b to the data using the interaction term `x*group` in the formula of the R-model.

```
> mod.two <- lm(y ~ x*group, data)
> summary(mod.two)
Call:
lm(formula = y ~ x * group.fac, data = data)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-0.8683 -0.2934 -0.1507  0.2955  1.0944

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.92758    0.18420   37.61   <2e-16 ***
x            -0.65701    0.04123  -15.94   <2e-16 ***
group.fac1    5.11112    0.27110   18.85   <2e-16 ***
x:group.fac1 -1.36200    0.06640  -20.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4786 on 59 degrees of freedom
Multiple R-squared:  0.9681,    Adjusted R-squared:  0.9665
F-statistic:   597 on 3 and 59 DF,  p-value: < 2.2e-16
```

The estimated lines are, cf. Fig. 9.3.i,

$$\begin{aligned} y &= \hat{\alpha} + \hat{\beta}x = 6.92758 - 0.65701x \\ y &= (\hat{\alpha} + \hat{\Delta\alpha}) + (\hat{\beta} + \hat{\Delta\beta})x \\ &= (6.92758 + 5.11112) + (-0.65701 - 1.36200)x = 12.0387 - 2.01901x. \end{aligned}$$

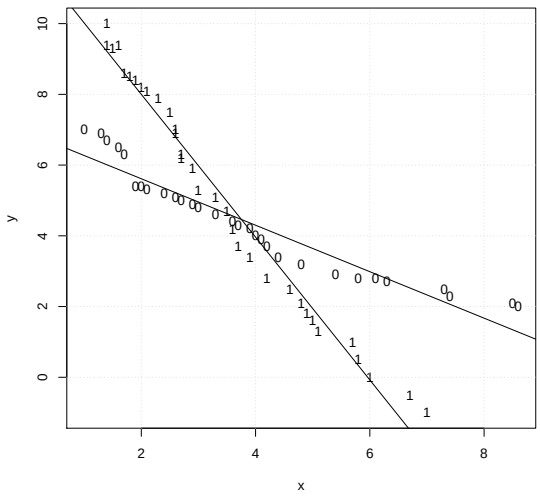


Figure 9.3.i: Data with the best lines fitted (first group: 0, second group: 1).

Now we want to test the hypothesis in Eqn. 9.3.c individually:

- $H_{\alpha,0} : \Delta\alpha = 0$ by looking at the t-statistic of the coefficient `group`. The P-value is very small and therefore the slopes are significantly different.

- $H_{\beta,0} : \Delta\beta = 0$ by looking at the t-statistic of the coefficient `x:group`. The P-value is very small and therefore the slopes are significantly different.

Testing both hypothesis $H_{\alpha,0} : \Delta\alpha = 0$ and $H_{\beta,0} : \Delta\beta = 0$ simultaneously can be done with the F-test to compare two models. The big model is given in Eqn. 9.3.b and the small model, which only fits one single straight line, is $y = \alpha + \beta x + \varepsilon$.

```
#   model with only one straight line
> mod <- lm(y ~ x, data)
#   F-test to to compare two models
> anova(mod.two, mod)
Analysis of Variance Table

Model 1: y ~ x * group.fac
Model 2: y ~ x
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     59  13.515
2     61 110.191 -2   -96.676 211.03 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conclude on the 5% level that the factor `group` is signifikant, therefore the two lines differ. These findings are not very surprising looking at Fig. 9.3.i.

9.4 Exercises

Problem 9.4.1 (**Rental Price Index of Zurich**, cf. [32], Reg3, Problem 1)**.** The data set `RPI-ZH.dat` contains the Rental Price Index of the city of Zurich (`RPI`), the Mortgage Interest of the Zürcher Kantonalbank (`MI`) and the Consumer Price Index of the canton of Zurich (`CPI`) for each quarter between 1994 and May 2005.
From legal regulations on rental rates, it is clear that the mortgage rate and the consumer price index strongly affect the rental price. Consider the following model

$$\texttt{RPI} = \beta_0 + \beta_1 \cdot \texttt{CPI} + \beta_2 \cdot \texttt{MI} + \varepsilon.$$

Assume that the errors ε are independent normally distributed with expected value 0 and variance σ^2.

a. Fit the model to the data. Report the estimated coefficients. Overall, do the explanatory variables influence the target values?

b. Interpret the coefficient of determination for this fit.

c. Calculate a 98% confidence interval on β_1.

d. Predict the rental price index for a mortgage interest of 4.5% and a consumer price index of 100.5. Report also a 95% prediction interval for this prediction.

Problem 9.4.2 (**Catheter**, cf. [32], Reg3, Problem 2)**.** Heart surgery in children requires a customised catheter. The optimal `length` of a catheter [in cm] depends on the `size` [in cm] and the `weight` of a child. The idea is to estimate the length of a catheter from the patient data. Consider the following model

$$\texttt{length} = \beta_0 + \beta_1 \cdot \texttt{size} + \beta_2 \cdot \texttt{weigth} + \varepsilon.$$

Assume that the errors ε are independent normally distributed with expected value 0 and variance σ^2. The data set `catheter.dat` contains recorded catheter lengths and patient data.

a. Fit the model to the data. Overall, do the explanatory variables influence the target values?

b. Test the null hypothesis $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Does the result contradict the overall test?

c. Check the assumptions on the errors with a normal q-q plot.

d. Remove the observation with the smallest residual and repeat the analysis. Compare the results and report your conclusions.

e. Report a table of 95% prediction intervals for all observations. In reality a prediction error of $\pm 2\,\text{cm}$ is acceptable. Is this model and the data therefore useful in practice?

f. Repeat question **e** with simpler reduced models. Report all your findings and conclusions.

g. Repeat question **e** with simpler reduced models and the reduced data set from question **d**. Report all your findings and conclusions.

Problem 9.4.3 (**Surface Finish**, cf. [23], Example 12.13)**.** A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed [in revolutions per minute] of the lathe. Two different types of cutting tools were used. The data set `surface-finish.dat` contains 20 measurements of the quality (`Quality`), the speed of the lathe (`RPM`) and the type of the cutting tool (`Type`).

a. Represent the data in a scatter diagram.

b. Fit a regression model with a linear relationship between the quality and the rotational speed and with the same slope for both types of the cutting tool but different intercepts. Add the best fits to the scatter diagram of question **a**.

c. Fit a regression model with a linear relationship between the quality and the rotational speed and with individual slopes and intercepts for both types of the cutting tool. Add the best fits to the scatter diagram of question **a**.

d. Test the hypothesis of equal slopes on the 5% level.

Problem 9.4.4 (**Experiment on a Dairy Product**, cf. [32], Reg3, Problem 4)**.** The data in the file `oxygen.dat` come from an experiment on a dairy product. We want to empirically model the oxygen uptake (`O2UP`) [in milligrams per minute] from the data collected using five chemical measurements. Chemical measurements include both biological (`BOD`) and chemical (`COD`) oxygen demand, Kjeldahl's total nitrogen content (`TKN`), total solids (`TS`) and volatile solids (`TVS`) that are part of the solids. All chemical measurements are given in milligrams per liter.

The measurements were carried out on a sample of a dairy product which was kept in a water suspension for 220 days. All observations were made on the same suspension over time. The aim of this analysis is a first modelling of the dependence of the logarithmic oxygen uptake on the chemical variables. We will omit observations 1 and 20 in the following analysis. The resulting model will then be checked later for its predictive suitability.

a. Fit the model

$$\log(\texttt{O2UP}) = \beta_0 + \beta_1 \cdot \texttt{BOD} + \beta_2 \cdot \texttt{TKN} + \beta_3 \cdot \texttt{TS} + \beta_4 \cdot \texttt{TVS} + \beta_5 \cdot \texttt{COD} + \varepsilon.$$

Report your findings on the P-values. Comment on the F-statistic in the summary-lm-output. What are the consequences for the fitted model?

b. Compare the full model in **a** with the reduced model

$$\log(\texttt{O2UP}) = \beta_0 + \beta_3 \cdot \texttt{TS} + \beta_5 \cdot \texttt{COD} + \varepsilon.$$

Report your conclusions.

Problem 9.4.5 (**From Multiple to Simple Linear Regression**)**.** By setting $m = 1$ in the least-squares estimator of the multiple linear regression problem with an intercept

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y},$$

cf. Eqn. 9.1.d, derive the solution of the least-squares estimator of the simple linear regression problem

$$\hat{\beta}_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2},$$

cf. Eqn. 7.2.e. Notice that in this case the design matrix $\mathbf{X}$ has only two columns.

Chapter 10

Sensitivity and Robustness

In this chapter, we come back to analyse the goodness of fit of a model, which was already started in Chap. 8. With simple linear regression, the scatter diagram of the response variable versus the explanatory variable was investigated to derive the presented graphical residual analysis tools. In multiple linear regression, too many dimensions (variables) are involved and we cannot visualise conventionally the regression model. However, the methods introduced in Chap. 8 are just the right tools to keep some clarity in modelling.

Furthermore, we examine the question of how individual observations can influence the estimation. From this understanding, further diagnostic tools can be provided. On the other hand, we wonder if there are no estimators that are less sensitive to outliers. Such estimators would allow us to perform the residual analysis even more efficiently.

10.1 Residual Analysis

The graphical tools **Tukey-Anscombe plot**, **scale-location plot** and **q-q plot** which were introduced in Chap. 8 to check goodness of fit for the simple linear regression model can also be used for multiple linear regression models. R is very efficient doing residual analysis, cf. Ex. 10.1.1.

Residual analysis consists of the following two main steps:

- Analyse Tukey-Anscombe plot, scale-location plot, q-q plot and the residuals versus leverage (will be introduced in Chap. 10.2). Compare with bootstrap-simulations. If possible check for time dependencies in the residuals.
- Draw conclusions. What is the biggest discrepancy and how can it be corrected?

We want to demonstrate residual analysis for multiple linear regression models with Ex. 9.1.1.

Example 10.1.1 (**Blasting**)**.** The model is

$$\log(\texttt{Vibration}) = \beta_0 + \beta_1 \cdot \log(\texttt{Distance}) + \beta_2 \cdot \log(\texttt{Charge}) + \varepsilon,$$

cf. Eqn. 9.1.f. In R we can create all the plots in Fig. 10.1.i with a few commands.

```
#   load R-function to do residual analysis (written by A. Ruckstuhl)
> source("RFn_Plot-lmSim.R")
```

```
#   read data
> file <- "blasting.dat"
> data <- read.table(file, header=TRUE)
#   define new log-transformed variables
> data$Charge.log    <- log10(data$Charge)
> data$Distance.log  <- log10(data$Distance)
> data$Vibration.log <- log10(data$Vibration)

#   fit model
> mod <- lm(Vibration.log ~ Distance.log + Charge.log, data)

#   diagnostic tools
> op <- par(mfcol=c(2,4))
>     #   Tukey-Anscombe plot
>     plot(mod, which=1, pch=20)
>     grid()
>     abline(h=0, lty=3)
>     plot.lmSim(mod, which=1, SEED=1)
>     grid()
>     abline(h=0, lty=3)
>
>     #   scale-location plot
>     plot(mod, which=3, pch=20)
>     grid()
>     abline(h=0, lty=3)
>     plot.lmSim(mod, which=3, SEED=1)
>     grid()
>     abline(h=0, lty=3)
>
>     #   q-q plot
>     plot(mod, which=2, pch=20)
>     grid()
>     plot.lmSim(mod, which=2, SEED=1)
>     grid()
>
>     #   residuals against leverages
>     plot(mod, which=5, pch=20)
>     grid()
>     abline(h=0, lty=3)
> par(op)
```

Looking at Fig. 10.1.i we report the following observations:

1. The smoothed residuals in the **Tukey-Anscombe plot** is slightly curved which is a sign of a non-constant expected value. The simulation shows that the original curve stands out from the 19 simulated curves. Therefore, the expected value is not constant zero. This could give us hints to further find appropriate transformations of the explanatory variables or to check for outliers. The observation with index 47 could for instance be omitted. Try it!

2. The smoothed curved in the **scale-location plot** has a very slight downward trend which is not extreme compared to the simulations. There is no hint that the scattering of the residuals is not constant.

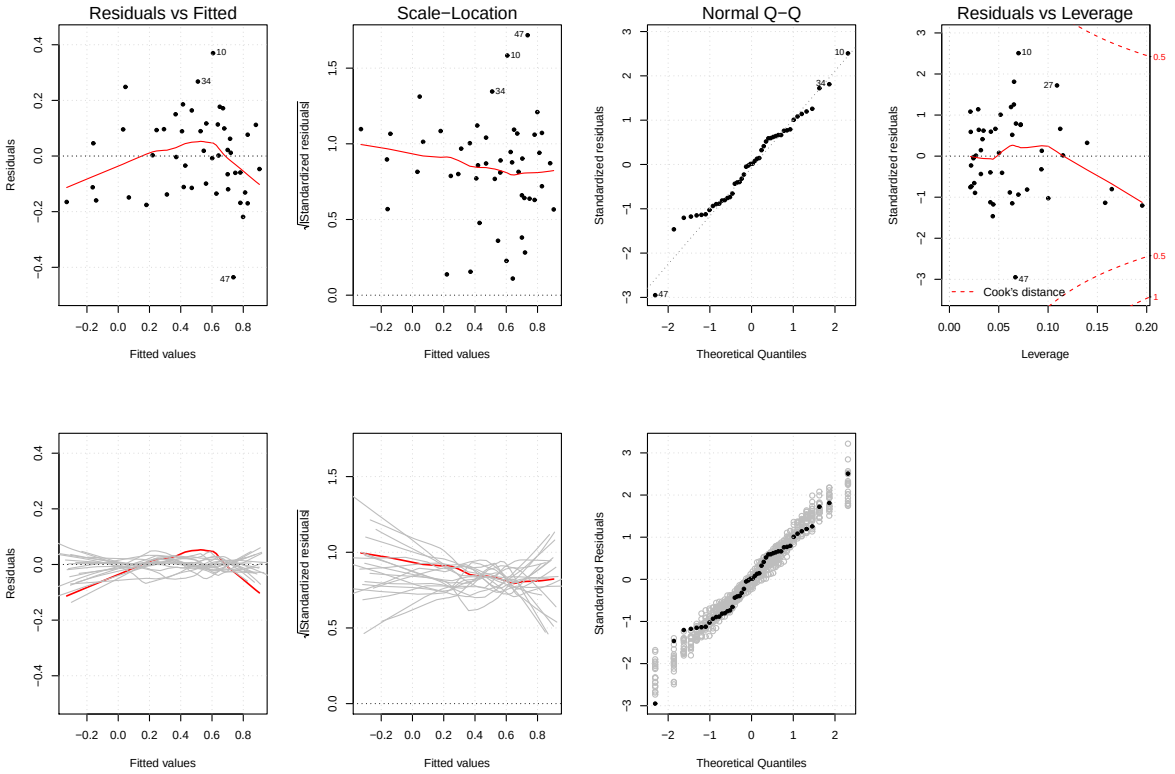


Figure 10.1.i: Residual analysis: Tukey-Anscombe plot, scale-location plot, q-q plot, residuals versus leverage: in R with `plot(mod)` (upper row), corresponding bootstrap-simulations: in R with `plot.lmSim(mod)` from `RFn_Plot-lmSim.R` written by A. Ruckstuhl (lower row).

3. In the **q-q plot** the residuals scatter nicely around the line. Moreover, almost all points lie within the simulated points. Observation with index 47 is borderline.

4. The diagram with the **residuals versus leverage** will be discussed in Chap. 10.2.

If the Tukey-Anscombe diagram shows deviations from the assumed form of the regression function, or if a transformation of the response variable is unsuccessful or has to be abandoned, then an appropriate transformation of one or more explanatory variables could help.

In multiple regression, however, it is no longer clear which explanatory variable might cause the deficit. To find out graphically, the residuals $e = y - \hat{y}$ are plotted versus an explanatory variable x_k used as the horizontal axis instead of the fitted values $\hat{y}$, cf. Fig. 10.1.ii and 10.1.iii.

```
#   residuals versus Distance.log
> plot(data$Distance.log, resid(mod), pch=20,
+      xlab="Distance.log", ylab="Residuals", ylim=c(-0.5,0.5),
+      main="Residuals versus log(Distance)")
> lines(loess.smooth(data$Distance.log, resid(mod), span=0.8), col="red")
```

```
> grid()
> abline(h=0, lty=3)

#   residuals versus Charge.log
> plot(data$Charge.log, resid(mod), pch=20,
+      xlab="Charge.log", ylab="Residuals", ylim=c(-0.5,0.5),
+      main="Residuals versus log(Charge)")
> lines(loess.smooth(data$Charge.log, resid(mod), span=0.8), col="red")
> grid()
> abline(h=0, lty=3)
```

Fig. 10.1.ii and 10.1.iii show these scatter diagrams for the example of blasting. For both explanatory variables we cannot report any violations of the assumptions of linearity and constant variance.

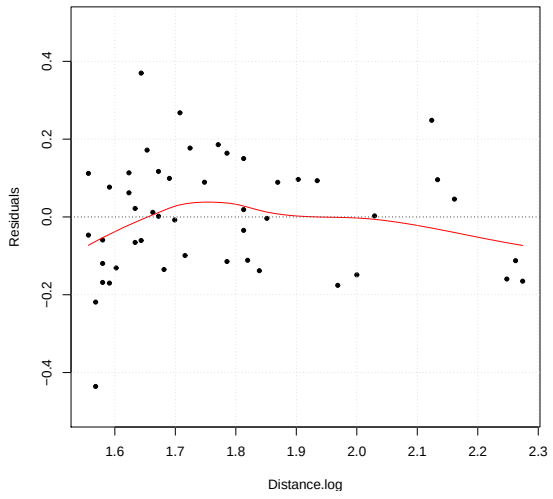


Figure 10.1.ii: Residuals versus log(Distance). No violation of assumptions of linearity and constant variance.

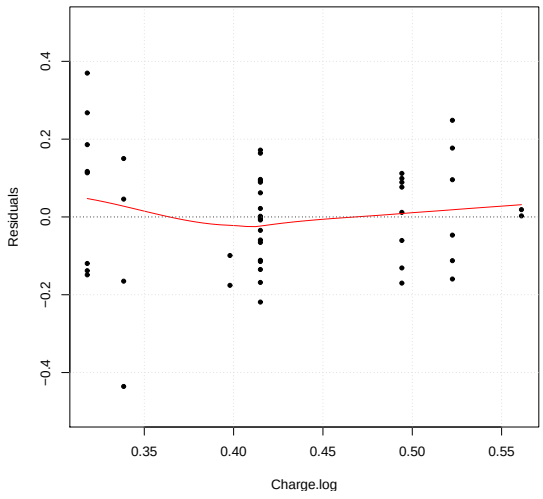


Figure 10.1.iii: Residuals versus log(Charge). No violation of assumptions of linearity and constant variance.

Often it is possible to observe the form of the smoothed values of the residuals versus explanatory variable x_k and use this information to appropriately transform x_k. If no transformation is fruitful, then an additional quadratic term x_k^2 might help. In R use `I(x^2)` instead of `x^2`.

It is also possible to observe in the diagram of residuals versus explanatory variable x_k that the variance of the residuals depends on x_k. In these situations a weighted linear regression might be appropriate, cf. Chap. 10.3.

10.2 Influential Observations

Finally, we want to come back to the topic of outliers. Residual analysis can lead to a more appropriate regression model, which can make extreme residuals small and therefore they are not considered as outliers. The answer to the question whether an observation is an outlier depends on the model.

If outliers remain in the data, the question arises of how strongly they influence the analysis, i.e. how sensitive is the estimation process to individual observations? Why is that important? If it concerns erroneous observations, the analysis is falsified. If they are correct observations and strongly influence the results, it is useful to know this. It can happen that outliers belong to a different population. I repeat myself: Often in science the unexpected observations are the most interesting and might show some new discovery.

The effect of an observation, especially an outlier, on the results can be examined by repeating the analysis without the observation in question. **Influence diagnostics** is based on the idea to remove one observation, repeat the analysis and measure the change.
Cook's distance measure is a very popular influence diagnostic tool which measures the change of the fitted values if one observation i is omitted.
Let $\hat{\boldsymbol{\beta}}$ the least-squares estimate based on all n points and $\hat{\boldsymbol{\beta}}_{(-i)}$ is obtained by deleting the ith point. Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the $n \times 1$ vector of the fitted values. Define $\hat{\mathbf{y}}_{(-i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(-i)}$ the $n \times 1$ vector of the fitted values obtained by deleting the ith point in the estimation process. Then for every measurement $i \in \{1, \ldots, n\}$ **Cook's distance measure** is

$$d_i = \frac{(\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})^t (\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})}{p\ \hat{\sigma}^2}, \tag{10.2.a}$$

where $p = m + 1$ is the number of estimated parameters in the case of a model with intercept and $p = m$ in the case without intercept.

Fortunately it is not necessary to repeat the analysis n times. A mathematical simplification (cf. [24], Chap. 6.3) leads to the equivalent form

$$d_i = \frac{e_i^2}{p\,\hat{\sigma}^2} \frac{h_{ii}}{(1 - h_{ii})^2} = \frac{\tilde{e}_{\mathrm{std},i}^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \tag{10.2.b}$$

where e_i is the ith residual and $\tilde{e}_{\mathrm{std},i}$ the standardised residual, cf. Eqn. 8.1.b.
Cook's distance measure is a function of the ith standardised residual $\tilde{e}_{\mathrm{std},i}$ and the so-called **leverages** h_{ii} of the ith observation. The leverages

$$h_{ii} = \mathbf{H}_{ii} = (\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)_{ii}$$

are the diagonal entries of the hat matrix $\mathbf{H}$, cf. Eqn. 9.1.e. Leverages h_{ii} satisfy $0 \leq h_{ii} \leq 1$ and the mean of all leverages is always $\frac{p}{n}$.

How to interpret the leverages h_{ii}?

- If a response y_i is changed by the amount Δy_i, then $h_{ii}\Delta y_i$ measures the change in the fitted value $\hat{y}_i$.

  A large leverage h_{ii} has a big influence on the fitted values.

- The variance of the ith residual is

$$\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2. \tag{10.2.c}$$

  A large leverage h_{ii} reduces the variance of the ith residual, and therefore the ith observation is close to regression hyper-plane[1].

[1] In simple linear regression a hyperplane is simply a straight line.

- Eqn. 10.2.c makes it clear, how the residuals can be scaled to equal variance. Define the **standardised residuals**

$$\tilde{e}_{\text{std},i} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad \text{for all} \quad i \in \{1,\ldots,n\}.$$

  For the simple linear regression with $p = m + 1 = 2$ we obtain again Eqn. 8.1.b. Observing the case of the simple linear regression[2] we conclude that the value $\frac{h_{ii}}{1-h_{ii}}$ is a measure of the distance of the point $\mathbf{x}_i$ to the centre of gravity $\bar{\mathbf{x}}$ in the space of the explanatory points.

  Leverage is a measure of how far away the explanatory values of an observation are from those of the other observations.

When is an observation a leverage point? There are no clear answers. We know that the mean of all leverages is

$$\bar{h} = \sum_{i=1}^{n} h_{ii} = \frac{p}{n},$$

where n is the number of observations and $p = m + 1$ is the number of estimated parameters in the case of a model with intercept and $p = m$ in the case without intercept.
So at least we have a clue. One rule says that leverages with $h_{ii} > 2\frac{p}{n}$ are potentially dangerous. Another rule by Huber classifies the observations in three groups:

- Observations with $h_{ii} \leq 0.2$ are harmless.
- Observations with $0.2 < h_{ii} \leq 0.5$ are potentially dangerous.
- Observations with $0.5 < h_{ii}$ should be avoided.

For large leverage, the observation must be almost perfectly described by the model, otherwise it has too much influence on the estimation of the model.

Let us look again at Cook's distance measure, cf. Eqn. 10.2.a. The larger the measure d_i is, the bigger is the influence of the corresponding observation with index i on the estimation of the coefficients. Also with Cook's distance measure no well defined limits exists. In practice a threshold of 1 is often used to decide:

- If $d_i > 1$ then the ith observation is dangerous.
- If $d_i \leq 1$ then the ith observation is harmless.

Cook's distance measure d_i is often plotted for each observation versus the index i.

Example 10.2.1 (**Blasting**). In Fig. 10.2.i we plotted Cook's distance measure for the blasting data set, cf. Ex. 10.1.1.

```
#   Cook's distance versus index
> plot(mod, which=4)
> abline(h=seq(0,1,by=.05), lty=3, col="gray")
> abline(h=0)
```

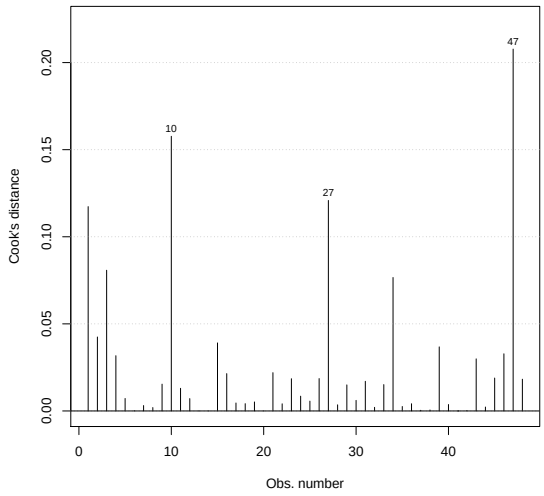


Figure 10.2.i: Cook's distance measure d_i versus index i for the blasting data set, cf. Ex. 10.1.1.

On the other hand, Cook's distance measure versus the index gives us no indication about the influence of the response value. Therefore another diagram is often used in residual analysis. In a scatter diagram the standardised residuals $\tilde{e}_{\mathrm{std},i}$ are plotted versus the leverages h_{ii}, cf. Fig. 10.2.ii. In addition a few contours of Cook's distance measure, cf. Eqn. 10.2.b, are added to the diagram to be able to judge the dangerousness of the observations. The contour of threshold 1 is highlighted.

Example 10.2.2 (**Standardised Residuals versus Leverages, Artificial Data**)**.** We created artificial data with the model $y = 5{+}3\,x{+}\varepsilon$ with $\mathrm{Var}(\varepsilon) = 10^2$ and $x \in \{0, 0.5, \ldots, 50\}$. Then we modified three observations, cf. Fig. 10.2.iii:

- Observation $i = 15$ is a leverage point in the explanatory variable $x_{15} = 130$.
- Observation $i = 37$ is an outlier with $x_{37} = 60$ and $x_{37} = 60$.
- Observation $i = 81$ is a harmless observation following the model with $x_{81} = 80$ and $x_{81} = 248$.

Now we can observe the behaviour of three modified measurements in the plot with the standardised residuals versus leverages, cf. Fig. 10.2.iv.

- Observation $i = 15$ has leverage $h_{15,15} > 0.3$ and a huge standardised residual, therefore Cook's distance measure is $d_{15} > 1$. This observation clearly attracts the best straight line.

[2] In case the design matrix $\mathbf{X}$ has only two columns, i.e. in simple linear regression we find

$$\frac{h_{ii}}{1-h_{ii}} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

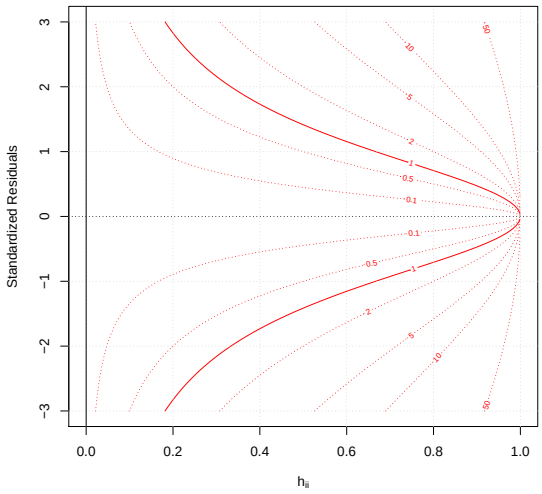


Figure 10.2.ii: Contour plot of Cook's distance measure $d(h, \tilde{e}_{\text{std}}) = \frac{\tilde{e}^2_{\text{std}}}{p} \frac{h}{1-h}$ with $p = 2$.

- Observation $i = 37$ has a moderate leverage $h_{37,37}$, but a huge standardised residual, therefore Cook's distance measure is $d_{37} > 1$. This observation clearly attracts the best straight line.
- Observation $i = 81$ is a harmless observation with Cook's distance measure $d_{81} < 1$.

Example 10.2.3 (**Blasting**)**.** In the last figure in the upper row in Fig. 10.1.i we plotted the standardised residuals versus leverages for the blasting data set, cf. Ex. 10.1.1.
We can observe the all observations have leverage $h_{ii} < 0.2$. Observation with index $i = 47$ has the largest Cook's distance measure, see also Fig. 10.2.i, which is much smaller than the critical threshold 1.
We conclude that there are no influential observations in the data of the blasting measurement.

The presented influence diagnostic measure like Cook's distance and leverages check the influence of only one observation. On the other hand, they do not show the effect if two or more observations are omitted. Omitting systematically more than one observations is combinatorially a very difficult task and not practiced in statistics.

10.3 Weighted linear regression

We saw in Ex. 8.1.2 of the dissociation pressure, cf. Fig. 8.2.vii and 8.2.viii, that the variance of the errors is not constant. It is possible to model the variance of the errors as a function of the explanatory variables. Let us introduce a new example.

Example 10.3.1 (**Fat Content in Fish**)**.** Is the mean fat content of four species of fish different? To answer this question three fish were randomly selected for each species. The fat content of four samples of each fish were measured.
Due to a mishap, it was not possible to determine the fat content of some of the samples. Can the remaining data still be analysed?

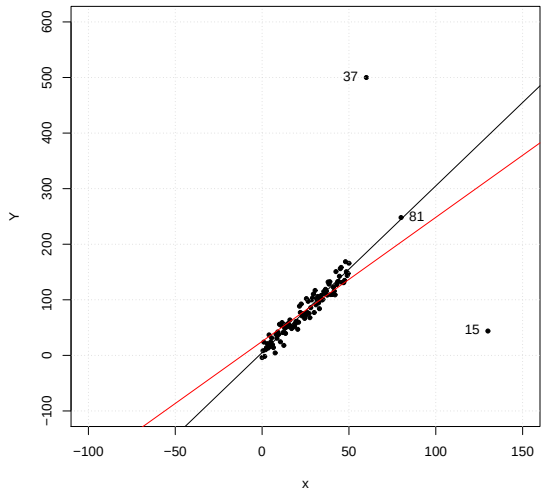


Figure 10.2.iii: Artificial data with two outliers, $i = 15$ and 37, and a harmless observation, $i = 81$). Original model (black) and estimated best fit (red).

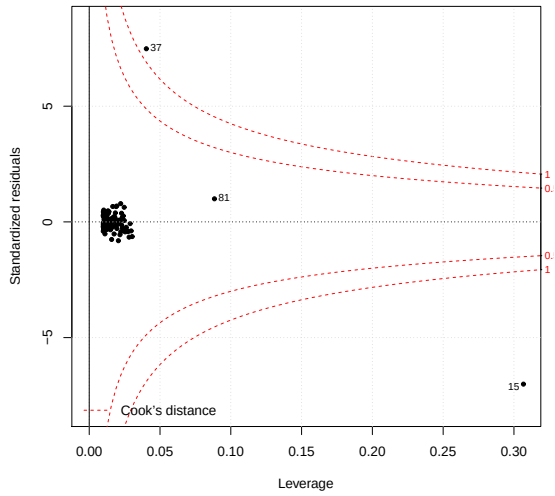


Figure 10.2.iv: Standardised residuals versus leverages with two outliers, $i = 15$ and 37, and a harmless observation, $i = 81$. Contours of Cook's distance measure (red).

It is clear that the samples of a fish are more similar than the samples between two fish. The measurement errors due to different sample sizes are therefore no longer identically distributed. One possible analysis is to use only the mean of the fat content per fish. However, it should be noted that the mean may have been calculated from a different number of good samples, therefore the precisions of the means is not the same.

For the sake of simplicity, we assume that the variance of the samples is independent of the fish and that the samples of different species of fish only differ by an additive shift.

In this case we can model the measurement y_{ijk} of the kth sample with $k \in \{1, 2, 3, 4\}$ of the jth fish with $j \in \{1, 2, 3\}$ of the ith species with $i \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$ as follows

$$y_{ijk} = \mu_i + \varepsilon_{ijk}$$

where ε_{ij} are independent and normally distributed with expected value zero and constant variance σ^2 for all samples $k \in \{1, 2, 3, 4\}$. The mean fat content μ_i depends only on the species with $i \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$.

The mean fat content of the jth fish of the ith species is

$$\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk},$$

where n_{ij} is the number of good samples of the jth fish of the ith species. Notice that for each fish the number of good samples n_{ij} is not constant. The variance of the mean fat content of

the jth fish of the ith species is therefore

$$\begin{aligned} \mathrm{Var}(\bar{y}_{ij}) &= \mathrm{Var}\left(\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}\right) = \mathrm{Var}\left(\frac{1}{n_{ij}}\sum_{j=1}^{n_{ij}}(\mu_i+\varepsilon_{ijk})\right) \\ &= \frac{1}{n_{ij}^2}\mathrm{Var}\left(\sum_{j=1}^{n_{ij}}(\mu_i+\varepsilon_{ijk})\right) = \frac{1}{n_{ij}^2}\sum_{j=1}^{n_{ij}}\mathrm{Var}(\mu_i+\varepsilon_{ijk}) \\ &= \frac{1}{n_{ij}^2}\sum_{j=1}^{n_{ij}}\mathrm{Var}(\varepsilon_{ijk}) = \frac{1}{n_{ij}}\sigma^2. \end{aligned} \tag{10.3.a}$$

We used the fact that the individual observations are assumed to be independent. In other words, Eqn. 10.3.a says that the variance of the mean gets smaller with increasing sample size and is therefore not constant.
Due to the mishap we use the fish as observation unit and therefore analyse the model

$$\bar{y}_{ij} = \mu_i + \bar{\varepsilon}_{ij},$$

where $\bar{\varepsilon}_{ij}$ are independent and normally distributed with expected value zero and non-constant variance $\mathrm{Var}(\bar{\varepsilon}_{ij}) = \frac{\sigma^2}{n_{ij}}$ depending on the jth fish of the ith species.

It certainly makes sense to give greater weight to the observations with smaller random error. That is, the more precise observations gain weight in the statistical analysis.

Therefore we need to find the least-squares estimate for **weighted multiple regression**. Define the $n \times n$ diagonal matrix of the **weights**

$$\mathbf{W} = \begin{pmatrix} w_1 & & 0 & \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix},$$

where $w_1, \ldots, w_n$ are the weights.
Now we want to find the vector of least-squares estimators $\boldsymbol{\beta}$, which minimise

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i\, \varepsilon_i^2 = \boldsymbol{\varepsilon}^t\mathbf{W}\boldsymbol{\varepsilon} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t\mathbf{W}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

With a bit of matrix algebra we can write

$$S(\boldsymbol{\beta}) = \mathbf{y}^t\mathbf{W}\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{W}\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{W}\mathbf{X}\boldsymbol{\beta}.$$

The least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ must now satisfy

$$\frac{\partial S}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = -2\mathbf{X}^t\mathbf{W}\mathbf{y} + 2\mathbf{X}^t\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = 0,$$

which simplifies to the so-called **weighted least-squares normal equations**

$$\mathbf{X}^t\mathbf{W}\mathbf{y} = \mathbf{X}^t\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{10.3.b}$$

The solution of the normal equations is the least-squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y}. \tag{10.3.c}$$

The estimate $\hat{\boldsymbol{\beta}}$ is unbiased with covariance

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}.$$

Therefore confidence intervals for the coefficients are readily available.

Example 10.3.2 (**Fat Content in Fish**)**.** Now we come back to Ex. 10.3.1. The aggregated data can be found in the following table.

Species	Fish	mean fat content per fish	number of good samples
i	j	$\bar{y}_{ij}$	n_{ij}
a	1	19.080	4
a	2	16.476	3
a	3	11.606	1
b	1	14.101	1
b	2	18.631	4
b	3	12.155	2
c	1	15.601	2
c	2	20.109	1
c	3	16.874	3
d	1	13.303	1
d	2	9.152	2
d	3	8.306	4

In the last column the number of good samples are recorded which will be our weights `NoSamples`.

```
#   read data
> file <- "fish-fat.dat"
> data <- read.table(file, header=TRUE)
#   weighted linear regression
> mod.w <- lm(Fat.mean ~ Species, data, weights=NoSamples)
> summary(mod.w)
Call:
lm(formula = Fat.mean ~ Species, data = data, weights = NoSamples)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.6265 -1.9802 -0.6998  3.2955  4.9949

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.1693     1.4862  11.552 2.86e-06 ***
Speciesb     -1.0357     2.1756  -0.476  0.64678
Speciesc     -0.1804     2.2703  -0.079  0.93861
Speciesd     -7.9077     2.1756  -3.635  0.00664 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.204 on 8 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.5602
F-statistic:  5.67 on 3 and 8 DF,  p-value: 0.02221
```

Notice that the intercept

$$\bar{y}_{\mathrm{w,a}} \quad = \quad 17.1693$$

is the weighted mean of the species a and the weighted means of the other species b, c and d can be calculated as follows:

$$\begin{aligned} \bar{y}_{\mathrm{w,b}} &= 17.1693 - 1.0357 = 16.1336 \\ \bar{y}_{\mathrm{w,c}} &= 17.1693 - 0.1804 = 16.9889 \\ \bar{y}_{\mathrm{w,d}} &= 17.1693 - 7.9077 = 9.2616 \end{aligned}$$

We want to test the null hypothesis that the fat content in all species is identical. We use the F-statistic in the above output and find a P-value of 0.02221 which is smaller than 5%. Therefore we reject the null hypothesis and conclude the fat content differs.

On the other hand, if the different sample sizes is not considered we cannot reject the null hypothesis on the 5% level.

```
> mod <- lm(Fat.mean ~ Species, data)
> summary(mod)
Call:
lm(formula = Fat.mean ~ Species, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1147 -1.9322 -0.7577  2.6981  3.6687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.7207     1.7792   8.836 2.12e-05 ***
Speciesb     -0.7583     2.5162  -0.301   0.7708
Speciesc      1.8073     2.5162   0.718   0.4930
Speciesd     -5.4670     2.5162  -2.173   0.0616 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.082 on 8 degrees of freedom
Multiple R-squared:  0.5325,    Adjusted R-squared:  0.3572
F-statistic: 3.038 on 3 and 8 DF,  p-value: 0.09282
```

10.4 Robust Methods

In multiple linear regression the model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,$$

where the random errors ε are independent and normally distributed with expected value zero and unknown variance σ^2.

In residual analysis, we check all assumptions as thoroughly as possible. But! *All diagnostic tools based on the least-squares method cannot always detect outliers.*

Among other things, we put a lot of effort in finding potential outliers and bad leverage points which we would like to remove from the analysis because they lead to unreliable fits. Is all this effort necessary? No, if we use robust methods. We can find outliers easier and the estimation of parameters can cope with a few doubtful measurements.

An outlier is a violation of the assumption about the normality of the errors. It can have devastating effects on the least-squares estimates. The goal of **robust statistics** is to find parameter estimates as if the outliers did not exist, cf. [19] and [30].

The oldest informal approach to robustness:

1. examine the data for obvious outliers
2. delete these outliers
3. apply the optimal inference procedure for the assumed model to the cleaned data set

However, this data analytic approach is not unproblematic:

- Even expert statisticians do not always screen the data. Why not?
- It can be difficult or even impossible to identify outliers, particularly in multivariate or highly structured data like multiple linear regression.
- Often the relationships in the data cannot be examined without first fitting a model.
- It can be difficult to formalise this process so that it can be automatised.
- Inference based on applying a standard procedure to the cleaned data will be based on distribution theory which ignores the cleaning process and hence will be inapplicable and possibly misleading.

The robustness of an estimator can be investigated by two measures: the **influence function** and the **breakdown point**. Both measures are based on the idea of studying the effect of an estimator under the influence of gross errors, i.e. arbitrary added data.
The **gross error sensitivity** is based on the influence function and measures the maximum effect of a single observation on the estimated value.
The **breakdown point** is the minimal proportion of incorrect observations which cause completely unrealistic estimates. An estimator with good robustness properties has a bounded sensitivity and a breakdown point of about $\frac{1}{2}$.

Example 10.4.1 (**Mean and Median**)**.** The **arithmetic mean** is an estimator with a breakdown point of $\frac{1}{n}$, because we can make

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$$

arbitrarily large just by changing any of the n observations $x_1, \ldots, x_n$.
On the other, hand the **median** is a robust estimator with a breakdown point of $\frac{1}{2}$ because it can handle almost have of incorrect data.

In the following we want to construct a robust estimator for the linear regression model. In least-squares regression the **loss function** $\rho(r) = r^2$ increases sharply with the size of the residuals r, cf. Fig. 10.4.i.a. Therefore we need to limit the influence of observations with large residuals with weights in a weighted linear regression analysis. Doing this, we will modify the loss function, cf. Fig. 10.4.ii.a and 10.4.v.a.
Let us look at the weighted least-squares normal equations, cf. Eqn. 10.3.b

$$\mathbf{X}^t\mathbf{W}\mathbf{y} = \mathbf{X}^t\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}$$

and write it in the form

$$\mathbf{X}^t\mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^t\mathbf{W}\mathbf{e} = 0,$$

where $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is the $n \times 1$ vector of residuals. Now we write the same $m + 1$ equations with a sum

$$\sum_{i=1}^{n} w_i e_i \, \mathbf{x}_i = 0, \tag{10.4.a}$$

where $e_i = y_i - \mathbf{x}_i^t\boldsymbol{\beta}$ and

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix}$$

is the ith row of the design matrix $\mathbf{X}$, cf. Eqn. 9.1.c.

In Eqn. 10.4.a the **influence of an outlier** in observation i is manifested in a large residual e_i. Therefore, the main idea of robust regression is to introduce weights w_i which weigh down this influence. If we want to limit the influence of large residuals the function

$$\psi(e_i) = w_i e_i$$

must be limited. This can be achieved, for instance by a ψ-function as in Fig. 10.4.ii.b.

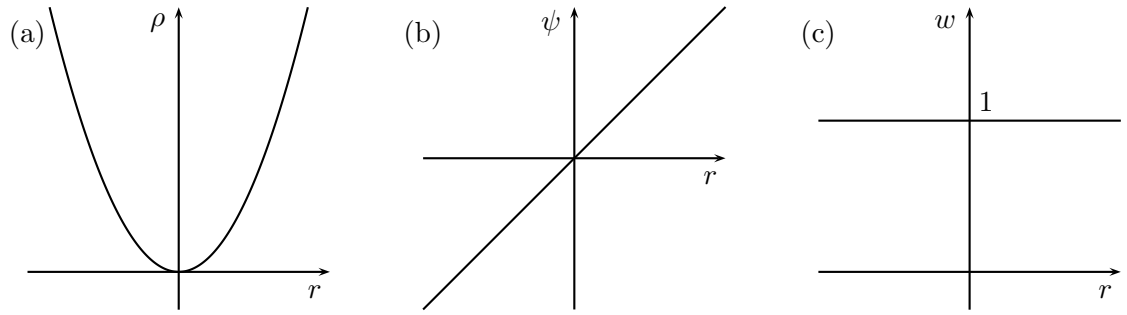


Figure 10.4.i: Loss function ρ of least-squares estimate (a), ψ-function (b) and corresponding weight function w (c).

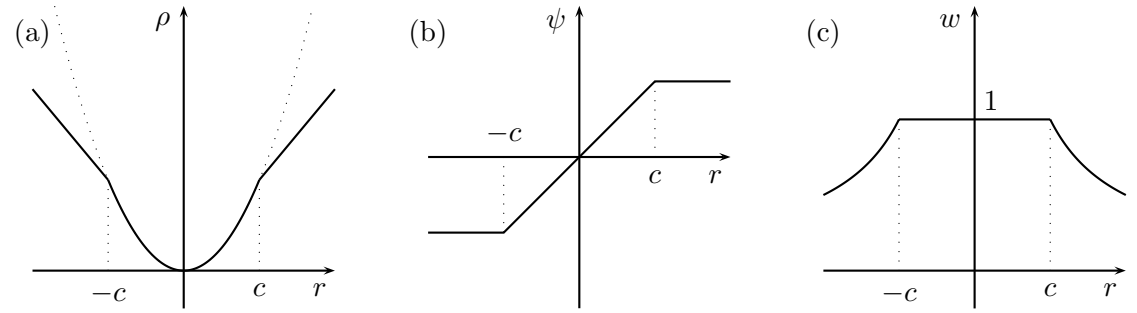


Figure 10.4.ii: Loss function ρ (a) of **Huber's ψ-function** (b) and corresponding weight function w (c).

The corresponding weights are then simply

$$w_i = \frac{\psi(e_i)}{e_i}.$$

The weights depend now on the residuals $e_i = y_i - \mathbf{x}_i^t\boldsymbol{\beta}$ which in turn depend on the parameter $\boldsymbol{\beta}$. This is a devils circle which does not allow us to derive explicit formula. The estimates have to be found with an iterative algorithm and the standard error of the estimates must be calculated differently than in weighted linear regression.

An estimator with such implicitly defined weights is called a robust **regression M-estimator**. If Huber's ψ-**function** in Fig. 10.4.ii.b is used, then we only need to specify the characteristic kink. The kink at c in the ψ-function determines from which point onwards extreme observations lose their influence. To do this we use the natural scale for the residuals which is $\frac{e_i}{\sigma}$.

The (robust) regression M-estimator $\hat{\boldsymbol{\beta}}_M$ is therefore defined by the implicit equations

$$\sum_{i=1}^{n} \psi(\tilde{e}_i)\,\mathbf{x}_i = 0 \quad \text{where} \quad \tilde{e}_i = \frac{y_i - \mathbf{x}_i^t\hat{\boldsymbol{\beta}}_M}{\sigma}, \tag{10.4.b}$$

or expressed with weights

$$\sum_{i=1}^{n} w_i e_i\,\mathbf{x}_i = 0 \quad \text{where} \quad w_i = \frac{\psi(\tilde{e}_i)}{\tilde{e}_i} \;\text{ and }\; \tilde{e}_i = \frac{y_i - \mathbf{x}_i^t\hat{\boldsymbol{\beta}}_M}{\sigma}. \tag{10.4.c}$$

The scale parameter σ is usually unknown. For the regression M-estimators however, it is necessary to determine the position c at which unsuitable observations will loose their influence. The standard deviation s as an estimator for σ is extremely non-robust. Therefore the **median of absolute values**

$$\hat{\sigma}_{\text{MAV}} = \frac{\text{median}_i(|e_i|)}{0.6745} \tag{10.4.d}$$

is used in the regression M-estimator. The correction factor $\frac{1}{0.6745}$ is needed to obtain a consistent estimate of the standard deviation in the case of a normal distribution. The

threshold c in Huber's ψ-function is set to be $c = 1.345$ which is based on theoretical efficiency considerations.

As in multiple linear regression it is possible to calculate confidence intervals for the regression M-estimator $\hat{\boldsymbol{\beta}}_M$. We will not do that here because R will calculate them for us.

Example 10.4.2 (**Waterflow Measurements of Kootenay River**)**.** The original data set is the waterflow in January of the Kootenay river, measured at two locations, namely, Libby (Montana) and Newgate (British Columbia) for 13 consecutive years, 1931-1943.
For didactical reasons we modify the original measurements $(77.6, 44.9)$ of the year 1934: once to the point $P_1(19.7, 44.9)$ and the other time to $P_2(77.6, 15.7)$.
In Fig. 10.4.iii the data with point P_1 and in Fig. 10.4.iv the data with point P_2 is shown. In addition the least-squares estimate of the straight line is added. It is obvious that one single outlier is enough to obtain a bad fit.

We want to test the regression M-estimator on the contaminated data. Therefore we need the load two R-packages `robustbase` and `MASS`.

```
#   load package
> require(robustbase)
> require(MASS)
```

The data set `waterflow.dat` contains the original data (marked with the suffix `0`), the data contaminated with the outlier P_1 (suffix `1`) and the data contaminated with the leverage point P_2 (suffix `2`).

```
#   read data
> file <- "waterflow.dat"
> data <- read.table(file, header=TRUE)
```

Then we fit with four different methods a straight line to the data.

```
#   1. Modified data
> plot(Newgate1 ~ Libby1, data, pch=20,
+                         xlab="Waterflow at Libby ",
+                         ylab="Waterflow at Newgate",
+                         main="1. Modified Data")
> text(data$Libby1[4], data$Newgate1[4], label="P1", pos=1)
> grid()
> abline(   lm(Newgate1 ~ Libby1, data),              lty=1, col="black")
> abline(   lm(Newgate1 ~ Libby1, data, subset=-4),   lty=1, col="red")
> abline(  rlm(Newgate1 ~ Libby1, data, method="M"),  lty=2, col="black")
> abline(lmrob(Newgate1 ~ Libby1, data, method="MM"), lty=3, col="black")
> legend("topright", lty=c(1,1,2,3), col=c("black", "red", "black", "black"),
+       legend=c("ls estimator",
+                "ls estimator without P1",
+                "regression M-estimator",
+                "regression MM-estimator"))

#   2. Modified data
> plot(Newgate2 ~ Libby2, data, pch=20,
+                         xlab="Waterflow at Libby ",
```

```
+                        ylab="Waterflow at Newgate",
+                        main="2. Modified Data")
> text(data$Libby2[4], data$Newgate2[4], label="P2", pos=3)
> grid()
> abline(   lm(Newgate2 ~ Libby2, data),              lty=1, col="black")
> abline(   lm(Newgate2 ~ Libby2, data, subset=-4),   lty=1, col="red")
> abline(  rlm(Newgate2 ~ Libby2, data, method="M"),  lty=2, col="black")
> abline(lmrob(Newgate2 ~ Libby2, data, method="MM"), lty=3, col="black")
> legend("topright", lty=c(1,1,2,3), col=c("black", "red", "black", "black"),
+        legend=c("ls estimator",
+                 "ls estimator without P2",
+                 "regression M-estimator",
+                 "regression MM-estimator"))
```

The point P_1 in Fig. 10.4.iii is an outlier. The least-squares estimator is attracted by the outlier and therefore a very bad fit. The regression M-estimator does a very good job and can cope with the outlier P_1. It performs equally well as the least-squares estimator without P_1.

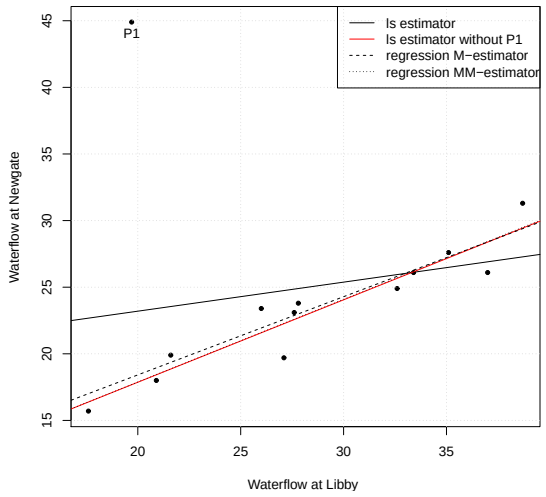


Figure 10.4.iii: The least-squares estimator (solid) fits very badly and the regression M-estimator (dashed) and the MM-estimator (dotted) perform equally well as the least-squares estimator without P_1 (red).

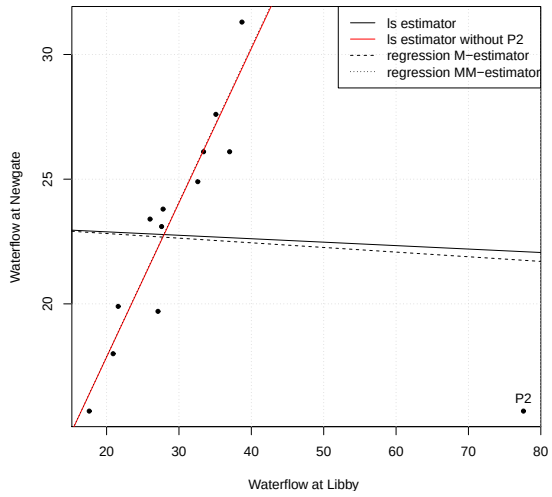


Figure 10.4.iv: The least-squares estimator (solid) and the regression M-estimator (dashed) fit very badly and the MM-estimator (dotted) performs very well as the least-squares estimator without P_2 (red).

On the other hand, the point P_2 in Fig. 10.4.iv is a leverage point. The least-squares estimator and the regression M-estimator are attracted by the leverage point and therefore fit very badly. Thus, it is obvious that the regression M-estimation limits only the influence of large residuals, but is still prone to leverage points and therefore has a break point of zero.

Ex. 10.4.2 showed clearly that we need a better robust regression estimator which can also cope with leverage points.
To be able to handle bad leverage points we need a redescending ψ-function such that observations with large absolute residuals are gradually ignored, cf. Fig. 10.4.v.b.

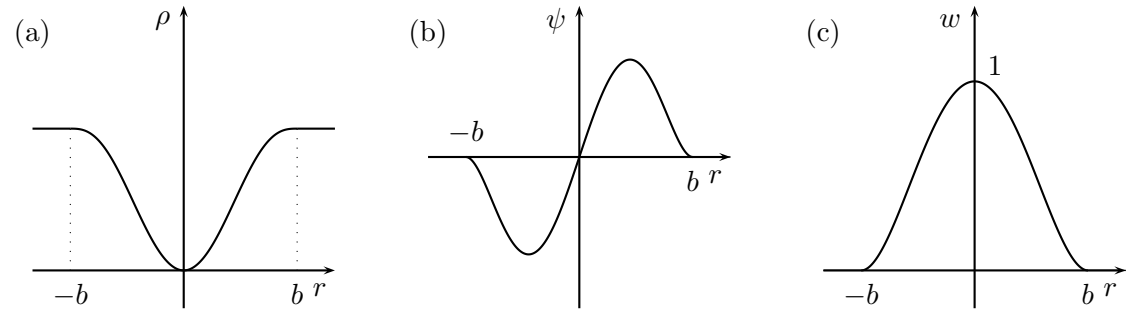


Figure 10.4.v: Loss function ρ (a) of **Tukey's bisquare** ψ-function (b) and corresponding weight function w (c).

The threshold b of Tukey's bisquare is set to be $b = 4.685$ which is based on theoretical efficiency considerations. Because the corresponding ψ-**function** drops back to the zero line, such regression M-estimators are called **redescending**. Since such an estimator has (many) local minima, the optimisation is difficult. Therefore we need a good initial value for the parameter $\boldsymbol{\beta}$. The iterative algorithm to find initial values is based on a random resampling algorithm.

This estimation procedure is called a **m**odified **M**-estimator or **regression MM-estimator** in short. The regression MM-estimator has a breakdown point of $\frac{1}{2}$ and an efficiency and an asymptotic distribution like the regression M-estimator.

Make a residual analysis also with robust estimators.

Example 10.4.3 (**Waterflow Measurements of Kootenay River**)**.** The regression MM-estimator in Ex.10.4.2 does a very good job and can cope with the leverage point P_2. It performs equally well as the least-squares estimator without P_2.

Remark. Robust procedures are implemented in many statistical software packages. The quality of the implementation is very heterogeneous. In R there are several implementations. The function `lmrob` in the package `robustbase` is a very good implementation which also works with factors, cf. Ex. 10.4.2-4 (Ex. 10.4.4 only as an R-File). Another implementation is `rlm` in the package `MASS`, [40].

10.5 Exercises

Problem 10.5.1 (**Rental Price Index of Zurich**, cf. [32], Reg4, Problem 1)**.** We come back to Prob. 9.4.1. Make a residual analysis of the fitted model and improve the model. Report all your findings and conclusions.

Problem 10.5.2 (**Experiment on a Dairy Product**, cf. [32], Reg4, Problem 2)**.** We come back to Prob. 9.4.4. Make a residual analysis of the fitted model on the complete data. What was the motivation to omit the observations with index $i = 1$ and 20? Report all your findings and conclusions.

Problem 10.5.3 (**Robust Regression**, cf. [32], Reg4, Problem 3)**.** Study `Example 10.4.4 Syntetic data to illustrate lmrob.R`. The data set is `lmrob-syntetic.dat`. You need the instal the R-packages `robustbase` and `rgl`.

Problem 10.5.4 (**Oxidation of Ammonia to Nitric Acid**, cf. [5], Ex. 12.12)**.** In a chemical plant, ammonia is processed by oxidation in nitric acid. The reaction is influenced by several factors. We want to study the dependence of the ammonia loss (`AmmoniaLoss`) on the airflow of the system (`AirFlow`), on the cooling water temperature (`WaterTemp`) and on the acid concentration (`AcidConc`). There are 19 measurements on consecutive days (`Day`). The data is given in the data frame `ammonia-oxidation.dat`.

a. Fit the model

$$\texttt{AmmoniaLoss} = \beta_0 + \beta_1 \cdot \texttt{AirFlow} + \beta_2 \cdot \texttt{WaterTemp} + \beta_3 \cdot \texttt{AcidConc} + \varepsilon$$

with least-squares. Make a residual analysis and analyse the Tukey-Anscombe and the normal q-q plot. Report your findings.

b. Fit the same model with a regression MM-estimator. Compare the summary-output with the output of question **a**. Make a residual analysis and compare it also with the analysis of question **a**.

c. Check the time dependence of the residuals.

Problem 10.5.5 (**Experiment on a Dairy Product**, cf. [32], Reg4, Problem 5)**.** We come back to Prob. 9.4.4 and 10.5.2. Fit the same models as in Prob. 9.4.4 with regression MM-estimators to the to complete data set. Compare the estimated coefficients with the coefficients of Prob. 10.5.2 and make a residual analysis and compare it too. Report your findings.

Chapter 11

Variable Selection and Modelling

In practice, regression is used in many situations. The approach depends on the previous knowledge and subject-specific questions.

1. Ideally, it is clear in advance that the response y depends linearly on the explanatory variables $x_1, \ldots, x_m$. The functional relationships are often based on physical, chemical or other engineering or scientific theory, i.e. they are derived from first principles. We call these **mechanistic models**. In this case we are most often interested in good estimates of the parameters, confidence and prediction intervals.

2. In the other extreme case, the study serves to investigate relationships between the response variable x and the explanatory variables. We do not know if and in what form the explanatory variables influence the response. Often in economics, management, life and biological sciences and social sciences, these functional relationships are unknown and so we have to find **empirical models**. In this case, data are often collected for a very large number of potential influencing factors.

3. Sometimes the approach is in between: If we are only interested in the influence of a single explanatory variable, but take into account the effects of other explanatory variables, or if a lot of previous studies and theoretical considerations is known and additional knowledge should be gained.

In the second and third case we need to select variables. Which explanatory variables should be used in which form in the model equation of linear regression?

Notice that the explanatory variables $x_1, \ldots, x_m$ in the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

can already be transformed observations u_k like $x_k = \log(u_k)$, $x_k = \frac{1}{u_k}$; or combinations of several observations u_k, u_l like $x_j = u_k \cdot u_l$. The same is true for the response variable y.

To illustrate the concepts in this chapter we introduce a new example from economics, cf. [6], Example G.

Example 11.0.1 (**Cost of Construction of Nuclear Power Plants**). The construction costs of 32 nuclear power plants, which were built in the years 1967-71 in the USA, were examined. It is required to predict the capital cost involved in the construction of further

light water reactor power plants. One question was whether a partial turnkey plant, i.e. a cost guarantee (PT) of the general contractor would lead to savings? Further explanatory data for the construction costs were collected and are described in the following table.

Variable	Description	Type	Proposed Transformation
C	Cost in dollars $\times 10^6$, adjusted to 1976 base	numeric	logarithm
D	Date construction permit issued	numeric	–
T1	Time between application for and issue of permit	numeric	–
T2	Time between issue of operating license and construction permit	numeric	–
S	Power plant net capacity [MWe]	numeric	logarithm
PR	Prior existence of an light water reactor on same site	binary	–
NE	Plant constructed in north-east region of USA	binary	–
CT	Use of cooling tower	binary	–
BW	Nuclear steam supply system manufactured by Babcock-Wilcox	binary	–
N	Cumulative number of power plants constructed by each architect-engineer	integer	square-root
PT	Partial turnkey plant, cost guarantee	binary	–

General considerations motivated us to apply the logarithmic transformations to the cost (C) and the construction date (D), and the square-root transformation to the number of power plants (N). The time between application (T1) and the time between issue of operating license and construction permit (T2) were not transformed, since we assume a linear influence of these times on the logarithm of the costs (derived from first principles in compound interest calculations).

Of course, many other transformations are conceivable and possible and should be tested on the basis of theoretical considerations and the results of the residual analysis.

The linear multiple regression model with all variables (full model) is

$$\begin{aligned}\log(\texttt{C}) = \beta_0 + \beta_1 \cdot \texttt{D} + \beta_2 \cdot \texttt{T1} + \beta_3 \cdot \texttt{T2} + \beta_4 \cdot \log(\texttt{S}) + \beta_5 \cdot \texttt{PR} \\ + \beta_6 \cdot \texttt{NE} + \beta_7 \cdot \texttt{CT} + \beta_8 \cdot \texttt{BW} + \beta_9 \cdot \sqrt{\texttt{N}} + \beta_{10} \cdot \texttt{PT} + \varepsilon\end{aligned}$$

We first read the data and transform the variables C, S and N

```
#   read data
> file <- "NPP.dat"
> data <- read.table(file, header=TRUE)
#   define new transformed variables and delete the original variables afterwards
> data$C.log  <- log10(data$C)
> data$C <- NULL
> data$S.log  <- log10(data$S)
> data$S <- NULL
> data$N.sqrt <-  sqrt(data$N)
> data$N <- NULL
#   order columns as in original data set
> data <- data[,c("C.log","D","T1","T2","S.log","PR","NE","CT","BW","N.sqrt","PT")]
```

and then we fit the full model in R.

```
#   linear regression with full model
> mod.full <- lm(C.log ~ ., data)
> summary(mod.full)
Call:
lm(formula = C.log ~ ., data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.129835 -0.044872  0.009197  0.039167  0.116093

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.025855   2.347286  -2.567  0.01796 *
D            0.095248   0.035797   2.661  0.01463 *
T1           0.002635   0.009550   0.276  0.78531
T2           0.002290   0.001982   1.155  0.26092
S.log        0.692542   0.137131   5.050 5.32e-05 ***
PR          -0.045734   0.035614  -1.284  0.21307
NE           0.110453   0.033907   3.257  0.00377 **
CT           0.053405   0.029700   1.798  0.08654 .
BW           0.012776   0.045370   0.282  0.78101
N.sqrt      -0.029973   0.017800  -1.684  0.10700
PT          -0.099511   0.055615  -1.789  0.08800 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07235 on 21 degrees of freedom
Multiple R-squared:  0.8684,    Adjusted R-squared:  0.8057
F-statistic: 13.85 on 10 and 21 DF,  p-value: 3.983e-07
```

If we look at the different P-values then we conclude that 5 to 7 variables are superfluous, including the cost guarantee `PT`. Has therefore the question already been answered?

The question is how can we decide which variables can be omitted?

11.1 Model Selection Methods

We will proceed the approach with **model selection methods** which are based on statistical measures of goodness of fit[1]. Such a measure should evaluate the **model accuracy** on the one hand and the **model complexity** on the other hand. Model accuracy means a good description of the data by the model. However, the model should also be as simple as possible, i.e. a small number of explanatory variables.

The goals of model complexity and model accuracy are contradictory: more explanatory variables always lead to a higher model accuracy.

In the following we will define three statistical measures of model accuracy which also take the model complexity into consideration. An obvious candidate for a model selection criterion

[1] We do not study the P-values of the individual variables. Because of multiple testing issues this is not a good idea. Besides, it is not state of the art.

is coefficient of determination R^2. However, the R^2 gets bigger and bigger as more variables are added[2] and is therefore not suitable for a variable selection criterion.

- The **adjusted R-squared** is

$$\begin{aligned} R^2_{\text{adj}} &= 1 - \frac{n-1}{n-p}(1-R^2) \\ &= R^2 - \frac{p-1}{n-p}(1-R^2), \end{aligned}$$

  where R^2 is the coefficient of determination, n the number of observations and $p = m+1$ the number of parameters in the model[3]. Notice that $R^2_{\text{adj}} \leq R^2$ and with increasing number of parameters p the adjusted R-squared gets smaller.

- **Akaike information criterion** is

$$\text{AIC} = n \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + 2\tilde{p} + \text{constant},$$

  where n is the number of observations, $\tilde{p}$ the number of model parameters (σ inclusive) and $e_1, \ldots, e_n$ are the n residuals. The Akaike information criterion is the most generalisable criterion and is also used in time series analysis.

- **Mallow's C_p statistic** is

$$\begin{aligned} C_p &= \frac{\text{SS}_E}{\hat{\sigma}^2_{p^*}} + 2p - n \\ &= (n-p)\left(\frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_{p^*}} - 1\right) + p, \end{aligned}$$

  where n is the number of observations, SS_E is the sum of squares of the errors and $\hat{\sigma}^2_p = \frac{1}{n-p}\text{SS}_E$ is the estimate of σ^2 of the model with p parameters. In addition, $\hat{\sigma}^2_{p^*}$ is the estimate of σ^2 of the full model with p^* parameters. Mallow's C_p statistic minimises the prediction error.

With these model selection criteria, three different **variable selection methods** can be formulated:

- **Forward selection.**

  1. In the first step choose the trivial model

$$y = \beta_0 + \varepsilon$$

     with an intercept only. Use `formula = y ~ 1` in R.
  2. In the following steps, the explanatory variable that leads to the largest improvement of the selected model selection criteria is added in the model.

[2] In the extrem case of interpolation, i.e. the number of parameters equals the number of observations, the coefficient of determination is one.

[3] $p = m$ parameters in the case of a model with no intercept

3. The procedure is stopped as soon as no improvement is possible by adding another explanatory variable.

- **Backward elimination.**

1. In the first step choose the full model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

with all possible explanatory variables. Use[4] `formula = y ~ .` in R.

2. In the following steps, the explanatory variable that leads to the largest improvement of the selected model selection criteria is removed from the model.
3. The procedure is stopped as soon as no improvement is possible by removing another explanatory variable.

- **Both** is a combination of forward selection and backward elimination.

1. In the first step choose any model.
2. The following steps will examine whether eliminating or adding an explanatory variable improves the selected model selection criteria. The action that leads to the largest improvement is carried out.
3. The procedure is stopped as soon as no improvement is possible.

Example 11.1.1 (**Cost of Construction of Nuclear Power Plants**)**.** We now come back to Ex. 11.0.1. In R we can carry out variable selection based on Akaike information criterion. First we fit the trivial model with only an intercept as the start of the search algorithm.

```
> mod.small <- lm(C.log ~ 1, data)
```

Any other model would also work. We use the R-function `step()` which chooses a model by AIC in a stepwise algorithm. With the argument `scope` we specify the smallest and the largest model between which the most suitable model should be found. In addition we specify the direction of our search algorithm.

```
> mod.both <- step(mod.small, scope=list(upper= ~ D + T1 + T2 + S.log + PR + NE
+                                                 + CT + BW + N.sqrt + PT,
+                                        lower= ~ 1),
+                                        direction="both")
Start:  AIC=-114.67
C.log ~ 1

         Df Sum of Sq     RSS     AIC
+ PT      1   0.37962 0.45556 -132.06
+ D       1   0.33055 0.50464 -128.79
+ T1      1   0.17238 0.66280 -120.06
+ S.log   1   0.14449 0.69070 -118.75
+ NE      1   0.12432 0.71086 -117.82
+ CT      1   0.05497 0.78022 -114.84
```

[4] Be careful, the dot `.` in a formula means a model in R with all columns in the data set not otherwise in the formula. So, if we added some transformed variables to the data set they and the original variable will be in the model.

```
<none>                0.83518 -114.67
+ N.sqrt  1   0.02839 0.80679 -113.77
+ BW      1   0.01675 0.81844 -113.31
+ PR      1   0.00959 0.82559 -113.04
+ T2      1   0.00110 0.83409 -112.71

Step:  AIC=-132.06
C.log ~ PT

Step:  AIC=-145.3
C.log ~ PT + S.log

Step:  AIC=-154.27
C.log ~ PT + S.log + D

Step:  AIC=-159.02
C.log ~ PT + S.log + D + NE

Step:  AIC=-161.63
C.log ~ PT + S.log + D + NE + CT

Step:  AIC=-163.73
C.log ~ PT + S.log + D + NE + CT + N.sqrt

         Df Sum of Sq     RSS     AIC
<none>                0.12388 -163.73
+ BW      1  0.003832 0.12005 -162.74
+ T2      1  0.003507 0.12038 -162.65
+ PR      1  0.003503 0.12038 -162.65
+ T1      1  0.001205 0.12268 -162.04
- N.sqrt  1  0.016964 0.14085 -161.63
- PT      1  0.021281 0.14517 -160.66
- CT      1  0.028750 0.15263 -159.05
- NE      1  0.050819 0.17470 -154.73
- D       1  0.102640 0.22653 -146.42
- S.log   1  0.174758 0.29864 -137.58
```

The above R-output is much longer and we kept only the first and the last step. According to the stepwise variable selection in both directions based on the AIC criterion the following model results.

$$\log(\texttt{C}) = \beta_0 + \beta_1 \cdot \texttt{D} + \beta_4 \cdot \log(\texttt{S}) + \beta_6 \cdot \texttt{NE} + \beta_7 \cdot \texttt{CT} + \beta_9 \cdot \sqrt{\texttt{N}} + \beta_{10} \cdot \texttt{PT} + \varepsilon$$

```
> summary(mod.both)
Call:
lm(formula = C.log ~ PT + S.log + D + NE + CT + N.sqrt, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14982 -0.03014  0.01074  0.03817  0.14014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -5.82981    1.48707  -3.920 0.000608 ***
PT          -0.10441    0.05038  -2.072 0.048700 *
S.log        0.72559    0.12218   5.939 3.37e-06 ***
D            0.09341    0.02053   4.551 0.000119 ***
NE           0.10449    0.03263   3.202 0.003694 **
CT           0.06523    0.02708   2.409 0.023707 *
N.sqrt      -0.03046    0.01646  -1.850 0.076133 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07039 on 25 degrees of freedom
Multiple R-squared:  0.8517,    Adjusted R-squared:  0.8161
F-statistic: 23.92 on 6 and 25 DF,  p-value: 3.183e-09
```

Now the cost guarantee `PT` remains in the model and is significant. But do not forget to perform a complete residual analysis of the optimal model.

A word of warning: It is in general quite hopeless to reliably discover 10 potential explanatory variables with only 32 (unplanned) observations.

Often, forward, backward and step by step choices lead to different final models. The reason is that none of these variable selection methods guarantees to find the globally optimal model according to the model selection criterion. This means that the optimum on the stepwise path does not necessarily have to correspond to the global optimum.
If we want to find the global minimum then we would have to calculate the model selection criteria for all possible models. This so-called **all subset selection** fits all 2^m models for the full model with m parameters. This procedure gets computationally very hard if many variables are involved.

Notice that the optimal model is not necessarily the right or true model. Simulations show that often models different from the true one are selected as optimal models. Therefore in explorative data analysis always consider several models and ask other non-statistical scientists involved in the project for their opinion.

11.2 Collinearity

Although high correlation coefficients between explanatory variables are allowed in the estimation procedure they lead to problems in the interpretation of the estimated coefficients. In such cases, it may be a matter of chance in forward and backward variable selection methods which of the correlated variables are first added, resp. omitted. If all the models are examined, then there are groups of similar models.

Collinearity means that an explanatory variable x_j can almost be represented by the others as a linear combination. That is, there exists an offset γ_0 and coefficients $\gamma_1, \ldots, \gamma_{k-1}$, γ_{k+1}, $\ldots$, γ_m such that

$$x_k \approx \gamma_0 + \sum_{j,\, j \neq k}^{m} \gamma_j x_j. \tag{11.2.a}$$

If this relation is exakt then the least-squares estimator is not uniquely defined and software implementations may have problems.

In the following we want to define a **measure for collinearity**. If Eqn. 11.2.a is only approximately satisfied we can consider it as a linear multiple regression model. For each explanatory variable x_k with $k \in \{1, \ldots, m\}$ define the coefficient of determination R_k^2 of the model in Eqn. 11.2.a. The **variance inflation factor** of variable x_k is then defined by

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}. \tag{11.2.b}$$

It is a good and easy to calculate measure of collinearity. A rule of thumb is that if this measure is greater than 5, there are problems with collinearity.

High collinearity leads to large standard errors of the estimated coefficients. It is possible to show, cf. [24], Chap. 9.4.1, that the standard error of the estimate of the coefficient β_k is increased by a factor of

$$\sqrt{\mathrm{VIF}_k}.$$

In an ideal situation with no collinearity we would have R_k^2 and therefore $\mathrm{VIF}_k = 1$ and hence no increase in the standard error of the estimated coefficient $\hat{\beta}_k$.

Example 11.2.1 (**Cost of Construction of Nuclear Power Plants**)**.** We come back to the data in Ex. 11.0.1. First, we graphically investigate pairwise correlations between all variables.

```
#   put (absolute) correlations on the lower panels
#   (font size is proportional to correlations)
> panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...){
+                 usr <- par("usr")
+                 on.exit(par(usr))
+                 par(usr=c(0, 1, 0, 1))
+                 r <- abs(cor(x, y, use="pairwise.complete.obs"))
+                 txt <- format(c(r, 0.123456789), digits=digits)[1]
+                 txt <- paste(prefix, txt, sep="")
+                 if(missing(cex.cor)){ cex.cor <- 0.8/strwidth(txt) }
+                 text(0.5, 0.5, txt, cex=cex.cor*r) }
> #   put histograms on the diagonal
> panel.hist <- function(x, ...){
+                 usr <- par("usr")
+                 on.exit(par(usr))
+                 par(usr=c(usr[1:2], 0, 1.5))
+                 h <- hist(x, plot=FALSE)
+                 breaks <- h$breaks
+                 nB <- length(breaks)
+                 y <- h$counts
+                 y <- y/max(y)
+                 rect(breaks[-nB], 0, breaks[-1], y, col="gray") }

#   pairs-plot of the transformed data
> pairs(data, upper.panel=panel.smooth,
+             lower.panel=panel.cor,
+             diag.panel=panel.hist,
+             pch=20,
+             col=3*data$PT+1)
```

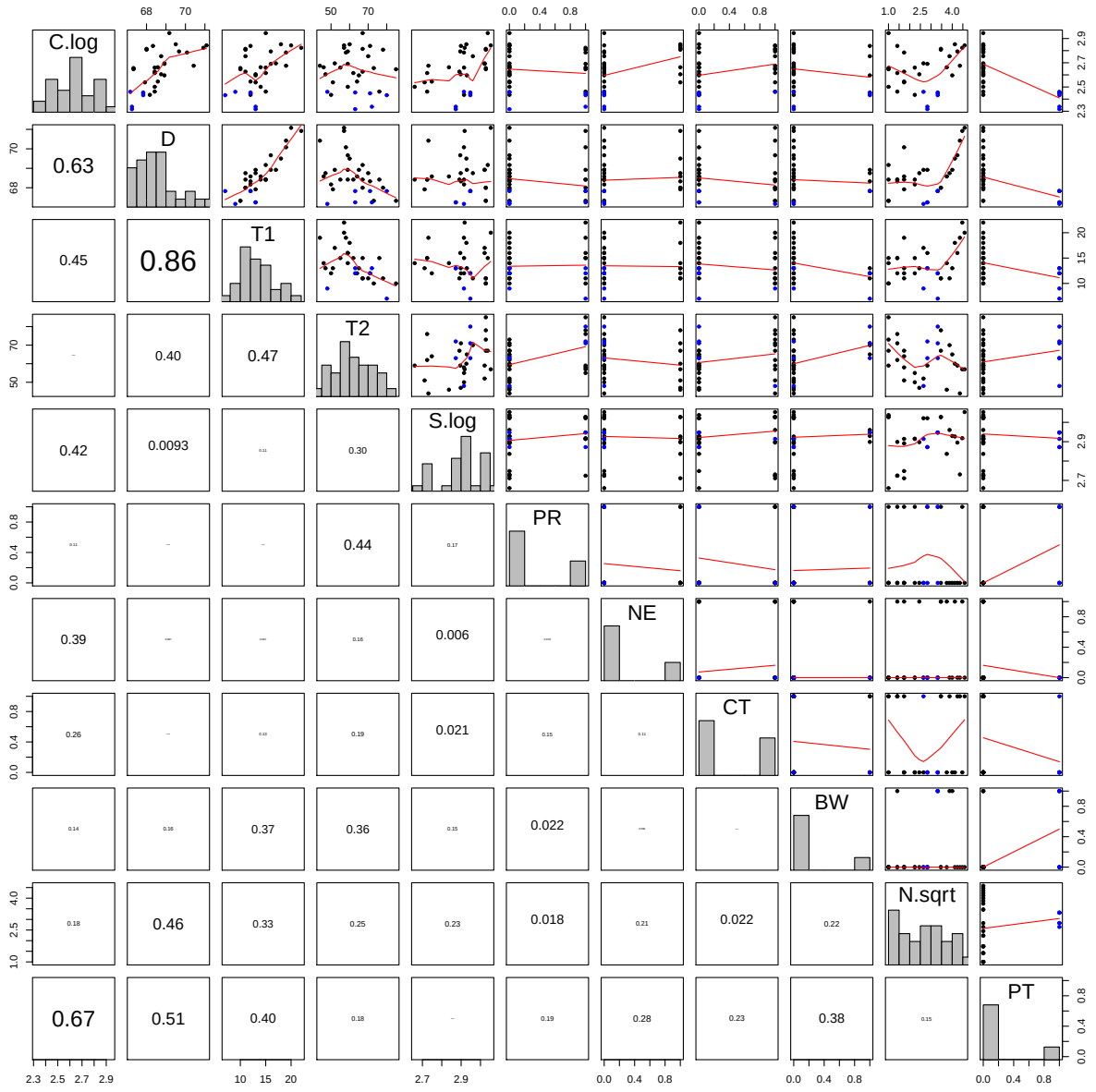


Figure 11.2.i: Pairs-plot of the complete transformed data set. Upper panel: all 55 possible combinations of scatter diagrams with smoothed curve of the data (red), cost guarantee `PT`=1 (blue points) and `PT`=0 (black points); diagonal: names and histograms of all variables; lower panel: pairwise correlation coefficients.

In Fig. 11.2.i we can immediately see that the variables `D` and `T1` are strongly correlated. In the scatter diagrams `C.log` versus all the other variables we observe that nuclear power plants with a cost guarantee are cheaper. This observation might have been the origin of the initial question in Ex. 11.0.1.

Second, we calculate the variance inflation factor for all explanatory variables used in the optimal model found in Ex. 11.1.1. We use the function `vif()` in the R-package `car`.

```
#   load package
> require(car)
#   fit optimal model
> mod.opt <- lm(C.log ~ PT + S.log + D + NE + CT + N.sqrt, data)
#   variance inflation factor
> vif(mod.opt)
      PT    S.log        D       NE       CT   N.sqrt
2.497443 1.089288 2.716501 1.289015 1.142437 2.248181
```

All variance inflation factors are smaller than 5. Therefore there is no problem with collinearity in the optimal model.

If possible, experiments should be designed such that collinearity is avoided right from the beginning. If this is not possible, we can introduce new explanatory variables like the sum or the difference of two collinear variables of the same type. There are always many possible linear transformations that lead to uncorrelated explanatory variables. But it is essential that new variables and thus their coefficients remain easily interpretable.

11.3 Modelling Strategies

In most cases modelling cannot be replaced to automated variable selection strategies. A satisfactory exploratory analysis is always based on an intelligent human interaction with the data and the other involved scientists. In the following we list some reasons:

- For several strategies the selection of the variables depends on chance. Therefore, it makes sense to consider not only the optimal model but also models that are nearly optimal in terms of the selected criteria.
- The optimal model does not necessarily have to meet all model requirements. It is therefore important to carry out a residual analysis.
- Automated variable selection started with a given set of variables, but all possible transformations and interactions are ignored. If all these possibilities have to be checked, then this is an impossible task because of the combinatorial complexity. Such transformed and additional terms must therefore be treated differently.
- It may happen that the automated variable selection provides models and coefficients that contradict the first principles, for example, the wrong sign of a parameter. Before a whole new theory is developed, more models should be tested.
- The concept of a best model may also depend on the purpose: In prediction, for example, it is important that the optimal model contains variables that are available.

In statistical regression modelling we have dealt with three important elements:

- **Residual analysis.** In the residual analysis, we check as well as possible the assumptions of independence, constant variance, linearity and normal distribution. We also try to identify influential observations. Robust methods are suitable for carrying out this analysis very efficiently.

- **Transformations.** Depending on the situation, it makes sense to transform the response as well as the explanatory variables, to add interactions or to expand the models with factor variables.
- **Variable selection.** Nowadays, it is easy to automatically select explanatory variables. That this is not in any case advisable has been explained at the beginning of this section.

In which order should you make these modelling steps? Or should some or all steps be repeated? When are you finished?

There are no conclusive answers. Statistical modelling is seen more as an art than a science.

Overall strategy.

1. Understand the problem. Are there already approaches for models available?
2. Collect and process data.
   - Clarify the coding of missing values (sometimes coded with -99 or 0). Define the treatment of missing values in the analysis.
   - Unify the meaning of the number 0 in different variables.
   - Treat data according to first-aid transformations unless there are valid reasons against it (eg. existing model).
   - Keep track of the data quality when merging multiple data sets. The overall quality is never higher than the weakest link.
3. First model fit (preferably with robust methods).
4. Residual analysis. Does the data help to solve the question? If not, repeat the first, second and third step.
5. Variable selection, treat collinearity if necessary.
6. Check goodness of fit.
   - Residual analysis with selected models.
   - Match models with first principles.
   - Validate models with data not used in modelling.

Final remarks. In this introduction, important elements of multiple linear regression analysis have been addressed. Lots of different generalisations exist and go in different directions.

- **Nonlinear regression**, cf. [2]. More complicated functions can be fitted.
- **Nonparametric regression**, cf. [35]. Less assumptions on the models are made (smoothing).
- **Generalised linear model**, cf. [20]. The response variable is allowed to have error distribution models other than the normal distribution (logistic, Poisson and gamma-regression).

- **Survival analysis**, cf. [39]. Censored data (reliability theory, survival and hazard function).
- **Time series analysis**, cf. [4]. The errors can be correlated.
- **Principal component** and **partial least squares regression** , cf. [42]. (Many) more variables than observations are available: large p, small n (chemometrics).

For all these generalisations there are very good R-packages. Use R!

11.4 Exercises

Problem 11.4.1 (**Experiment on a Dairy Product**, cf. [32], Reg5, Problem 1)**.** We come back to Ex. 9.4.4.

a. Fit the model

$$\log(\texttt{O2UP}) = \beta_0 + \beta_1 \cdot \texttt{BOD} + \beta_2 \cdot \texttt{TKN} + \beta_3 \cdot \texttt{TS} + \beta_4 \cdot \texttt{TVS} + \beta_5 \cdot \texttt{COD} + \varepsilon.$$

Perform a variable selection with a backward elimination using AIC.

b. Fit the trivial model

$$\log(\texttt{O2UP}) = \beta_0 + \varepsilon.$$

Perform a variable selection with a forward selection using AIC.

c. Fit the model

$$\log(\texttt{O2UP}) = \beta_0 + \beta_1 \cdot \texttt{BOD} + \beta_3 \cdot \texttt{TS} + \varepsilon.$$

Perform a variable selection with a search in both directions using AIC.

d. Compare your results. Which model(s) would you like to investigate further?

e. Check for collinearity in the data.

f. Repeat the above without the observations with index $i = 2$, 3 and 17. How sensitive is variable selection with respect to these observations?

Problem 11.4.2 (**Sniffer Data**, cf. [43], Chap. 8.3)**.** When gasoline is pumped into a tank, hydrocarbon vapors are forced out of the tank and into the atmosphere. To reduce this significant source of air pollution, devices are installed to capture the vapor. In testing these vapor recovery systems, a sniffer measures the amount recovered. To estimate the efficiency of the system, some method of estimating the total amount given off must be used. To this end, a laboratory experiment was conducted in which the amount of vapor given off was measured under controlled conditions. Four predictors are relevant for modelling:

`TankTemp`	initial tank temperature [in °F]
`GasTemp`	temperature of the dispensed gasoline [in °F]
`TankPres`	initial vapor pressure in the tank [in psi]
`GasPres`	vapor pressure of the dispensed gasoline [in psi]

The response is the hydrocarbons (`Hydrocarbons`) emitted in grams. The data of 32 measurements is given in the data frame `sniffer.dat`.

a. Draw a pairs-plot of the complete data set. Report all your findings which seem to be important with respect to regression modelling of the hydrocarbons.

b. Check for collinearity in the data.

c. Fit the model

$$\begin{aligned}\texttt{Hydrocarbons} = \beta_0 + \beta_1 \cdot \texttt{TankTemp} + \beta_2 \cdot \texttt{GasTemp} \\ + \beta_3 \cdot \texttt{TankPres} + \beta_4 \cdot \texttt{GasPres} + \varepsilon.\end{aligned}$$

Report all your findings studying the summary-lm-output.

d. Perform a variable selection using AIC.

e. Which model(s) would you like to investigate further?

Problem 11.4.3 (**Thrust of Jet-Turbine Engines**, cf. [23], Problem 12-108)**.** Thrust of jet-turbine engines `y` depends on the following variables:

`x1`	=	primary speed of rotation
`x2`	=	secondary speed of rotation
`x3`	=	fuel flow rate
`x4`	=	pressure
`x5`	=	exhaust temperature
`x6`	=	ambient temperature at time of test

Since the exact relationship is unknown, it must be modeled empirically based on the data in the data set `jet-engines.dat`. Develop models which can then be discussed with the experts.

Part III

Design of Experiment

Chapter 12

Design of Experiments

Scientific studies collect empirical data either through experiments or surveys. Only in rare cases it is possible to choose freely between these two forms of data collection. However, a distinction is important because the two forms lead to different interpretations of the results. In experiments, we create situations which are clearly defined. The analysed variables are systematically varied and the other influences eliminated as far as possible. By means of measurements, surveys or observations the events are recorded. The resulting data forms the basis for answering the actual questions that led to the experiment.

In a survey, subjects or objects are observed in the context of an existing situation. The influencing variables to be analysed are not directly adjustable. Surveys may be observational studies of natural processes, such as the noise levels in railway stations, gymnastic halls, etc. or sample surveys such as in opinion polls or in quality control. In surveys, people record systematically potential influencing factors, so that their effects can be estimated from these factors.

Many scientific studies are concerned with the understanding of causal relationships. In general, it is much easier to investigate causal relationships with experiments than surveys:

- In an experiment at least some of the influencing factors can be specifically and optimally adjusted. Surveys rely on the conditions that are currently present and we are often confronted with situations in which some variables remain practically constant.
- In a survey often (too) many variables are recorded and we will be confronted with random correlations which are very difficult to interpret.

To prove a causal relationship between the response and the explanatory variables we prefer experiments. On the other hand, in many situations only surveys are feasible, because experiments are excluded due to ethical, financial or technical reasons.

Statistical **design of experiment** (DoE) refers to the process of planning an experiment so that the data can be evaluated with statistical methods in a convenient way. However, the evaluation as well as the quality of the results essentially depends on the chosen test plan. In a scientific environment, it is important to interpret the planning and the analysis as a whole, so that the results of the experiment are sound and unbiased. In order to better understand the considered processes and systems, potentially influential input quantities are systematically varied so that the reasons for the changes in the output quantities can be identified and understood.

In this third part we will be interested in the following aspects.

- **Comparison of treatments.** The most common aim in an experiment is to compare several treatments and to choose the best one.
- **Variable screening.** If there are lot of potentially influential factors in a process it is common to try to find the most important factors with a systematic screening experiment.
- **Response surface methodology.** If once the most important factors of a process are determined the optimal setting of these factors is often searched.

12.1 Terminology and Concepts

The response variable is a continuous measurable variable. The explanatory variables can be either continuous or discrete, but many designs rely on explanatory factor variables. In both cases we use regression analysis as our main tool.

Design of experiment is based on a few basic elements. It is important that the bias, i.e. systematic error, and the variance of the measurements are small. Moreover, the repeatability of measurements with equal test conditions should be guaranteed. To counteract systematic errors often **randomised experiments** are designed. We will learn about these things with the following example.

Example 12.1.1 (**Strength Testing of Concrete**, cf. [9], Example 4.3)**.** It has been shown that the strength of concrete can vary from one batch to another. Therefore, concrete samples are taken from each batch, formed into small test cylinders, and then dried in a climatic chamber, where the temperature and humidity can be controlled. The batches can stay up to 28 days in such a climate chamber. Then the strength is measured[1].

We want to compare three different drying methods. From ten different batches, we take three samples each and form them into cylinders. Each of the three cylinder is dried out according to one of the three methods.

The question of whether the three methods differ appears to be answered by a simple design. But the problem is now that the strength of concrete varies from one batch to another and therefore measurements within one batch can differ systematically to another one. Therefore it might be better to also control the influence of the batch.

In general the response variable depends on many explanatory variables which can be separated in two groups.

- **Primary variables** are directly connected with the study.
- **Secondary variables** are also controllable with a reasonable effort.

In an experiment, primary and secondary variables can be varied in a prescribed controlled manner. Further variables are unknown and will increase the random error in the model. The random error is nothing else than the sum of all unknown influences.

[1] ASTM C 143, Standard Method for Slump of Portland Cement Concrete

How should we choose the experimental conditions so that the primary questions can be answered as accurately as possible? First, we have to recognise all kinds of influences. One possible tool is the **cause-and-effect diagram**, cf. Chap. 1.2.4.
Second, we recapitulate a few facts from regression analysis.

Remark 12.1.1 (**Design of Experiment in Regression**)**.** If a primary variable x has a continuous range, i.e. a concentration, temperature, length, etc., then we can choose the values x_i at which we will make our measurements without any condition. The variance of the slope is minimal if the sum of squares S_{xx}, cf. Eqn. 7.2.g, in the denominator is maximal. Therefore, to estimate as precisely as possible the slope of a linear relationship it is important to choose a wide range of x-values and as many y-values as possible (for each x-value the same number). On the other hand, linearity is often only approximately true and becomes less accurate as the range of x-values becomes larger. In practice the x-values are within some reasonable bounds. Within these bounds we should choose the x_i equally spaced to obtain a good estimate of the corresponding slope and to be able to detect a deviation from linearity or other assumed relationships. If x is a factor, then the same number of y values should be obtained for all possible levels.
If several primary variables are involved, then choose them as independent as possible. In other words, choose a design with uncorrelated explanatory variables. We denote such a design **orthogonal**. In the case of factors we have a **balanced design**. In the simplest case, choose the same number of observations for all combinations of levels of factors. In a balanced design a factor can be omitted without changing the estimate of the effect of the others, but with increasing variance of the errors and loosing power.

To increase the accuracy and the interpretability of the results it is a good idea to record also secondary variables. In a model with only primary variables, the variance of the random errors is larger because it also contains the influences of the secondary variables. Most confidence intervals include an estimate of random errors as a factor and are therefore larger. If we include secondary variables, test statistics will be more powerful and confidence intervals narrower. But, if we include non-significant secondary variables then we will have the opposite effect of loosing only degrees of freedom, i.e. precision. We should not forget that if variables with significant influence on the response are unconsidered, then the interpretation of the results for the primary variables is questionable.
Another possibility to reduce the random error is the use **block designs**. In Ex. 12.1.1 the blocks are the different batches. The block effect in the model explains the differences between the batches and therefore reduces the remaining random error. In many applications we are confronted with examination units which can be used for block designs, i.e. production lot, origin of raw materials, age groups of patients, etc.
As a consequence we should try to keep the secondary variables as constant as possible. This is of course easier in laboratory than in field studies. On the other hand, the findings with constant secondary variables might not be generalisable to other situations with different settings of the secondary variables. Therefore block designs with as little variance as possible within the blocks and as large variance as possible between the blocks are optimal.

Too often, the **chronological order** of the observations has an impact on the response variable. This leads to temporal trends and correlations between the random errors of the model. The order is a secondary variable that always exists and should always be used to validate the assumptions. To avoid difficulties, it can be used as a block factor. **Randomisation** is an

other possibility to avoid the effect of the order. After the list of all experimental conditions has been defined, use random numbers to select which condition will be realised first, which will be the second, etc. This makes the chronological order independent of all explanatory variables and therefore cannot lead to misinterpretations. The same consideration applies to the assignment of experimental conditions to units of investigation.

It is useful to perform multiple measurements for the same experimental conditions, so-called **replicates**. Like any increase in the number of observations this leads to an improved accuracy of the statements. With replicates an estimate of the variation of the random errors can be obtained and in the analysis of variance the existence of interactions can be tested. Replicates are not allowed to be measured consecutively, otherwise they are not independent. They also need to be randomised despite the increased experimental effort. Often, instead of increasing the number of measurements, the experimenter chooses additional explanatory variables. But, many studies fail because of too extensive investigations with too few observations.

Design of Experiment - Basic Principles

- The **main question** defines the response variable and the primary explanatory variables and their relevant value range.
- Record as many variables which might have an influence on the response as possible. These secondary variables should, if possible, be kept constant or used for **blocking**. If this is not possible, they should still be considered in the model.
- All explanatory variables should be **orthogonal** and the factors **balanced**.
- The assignment of experimental conditions to study units should be **randomised**.
- Independent **replicates** allow additional model validation.

12.2 Experimental Design

In order to draw the most sound and unbiased conclusions from an experiment, it is advisable to choose the settings of the various factors as well as possible. The list of experimental conditions that determine how each factor is to be varied at which levels is called the **experimental design**. We will limit ourselves to **factorial design** were the explanatory variables are factor variables.

The **completely randomised design** applied to one multi-level primary factor variable is the simplest possible design. One or more (but always the same number) measurements are made per level. The design is processed in random order so that systematic effects can be avoided.

A block design is used when, in addition to the primary factor, a secondary factor influences the target variable. If every level of the primary factor (eg. treatment) is used at least once in each block, it is called a **complete block design**.

If k factors with L levels each are examined the **complete factorial design** which contains all combinations of experimental conditions is used. Already with $k = 3$ factors with $L = 5$

levels a design with $5^3 = 125$ measurements is needed. Unfortunately, these are often already too many measurements.
If we assume that there are no interactions, we can also estimate primary effects with fewer combinations. They are called **incomplete designs**. Incomplete designs which are balanced are called **fractional factorial designs**.
In screening experiments, many potential factors are investigated for their influence on the response variable. To keep the effort as low as possible, only two levels (*high* and *low*) per factor are examined. They are called a 2^k **factorial designs**. To save even more resources, often only a 2^k **fractional factorial designs**, i.e. a balanced part of the complete 2^k design is realised.

12.3 Planning and Conducting a Study

Most scientific studies that deal with empirical data show similar work processes and often also similar difficulties. The following enumeration contains a lot of self-evident information. Experience shows that it is nevertheless worthwhile to consciously go through the points.

1. **Scientific question.** The essence of each study is, of course, the scientific question. The ideal of a potentially important question is often opposed to the reality that it cannot be directly and comprehensively investigated in an empirical study. Compromises are necessary. But it should not happen that experiments and observations are carried out without precise questions and hypotheses. The necessary study of the specialist literature is worthwhile, it saves a lot of inefficient practical work.

2. **Response and explanatory variables.** Due to the question, it is most often more or less clearly defined which is the response variable. The important explanatory variables may be determined with a cause-and-effect diagram. It may be more difficult to choose appropriate factor levels for continuous variables. On the one hand, they should be set to produce measurable effects in the response variable if the variable is significant. On the other hand, the levels should not be so far apart that there is a potential optimum in between.

3. **Preliminary experiments.** Own experience with the topic and the measurement equipment is important. In addition, in this step it can be verified whether the response variable has been chosen appropriately, i.e. with regard to the measurability and reproducibility.

4. **Planning.** The question must be stated more precisely in relation to the type of observable data so that possible statistical models (such as analysis of variance and regression models) can be formulated. How should the data look for the hypothesis under investigation to be rejected or not? It is worth going through this question, eg. with simulated data. In relation to the effort a design is determined. As a result of the planning, the evaluation of the data in relation to the main questions should be clear. Of course, the actual data collection has to be well planned.

5. **Sample size.** In principle, it should be stated here that the required sample size must be calculated using statistical methods. This always requires an approximate magnitude of the random fluctuations and the size of the expected effects. If both are available

from the literature or previous experiments, the required sample size can be calculated. Preliminary experiments are usually only suitable for estimating an order of magnitude of random fluctuations. From this it can be roughly calculated how large, given a number of observations, the effects must be, so that they can be detected with high probability. If, as usual, the number of observations is limited by the possible effort, this rough estimate is always worthwhile in order to assess the chances of success of the entire study. For reliable estimations of the sample size, however, more accurate estimates of the random variation than can be obtained from preliminary experiments are necessary. Often, these considerations lead to the insight that the question must be downsized.

6. **Data.** Preliminary experiments should ensure that the quality of the data is high from the start. During the main phase, the measuring methods should only be changed in an emergency. It is important to keep a journal that documents the process of data acquisition and highlights any peculiarities and unforeseen circumstances that may later explain unexpected data discrepancies. At the same time the data acquisition, control and storage on the computer should be started. At this stage, it goes without saying that the statistical methods and the software used, are well understood.

7. **Data cleanup.** Checking the plausibility of the data and visualising it saves a lot of effort. This in-depth data control can be started in parallel to the data acquisition, but not completed. It becomes difficult when there are still shortcomings in the experiment and the data acquisition: Care must be taken to see whether improving quality is more valuable than homogenising the data and adhering to the design exactly.

8. **Evaluation, interpretation.** The methodology for evaluating the data in relation to the main question should already be clear. Careful review of model assumptions and necessary model adjustments, however, takes some time. Finally, new questions and hypotheses often emerge, which can be examined more precisely in the sense of exploratory analysis.
   The interpretation of the results in a technical context can be clear – especially if the study was carefully planned. However, several interpretations and several models can also be compatible with the data. The inferences must be clearly formulated and, as far as possible, clearly justified by the individual results of the data analysis.

9. **Report, presentation.** This brings us to the writing of the report, which many consider the most humble part of a study. However, it is clear that the usefulness of the research is limited by the readability of reports and the comprehensibility of presentations.
   Many useful guidelines for writing reports exist. Before writing a report ask yourself the question: For whom is the report mainly intended? The main part of the report should be easy to understand for the target audience.

10. **Completion.** It is worthwhile to collect and discuss the experiences made with all participants.

12.4 Exercises

Problem 12.4.1 (**Experiment on Boys' Shoes**, cf. [3], Chap. 3.2)**.** A shoe manufacturer wants to investigate the wear of shoe soles made of different materials. He wants to test the soles of the shoes of sixteen boys.
What are your recommendations, if he wants to test two different materials?

Solutions

Solution 12.4.1. Since one can assume that the boys use the shoes differently, it makes sense to let each boy test both materials of soles at the same time. This eliminates the large variance between the individual boys.
In order to eliminate all effects from left and right as well, the soles of the shoes are randomly assigned to the right and left feet. It is only important that every boy gets both materials.
The data collected with this experiment can be evaluated with a two way analysis of variance, cf. Chap. 13.2, using the sole as a factor and the boy as a block factor. The effects of a different wear on the left and right shoe are described by the random error.

Chapter 13

Factorial Designs and Analysis of Variance

In regression and analysis of variance we are interested in situations with only one response variable which is directly measurable or observable. In general this response variable is influenced by several explanatory variables or factors. The knowledge of the explanatory variables and factors is essential to be able to interpret the response correctly.

In these situations, we represent the response variable as a function of explanatory variables superimposed by a random error. This approach allows us to investigate the influence of one explanatory variable, taking into account the other influences. Careful planning is essential in order to successfully complete the study that will investigate such dependency.

In order to evaluate the data of an experiment, models are generally used in which the response variable is expressed as a function of nominal explanatory variables, so-called factors. In Part II, Chap. 9.2.4 we showed how such factors can be used in a regression model. Part of the analysis of variance can therefore be regarded as a special case of the regression analysis. Since only factor variables are involved it is worthwhile to consider this special case separately.

The generic term for analysis of variance and regression are **linear models**.

13.1 One Factor Analysis of Variance

The comparison of two groups is a basic technique in statistics and is dealt with in each introductory course. But often in an experiment more than two groups are examined.
Let us look at the following example of the tensile strength of paper.

Example 13.1.1 (**Tensile Strength of Paper**, cf. [23], Example 13-2.1)**.** A manufacturer of paper which is used for making paper bags is interested in improving the tensile strength of the paper.

Product engineering believes that tensile strength depends on the hardwood concentration in the pulp. The range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration 5%, 10%, 15% and 20%. They decide to make up six test specimens at each concentration level by using a pilot plant. All 24 specimens are tested on a laboratory tensile tester in random order.

The measured tensile strength [in psi] from this experiment are reported in the following table.

Hardwood concentration [%]	Observations 1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

The question of the effects of different treatments can first be examined by comparing each group to each other by a **two-sample test**, eg. by the Wilcoxon rank sum test, Student's t-test or the sign test. The results for a given test are often summarised in a symmetric matrix of P-values.

```
#   read data
> file <- "paper.dat"
> data <- read.table(file, header=TRUE)
> data
   Strength Conc
1         5    5
2         7    5
3         8    5
4        15    5
5        11    5
6         9    5
7        10    5
...
28       20   20
#   Hardwood concentration as a factor
> data$Conc <- as.factor(data$Conc)
> str(data)
'data.frame':   28 obs. of  3 variables:
 $ Strength: int  5 7 8 15 11 9 10 10 12 17 ...
 $ Conc    : Factor w/ 4 levels "5","10","15",..: 1 1 1 1 1 1 1 2 2 2 ...

#   Pairwise comparisons using t tests with non-pooled standard deviations
> pairwise.t.test(x=data$Strength,
+                 g=data$Conc,
+                 paired=FALSE,
+                 pool.sd=FALSE,
+                 alternative="two.sided",
+                 p.adj="none")

Test:  Pairwise comparisons using t tests with non-pooled SD
data:  data$Strength and data$Conc

   5       10      15
10 0.00782 -       -
15 0.00039 0.22659 -
20 8.4e-06 0.00237 0.00331

P value adjustment method: none
```

We conclude that there are significant differences between the different treatments.

Notched box plots allow us to graphically determine which pairs of groups differ significantly according to the corresponding test: If the notches do not intersect then the corresponding groups are significantly different, cf. Fig. 13.1.i.

```
#   notched box-plot
> boxplot(Strength ~ Conc, data,
+         notch=T,
+         col="gray",
+         ylim=c(0,30),
+         xlab="Hardwood Concentration [%]",
+         ylab="Tensile Strength of Paper [psi]",
+         main="Tensile Strength of Paper versus Hardwood Concentration")
> grid()
> abline(h=0)
```

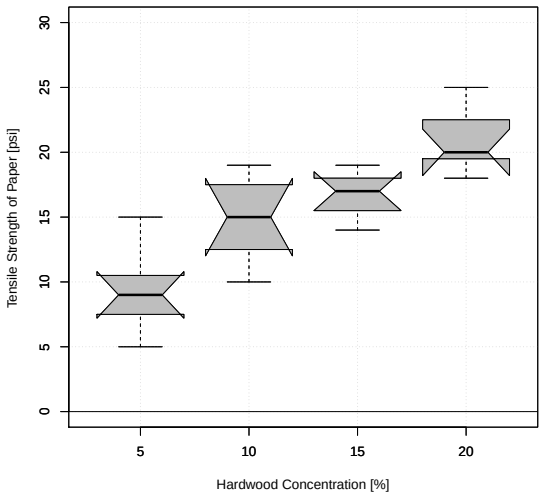


Figure 13.1.i: Notched box plots. With small samples sizes, it can happen that the notch protrudes beyond the actual box.

The two types of evaluation lead in this case to the same result that the tensile strength does not differ significantly at hardwood concentrations of 10% and 15%.

In general, however, the two methods do not have to agree, since they are just two different methods.

Such a large number of pairwise statistical hypothesis tests has serious drawbacks. If we want to compare, eg. seven groups, then already $\frac{6 \cdot 7}{2} = 21$ paired comparison tests result. If a significance level of 5% is assumed, then it is clear that from time to time under 21 tests the null hypothesis is erroneously rejected and a type I error results.
Assume the null hypothesis that all groups obey the same model. It is rejected if the most extreme difference is significant at the five percent level. But this is much too often. How can this be avoided? The consequent answer is to ask only one single question which is answered

by one statistical test. In practice, we will ask if there is a difference between one group and another group. The corresponding null hypothesis is:

$$H_0 : \text{all groups obey the same model}$$

We describe the situation with a **linear model**. We want to compare g groups with m measurements each. The model where individual observations within the ith group scatter randomly around a common value μ_i is described by

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{with} \quad i \in \{1, 2, \ldots, g\} \quad \text{and} \quad j \in \{1, 2, \ldots, m\}, \tag{13.1.a}$$

where ε_{ij} is a random error component. We assume that the errors ε_{ij} are normally and independently distributed with mean zero and variance σ^2. The corresponding hypothesis are

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_g, \\ &H_1 : \mu_i \neq \mu_j \text{ for at least one pair } i, j \text{ with } i \neq j. \end{aligned}$$

Notice that the model can also be written as

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{with} \quad i \in \{1, 2, \ldots, g\} \quad \text{and} \quad j \in \{1, 2, \ldots, m\} \tag{13.1.b}$$

with corresponding hypothesis

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0, \\ &H_1 : \tau_i \neq 0 \text{ for at least one } i. \end{aligned}$$

The parameter μ is common to all treatments called the **overall mean**, τ_i is a parameter associated with the ith treatment called the ith **treatment effect** and ε_{ij} is a random error component. Note that if the alternative hypothesis is true, the observations will differ and not follow the same distribution.

We are looking for a test statistic that takes extreme values if the means[1]

$$\bar{y}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} y_{ij} \quad \text{for all groups} \quad i \in \{1, 2, \ldots, g\}$$

of the g groups differ. Let $n = g \cdot m$ denote the total number of observations. If the variance of the group means

$$S_{\mathrm{b}}^2 = \frac{1}{g-1} \sum_{i=1}^{g} m(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

compared to the variance of the observations within the groups

$$S_{\mathrm{w}}^2 = \frac{1}{n-g} \sum_{i=1}^{g} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot})^2,$$

[1] The . in the index of the ith group mean $\bar{y}_{i\cdot}$ indicates that we take a sum over this missing index.

is large, then we reject the null hypothesis. Therefore the test statistics for the **analysis of variance test** (ANOVA) is

$$\begin{aligned} F &= \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{S_{\text{b}}^2}{S_{\text{w}}^2} \\ &= \frac{\dfrac{1}{g-1}\displaystyle\sum_{i=1}^{g} m(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\dfrac{1}{n-g}\displaystyle\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2}, \end{aligned} \tag{13.1.c}$$

where

$$\bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{m} y_{ij}$$

is the estimated overall mean.

The terms leading to the test statistic F are summarised in the following **analysis of variance table**.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Treatments (groups)	$\text{SS}_G = \displaystyle\sum_{i=1}^{g} m(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\text{df}_G = g - 1$	$\text{MS}_G = \text{SS}_G/\text{df}_G$
Error (residuals)	$\text{SS}_E = \displaystyle\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2$	$\text{df}_E = n - g$	$\text{MS}_E = \text{SS}_E/\text{df}_E$
Total	$\text{SS}_T = \displaystyle\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$\text{df}_T = n - 1$	

It is possible to show that

$$\text{SS}_T = \text{SS}_G + \text{SS}_E. \tag{13.1.d}$$

The identity in Eqn. 13.1.d shows that the total variability in the data, measured by the **total corrected sum of squares** SS_T, can be partitioned into a sum of squares of differences between treatment means and the grand mean called the **treatment sum of squares** SS_T and a sum of squares of differences of observations within a treatment from the treatment mean called the **error sum of squares** SS_E. This is where the name **analysis of variance** comes from.

If we divide the sum of squares by the corresponding degrees of freedom we obtain two **mean squares**

$$\text{MS}_G = \frac{\text{SS}_G}{\text{df}_G} \quad \text{and} \quad \text{MS}_E = \frac{\text{SS}_E}{\text{df}_E}.$$

If we assume the null hypothesis to be true, i.e. all y_{ij} come from the same normal distribution with mean $\mu = \mu_1 = \cdots = \mu_g$ and variance σ^2, then MS_G and MS_E are unbiased estimators for σ^2.

The test statistic of the **analysis of variance test** in Eqn. 13.1.c can also be written as

$$F = \frac{\text{MS}_G}{\text{MS}_E}. \tag{13.1.e}$$

This test statistic compares the two estimates of the variance σ^2. Under the null hypothesis the F-statistic follows an F-distribution with $(\mathrm{df}_G, \mathrm{df}_E) = (g-1, n-g)$ degrees of freedom. Critical values of the F-distribution can be found in T.6–T.14. This is in fact the same test as the F-Test in Chap. 9.3.

Example 13.1.2 (**Tensile Strength of Paper**, cf. [23], Example 13-2.1)**.** We come back to Ex. 13.1.1 and describe the situation with a linear model of the form in Eqn. 13.1.b. First we have to transform the numeric variable hardwood concentration (`Conc`) into a factor. This was already necessary in Ex. 13.1.1 to draw the boxplot. We repeat this very important step here again for the sake of completeness.
Then we perform an analysis of variance with R with the function `aov()` which has the same inputs as `lm()`.

```
#   Hardwood concentration as a factor
> data$Conc <- as.factor(data$Conc)

#   ANOVA
> mod <- aov(Strength ~ Conc.fac, data)
> summary(mod)
            Df Sum Sq Mean Sq F value   Pr(>F)
Conc         3  495.3  165.08   21.57 5.37e-07 ***
Residuals   24  183.7    7.65
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the first column we find the degrees of freedom, in the second the sum of squares and in the third the mean squares. We can also read off the test statistic and its P-value. The P-value of $5.37 \cdot 10^{-7}$ implies that the effect of different hardwood concentration is significant on the 5% level. This corresponds well with the notched box-plots in Fig. 13.1.i and the pairwise comparisons using t-tests.

If a significant test result has been found, the immediate question is whether each group is different from each other, or if only some of the $\frac{k(k-1)}{2}$ possible differences are really detectable. The problem can be found in the literature under the name of **multiple comparisons** or **multiple contrasts**, cf. [21]. Several methods that solve this problem are known, eg. the method according to Bonferroni.

13.2 Two Factor Analysis of Variance

The one factor analysis of variance is the generalisation of the comparison of two independent samples. We can further generalise to the analysis of variance with several factors.

Example 13.2.1 (**Strength Testing of Concrete**, cf. [9], Example 4.3)**.** In Ex. 12.1.1 the strength of concrete is potentially dependent on the drying method and the batch. The measured data from the blocked experiment is reported in the following table.

	Strength [MPa]		
Batch	Method 1	Method 2	Method 3
1	30.7	33.7	30.5
2	29.1	30.6	32.6
3	30.0	32.2	30.5
4	31.9	34.6	33.5
5	30.5	33.0	32.4
6	26.9	29.3	27.8
7	28.2	28.4	30.7
8	32.4	32.4	33.6
9	26.6	29.5	29.2
10	28.6	29.4	33.2

The data consists of two **factors** `Method` and `Batch` with 3 and 10 **levels**.

A graphical representation of the data, cf. Fig. 13.2.i, shows that Method 1 results in the lowest strength for all batches. However, Method 1 sometimes leads to values in one batch that would be highest in other batches.

```
#   read data
> file <- "concrete.dat"
> data <- read.table(file, header=TRUE)
> data
   Strength Method Batch
1      30.7      1     1
2      29.1      1     2
3      30.0      1     3
...
30     33.2      3    10

#   Method and Batch as a factor
> data$Method <- as.factor(data$Method)
> data$Batch <- as.factor(data$Batch)
> str(data)
'data.frame':   30 obs. of  3 variables:
 $ Strength: num  30.7 29.1 30 31.9 30.5 26.9 28.2 32.4 26.6 28.6 ...
 $ Method  : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ Batch   : Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...

#   interaction plot
> interaction.plot(x.factor=data$Batch,
+                  trace.factor=data$Method,
+                  response=data$Strength,
+                  fun=mean,
+                  ylim=c(25,37),
+                  xlab="Batch",
+                  trace.label="Method",
+                  ylab="Mean of Strength",
+                  main="Interaction Plot")
> abline(v=data$Batch, lty=3, col="gray")
> text(data$Batch, data$Strength, label=data$Method, pos=3)
```

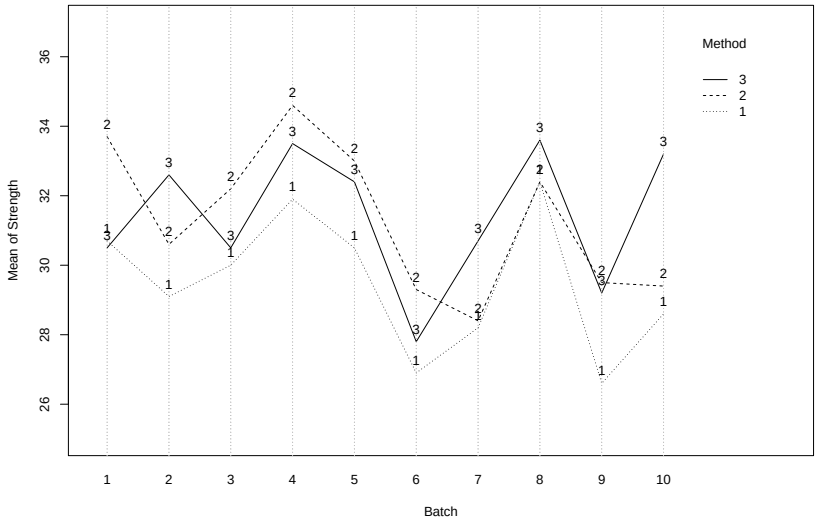


Figure 13.2.i: Interaction plot. Observations with the same drying method are connected with polygonal lines.

This observation leads to a model which allows us to statistically test the expected effect. We consider strength measurements for each method. If there are strong effects of the batch, all measurements will move up or down.

Suppose that we have two factors A and B with a and b levels. A model which is capable to separate real effects from random fluctuations can be built as follows

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \quad \text{with} \quad i \in \{1, 2, \ldots, a\} \quad \text{and} \quad j \in \{1, 2, \ldots, b\}. \tag{13.2.a}$$

In this model every observation y_{ij} has a random deviation ε_{ij} from an ideal value μ_{ij} which depends on the method i and the batch j. More precisely the ideal value

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

is based on an overall value μ, an effect α_i for all $i \in \{1, 2, \ldots, a\}$ which describes the factor A and an effect β_j for all $j \in \{1, 2, \ldots, b\}$ which describes the factor B.
To be able to estimate the parameters uniquely we need to assume two constraints, cf. the discussion in Ex. 9.2.3,

$$\sum_{i=1}^{a} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0. \tag{13.2.b}$$

The parameters μ, $\alpha_1, \alpha_2, \ldots, \alpha_a$ and $\beta_1, \beta_2, \ldots, \beta_b$ can be estimated from the model, cf. Eqn. 13.2.a,

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{with} \quad i \in \{1, 2, \ldots, a\} \quad \text{and} \quad j \in \{1, 2, \ldots, b\}.$$

For every level $i \in \{1, 2, \ldots, a\}$ we obtain

$$\begin{aligned} \bar{y}_{i\cdot} &= \frac{1}{b}\sum_{j=1}^{b}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}) \\ &= \frac{1}{b}\sum_{j=1}^{b}\mu + \frac{1}{b}\sum_{j=1}^{b}\alpha_i + \frac{1}{b}\sum_{j=1}^{b}\beta_j + \frac{1}{b}\sum_{j=1}^{b}\varepsilon_{ij} \\ &= \mu + \alpha_i, \end{aligned}$$

where we used the constraint in Eqn. 13.2.b and the fact that the errors have mean zero. For every level $j \in \{1, 2, \ldots, b\}$ we obtain

$$\begin{aligned} \bar{y}_{\cdot j} &= \frac{1}{a}\sum_{i=1}^{a}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}) \\ &= \frac{1}{a}\sum_{i=1}^{a}\mu + \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \frac{1}{a}\sum_{i=1}^{a}\beta_j + \frac{1}{a}\sum_{i=1}^{a}\varepsilon_{ij} \\ &= \mu + \beta_j, \end{aligned}$$

where we used the constraint in Eqn. 13.2.b and the fact that the errors have mean zero. Moreover, we obtain

$$\bar{y}_{\cdot\cdot} = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{b}\sum_{j=1}^{b}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}) = \mu.$$

All together we obtain the estimates[2]:

$$\begin{aligned} \hat{\mu} &= \bar{y}_{\cdot\cdot} \\ \hat{\alpha}_i &= \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \quad \text{for all } \; i \in \{1, 2, \ldots, a\} \\ \hat{\beta}_j &= \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \quad \text{for all } \; j \in \{1, 2, \ldots, b\} \end{aligned}$$

These estimates can also be derived from the least-squares criterion.

Like this we obtain all the **fitted values**

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \quad \text{with } \; i \in \{1, 2, \ldots, a\} \; \text{ and } \; j \in \{1, 2, \ldots, b\}.$$

The differences between the observed and the fitted values

$$e_{ij} = y_{ij} - \hat{y}_{ij} \quad \text{with } \; i \in \{1, 2, \ldots, a\} \; \text{ and } \; j \in \{1, 2, \ldots, b\}.$$

are again called **residuals**.

Example 13.2.2 (**Strength Testing of Concrete**, cf. [9], Example 4.3)**.** We now come back to Ex. 12.1.1. We can add the fitted values to the interaction plot in Fig. 13.2.i. The fitted model in Eqn. 13.2.a has no interactions. Therefore the differences between, eg. the first and second, polygonal lines

$$\hat{y}_{1j} - \hat{y}_{2j} = (\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_j) - (\hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_j) = \hat{\alpha}_1 - \hat{\alpha}_2,$$

[2]The . in the index of the ith group mean $\bar{y}_{i\cdot}$ indicates that we take a sum over this missing index.

are independent of the index j and therefore do not depend on the batch. The corresponding fitted polygonal lines in Fig. 13.2.ii are thus parallel.

With R we can easily estimate the fitted values with the function `aov()`. In this model both factors are treated equally, even if we are mainly interested in the factor `Method`.

```
#   parameter estimation
> mod <- aov(Strength ~ Method + Batch, data)
> data$fit <- fitted.values(mod)

#   interaction plot with fitted values
> interaction.plot(x.factor=data$Batch,
+                  trace.factor=data$Method,
+                  response=data$Strength,
+                  fun=mean,
+                  ylim=c(25,37),
+                  xlab="Batch",
+                  trace.label="Method",
+                  ylab="Mean of Strength",
+                  main="Interaction Plot")
> abline(v=data$Batch, lty=3, col="gray")
> text(data$Batch, data$Strength, label=data$Method, pos=3)
> interaction.plot(x.factor=data$Batch,
+                  trace.factor=data$Method,
+                  response=data$fit,
+                  col="red",
+                  trace.label="Fitted Values",
+                  add=T)
```

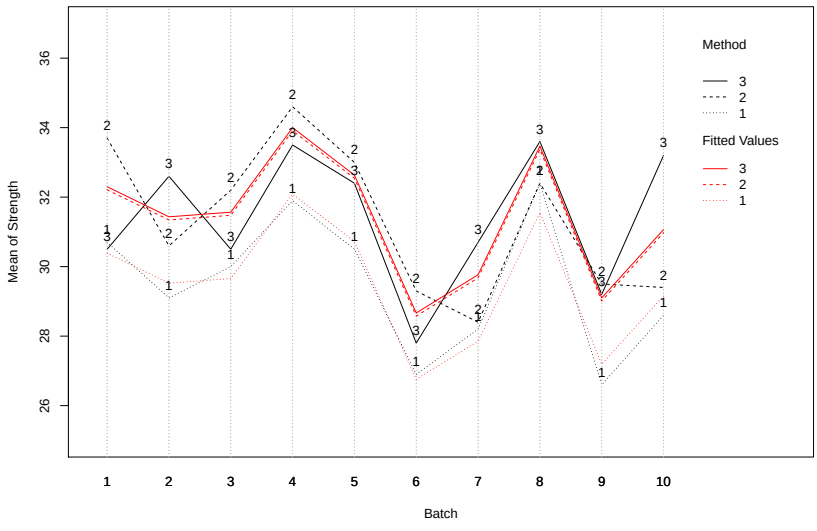


Figure 13.2.ii: Interaction plot (black) with fitted values (red).

We asked the question whether the methods differ in their strength measurements. As a basis

for statistical tests we set up a variance analysis table for two factors which is an extension of the one factor analysis of variance table.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test Statistic
Factor A	SS_A	$\mathrm{df}_A = a-1$	MS_A	$\mathrm{MS}_A/\mathrm{MS}_E$
Factor B	SS_B	$\mathrm{df}_B = b-1$	MS_B	$\mathrm{MS}_B/\mathrm{MS}_E$
Error	SS_E	$\mathrm{df}_E = n-a-b+1$	MS_E	
Total	SS_T	$\mathrm{df}_T = n-1$		

The sums of squares are defined by

$$\mathrm{SS}_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\hat{\alpha}_i^2 = b\sum_{i=1}^{a}\hat{\alpha}_i^2$$

$$\mathrm{SS}_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\hat{\beta}_j^2 = a\sum_{j=1}^{b}\hat{\beta}_j^2$$

$$\mathrm{SS}_E = \sum_{i=1}^{a}\sum_{j=1}^{b}e_{ij}^2$$

$$\mathrm{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2$$

We test the alternative hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0, \tag{13.2.c}$$

$$H_1 : \alpha_i \neq 0 \text{ for at least one } i, \tag{13.2.d}$$

i.e. all effects of the factor A are zero, with the test statistic

$$F_A = \frac{\mathrm{MS}_A}{\mathrm{MS}_E} = \frac{\dfrac{\mathrm{SS}_A}{\mathrm{df}_A}}{\dfrac{\mathrm{SS}_E}{\mathrm{df}_E}}. \tag{13.2.e}$$

Under the null hypothesis the statistic follows an F-distribution with

$$(\mathrm{df}_A, \mathrm{df}_E) = (a-1, n-a-b+1) = (a-1, (a-1)(b-1))$$

degrees of freedom. Critical values of the F-distribution can be found in T.6–T.14. This is in fact the same test as the F-Test in Chap. 9.3. Analogously, we can test the factor B.

Example 13.2.3 (**Strength Testing of Concrete**, cf. [9], Example 4.3)**.** We now come back the Ex. 12.1.1 and want to test the factors `Method` and `Batch` for significance.

First, we use explicit formula to test the effect of each factor.

```
#   estimate coefficients (explicit formulae)
> mu.hat <- mean(data$Strength)
> alpha.hat <- aggregate(data$Strength ~ data$Method, FUN=mean)[,2] - mu.hat
```

```
> beta.hat  <- aggregate(data$Strength ~ data$Batch,  FUN=mean)[,2] - mu.hat
#   number of levels
> a <- nlevels(data$Method)
> b <- nlevels(data$Batch)
#   sum of squares
> SS.Method <- b*sum(alpha.hat^2);   SS.Method
23.22867
> SS.Batch  <- a*sum(beta.hat^2);   SS.Batch
86.79333
> SS.E <- sum(residuals(mod)^2);   SS.E
24.04467
#   test statistic for each factor
> F.Method <- (SS.Method/(a-1)) / (SS.E/(a*b-a-b+1));   F.Method
8.694568
> F.Batch <- (SS.Batch/(b-1)) / (SS.E/(a*b-a-b+1));   F.Batch
7.219342
#   P-values
> 1-pf(F.Method, df1=a-1, df2=a*b-a-b+1)
0.002278362
> 1-pf(F.Batch, df1=b-1, df2=a*b-a-b+1)
0.0002021506
```

Second, we can easily obtain the above values with R.

```
#   analysis of variance table for two factors (aov)
> mod <- aov(Strength ~ Method + Batch, data)
> summary(mod)
            Df Sum Sq Mean Sq F value   Pr(>F)
Method       2  23.23  11.614   8.695 0.002278 **
Batch        9  86.79   9.644   7.219 0.000202 ***
Residuals   18  24.04   1.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the factor `Method` we find a P-value of 0.002278. Therefore the effect is significant on the 5% level. Apparently, also the effect of the blocking factor `Batch` is significant.

Often the main factor is called **treatment** and the other factor collects the observations in **blocks**. Blocks should be as homogeneous as possible, so that any differences within the block are as exclusively as possible due to the effect of the different treatments.

On the other hand, in a **two way analysis of variance** both factors are treated equally. Classic example in agriculture: the yield depends on the type of wheat and fertilizer.

13.3 Two Factor Analysis of Variance with Interaction

So far, it has been assumed that for each combination (i, j) of one level of factor A with one level of factor B, only one observation y_{ij} is recorded. In the simplest model with c observations per combination (i, j), one more index k has to be added to the previous model in Eqn. 13.2.a

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

with $i \in \{1, 2, \ldots, a\}$, $j \in \{1, 2, \ldots, b\}$ and $k \in \{1, 2, \ldots, c\}$. The observations on the same combination (i, j) are called **replicates**.

If the polygonal lines in the interaction plot, cf. 13.2.i, are not almost parallel, then the additive model in Eqn. 13.2.a has to be modified too.

The model with two factors and an **interaction** is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{13.3.a}$$

with $i \in \{1, 2, \ldots, a\}$, $j \in \{1, 2, \ldots, b\}$ and $k \in \{1, 2, \ldots, c\}$. To be able to properly estimate the coefficients we have to add further constraints on the interaction term γ_{ij}

$$\sum_{i=1}^{a} \gamma_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{b} \gamma_{ij} = 0.$$

It is very important to know that the parameter estimate of the model with interaction, cf. Eqn. 13.3.a, needs more than one observation per combination (i, j).
The estimates for the parameters can be found analogously as in Chap. 13.2, for the model with two factors. We obtain the estimates[3]:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\ldots} \\
\hat{\alpha}_i &= \bar{y}_{i\,.\,.} - \bar{y}_{\ldots} \quad \text{for all } \; i \in \{1, 2, \ldots, a\} \\
\hat{\beta}_j &= \bar{y}_{.\,j\,.} - \bar{y}_{\ldots} \quad \text{for all } \; j \in \{1, 2, \ldots, b\} \\
\hat{\gamma}_{ij} &= \bar{y}_{ij\,.} - \bar{y}_{i\,.\,.} - \bar{y}_{.\,j\,.} + \bar{y}_{\ldots} \quad \text{for all } \; i \in \{1, 2, \ldots, a\}, j \in \{1, 2, \ldots, b\}
\end{aligned}$$

As a basis for statistical tests we set up a variance analysis table for two factors with an interaction:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test Statistic
Factor A	SS_A	$\mathrm{df}_A = a - 1$	MS_A	$\mathrm{MS}_A/\mathrm{MS}_E$
Factor B	SS_B	$\mathrm{df}_B = b - 1$	MS_B	$\mathrm{MS}_B/\mathrm{MS}_E$
Interaction I	SS_I	$\mathrm{df}_I = (a-1)(b-1)$	MS_I	$\mathrm{MS}_I/\mathrm{MS}_E$
Error	SS_E	$\mathrm{df}_E = ab(c-1)$	MS_E	
Total	SS_T	$\mathrm{df}_T = n - 1$		

The sums of squares are defined by

$$\mathrm{SS}_A = bc \sum_{i=1}^{a} \hat{\alpha}_i^2$$

$$\mathrm{SS}_B = ac \sum_{j=1}^{b} \hat{\beta}_j^2$$

$$\mathrm{SS}_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} e_{ijk}^2$$

$$\mathrm{SS}_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (y_{ijk} - \bar{y}_{\ldots})^2$$

[3]The . in the index of the ith group mean $\bar{y}_{i\,.}$ indicates that we take a sum over this missing index.

and

$$\mathrm{SS}_I = \mathrm{SS}_T - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_E.$$

We test the alternative hypothesis

$$H_0 : \gamma_{11} = \gamma_{12} = \gamma_{13} = \cdots = \gamma_{ab} = 0, \tag{13.3.b}$$
$$H_1 : \gamma_{ij} \neq 0 \text{ for at least one combination } (i, j), \tag{13.3.c}$$

i.e. all effects of the interaction I are zero, with the test statistic

$$F_I = \frac{\mathrm{MS}_I}{\mathrm{MS}_E} = \frac{\dfrac{\mathrm{SS}_I}{\mathrm{df}_I}}{\dfrac{\mathrm{SS}_E}{\mathrm{df}_E}}. \tag{13.3.d}$$

Under the null hypothesis the statistic follows an F-distribution with

$$(\mathrm{df}_I, \mathrm{df}_E) = (ab(c-1), (a-1)(b-1))$$

degrees of freedom. Critical values of the F-distribution can be found in T.6–T.14.

Since a block cannot interact with the treatment, no interactions with block designs are allowed.

In R it is easy to define models with interactions. Assume that we have two factors `A` and `B` and the response `Y`, then the simplest model with an interaction term is

$$\texttt{Y} \sim \texttt{A} + \texttt{B} + \texttt{A*B}$$

or even simpler:

$$\texttt{Y} \sim \texttt{A*B}$$

13.4 Residual Analysis

Analysis of variance and regression are very similar. It is also an essential step in analysis of variance to check the model assumptions with the help of residual analysis.
Almost all the tools in Chap. 10.1 can also be used in analysis of variance: Tukey-Anscombe plot, scale-location plot, q-q plot, residuals versus explanatory variables and residuals versus time.

In analysis of variance the residuals versus leverage is not useful, since all explanatory variables have the same leverage.

Example 13.4.1 (**Tensile Strength of Paper**, cf. [23], Example 13-2.1)**.** We want to check the model assumptions of Ex. 13.1.2. To be able to use the function `plot.lmSim()`, written by A. Ruckstuhl, we need to fit the model with `lm()`.

```
#   fit with lm
> mod.lm <- lm(Strength ~ Conc, data)
> summary(mod.lm)
Call:
lm(formula = Strength ~ Conc, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.8571 -1.8929 -0.0714  1.7857  5.7143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.286      1.046   8.880 4.74e-09 ***
Conc10         5.571      1.479   3.767 0.000947 ***
Conc15         7.429      1.479   5.023 3.92e-05 ***
Conc20        11.714      1.479   7.921 3.76e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.767 on 24 degrees of freedom
Multiple R-squared:  0.7294,    Adjusted R-squared:  0.6956
F-statistic: 21.57 on 3 and 24 DF,  p-value: 5.375e-07
```

The P-value of the F-test is indeed the same!

```
#   diagnostic tools
> source("RFn_Plot-lmSim.R")
> op <- par(mfcol=c(1,3))
>     #   Tukey-Anscombe plot
>     plot(mod.lm, which=1, pch=20)
>     grid()
>     abline(h=0, lty=3)
>     plot.lmSim(mod.lm, which=1, SEED=4)
>     grid()
>     abline(h=0, lty=3)
>
>     #   scale-location plot
>     plot(mod.lm, which=3, pch=20)
>     grid()
>     abline(h=0, lty=3)
>     plot.lmSim(mod.lm, which=3, SEED=4)
>     grid()
>     abline(h=0, lty=3)
>
>     #   q-q plot
>     plot(mod.lm, which=2, pch=20)
>     grid()
>     plot.lmSim(mod.lm, which=2, SEED=4)
>     grid()
> par(op)
```

There is nothing noticeable in the diagrams of Fig. 13.4.i. Therefore we can assume that the conclusions are valid.

13.5 Exercises

Problem 13.5.1 (**Compression of Single-Wall Containers**, cf. [9], Example 2.14)**.** The article *Compression of Single-Wall Corrugated Shipping Containers Using Fixed and Floating*

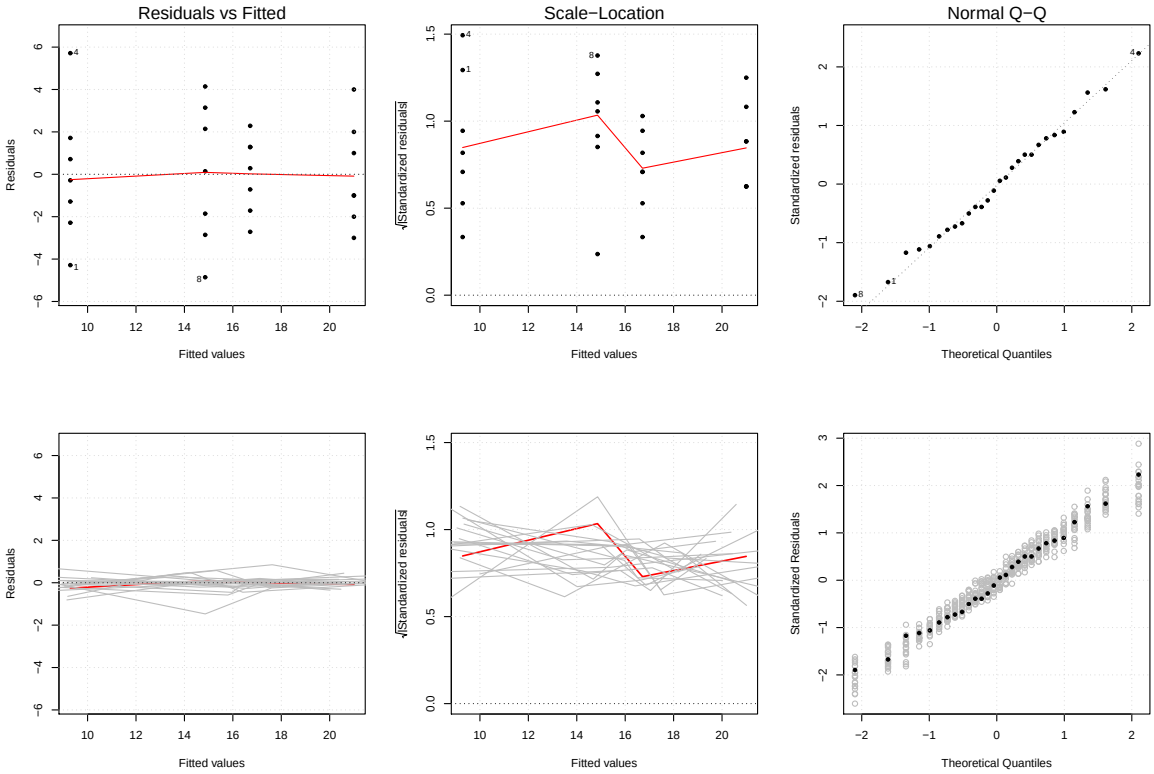


Figure 13.4.i: Residual analysis: Tukey-Anscombe plot, scale-location plot and q-q plot (upper row), corresponding bootstrap-simulations (lower row).

Test Platens in the Journal of Testing and Evaluation, 1992, p. 318-320, describes an experiment in which several different types of boxes were compared with respect to compression strength.

Type of Box	Compression Strength [lb]					
1	655.5	788.3	734.3	721.4	679.1	699.4
2	789.2	772.5	786.9	686.1	732.1	774.8
3	737.1	639.0	696.3	671.7	717.2	727.1
4	535.1	628.7	542.4	559.0	586.9	520.0

a. Enter the data yourself in R and represent it with notched box-plots. You need to define a factor (`Box`) for the type of box.

b. Is there a difference between different types of boxes? Perform a statistical hypothesis test on the 5% level.

Problem 13.5.2 (**Energy Consumption of Dehumidifier**, cf. [32], DoE2, Problem 3)**.** A consumer protection organisation compares the annual energy consumption of five different brands of dehumidifiers. Because energy consumption depends on the current humidity,

each brand has been tested at four different humidity levels ranging from moderate to high humidity. For each brand, therefore, four devices were randomly assigned to humidity levels. The mean energy consumption resulting from the experiment [in kWh] is recorded in the following table:

	Humidity Level			
Brand	1	2	3	4
1	685	792	838	875
2	722	806	893	953
3	733	802	880	941
4	811	888	952	1005
5	828	920	978	1023

a. Enter the data yourself in R. You need to define two factors `Brand` and `Humidity`.

b. Is there a difference in energy consumption between the five different brands? Perform a statistical hypothesis test on the 5% level.

c. Is there a difference in the different humidity levels? Perform a statistical hypothesis test on the 1% level. Do your findings support the idea to use the block factor `Humidity`?

d. Perform a residual analysis.

e. Check the additivity of the model with an interaction plot.

Problem 13.5.3 (**Fuel Consumption of Cars**, cf. [10], p. 191)**.** In three US cities (`Town`) the fuel consumption of five types of cars was measured (`Car`). For each combination, three test drives were performed. The response variable (`KMP4L`) is the distance [km] the car was able to drive with 4 liters fuel. The data is given in the data frame `car-fuel.dat`.

a. Display the variables `Car` and `KMP4L` graphically.

b. Analyse the response variable `KMP4L` with a two factor analysis of variance. Use the full model with interaction. Are the interactions necessary at the a 5% level?

c. Perform a residual analysis.

d. Represent the means with an interaction plot. How can the significant interaction be explained?

e. Perform a one-way analysis of variance for each city individually.

f. Repeat subtask **b.** without the values of San Francisco.

Problem 13.5.4 (**Analysis of Variance**)**.** Prove Eqn. 13.1.d.

Solutions

Solution 13.5.4. See Chap. 10.1 in [36].

Chapter 14

Screening Experiments

A production process is usually influenced by many factors. The factors that can be actively varied make it possible to control the process. In contrast to the influences of the other factors (which are beyond our control), the production process or the product should be as insensitive or robust as possible. We are now concentrating on an essential factor of the product that should become the target of our investigations. This target may be modified (improved or worsened) by a change in the production condition. In order to find the best production condition, we have to know how the essential factors interact.

If we want to experimentally investigate the interaction of several factors on the response variable we must observe the process under different production conditions.

14.1 2^k Factorial Designs

The number of experiments, often referred to as **runs**, rapidly increases to an unrealistically large number the more factors are involved. To reduce the number of runs we only use two levels (*low* and *high*) per factor. With k factors we obtain 2^k runs. Such a design is commonly denoted a 2^k factorial design.

With a 2^k factorial design we can efficiently code the different runs by two coding schemes. One is based on $+$ and $-$ signs and the other on lowercase letters. In the plus-minus coding scheme $+1$ is used to denote one level of a factor, often called the high level, whereas -1 is used to denote the low level.
The plus-minus coding scheme allows for a quick, complete listing of all 2^k single tests on k factors. The plus-minus coding scheme consist of a table with k columns with plus and minus signs which is constructed with the following rules:

1. column: Starting with -1, create a column of length 2^k by alternating -1 and $+1$ values.

2. column: Create a column of alternating blocks of two -1 values and two $+1$ values.

3. column: Create a column of alternating blocks of four -1 values and four $+1$ values.

4. column: Create a column of alternating blocks of eight -1 values and eight $+1$ values.

Continue in this manner, using block sizes that are successive powers of 2, until all k columns have been formed.
Placing the columns side by side we obtain the **design matrix** of the experiment, in which each row specifies a particular combination of factor settings. In other words, each row is one of the 2^k experimental test runs. The order in which these runs are listed in the design matrix is called **Yates standard order** after Frank Yates, a colleague of R. A. Fisher who helped develop the methodology of factorial designs.

The eight experimental runs for a 2^3 design based on three factors A, B and C are listed in the following table.

Run	Factors			Letters
	A	B	C	
1	-1	-1	-1	(1)
2	$+1$	-1	-1	a
3	-1	$+1$	-1	b
4	$+1$	$+1$	-1	ab
5	-1	-1	$+1$	c
6	$+1$	-1	$+1$	ac
7	-1	$+1$	$+1$	bc
8	$+1$	$+1$	$+1$	abc

In the last column we added the alternative coding scheme with lowercase letters which indicates only the high levels of the corresponding factors A, B and C with the lowercase letters a, b and c. If the lower level has been chosen for the corresponding factor, the lowercase letter is simply omitted.

Note that for the order of the experiments the randomised order rather than Yates standard order should be used.

Example 14.1.1 (**CD/DVD Production**, cf. [9], Example 10.1)**.** Compact discs (CDs) and digital video discs (DVDs) are produced by the same process. A master disc is created by baking a photosensitive material on a round glass plate. Then, timed pulses from a laser beam etch digital signals in a tight spiral on the plate. The plate is then developed to reveal a sequence of surface pits that encode the digital information. Master plates are electroplated to produce metal stampers, which, when placed in plastic injection molding machines, press thousands of copies of the final disc.

Some of the factors that influence the mastering stage are listed here, along with the factor levels that were used in a 2^3 experiment. The aim of the experiment was to minimise jitter, which is a measure of how well the CD can be read by a CD/DVD player. The factor linear velocity is a measure of the speed with which the laser travels in a slowly increasing spiral path as it burns the pits in the photosensitive material.

Factor		Low Level	High Level
laser power	A	90%	110%
developing time	B	20 s	30 s
linear velocity	C	1.20	1.30

Data from the 2^3 design is given in the following table.

Run	Factors			Jitter
	A	B	C	
1	-1	-1	-1	34
2	$+1$	-1	-1	26
3	-1	$+1$	-1	33
4	$+1$	$+1$	-1	21
5	-1	-1	$+1$	24
6	$+1$	-1	$+1$	23
7	-1	$+1$	$+1$	19
8	$+1$	$+1$	$+1$	18

The 8 test runs were conducted in random order but they are presented in Yates standard order. With a 2^k plan it is possible to estimate both the main effects as well as interactions.

- The **main effect of a factor** is the average response value for all test runs at the high level of the factor minus the average response value for runs at the low level of the factor.
- The **two factor interaction effect** is one-half of the difference between the main effects of one factor calculated at the two levels of the other factor.

The interaction effect between the two factors A and B is denoted by $A{:}B$ and the interaction effect between the three factors A, B and C with $A{:}B{:}C$.

These definitions are best illustrated by an example.

Example 14.1.2 (**CD/DVD Production**, cf. [9], Example 10.1)**.** We come back to Ex. 14.1.1 and want to estimate the main effect of the factor laser power

$$\begin{aligned} \text{main effect for } A &= \frac{1}{4}(26+21+23+18) - \frac{1}{4}(34+33+24+19) \\ &= 22.0 - 27.5 = -5.5. \end{aligned}$$

This implies that a change of laser power from the low level of 90% to the high level of 110% reduces the jitter by 5.5 units.
We want to estimate the interaction effect between factor A and C. Fig. 14.1.i shows the interaction plot for the mean values of the response variable for all four combinations of the factor levels of A and C. That is, each point is a mean of two response values. The point where both the factor A and the factor C are set at the lower level is the mean of 34 and 33, i.e. 33.5.
If there was no interaction, the two lines in Fig. 14.1.i should be approximately parallel. The more the two straight lines deviate from the parallelism, the greater the interaction effect. Based on Fig. 14.1.i we now have the following estimation of the interaction between factor A and C

$$\begin{aligned} \text{interaction effect } A{:}C &= \frac{1}{2}((33.5-21.5)-(23.5-20.5)) \\ &= \frac{1}{2}(12-3) = 4.5. \end{aligned}$$

The estimation of the interaction effects is quite cumbersome, especially when higher order interactions have to be estimated. Fortunately there is a much simpler calculation scheme.

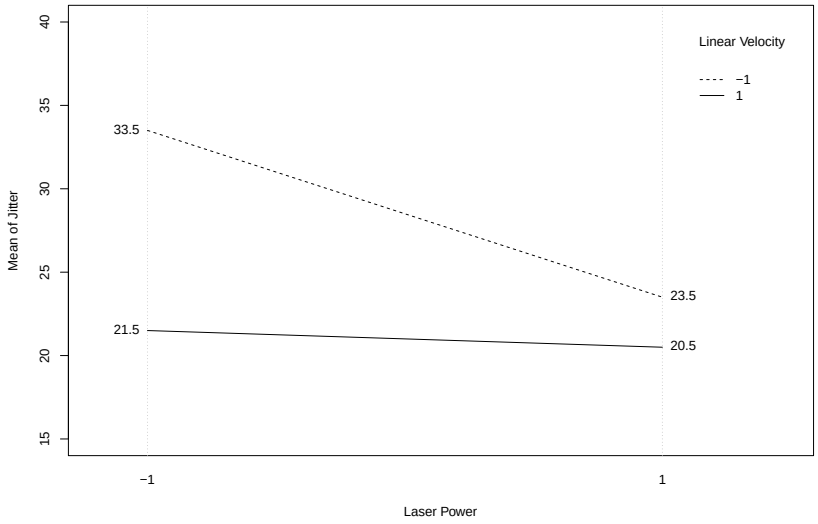


Figure 14.1.i: Interaction plot for both factors A (laser power) and B (linear velocity).

This scheme relies on expanding the matrix with additional columns by adding all possible products between the columns of the matrix. The resulting expanded **design matrix** of Ex. 14.1.1 can be found in the following table.

Run	Factors			Interactions				Jitter
	A	B	C	$A{:}B$	$A{:}C$	$B{:}C$	$A{:}B{:}C$	
1	-1	-1	-1	$+1$	$+1$	$+1$	-1	34
2	$+1$	-1	-1	-1	-1	$+1$	$+1$	26
3	-1	$+1$	-1	-1	$+1$	-1	$+1$	33
4	$+1$	$+1$	-1	$+1$	-1	-1	-1	21
5	-1	-1	$+1$	$+1$	-1	-1	$+1$	24
6	$+1$	-1	$+1$	-1	$+1$	-1	-1	23
7	-1	$+1$	$+1$	-1	-1	$+1$	-1	19
8	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	18

For instance, the column of the interaction $A:B$ is obtained by multiplying the columns of the factors A and B. Now we can estimate the corresponding effects of the interactions by taking the scalar product of the response variable with the corresponding column of the design matrix and dividing this product by half of the number of rows in the design matrix. This is a kind weighted mean. We want to estimate again the effect of the interaction between the factors A and C:

$$\begin{aligned} \text{interaction effect } A{:}C &= \frac{1}{2^{3-1}}(34 - 26 + 33 - 21 - 24 + 23 - 19 + 18) \\ &= \frac{18}{4} = 4.5. \end{aligned}$$

This is indeed the same value as above.

In general software implementations give values which are exactly one half of the effect.

Example 14.1.3 (**CD/DVD Production**, cf. [9], Example 10.1)**.** We now come back again the Ex. 14.1.1. The number of 8 runs with a 2^3 design only allows us to calculate point estimates of the effects. There are no degrees of freedom left for the residuals and it is not possible to perform statistical tests. Therefore, two measurements were made in each test situation. Of course, all experiments were done in random order, so no measurements were taken for the same experimental situation in a row. The following table shows the complete design matrix and the data from two replicated runs of the 2^3 experiment.

Run	Factors			Interactions				Jitter	
	A	B	C	$A:B$	$A:C$	$B:C$	$A:B:C$	1.	2.
1	-1	-1	-1	$+1$	$+1$	$+1$	-1	34	40
2	$+1$	-1	-1	-1	-1	$+1$	$+1$	26	29
3	-1	$+1$	-1	-1	$+1$	-1	$+1$	33	35
4	$+1$	$+1$	-1	$+1$	-1	-1	-1	21	22
5	-1	-1	$+1$	$+1$	-1	-1	$+1$	24	23
6	$+1$	-1	$+1$	-1	$+1$	-1	-1	23	22
7	-1	$+1$	$+1$	-1	-1	$+1$	-1	19	18
8	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	18	18

Now we want to analyse the data with R. First, we have to define the part of the design matrix for the three factors A, B and C and add the data.

```
#   data with replicates
> data2 <- data.frame(A=rep(rep(c(-1,1),4),2),
                      B=rep(rep(c(-1,1),c(2,2)),4),
                      C=rep(rep(c(-1,1),c(4,4)),2),
                      y=c(34,26,33,21,24,23,19,18,40,29,35,22,23,22,18,18))
> data2
    A  B  C  y
1  -1 -1 -1 34
2   1 -1 -1 26
3  -1  1 -1 33
4   1  1 -1 21
5  -1 -1  1 24
6   1 -1  1 23
7  -1  1  1 19
8   1  1  1 18
9  -1 -1 -1 40
10  1 -1 -1 29
11 -1  1 -1 35
12  1  1 -1 22
13 -1 -1  1 23
14  1 -1  1 22
15 -1  1  1 18
16  1  1  1 18
```

Second, we want to analyse the data with R.

```
#   ANOVA
> mod <- aov(y ~ A*B*C, data2)
> summary(mod)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
A            1  138.1   138.1  41.679 0.000197 ***
B            1   85.6    85.6  25.830 0.000950 ***
C            1  351.6   351.6 106.132 6.79e-06 ***
A:B          1    1.6     1.6   0.472 0.511620
A:C          1  105.1   105.1  31.717 0.000492 ***
B:C          1    0.1     0.1   0.019 0.894140
A:B:C        1    3.1     3.1   0.925 0.364446
Residuals    8   26.5     3.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.lm(mod)
Call:
aov(formula = y ~ A * B * C, data = data)

Residuals:
   Min     1Q Median     3Q    Max
  -3.0   -0.5    0.0    0.5    3.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.3125     0.4550  55.631 1.21e-11 ***
A            -2.9375     0.4550  -6.456 0.000197 ***
B            -2.3125     0.4550  -5.082 0.000950 ***
C            -4.6875     0.4550 -10.302 6.79e-06 ***
A:B          -0.3125     0.4550  -0.687 0.511620
A:C           2.5625     0.4550   5.632 0.000492 ***
B:C          -0.0625     0.4550  -0.137 0.894140
A:B:C         0.4375     0.4550   0.962 0.364446
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.82 on 8 degrees of freedom
Multiple R-squared:  0.9628,    Adjusted R-squared:  0.9302
F-statistic: 29.54 on 7 and 8 DF,  p-value: 4.188e-05
```

Since every factor has two levels the analysis with a t-test, i.e. with `summary.lm()`, is equivalent to a corresponding analysis with `summary()` of the `aov()` object. Therefore, the P-values in the first part of the R output agree with those in the second part. In the second part, the estimates for the main effects and the interactions can be found. The estimates in R, however, correspond to exactly half of the values in the manual calculation.
The effect of $A+$ is -2.9375 and the effect of $A-$ is 2.9375, together

$$\text{main effect for } A = 2(-2.9375) = -5.875.$$

Note that this value is not identical to the hand calculation in Ex. 14.1.2, where only the first 8 of the total 16 runs were considered.

From the analysis it can be concluded that the three main effects A, B and C as well as the interaction effect $A:C$ are significant. Because the design is balanced, the effects are orthogonal and we can directly draw this conclusion, in contrast to the usual situation in the regression analysis.

If only one measurement is made in each test situation, then we can estimate all the main and interaction effects, but not test them for significance (i.e. test the null hypothesis that the individual effects are zero). Nevertheless, there is a possibility to identify significant effects.

If most of the effects are zero, then the estimated effects will only detect noise due to the measurement error and therefore vary randomly around zero. If all the estimated effects are considered as realisations of the same noise, then we can graphically analyse them with a q-q plot. The significant effects, which are not noise at all, will then show up as outliers. To make the analysis even more efficient, instead of the q-q plot we use the **half-normal plot**. The half-normal plot shows the absolute values of the effects against the corresponding quantiles of the normal distribution, The quantile of the ith ordered observation can be calculated according to the formula

$$q_i = \frac{1}{2} + \frac{1}{2}\frac{i - \frac{3}{8}}{e + \frac{1}{4}},$$

where e denotes the number of estimated effects.

Example 14.1.4 (**Chemical Process Experiment**, cf. [16], Chapter 3.7.5)**.** Byproducts in a chemical process pollute the reactor, so that the production process must be interrupted from time to time to clean the reactor. To keep the pollution process under control, a 2^4 factorial design was conducted. The factors that might affect the pollution are listed in the following table:

Symbol	Factor Name
A	excess of reactant
B	catalyst concentration
C	pressure in the reactor
B	temperature of the coated mixing

The data can be found in the R-package `daewr`[1].

```
#   load package
> require(daewr)
> help(daewr)
> chem
    A  B  C  D  y
1  -1 -1 -1 -1 45
2   1 -1 -1 -1 41
3  -1  1 -1 -1 90
4   1  1 -1 -1 67
5  -1 -1  1 -1 50
6   1 -1  1 -1 39
7  -1  1  1 -1 95
8   1  1  1 -1 66
9  -1 -1 -1  1 47
10  1 -1 -1  1 43
11 -1  1 -1  1 95
12  1  1 -1  1 69
```

[1]This package contains the data sets and functions from the book [16].

```
13 -1 -1  1  1 40
14  1 -1  1  1 51
15 -1  1  1  1 87
16  1  1  1  1 72
```

If we estimate all the main effects and interactions with these 16 measurements, then we can see that no degrees of freedom are left to determine the variance of the estimated effects. That is, all the standard errors, t-values and P-values are `NA` and the residual standard error, the adjusted R-squared and the F-statistics are `NaN`.

```
#   ANOVA
> mod <- lm(y ~ A*B*C*D, data=chem)
> summary.lm(mod)
Call:
lm(formula = y ~ A * B * C * D, data = chem)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.3125         NA      NA       NA
A            -6.3125         NA      NA       NA
B            17.8125         NA      NA       NA
C             0.1875         NA      NA       NA
D             0.6875         NA      NA       NA
A:B          -5.3125         NA      NA       NA
A:C           0.8125         NA      NA       NA
B:C          -0.3125         NA      NA       NA
A:D           2.0625         NA      NA       NA
B:D          -0.0625         NA      NA       NA
C:D          -0.6875         NA      NA       NA
A:B:C        -0.1875         NA      NA       NA
A:B:D        -0.6875         NA      NA       NA
A:C:D         2.4375         NA      NA       NA
B:C:D        -0.4375         NA      NA       NA
A:B:C:D      -0.3125         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

The significant effects are now identified with the half-normal plot, cf. Fig. 14.1.ii. In this plot we can see that the two main effects A and B and the interaction effect $A:B$ are marked as outliers and are therefore significant.

```
#   detect significant effects in unreplicated fractional factorials
#   (without intercept)
> LGB(coefficients(modf)[-1])
Effect Report

Label     Half Effect    Sig(.05)
```

```
A          -6.3125       yes
B          17.8125       yes
C           0.1875       no
D           0.6875       no
A:B        -5.3125       yes
A:C         0.8125       no
B:C        -0.3125       no
A:D         2.0625       no
B:D        -0.0625       no
C:D        -0.6875       no
A:B:C      -0.1875       no
A:B:D      -0.6875       no
A:C:D       2.4375       no
B:C:D      -0.4375       no
A:B:C:D    -0.3125       no

Lawson, Grimshaw & Burt Rn Statistic =  3.006953
95th percentile of Rn =  1.201
```

Note that for significant interaction effects, the corresponding main effects are always added in the reduced model.

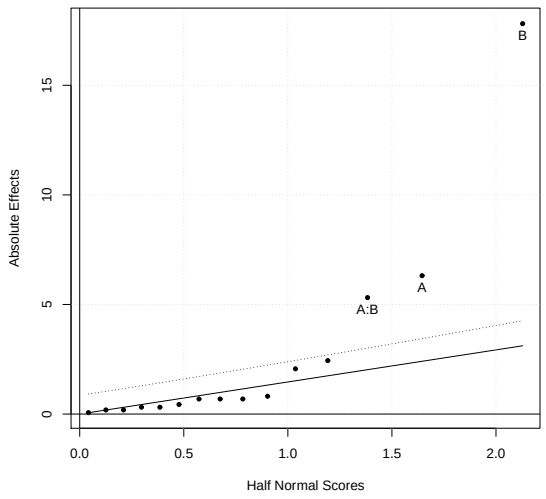


Figure 14.1.ii: Half normal plot of the estimated main and interaction effects. The non-significant effects vary around the diagonal solid line and are below the dotted line. The significant effects are labeled with the name of the factor.

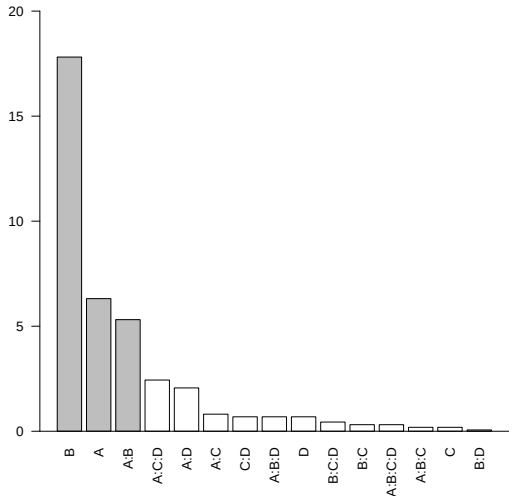


Figure 14.1.iii: Pareto Chart. The significant main effects A and B and the interaction effect $A:B$ (gray) clearly dominate the non-significant effects (white).

Statistical significance does not always have practical relevance. Use for instance a Pareto chart, cf. Chap. 1.2.2, to list the estimated effects according to their absolute values, cf. Fig. 14.1.iii. Like this it can be visually assessed whether the effects are relevant amounts. The effects are divided into two groups: significant and negligible effects.
From the technical context, it must now be judged whether the significant effects are are also

relevant effects. This can most often not be decided by the statistician and needs experts from the involved scientific fields.

14.2 2^k Fractional Factorial Designs

In Ex. 14.1.2 we came across a phenomena which often occurs in screening experiments. Few major effects and few interaction effects are significant. Therefore, higher order interactions are often combined with the experimental error. That is, we choose the smaller model to analyse the data. This empirical fact is used in screening experiments to keep the number of experiments within manageable limits. If, as we will see in Ex. 14.2.3, six factors and their interaction are examined for their influence on the response variable, then a full 2^k design would require a multiple of $2^6 = 64$ individual experiments. Often you cannot afford such an effort. The effort is mainly necessary to estimate the interactions. The more factors examined, the smaller the ratio between the number of major effects and the number of interactions. With six factors, it has one intercept, six major effects and 57 interaction effects, i.e. only 9.4% major effects. The remaining 90.6% of the effort is needed to estimate the interaction effects, many of which are not significant or relevant. Thus, in screening experiments, complete 2^k designs are rather inefficient for large k, which is why so-called 2^k **fractional factorial designs** are used.

Fractional factorial designs are created by replacing some of the higher-order interaction with additional experimental factors. For example, consider the case where four factors A, B, C and D are to be examined, but instead of the full 2^4 design, a design with only 8 runs is to be used. We take a 2^3 design. Because the interaction effect of the highest order is most likely to be insignificant, the corresponding column in the following table, i.e. the column $A:B:C$ is now assigned to a new main effect D.

Run	Factors			
	A	B	C	D
1	-1	-1	-1	-1
2	$+1$	-1	-1	$+1$
3	-1	$+1$	-1	$+1$
4	$+1$	$+1$	-1	-1
5	-1	-1	$+1$	$+1$
6	$+1$	-1	$+1$	-1
7	-1	$+1$	$+1$	-1
8	$+1$	$+1$	$+1$	$+1$

Because only half of the runs need to be measured compared to a full 2^4 design, this is called a 2^{k-1} fractional **factorial design**. The advantage of using fractional factorial designs is a significant reduction of the number of runs required. However, it is not surprising that a price has to be paid for it. A 2^{4-1} design with 8 runs cannot contain the same amount of information about four factors as a full 2^4 design with 16 runs. With a fractional factorial design we lose the ability to differentiate between the influence of the factors and their interactions with the response values. To illustrate this, let us take another look at the 2^{4-1} design above with the new factor $D = A:B:C$. It is obvious that the effect D and the interaction effect $A:B:C$ cannot be distinguished because the same column of plus and minus signs is used in the design matrix to estimate these effects. The effects D and $A:B:C$ are called **confounded**

or **aliased** in such a situation. The reason to alias D with $A:B:C$ is first of all that we speculate that the effect $A:B:C$ is negligible. If so, we get an estimate of the main effect D with only 8 runs.

Unfortunately, aliasing D with $A:B:C$ leads to even more alias pairs that are not immediately obvious from the start. For example, consider the interactions $A:B$ and $C:D$. The columns of $A:B$ and $C:D$ are identical in the design matrix. Therefore not only D with $A:B:C$ are aliased but also $A:B$ and $C:D$. There are even more aliased pairs. The entire set of aliases in a fractional factorial design is called **alias structure** of the design.

As is the case with all the other aspects of 2^k designs, there is a fairly easy method for writing down the alias structure of a fractional factorial design. This method depends on some simple observations about multiplying columns of plus and minus signs:

1. First, the letter I denotes the column consisting entirely of plus signs. This is called the **identity**.
2. Note that any column multiplied by itself yields column I. For example, $A \cdot A = A^2 = I$, $B \cdot B = B^2 = I$ and so on.
3. Multiplying column I by any other column does not change the column. For example, $A \cdot I = I \cdot A = A$.

Using these facts, we can obtain the alias structure of any fractional factorial as follows:

1. First, write the p assignments of additional factors in equation form. These p equations are called the **design generators**.
2. Multiply each generator from Step 1 by its left side to put each generator into the form $I = w$, where w is a word composed of several letters representing particular experimental factors (e.g. $D = ABC$ becomes $I = ABCD$). It is also possible to create words with minus signs, such as $D = -ABC$. If this is done, the resulting design will use a different fraction of the runs from the full 2^k design.
3. Denote $I = w_1$, $I = w_2, \ldots, I = w_p$ the p design generators from Step 2, form all possible products of the words w_i (one at a time, two at a time, three at a time, etc.). Use the fact that squares of factors can be eliminated (e.g. $A^2 = I$ and multiplying by I does not change anything). There will be a total of 2^p words formed. This collection is called the **defining relation** of the design.
4. Multiply each word in the defining relation by all $2^k - 1$ effects based on k factors. Use the fact that squares of factors cancel out to simplify the products. The result is called the **alias structure** of the design.

Example 14.2.1 (**Alias Structure of the 2^{4-1} Design**, cf. [9], Example 10.13)**.** Let us determine the alias structure of the 2^{4-1} design where D is aliased with ABC.

1. The only generator ($p = 1$) of this design is $D = ABC$.
2. Multiplying both sides by D gives $D \cdot D = D(ABC)$ or $I = ABCD$.
3. Since there is only one word in this equation, the defining relation is also of the form $I = ABCD$.

4. Multiplying each of the $2^4 - 1 = 15$ effects by the relation $I = ABCD$ yields to the following alias structure:

Effect	Alias	Effect	Alias
A	$= BCD$	BD	$= AC$
B	$= ACD$	CD	$= AB$
C	$= ABD$	ABC	$= D$
D	$= ABC$	ABD	$= C$
AB	$= CD$	ACD	$= B$
AC	$= BD$	BCD	$= A$
AD	$= BC$	$ABCD$	$= I$
BC	$= AD$		

There is a lot of repetition in this list. Eliminating duplicate equations, we can summarise the alias structure of the design as follows:

$A = BCD$	$AB = CD$
$B = ACD$	$AC = BD$
$C = ABD$	$AD = BC$
$D = ABC$	$ABCD = I$

The alias structure can be summarised as follows:

- Each main effect is aliased with a three-factor interaction.
- All two factor interactions are aliased with one another.
- The single four-factor interaction is aliased with the grand mean of the data.

Example 14.2.2 (**Alias Structure of the 2^{5-1} Design**, cf. [9], Examples 10.12 and 14)**.** The 2^{5-2} design provides a better illustration of how the defining relation is formed than Ex. 14.2.1.
Suppose that you want to study five factors using only 8 test runs. How do you create a fractional factorial design to accomplish this? First, start with the full 2^3 design, cf. Ex. 14.1.3. Write the column headings of the extended design matrix: A, B, C, AB, AC, BC and ABC. Finally, choose two of the interaction columns, say, ABC and AC, and assign the additional two factors, D and E, to these columns. Denote this column assignment by writing $D = ABC$ and $E = AC$. Because we are adding two factors D and E to a full design based on three factors A, B and C, this design is called a 2^{5-2} fractional factorial design. To create the design matrix for this particular 2^{5-2} experiment, first write the design matrix in Yates standard order for the 2^3 experiment with factors A, B and C. Then append columns D and E. The entries in D are found by multiplying the entries of columns A, B and C. Similarly, the entries in column E are found by multiplying the entries of columns A and C.
Recall that the design generators are $D = ABC$ and $E = AC$. Writing these in the form $I = ABCD$ and $I = ACE$, we can see that the defining relation is formed from the words $ABCD$ and ACE and all possible products of these words. Since there is only one such product

$$(ABCD)(ACE) = A^2BC^2DE = BDE,$$

the defining relation is

$$I = ACE = BDE = ABCD.$$

Multiplying each of the 2^5-1 effects by the defining relation gives the following alias structure:

$$\begin{array}{llllllll} I &= ACE &= BDE &= ABCD \\ A &= CE &= BCD &= ABDE \\ B &= DE &= ACD &= ABCE \\ C &= AE &= ABD &= BCDE \\ D &= BE &= ABC &= ACDE \\ E &= AC &= BD &= ABCDE \\ AB &= CD &= ADE &= BCE \\ BC &= AD &= ABE &= CDE \end{array}$$

Notice that each main effect is now aliased with at least one two factor interaction as well as higher-order interactions in this design.

Example 14.2.3 (**Antibody Yield**, cf. [12], Chapter 2.2)**.** Larger amounts of antibodies can be obtained in biotechnologically altered cells. These cells are injected in host animals (e.g. mice). After a while, these cells begin to produce and excrete antibodies. The excreted liquid is collected and further processed.
According to experience, the cells can only produce antibodies if the immune system of the host animals is weakened. There are four factors to this. It is also believed that the amount of cells injected and their level of development affects antibody production. Because there are no theoretical models for such complex biological processes, the relevant process factors are determined by an experiment. With an experiment, the most important factors should be determined with the least possible effort so that, among other things, not too many host animals are needed.
In order to weaken the immune system of the host animal, the animal is first irradiated with a non-iodine dose of ^{60}Co gamma rays (`RadDos`). Then an oil is injected into the animal. The amount (`VolPrs`) and time (`Prime1`) of injection could influence the yield of antibodies. Two weeks after the injection of the oil, the host animal is injected with the antibody-producing cells which have previously been cultured in vitro. The number of cells (`CelNum`) and their development status (`Growth`) could also influence the production. It is further suggested that a second injection (`Prime2`) of oil just before vaccination with antibody-producing cells may affect the yield (`TtrVol`). The yield is proportional to the number of molecules of antibody produced.
Each of these six factors will be examined in two levels. How they were set in the experiment is shown in the following table.

Factor Name	Low Level	High Level
`CelNum`	10^6 cells	10^7 cells
`VolPrs`	0.1 ml	0.5 ml
`Prime1`	1 week	3 weeks
`RadDos`	250 rads	500 rads
`Growth`	log state	saturated
`Prime2`	no	yes

It has been decided to use a 2^{6-2} fractional factorial design. The triple interaction $A:B:C$ was assigned to the main effect `Growth` and the triple interaction $B:C:D$ to the main effect `Prime2`. The corresponding design and the obtained yield values is shown in the following table.

Run	Experimental Factors						Yield
	CelNum A	VolPrs B	Prime1 C	RadDos D	Growth $A{:}B{:}C$	Prime2 $B{:}C{:}D$	TtrVol
1	−1	−1	−1	−1	−1	−1	70
2	+1	−1	−1	−1	+1	−1	150
3	−1	+1	−1	−1	+1	+1	34
4	+1	+1	−1	−1	−1	+1	32
5	−1	−1	+1	−1	+1	+1	137.5
6	+1	−1	+1	−1	−1	+1	56
7	−1	+1	+1	−1	−1	−1	123
8	+1	+1	+1	−1	+1	−1	225
9	−1	−1	−1	+1	−1	+1	50
10	+1	−1	−1	+1	+1	+1	2.7
11	−1	+1	−1	+1	+1	−1	1.2
12	+1	+1	−1	+1	−1	−1	12
13	−1	−1	+1	+1	+1	−1	90
14	+1	−1	+1	+1	−1	−1	2.1
15	−1	+1	+1	+1	−1	+1	4
16	+1	+1	+1	+1	+1	+1	15

The effort amounts to only a quarter of the complete design. However, to obtain sufficient antibodies for the analysis, five host animals must be used for each experimental condition. To avoid further systematic influences, the runs are processed in random order.

Now we want to analyse this design with R.

```
#   read data
> file <- "antibody1.dat"
> data <- read.table(file, header=TRUE)
> str(data)
'data.frame':   16 obs. of  7 variables:
 $ CelNum: int  -1 1 -1 1 -1 1 -1 1 -1 1 ...
 $ VolPrs: int  -1 -1 1 1 -1 -1 1 1 -1 -1 ...
 $ Prime1: int  -1 -1 -1 -1 1 1 1 1 -1 -1 ...
 $ RadDos: int  -1 -1 -1 -1 -1 -1 -1 -1 1 1 ...
 $ Growth: int  -1 1 1 -1 1 -1 -1 1 -1 1 ...
 $ Prime2: int  -1 -1 1 1 1 1 -1 -1 1 1 ...
 $ TtrVol: num  70 150 34 32 138 ...

#   ANOVA
> mod <- aov(TtrVol ~., data)
> anova(mod)
Analysis of Variance Table

Response: TtrVol
          Df  Sum Sq Mean Sq F value   Pr(>F)
CelNum     1    13.9    13.9  0.0065 0.937543
VolPrs     1   785.4   785.4  0.3675 0.559354
Prime1     1  5651.3  5651.3  2.6442 0.138371
RadDos     1 26446.9 26446.9 12.3745 0.006539 **
Growth     1  5863.7  5863.7  2.7436 0.132024
Prime2     1  7314.5  7314.5  3.4225 0.097353 .
```

```
Residuals  9 19234.9  2137.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.lm(mod)
Call:
aov(formula = TtrVol ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-47.131 -26.559  -1.781  10.853  79.256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.7812    11.5575   5.432 0.000415 ***
CelNum       -0.9313    11.5575  -0.081 0.937543
VolPrs       -7.0062    11.5575  -0.606 0.559354
Prime1       18.7938    11.5575   1.626 0.138371
RadDos      -40.6562    11.5575  -3.518 0.006539 **
Growth       19.1437    11.5575   1.656 0.132024
Prime2      -21.3812    11.5575  -1.850 0.097353 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.23 on 9 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.5091
F-statistic: 3.593 on 6 and 9 DF,  p-value: 0.04219
```

Now we perform a model selection by the AIC criterion with a stepwise algorithm in both directions.

```
> mod.both <- step(mod, direction="both")
Start:  AIC=127.47
TtrVol ~ CelNum + VolPrs + Prime1 + RadDos + Growth + Prime2
Step:  AIC=125.48
TtrVol ~ VolPrs + Prime1 + RadDos + Growth + Prime2
Step:  AIC=124.12
TtrVol ~ Prime1 + RadDos + Growth + Prime2

> summary.lm(mod.both)
Call:
aov(formula = TtrVol ~ Prime1 + RadDos + Growth + Prime2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-48.406 -24.188   0.069  10.794  87.194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    62.78      10.67   5.884 0.000105 ***
Prime1         18.79      10.67   1.762 0.105885
RadDos        -40.66      10.67  -3.811 0.002890 **
Growth         19.14      10.67   1.794 0.100263
```

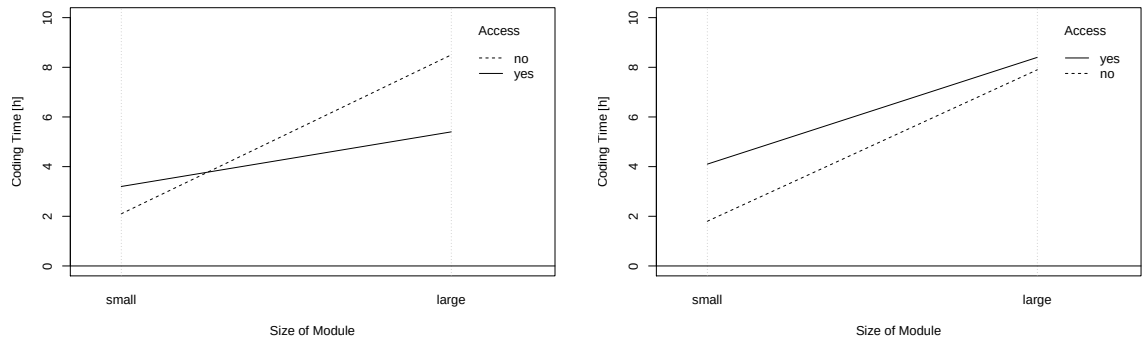
```
Prime2        -21.38      10.67  -2.004 0.070317 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.68 on 11 degrees of freedom
Multiple R-squared:  0.6932,    Adjusted R-squared:  0.5817
F-statistic: 6.215 on 4 and 11 DF,  p-value: 0.00724
```

The variables `CelNum` and `VolPrs` are not important. Of course, it is important to inject enough cells, but obviously both levels are sufficient.
There is a positive effect on the variable `Growth`, and therefore fully developed cells will be used in future. Likewise, the negative effect of a second oil injection `Prime2` should be avoided in future.
The two important factors `RadDos` and `Prime1` will be further investigated in Ex. 15.2.3.

A significant main effect for an ordered explanatory variable means that the response for the low and the high level of the explanatory variables is noticeably different.
However, for an ordered variable, it often happens that the response is high in a medium range, but drops off for low and high levels of the explanatory variable, or vice versa. Examing only two levels we run the risk of missing such a relationship.

14.3 Exercises

Problem 14.3.1 (**Experimental Design in Software Engineering**, cf. [27])**.** Factorial designs have also been used to study the productivity in software engineering. The goal of the study was to reduce the coding time of a software module.
Suppose that an experiment is performed to investigate the time it takes to code a software module. Factors that can affect coding time are the sizes of the module and whether or not the programmer has access to a library of previously encoded submodules. The module sizes are studied on two levels, small and large.
The analysis of a sample based on a two factor design showed that the interaction between module sizes (`Size`) and library access (`Access`) is significant.



Describe what you conclude from the interaction plot

a. on the left and

b. on the right.

Problem 14.3.2 (**Filling Variability**, cf. [23], Problem 14-51)**.** A factorial experiment is used to study the filling variability of dry soup mix packages. The factors are

- the number of mixing ports through which the vegetable oil was added (`A`) with levels one or two ports,
- the temperature surrounding the mixer (`B`) with levels cooled and ambient,
- the mixing time (`C`) with levels 60 or 80 seconds,
- the batch weight (`D`) with levels 1500 or 2000 lb and
- the number of days of delay between mixing and packaging (`E`) with levels one or seven days.

Between 125 and 150 packages of soup were sampled over an eight-hour period for each run in the design. The standard deviation of package weight was used as the response variable y. The data and the design is summarised in the following table:

Run	A	B	C	D	E	y
1	-1	-1	-1	-1	-1	1.13
2	$+1$	-1	-1	-1	$+1$	1.25
3	-1	$+1$	-1	-1	$+1$	0.97
4	$+1$	$+1$	-1	-1	-1	1.70
5	-1	-1	$+1$	-1	$+1$	1.47
6	$+1$	-1	$+1$	-1	-1	1.28
7	-1	$+1$	$+1$	-1	-1	1.18
8	$+1$	$+1$	$+1$	-1	$+1$	0.98
9	-1	-1	-1	$+1$	$+1$	0.78
10	$+1$	-1	-1	$+1$	-1	1.36
11	-1	$+1$	-1	$+1$	-1	1.85
12	$+1$	$+1$	-1	$+1$	$+1$	0.62
13	-1	-1	$+1$	$+1$	-1	1.09
14	$+1$	-1	$+1$	$+1$	$+1$	1.10
15	-1	$+1$	$+1$	$+1$	$+1$	0.76
16	$+1$	$+1$	$+1$	$+1$	-1	2.10

a. The design generator is $E = -ABCD$. Describe the design used in the experiment:

- Name of the design?
- Defining relations?
- Alias structure?

b. Calculate by hand the estimate for the main effect A as well as for the interactions $C:D$ and $A:B:E$.

c. Use a model without interactions. Which effects are statistically significant?

d. Perform an analysis of residuals. Report your findings.

e. Fit a model with all main effects and all second-order interaction effects. Find all significant effects and fit a reduced model. Note that for significant interaction effects, the corresponding main effects are always added in the reduced model.

f. Report your results of your analysis of the data from the experiment.

Solutions

Solution 14.3.1.

a. Since the interaction is significant, the library access allows a lower programming time than without. On the other hand, for small modules, it is a disadvantage for programmers to have access to a library.

b. Since the interaction is significant, we must assume that the disadvantage of accessing libraries in terms of programming time is much greater for small modules than for large modules.

Chapter 15

Response Surface Methodology

Response surface methodology (RSM) is a statistical technique which is used to model and analyse processes and data in applications in which a response of interest is influenced by several explanatory variables and the aim is to optimise this response.

To find the optimum settings of the explanatory variables, it may be useful to proceed in three steps:

1. A 2^k factorial design (possibly fractional) is used to detect the important influencing factors, cf. Ex. 14.2.3.

2. For the most important influencing factors, the main effects are estimated in order to determine the direction of the steepest ascent.

3. Now we can hope to find the global optimum in the determined direction with a suitable design and additional experiments.

This kind of procedure is a response surface exploration which is based on a sequence of experimental design.

15.1 Response Surface

In the search for relevant influencing factors, we assumed two (or more) levels per factor and used the analysis of variance to estimate and test main effects (and interactions). The model for two factors A and B with two levels each and without interaction is

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

with $i \in \{1, 2\}$ and $j \in \{1, 2\}$.

If both explanatory variables x_1 and x_2 are continuous then the model is

$$y = f(x_1, x_2) + \varepsilon, \tag{15.1.a}$$

where ε represents the noise or error observed in the response y. If we denote the expected response by $\mathrm{E}(y) = f(x_1, x_2) = \eta$, then the surface represented by

$$\eta = f(x_1, x_2)$$

is called a **response surface**.

For two explanatory variables we can represent the response surface graphically by plotting η versus the levels of x_1 and x_2. The response is represented as a surface plot in a three-dimensional space. We often plot the contours of the response surface. Lines of constant response are drawn in the x_1x_2-plane. Each contour corresponds to a particular constant height of the response surface.

The form of the relationship in Eqn. 15.1.a between the response and the independent variables is unknown in response surface methodology problems. Thus, the first step is to find a suitable approximation for the true relationship between the response and the explanatory variables. In general a low-order polynomial in some region of the explanatory variables is fitted to the data.

The analysis of the response surface is then based on the fitted low-order polynomial surface. If this fitted surface is a good approximation of the true response function, then the analysis of the fitted surface will be approximately equivalent to the analysis of the actual data.

Remark 15.1.1 (**Do not vary one variable after another**)**.** An obvious, but wrong, approach to finding the optimum is to vary one variable after another, cf. Fig. 15.1.i. Unfortunately, this wrong approach is very common among non-statisticians.

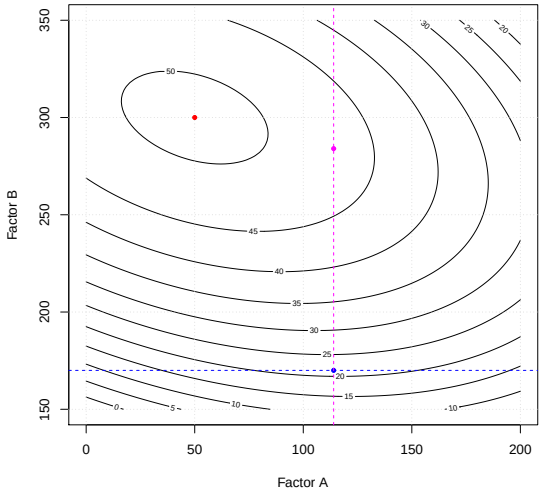


Figure 15.1.i: The maximum of the response surface at the red point cannot be found by varying one variable after another which gives the magenta point.

Although initially the setting of the process variables may be far away from optimum, an appropriate sequence of well-chosen experimental setups will result in the optimum. For this we start with a so-called **first-order design**, a 2^k design with additional measurements in the center of the test conditions. Experiences from previous experiments should help us to choose appropriate levels of the factors.

If the response is well modeled by a linear function of the k explanatory variables, the approximating function is the **first-order model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{15.1.b}$$

We want to illustrate this approach with the following example from chemistry.

Example 15.1.1 (**Chemical Reaction Analysis**, idea from [23], Example 14-6)**.** A chemical engineer is studying the percentage of conversion or yield of a process. There are two variables of interest, reaction time t and reaction temperature T.
From previous studies, it is known that a reaction temperature of 140° C and a reaction time of 60 minutes gives good results. In the experiment, the reaction time should now be varied by 10 minutes and the reaction temperature by 20° C on both sides.
The corresponding first-order test design and the results of the measurement are reported in the following table.

Run	Variables in original units		Variables in coded units		Yield
	T [° C]	t [min]	`Temp`	`Time`	y [g]
1	120	50	−1	−1	52.3
2	160	50	+1	−1	62.2
3	120	70	−1	+1	60.1
4	160	70	+1	+1	70.7
5	140	60	0	0	63.5
6	140	60	0	0	65.6

The design is given in coded form and in original units. The transformations from the original to the coded variables are

$$\texttt{Temp} = \frac{T - 140}{20} \quad \text{and} \quad \texttt{Time} = \frac{t - 60}{10} \tag{15.1.c}$$

and the inverse transformations from the coded to the original variables are

$$T = 20 \cdot \texttt{Temp} + 140 \quad \text{and} \quad t = 10 \cdot \texttt{Time} + 60. \tag{15.1.d}$$

The randomised order of the experiments was runs $5, 4, 2, 6, 1, 3$.

The first-order linear model in coded variables is

$$y = \beta_0 + \beta_1 \cdot \texttt{Temp} + \beta_2 \cdot \texttt{Time} + \varepsilon.$$

Now we fit this model to the data of the 2^2 design with two additional measurements (runs 5 and 6) in the centre of the test conditions.
Since the experiment consists of three parts it is practical to use an Excel-file to store the data in three different sheets. We read the `xlsx`-File with the R-package `openxlsx`.

```
#   load package
> require(openxlsx)
> file <- "reaction.xlsx"
> data1 <- read.xlsx(file, sheet="experiment 1")
> mod1 <- lm(y ~ Temp+ Time, data1)
> summary(mod1)
Call:
lm(formula = y ~ Temp + Time, data = data1)

Residuals:
```

```
    1     2     3     4     5     6
-0.90 -1.25 -1.25 -0.90  1.10  3.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4000     0.9485  65.786 7.74e-06 ***
Temp          5.1250     1.1617   4.412   0.0216 *
Time          4.0750     1.1617   3.508   0.0393 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.323 on 3 degrees of freedom
Multiple R-squared:  0.9137,    Adjusted R-squared:  0.8562
F-statistic: 15.88 on 2 and 3 DF,  p-value: 0.02535
```

The gradient of the response function

$$f(\texttt{Temp}, \texttt{Time}) = \beta_0 + \beta_1 \cdot \texttt{Temp} + \beta_2 \cdot \texttt{Time}$$

is

$$\operatorname{grad}(f(\texttt{Temp}, \texttt{Time})) = \begin{pmatrix} \dfrac{\partial}{\partial \texttt{Temp}} f(\texttt{Temp}, \texttt{Time}) \\ \dfrac{\partial}{\partial \texttt{Time}} f(\texttt{Temp}, \texttt{Time}) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Therefore the estimated gradient

$$\widehat{\operatorname{grad}}(f) = \begin{pmatrix} 5.125 \\ 4.075 \end{pmatrix}$$

points into the direction of the largest change in the coded Temp-Time-domain, cf. Fig. 15.1.ii. Notice that this direction leads us to a maximal increase of the response. This method is called **steepest ascent**.
If we are interested in the gradient in the original domain, then we have to transform it with the gradient of inverse affine function in Eqn. 15.1.c.

Observations that lie in the centre of a 2^k design have no effect on the estimates of the parameters and therefore do not affect the gradient.

```
> lm(y ~ Temp+ Time, data1[1:4,])
Call:
lm(formula = y ~ Temp + Time, data = data1[1:4, ])

Coefficients:
(Intercept)         Temp         Time
     61.325        5.125        4.075
```

The reader might ask why we extended the 2^2 design in Ex. 15.1.1. with two additional measurements in the centre of the test conditions. There are mainly two reasons for doing this:

- If several measurements for that point are made, then the measurement error can be determined without assuming that the first-order linear model approximates the true response sufficiently well.

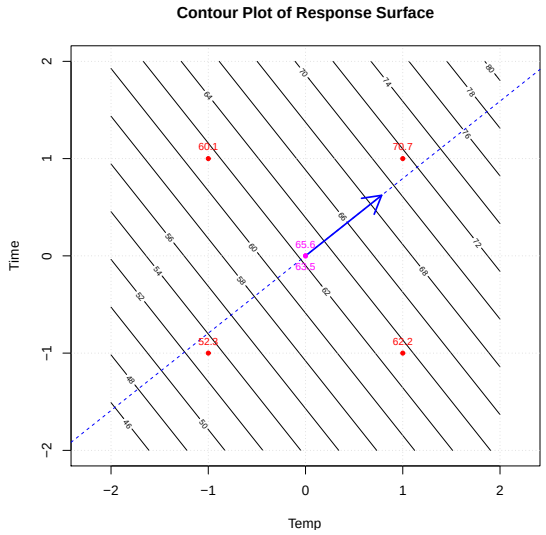


Figure 15.1.ii: Contour plot of the estimated response surface (plane) with measurements in the corners of the 2^2 factorial design (red) and in the centre (magenta) in coded variables. The estimated normalised gradient (blue) is orthogonal on the contours of the response surface.

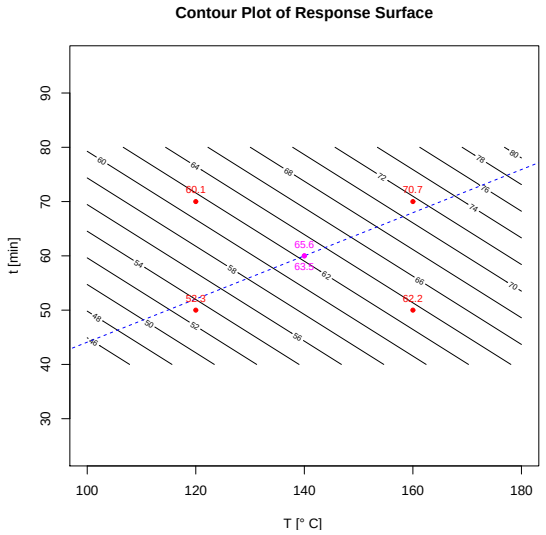


Figure 15.1.iii: Contour plot of the estimated response surface (plane) with measurements in the corners of the 2^2 factorial design (red) and in the centre (magenta) in original units. Search for an optimum along the blue dashed line.

- It is possible to detect deviations (curvature) from the first-order linear model. If no curvature is present and if the plane is a sufficient approximation of the response within the experimental domain, then the mean of the central observations $\bar{y}_c$ and the mean of all the other observations in the 2^k factorial design $\bar{y}_f$ are almost the same, cf. Fig. 15.1.ii and 15.1.iii. If, on the other hand, this difference is significantly different from zero then this is an indication that a model with curvature is more appropriate.

The corresponding hypothesis to test the presence of curvature in a 2^k design are

$$H_0 : \text{no curvature in the data,}$$
$$H_1 : \text{curvature in the data.}$$

The empirical variance s_c^2 which is calculated from the n_c repetitions in the centre of the design can be used to find the standard deviation of the difference between $\bar{y}_c$ and $\bar{y}_f$. The variance of $\bar{y}_c$ is $\frac{s_c^2}{n_c}$ and the variance of $\bar{y}_f$ is $\frac{s_c^2}{2^k}$. Since $\bar{y}_c$ and $\bar{y}$ are independent the variance of the difference is simply the sum of the variances of $\bar{y}_c$ and $\bar{y}_f$. Therefore, the test **statistic for the curvature** in a 2^k factorial design is

$$t_{\text{curv}} = \frac{\bar{y}_c - \bar{y}_f}{\sqrt{s_c^2 \left(\frac{1}{n_c} + \frac{1}{2^k}\right)}}.$$

Under the null hypothesis the test statistic t_{curv} follows a t-distribution with $n_c - 1$ degrees of freedom. Critical values of the t-distribution can be found in T.3.

If the experiments for the central test conditions are performed consecutively and not randomised, then there is a great risk that their variance, measured by s_c^2, will not capture the variance between independent experiments. In this case, a significant curvature is often erroneously assumed by the above test.

If the first-order linear model is fits the data well then we can search an optimum along the straight line through the centre of the design and along the estimated gradient.

Example 15.1.2 (**Chemical Reaction Analysis**)**.** The estimated first-order linear model in Ex. 15.1.1 is

$$\hat{f}(\texttt{Temp}, \texttt{Time}) = 61.325 + 5.125 \cdot \texttt{Temp} + 4.075 \cdot \texttt{Time},$$

where `Temp` and `Time` are coded variables.

- In coded variables `Temp` and `Time` the steepest ascent is on the line

$$r(\tau) = 0 + \tau \cdot \widehat{\operatorname{grad}}(f) = \begin{pmatrix} 5.125\,\tau \\ 4.075\,\tau \end{pmatrix},$$

  where τ is a parameter of the line, cf. Fig. 15.1.ii.

- In original variables T and t the above line has to be transformed with the gradient of the inverse transformations, cf. Eqn. 15.1.d. The steepest ascent is therefore on the line

$$r(\tau) = \begin{pmatrix} 140 + 102.50\,\tau \\ 60 + \ 40.75\,\tau \end{pmatrix},$$

  where τ is the parameter of the line, cf. Fig. 15.1.iii.

To find on optimum five more experiments were conducted with the experimental conditions on the estimated line in original variables T and t. The settings and the yield are reported in the following table.

Run	Variables in original units		Yield
	T [° C]	t [min]	y [g]
7	166	70	72.3
8	191	80	77.4
9	217	91	79.9
10	242	101	76.1
11	268	111	70.8

The temperatures and the times in original units of the new runs $7, 8, 9, 10, 11$ are rounded to the next integer to make the design easier to realise in the laboratory. The variables in the coded units are not useful here. Studying the profile of the five new measurements we find an optimum in the vicinity of 217° C and 91 min.

If the above test detects curvature in the data then the first-order linear model for the response is not suited. This could be an indication that the optimum is nearby or that we have to choose a response surface with interactions.

15.2 Second-Order Response Surface with Interactions

If we are close to the optimum then the estimated response plane is almost horizontal, i.e. the estimated coefficients are almost zero. In such a situation the test for the curvature will reject the null hypothesis.
If there is curvature in the system, then a polynomial of higher degree must be used, such as the **second-order model**

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon. \tag{15.2.a}$$

In the case of two explanatory variables the model simplifies to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon. \tag{15.2.b}$$

The 2^k design is not enough to estimate the parameters in the second-order model, since at least three levels are necessary per factor for the estimation. An appropriate design can be obtained in many different ways. The more levels available per factor, the better the curvature is estimated. To keep the effort within reasonable limits, so-called **rotatable central composite design** or **second-order central composite design** are very popular. This design can be obtained from a 2^k design by adding more experimental conditions.

For the case of two explanatory variables all experimental conditions in coded variables (except the centre) are equidistant from the center $(0, 0)$. Each factor is measured at five levels. Repeating the experiment several times at the centre gives a more accurate estimate for the quadratic part of the model.

Example 15.2.1 (**Chemical Reaction Analysis**)**.** The chemical reaction in Ex. 15.1.1 will be investigated by a second-order central composite design in the vicinity of the optimum 217° C and 91 min found in Ex. 15.1.2.
The corresponding second-order test plan and the results of the measurements are reported in the following table.

Run	Variables in original units		Variables in coded units		Yield
	T [° C]	t [min]	`Temp`	`Time`	y [g]
12	195	80	-1	-1	78.7
13	235	80	$+1$	-1	76.2
14	195	100	-1	$+1$	72.6
15	235	100	$+1$	$+1$	75.5
16	187	90	$-\sqrt{2}$	0	74.5
17	243	90	$\sqrt{2}$	0	76.2
18	215	76	0	$-\sqrt{2}$	77.8
19	215	104	0	$\sqrt{2}$	72.2
20	215	90	0	0	80.7

The temperatures and the times in original units of the runs 16, 17, 18 and 20 are rounded to the next integer to make the design easier to realise in the laboratory. The design is given in

coded form, cf. Fig. 15.2.i, and in original units, cf. Fig. 15.2.ii. The transformations which map the original to the coded variables are now

$$\texttt{Temp} = \frac{T - 215}{20} \quad \text{and} \quad \texttt{Time} = \frac{t - 90}{10}. \tag{15.2.c}$$

The second-order linear model in original variables is

$$y = \beta_0 + \beta_1\, T + \beta_2\, t + \beta_{11}\, T^2 + \beta_{22}\, t^2 + \beta_{12}\, T \cdot t + \varepsilon.$$

Now we fit this model with the interaction term to the data in original units of the rotatable central composite design of experiment 3.

```
> data3 <- read.xlsx(file, sheet="experiment 3")
> mod3 <- lm(y ~ T + t + I(T^2) + I(t^2) + T * t, data3)
> summary(mod3)
Call:
lm(formula = y ~ T + t + I(T^2) + I(t^2) + T * t, data = data3)

Residuals:
       1        2        3        4        5        6        7        8       9
 0.43918 -0.06284  0.73615  0.23413 -0.70211  0.01506 -0.13141 -0.55565 0.02748

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.896e+02  8.465e+01  -3.422  0.04179 *
T            2.141e+00  4.960e-01   4.316  0.02291 *
t            3.258e+00  8.770e-01   3.715  0.03393 *
I(T^2)      -6.351e-03  1.087e-03  -5.840  0.01000 *
I(t^2)      -2.719e-02  4.350e-03  -6.251  0.00826 **
T:t          6.750e-03  1.833e-03   3.682  0.03472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7334 on 3 degrees of freedom
Multiple R-squared:  0.9736,    Adjusted R-squared:  0.9297
F-statistic: 22.15 on 5 and 3 DF,  p-value: 0.0142
```

The estimated second-order response function in original variables is

$$\hat{f}(T, t) = -289.6 + 2.141\, T + 3.258\, t - 0.006351\, T^2 - 0.02719\, t^2 + 0.00675\, T \cdot t.$$

The estimated response surface is visualised as a contour plot in Fig. 15.2.ii. We could easily find the optimum graphically in Fig. 15.2.ii, but we want to get it analytically.

A second-order response surface can describe different types of surfaces depending on its parameters. The three most common non-degenerated surfaces are a minimum, a maximum or a saddle. The surfaces with maximum (or minimum) hardly require any explanation: leaving the optimum in any direction decreases (or increases) the yield y.

In the saddle, the yield y may increase or decrease depending on the direction in which the experimental conditions are moving. A fitted saddle is useless to find an optimum.

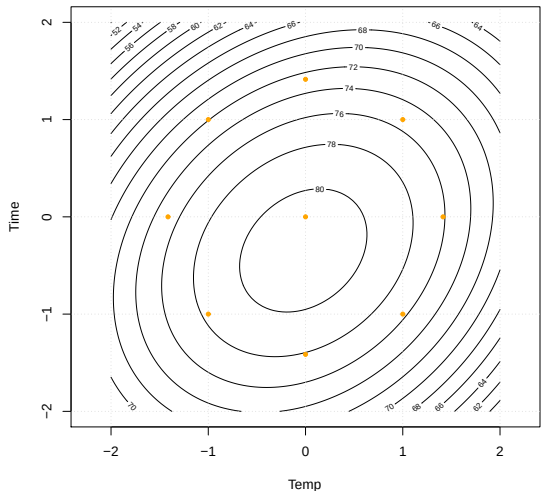


Figure 15.2.i: Contour plot of the estimated second-order response surface with measurements (orange) in coded variables.

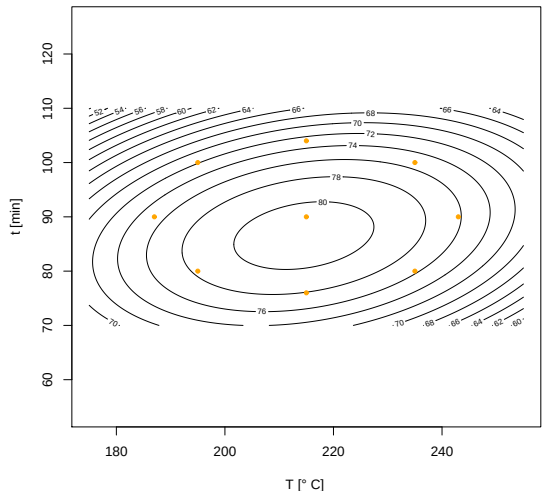


Figure 15.2.ii: Contour plot of the estimated second-order response surface with measurements (orange) in original units.

The optimum of the response function, cf. Eqn. 15.2.a,

$$f(x_1, \ldots, x_k) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j$$

can be calculated analytically. The stationary point can be found by setting all k partial derivatives equal to zero

$$\frac{\partial}{\partial x_1} f(x_1, \ldots, x_k) = 0, \quad \ldots, \quad \frac{\partial}{\partial x_k} f(x_1, \ldots, x_k) = 0.$$

This is a system of k linear equations in k unknowns $x_1, \ldots, x_k$ which is easy to solve. There are a few pitfalls:

- The solution can be a saddle point. This gets more likely the more factors k are involved. Check if the $k \times k$ Hessian matrix containing all second partial derivatives is positive definite.
- The maximum (or minimum) is unique: The probability to obtain a degenerate response surface (i.e. indefinit Hessian matrix) is zero.

In the case of two explanatory variables we obtain

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x_1, x_2) &= \beta_1 + 2\beta_{11}\, x_1 + \ \beta_{12}\, x_2 = 0 \\ \frac{\partial}{\partial x_2} f(x_1, x_2) &= \beta_2 + \ \beta_{12}\, x_1 + 2\beta_{22}\, x_2 = 0 \end{aligned}$$

and the estimated stationary point is

$$\hat{x}_1 = \frac{\hat{\beta}_{12}\hat{\beta}_2 - 2\hat{\beta}_{22}\hat{\beta}_1}{4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2} \quad \text{and} \quad \hat{x}_2 = \frac{\hat{\beta}_{12}\hat{\beta}_1 - 2\hat{\beta}_{11}\hat{\beta}_2}{4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2}.$$

If the estimated determinant of the Hessian matrix is

$$4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2 > 0$$

then we have a maximum (or a minimum) otherwise a useless saddle point.

Example 15.2.2 (**Chemical Reaction Analysis**)**.** We come back to Ex. 15.2.1 and want to estimate the maximum of the fitted second-order response function

$$\hat{f}(T,t) = -289.6 + 2.141\,T + 3.258\,t - 0.006351\,T^2 - 0.02719\,t^2 + 0.00675\,T\,t.$$

We find the estimated stationary point at

$$\hat{T} = 214.5^\circ\,\mathrm{C} \quad \text{and} \quad \hat{t} = 86.55\,\mathrm{min}$$

with the positive estimated determinant of the Hessian matrix 0.000645. At this stationary point we can expect a yield of

$$\hat{y}_{\max} = \hat{f}(\hat{T},\hat{t}) = 81.0\,\mathrm{g}.$$

The contour plot in original units of the estimated second-order response surface with the measurements of all three experiments can be seen in Fig. 15.2.iii.

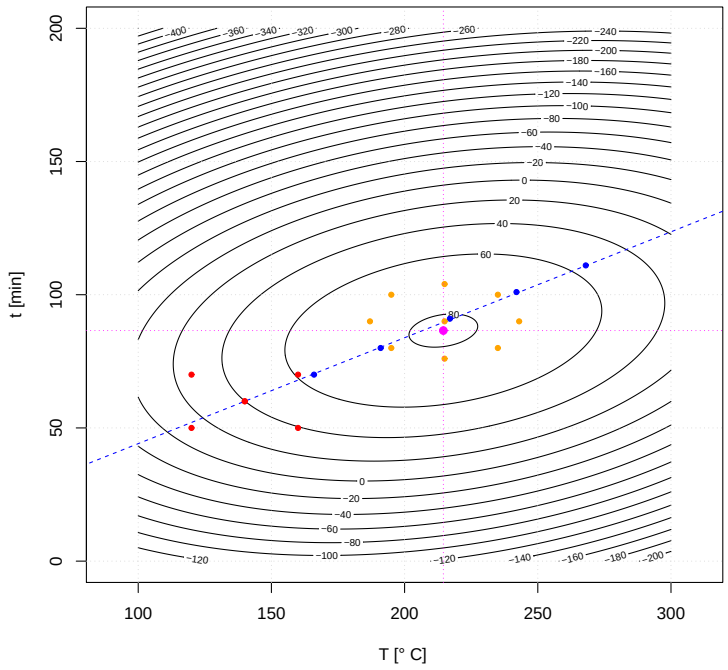


Figure 15.2.iii: Contour plot in original units of the estimated second-order response surface with measurements of the three experiments: Experiment 1 is a 2^2 factorial design with additional doubled measurement in the centre (red), experiment 2 is an optimisation along the gradient (blue) and experiment 3 is a rotatable central composite design (orange). Optimal settings (magenta).

Example 15.2.3 (**Antibody Yield**, cf. [12], Chapter 2.2)**.** We saw in Ex. 14.2.3 that only two factors `RadDos` and `Prime1` have a significant effect on the yield `TtrVol`. Therefore we want to find the settings of the variables `RadDos` and `Prime1` with maximal yield. In the first experiment the factors were set on the levels in the following table:

Factor Name	Low Level	High Level
`Prime1`	1 week	3 weeks
`RadDos`	250 rads	500 rads

The estimated effect of `Prime1` was positive and of `RadDos` negative, therefore we will setup a new experiment with `Prime1` at 7 and 21 days and `RadDos` at 100 rads and 300 rads. The reason for keeping the high level of `Prime1` at 21 days (as in the first experiment in Ex. 14.2.3) is to reduce cost.
We define a 2^2 factorial design (runs 17, 18, 19 and 20) with three repeated measurements in the centre (runs 21, 22 and 23).

Run	Variables in original units		Variables in coded units		Yield
	`RadDos.orig` [rads]	`Prime1.orig` [days]	`RadDos`	`Prime1`	`TtrVol` [g]
17	100	7	−1	−1	207
18	100	21	−1	+1	257
19	300	7	+1	−1	306
20	300	21	+1	+1	570
21	200	14	0	0	630
22	200	14	0	0	528
23	200	14	0	0	609

Now we analyse this design and the observed yields with R.

```
#   read data of experiment 2 and 3
> file <- "antibody2.dat"
> data <- read.table(file, header=TRUE)
#   only 2nd experiment
> data2 <- data[data$Run>=17 & data$Run<=23,]
```

We want to test for curvature in the centre to see if we are already close to the optimum.

```
#   means, variance and number of replicates
> TtrVol.f.bar <- mean(data2$TtrVol[data2$Run<=20])
> TtrVol.c.bar <- mean(data2$TtrVol[data2$Run>20])
> sc2 <- var(data2$TtrVol[data2$Run>20])
> sqrt(sc2)
53.86093
> nc <- 3
#   test statistic
> t.curv <- (TtrVol.c.bar-TtrVol.f.bar)/sqrt(sc2*(1/nc+1/2^2))
#   P-value
> 2*(1-pt(t.curv, df=nc-1))
0.02524105
```

From the three measurements in the centre (runs 21, 22 and 23) we can estimate the standard deviation $\hat{\sigma} = 53.861$ g. Since the P-value is $0.0252 < 0.05$ we conclude that there is curvature present. Comparing the yield of the second experiment in the centre with the corners we are not surprised.

Thus we conclude that the optimum is somewhere within the experimental settings of the second experiment. Therefore we supplement it with a third experiment, i.e. a rotatable central composite design (runs 24, 25, 26 and 27).

Run	Variables in original units		Variables in coded units		Yield
	`RadDos.orig` [rads]	`Prime1.orig` [days]	`RadDos`	`Prime1`	`TtrVol` [g]
24	200	4	0	$-\sqrt{2}$	315
25	200	24	0	$\sqrt{2}$	154
26	59	14	$-\sqrt{2}$	0	100
27	341	14	$\sqrt{2}$	0	513

We now estimate the coefficients of the second-order response surface in original variables `RadDos.orig` and `Prime1.orig` of the second and third experiments.

```
> mod3 <- lm(TtrVol ~ RadDos.orig + Prime1.orig
+                     + I(RadDos.orig^2) + I(Prime1.orig^2)
+                     + RadDos.orig*Prime1.orig, data=data)
> summary(mod3)
Call:
lm(formula = TtrVol ~ RadDos.orig + Prime1.orig + I(RadDos.orig^2) +
    I(Prime1.orig^2) + RadDos.orig * Prime1.orig, data = data)

Residuals:
       1         2         3         4         5         6         7         8
 -15.3301  120.8820  -58.6543   77.5578   40.7093  -61.2907   19.7093   64.8569
       9        10        11
-125.8401  -62.0265   -0.5737

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -6.084e+02  3.279e+02  -1.856   0.1226
RadDos.orig             5.237e+00  2.076e+00   2.522   0.0531 .
Prime1.orig             7.700e+01  2.920e+01   2.637   0.0462 *
I(RadDos.orig^2)       -1.265e-02  4.399e-03  -2.876   0.0348 *
I(Prime1.orig^2)       -3.243e+00  8.786e-01  -3.691   0.0141 *
RadDos.orig:Prime1.orig 7.643e-02  7.423e-02   1.030   0.3504
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 103.9 on 5 degrees of freedom
Multiple R-squared:  0.856,     Adjusted R-squared:  0.7119
F-statistic: 5.943 on 5 and 5 DF,  p-value: 0.03636
```

The estimated response function is

$$\begin{aligned}\hat{f}(\texttt{RadDos.orig}, \texttt{Prime1.orig}) &= -608.400 \\ &\quad +5.237 \cdot \texttt{RadDos.orig} + 77.000 \cdot \texttt{Prime1.orig} \\ &\quad -0.013 \cdot \texttt{RadDos.orig}^2 - 3.243 \cdot \texttt{Prime1.orig}^2 \\ &\quad +0.076 \cdot \texttt{RadDos.orig} \cdot \texttt{Prime1.orig}.\end{aligned}$$

To find the stationary point we calculate both partial derivatives and set them zero

$$\begin{aligned}&\frac{\partial}{\partial \texttt{RadDos.orig}} \hat{f}(\texttt{RadDos.orig}, \texttt{Prime1.orig}) \\ &\qquad = 5.237 - 2 \cdot 0.013 \cdot \texttt{RadDos.orig} + 0.076 \cdot \texttt{Prime1.orig} \\ &\qquad = 0, \\ &\frac{\partial}{\partial \texttt{Prime1.orig}} \hat{f}(\texttt{RadDos.orig}, \texttt{Prime1.orig}) \\ &\qquad = 77.000 - 2 \cdot 3.243 \cdot \texttt{Prime1.orig} + 0.076 \cdot \texttt{RadDos.orig} \\ &\qquad = 0.\end{aligned}$$

The estimated stationary point is

$$\widehat{\texttt{RadDos}} = 251.81\,\text{rads} \quad \text{and} \quad \widehat{\texttt{Prime1}} = 14.84\,\text{days}$$

with the positive estimated determinant of the Hessian matrix 0.1582514. At this stationary point we can expect a yield of

$$\texttt{TtrVol}_{\max} = \hat{f}(\widehat{\texttt{RadDos}}, \widehat{\texttt{Prime1}}) = 622.21\,\text{g}.$$

The calculation in R is as follows.

```
#   stationary point
> v <- coefficients(mod3)[c("RadDos.orig","Prime1.orig")]
> H <- matrix(c(2,1,1,2)*
+             coefficients(mod3)[c("I(RadDos.orig^2)","RadDos.orig:Prime1.orig",
+                                  "RadDos.orig:Prime1.orig","I(Prime1.orig^2)")],
+             nrow=2,ncol=2)
> det(H)
0.1582514
> RadDosPrime1.opt <- solve(H,-v)
> RadDosPrime1.opt
251.81027  14.83945
#   maximal yield
> predict(mod3, newdata=list(RadDos.orig=RadDosPrime1.opt[1],
+                            Prime1.orig=RadDosPrime1.opt[2]))
622.2078
```

The contour plot in original units of the estimated second-order response surface with the measurements of the second and third experiments can be seen in Fig. 15.2.iv.

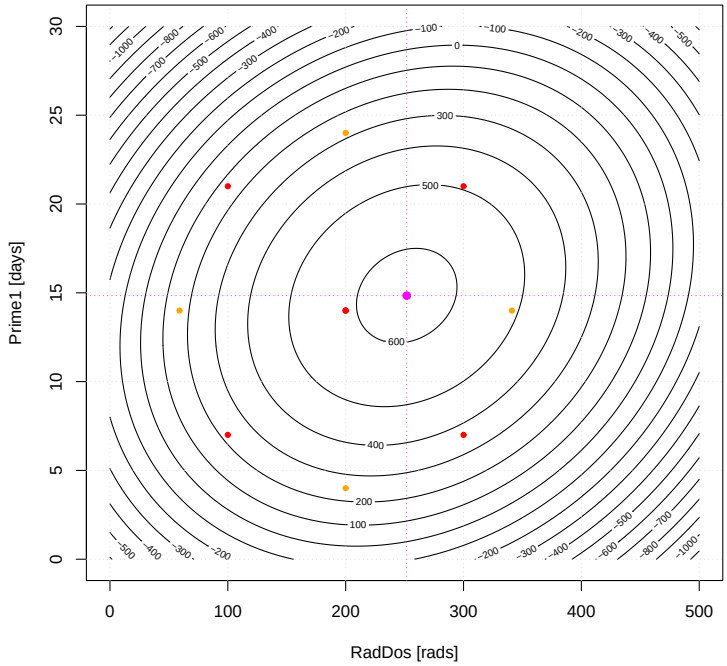


Figure 15.2.iv: Contour plot in original units of the estimated second-order response surface with measurements of two experiments: Experiment 2 is a 2^2 fractional factorial design with three replicated measurements in the centre (red), experiment 3 is a rotatable central composite design (orange). Optimal settings (magenta).

Summary. In order to find the settings of the explanatory variables that lead to an optimal yield of the response variable, iterative experiments are performed according to special test designs.

1. If the actual settings are far away from the optimum then use a first-order response surface on a 2^k fractional factorial design to obtain an overview of the situation.
2. Use the method of steepest ascent to obtain further measurements along the gradient till the response gets smaller (larger) again.
3. Repeat Step 1 and 2, if necessary, in a vicinity of the optimum found in Step 2.
4. Use a rotatable central composite designs in the vicinity of the optimum found in Step 3. Estimate the second-order response surface. Find analytically an estimate of the stationary point.

Literature. For further information and other designs consult the book of R. H. Myers and D. C. Montgomery about *Response Surface Methodology*, [25].

15.3 R-Package `rsm`

In the following I refer to the papers [17] and [18], which describe the package perfectly. Try to solve Prob. 15.4.2 with the package `rsm`.

15.4 Exercises

Problem 15.4.1 (**Laboratory Experiment**, cf. [3], Chapter 12.4[1]**).** With laboratory experiments it should be found out, how long and at which temperature a process should run, so that the yield becomes as high as possible. From experience, it was known that the best setting is at about 75 minutes and 130° C.
First, a 2^2 factorial design was applied with three additional measurements at the centre of the experimental conditions. For the time, the two stages 70 and 80 minutes were selected and for the temperature the levels 127.5 and 132.5° C. The data in sheet `experiment 1` of the Excel-file `laboratry.xlsx` contains the variables `Time`, `Temp`, `Yield` and `A`, `B`. The variable `A` corresponds to the time and `B` to the temperature in coded units.

a. How were the variables `A` and `B` coded? Report the formula.

b. Fit a first-order response surface to the data in coded units and present it graphically. Determine the direction of the steepest ascent. Add the gradient and all seven laboratory settings to the graphic.

c. Is a first-order response surface enough? Argue with a statistical test on the 5% level.

Second, with the method of steepest ascent three new measurements were made, cf. the data in sheet `experiment 2` of the file `laboratry.xlsx`. The order of the measurements was run 7, 9 and 8. In the vicinity of the settings of run 8 (with the largest yield) a third experiment with a 2^2 factorial design with three additional measurements in the centre was conducted, cf. the data in sheet `experiment 3`.

d. Fit a first-order response surface the data of the third experiment (runs 11 to 17).

e. Test for significant curvature on the 5% level.

f. After showing that the true response surface is curved in the vicinity of run 8, the measurements were supplemented to a rotatable central composite design (runs 18 to 23). Fit a second-order response surface to the merged data of the third and fourth experiment in sheet `experiment 4`.

g. Present the fitted response surface graphically.

h. Determine analytically the experimental conditions with maximum yield.

Problem 15.4.2 (**Laboratory Experiment with `rsm`**)**.** Install the package `rsm` in R and read [17] and [18]. Load the package with `require(rsm)`. Solve Prob. 15.4.1.**d**.-**h**. with this package. Read only original variables of the third and fourth experiment

```
file <- "laboratry.xlsx"
#   read data of experiment 3 and 4 (without coded variables A and B)
data3 <- read.xlsx(file, sheet="experiment 3", cols=1:4)
data4 <- read.xlsx(file, sheet="experiment 4", cols=1:4)
#   merge data
data <- rbind(data3, data4)
```

and create coded variables with `coded.data()`.

[1] Data slightly modified.

Appendix

Tables

T.1 Distribution function $\Phi(z, 0, 1)$ of the standard normal distribution 230
T.2 Quantiles z_p for standard normal distribution 231
T.3 Quantiles $t_{n,p}$ for Student's t-distribution with n degrees of freedom 232
T.4 Quantiles $\chi^2_{n,p}$ (lower) for χ^2-distribution with n degrees of freedom 233
T.5 Quantiles $\chi^2_{n,p}$ (upper) for χ^2-distribution with n degrees of freedom 234
T.6 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 235
T.7 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 236
T.8 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 237
T.9 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 238
T.10 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 239
T.11 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 240
T.12 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 241
T.13 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 242
T.14 Quantiles $F_{m,n,p}$ for F-distribution with (m, n) degrees of freedom 243
T.15 Factors for constructing control charts 244

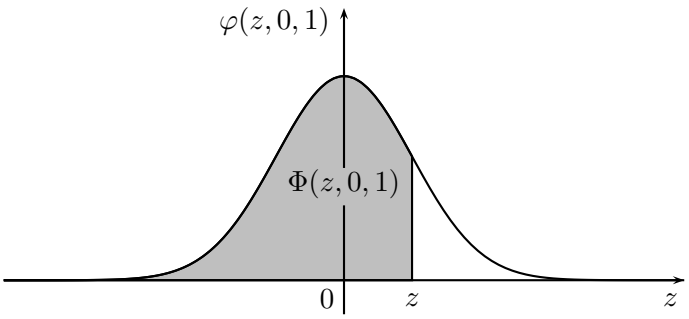


Table T.1: Distribution function $\Phi(z,0,1)$ of the **standard normal distribution** $\mathcal{N}(0,1)$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

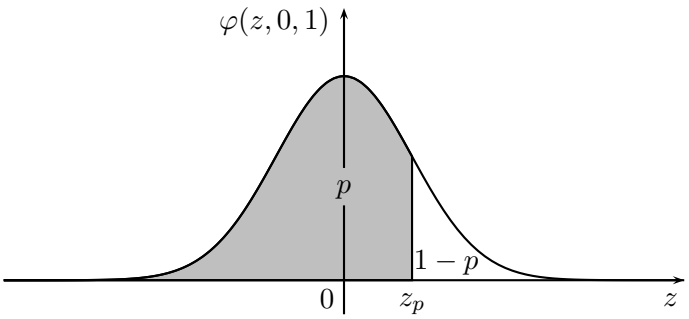


Table T.2: p-quantiles z_p for **standard normal distribution** $\mathcal{N}(0,1)$ with $z_{1-p} = -z_p$.

p	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.000	0.003	0.005	0.008	0.010	0.013	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.105	0.108	0.111	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.181	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.259	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.285	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.305	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.335	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.375	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.457	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.485	0.487	0.490	0.493
0.69	0.496	0.499	0.502	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.693	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.765	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.871	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.063	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.227	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.317	1.323	1.329	1.335
0.91	1.341	1.347	1.353	1.359	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.433	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.499	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

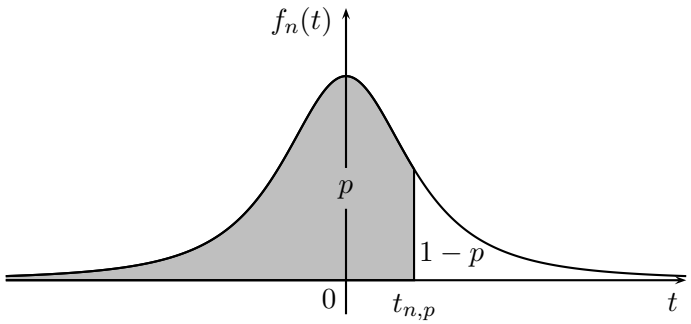


Table T.3: p-quantile $t_{n,p}$ for **Student's t-distribution** with n degrees of freedom. The density function is symmetric, therefore $t_{n,1-p} = -t_{n,p}$.

	p							
n	0.9000	0.9500	0.9750	0.9900	0.9950	0.9990	0.9995	n
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578	1
2	1.886	2.920	4.303	6.965	9.925	22.328	31.600	2
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924	3
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869	5
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	13
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140	14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073	15
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015	16
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965	17
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922	18
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883	19
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850	20
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819	21
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792	22
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768	23
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745	24
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725	25
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707	26
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689	27
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674	28
29	1.311	1.699	2.045	2.462	2.756	3.396	3.660	29
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646	30
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551	40
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496	50
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460	60
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435	70
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416	80
90	1.291	1.662	1.987	2.368	2.632	3.183	3.402	90
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390	100
150	1.287	1.655	1.976	2.351	2.609	3.145	3.357	150
200	1.286	1.653	1.972	2.345	2.601	3.131	3.340	200
300	1.284	1.650	1.968	2.339	2.592	3.118	3.323	300
400	1.284	1.649	1.966	2.336	2.588	3.111	3.315	400
500	1.283	1.648	1.965	2.334	2.586	3.107	3.310	500
600	1.283	1.647	1.964	2.333	2.584	3.104	3.307	600
800	1.283	1.647	1.963	2.331	2.582	3.100	3.303	800
1000	1.282	1.646	1.962	2.330	2.581	3.098	3.300	1000
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291	∞

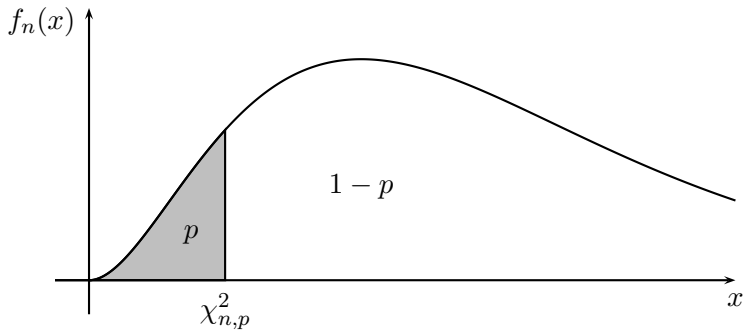


Table T.4: p-quantiles $\chi^2_{n,p}$ for χ^2-**distribution** with n degrees of freedom.

	p							
n	0.005	0.010	0.025	0.050	0.100	0.250	0.500	n
1	0.000	0.000	0.001	0.004	0.016	0.102	0.455	1
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	3
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	4
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	5
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	6
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	7
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	8
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	9
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	10
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	11
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	12
13	3.565	4.107	5.009	5.892	7.041	9.299	12.340	13
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	14
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	15
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	16
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	17
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	18
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	19
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	20
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	21
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	22
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	23
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	24
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	25
26	11.160	12.198	13.844	15.379	17.292	20.843	25.336	26
27	11.808	12.878	14.573	16.151	18.114	21.749	26.336	27
28	12.461	13.565	15.308	16.928	18.939	22.657	27.336	28
29	13.121	14.256	16.047	17.708	19.768	23.567	28.336	29
30	13.787	14.953	16.791	18.493	20.599	24.478	29.336	30
31	14.458	15.655	17.539	19.281	21.434	25.390	30.336	31
32	15.134	16.362	18.291	20.072	22.271	26.304	31.336	32
33	15.815	17.073	19.047	20.867	23.110	27.219	32.336	33
34	16.501	17.789	19.806	21.664	23.952	28.136	33.336	34
35	17.192	18.509	20.569	22.465	24.797	29.054	34.336	35
36	17.887	19.233	21.336	23.269	25.643	29.973	35.336	36
37	18.586	19.960	22.106	24.075	26.492	30.893	36.336	37
38	19.289	20.691	22.878	24.884	27.343	31.815	37.335	38
39	19.996	21.426	23.654	25.695	28.196	32.737	38.335	39
40	20.707	22.164	24.433	26.509	29.051	33.660	39.335	40
50	27.991	29.707	32.357	34.764	37.689	42.942	49.335	50
60	35.534	37.485	40.482	43.188	46.459	52.294	59.335	60
70	43.275	45.442	48.758	51.739	55.329	61.698	69.334	70
80	51.172	53.540	57.153	60.391	64.278	71.145	79.334	80
90	59.196	61.754	65.647	69.126	73.291	80.625	89.334	90
100	67.328	70.065	74.222	77.929	82.358	90.133	99.334	100

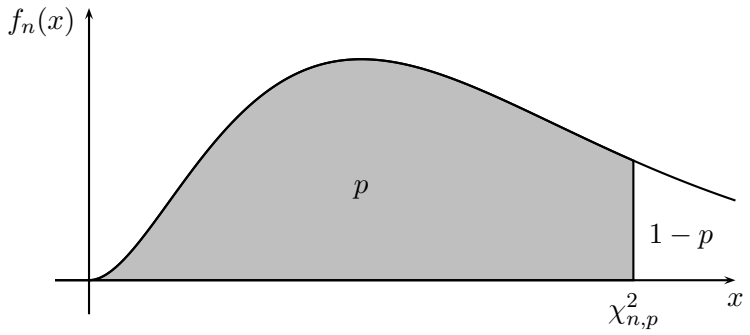


Table T.5: p-quantiles $\chi^2_{n,p}$ for χ^2-**distribution** with n degrees of freedom.

	p							
n	0.750	0.900	0.950	0.975	0.990	0.995	0.999	n
1	1.323	2.706	3.841	5.024	6.635	7.879	10.827	1
2	2.773	4.605	5.991	7.378	9.210	10.597	13.815	2
3	4.108	6.251	7.815	9.348	11.345	12.838	16.266	3
4	5.385	7.779	9.488	11.143	13.277	14.860	18.466	4
5	6.626	9.236	11.070	12.832	15.086	16.750	20.515	5
6	7.841	10.645	12.592	14.449	16.812	18.548	22.457	6
7	9.037	12.017	14.067	16.013	18.475	20.278	24.321	7
8	10.219	13.362	15.507	17.535	20.090	21.955	26.124	8
9	11.389	14.684	16.919	19.023	21.666	23.589	27.877	9
10	12.549	15.987	18.307	20.483	23.209	25.188	29.588	10
11	13.701	17.275	19.675	21.920	24.725	26.757	31.264	11
12	14.845	18.549	21.026	23.337	26.217	28.300	32.909	12
13	15.984	19.812	22.362	24.736	27.688	29.819	34.527	13
14	17.117	21.064	23.685	26.119	29.141	31.319	36.124	14
15	18.245	22.307	24.996	27.488	30.578	32.801	37.698	15
16	19.369	23.542	26.296	28.845	32.000	34.267	39.252	16
17	20.489	24.769	27.587	30.191	33.409	35.718	40.791	17
18	21.605	25.989	28.869	31.526	34.805	37.156	42.312	18
19	22.718	27.204	30.144	32.852	36.191	38.582	43.819	19
20	23.828	28.412	31.410	34.170	37.566	39.997	45.314	20
21	24.935	29.615	32.671	35.479	38.932	41.401	46.796	21
22	26.039	30.813	33.924	36.781	40.289	42.796	48.268	22
23	27.141	32.007	35.172	38.076	41.638	44.181	49.728	23
24	28.241	33.196	36.415	39.364	42.980	45.558	51.179	24
25	29.339	34.382	37.652	40.646	44.314	46.928	52.619	25
26	30.435	35.563	38.885	41.923	45.642	48.290	54.051	26
27	31.528	36.741	40.113	43.195	46.963	49.645	55.475	27
28	32.620	37.916	41.337	44.461	48.278	50.994	56.892	28
29	33.711	39.087	42.557	45.722	49.588	52.335	58.301	29
30	34.800	40.256	43.773	46.979	50.892	53.672	59.702	30
31	35.887	41.422	44.985	48.232	52.191	55.002	61.098	31
32	36.973	42.585	46.194	49.480	53.486	56.328	62.487	32
33	38.058	43.745	47.400	50.725	54.775	57.648	63.869	33
34	39.141	44.903	48.602	51.966	56.061	58.964	65.247	34
35	40.223	46.059	49.802	53.203	57.342	60.275	66.619	35
36	41.304	47.212	50.998	54.437	58.619	61.581	67.985	36
37	42.383	48.363	52.192	55.668	59.893	62.883	69.348	37
38	43.462	49.513	53.384	56.895	61.162	64.181	70.704	38
39	44.539	50.660	54.572	58.120	62.428	65.475	72.055	39
40	45.616	51.805	55.758	59.342	63.691	66.766	73.403	40
50	56.334	63.167	67.505	71.420	76.154	79.490	86.660	50
60	66.981	74.397	79.082	83.298	88.379	91.952	99.608	60
70	77.577	85.527	90.531	95.023	100.425	104.215	112.317	70
80	88.130	96.578	101.879	106.629	112.329	116.321	124.839	80
90	98.650	107.565	113.145	118.136	124.116	128.299	137.208	90
100	109.141	118.498	124.342	129.561	135.807	140.170	149.449	100

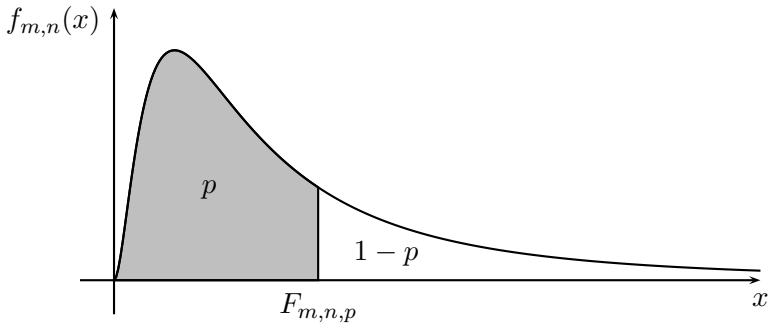


Table T.6: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	1	2	3	4	5	6	7	8	9	10
1	0.990	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056
	0.975	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3	968.6
	0.950	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	0.900	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	0.990	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40
	0.975	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
	0.950	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	0.900	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
3	0.990	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23
	0.975	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
	0.950	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785
	0.900	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
4	0.990	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	0.975	12.22	10.65	9.979	9.604	9.364	9.197	9.074	8.980	8.905	8.844
	0.950	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
	0.900	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
5	0.990	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	0.975	10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619
	0.950	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
	0.900	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
6	0.990	13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
	0.975	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461
	0.950	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
	0.900	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
7	0.990	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
	0.975	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761
	0.950	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
	0.900	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
8	0.990	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
	0.975	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295
	0.950	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347
	0.900	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
9	0.990	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
	0.975	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964
	0.950	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
	0.900	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
10	0.990	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
	0.975	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717
	0.950	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
	0.900	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323

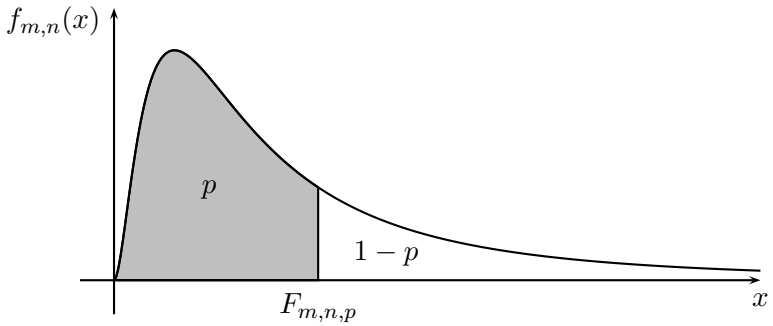


Table T.7: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	11	12	13	14	15	16	17	18	19	20
1	0.990	6083	6107	6126	6143	6157	6170	6181	6191	6201	6209
	0.975	973.0	976.7	979.8	982.5	984.9	986.9	988.7	990.3	991.8	993.1
	0.950	243.0	243.9	244.7	245.4	245.9	246.5	246.9	247.3	247.7	248.0
	0.900	60.47	60.71	60.90	61.07	61.22	61.35	61.46	61.57	61.66	61.74
2	0.990	99.41	99.42	99.42	99.43	99.43	99.44	99.44	99.44	99.45	99.45
	0.975	39.41	39.41	39.42	39.43	39.43	39.44	39.44	39.44	39.45	39.45
	0.950	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
	0.900	9.401	9.408	9.415	9.420	9.425	9.429	9.433	9.436	9.439	9.441
3	0.990	27.13	27.05	26.98	26.92	26.87	26.83	26.79	26.75	26.72	26.69
	0.975	14.37	14.34	14.30	14.28	14.25	14.23	14.21	14.20	14.18	14.17
	0.950	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660
	0.900	5.222	5.216	5.210	5.205	5.200	5.196	5.193	5.190	5.187	5.184
4	0.990	14.45	14.37	14.31	14.25	14.20	14.15	14.11	14.08	14.05	14.02
	0.975	8.794	8.751	8.715	8.684	8.657	8.633	8.611	8.592	8.575	8.560
	0.950	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803
	0.900	3.907	3.896	3.886	3.878	3.870	3.864	3.858	3.853	3.848	3.844
5	0.990	9.963	9.888	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55
	0.975	6.568	6.525	6.488	6.456	6.428	6.403	6.381	6.362	6.344	6.329
	0.950	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558
	0.900	3.282	3.268	3.257	3.247	3.238	3.230	3.223	3.217	3.212	3.207
6	0.990	7.790	7.718	7.657	7.605	7.559	7.519	7.483	7.451	7.422	7.396
	0.975	5.410	5.366	5.329	5.297	5.269	5.244	5.222	5.202	5.184	5.168
	0.950	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874
	0.900	2.920	2.905	2.892	2.881	2.871	2.863	2.855	2.848	2.842	2.836
7	0.990	6.538	6.469	6.410	6.359	6.314	6.275	6.240	6.209	6.181	6.155
	0.975	4.709	4.666	4.628	4.596	4.568	4.543	4.521	4.501	4.483	4.467
	0.950	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445
	0.900	2.684	2.668	2.654	2.643	2.632	2.623	2.615	2.607	2.601	2.595
8	0.990	5.734	5.667	5.609	5.559	5.515	5.477	5.442	5.412	5.384	5.359
	0.975	4.243	4.200	4.162	4.130	4.101	4.076	4.054	4.034	4.016	3.999
	0.950	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150
	0.900	2.519	2.502	2.488	2.475	2.464	2.454	2.446	2.438	2.431	2.425
9	0.990	5.178	5.111	5.055	5.005	4.962	4.924	4.890	4.860	4.833	4.808
	0.975	3.912	3.868	3.831	3.798	3.769	3.744	3.722	3.701	3.683	3.667
	0.950	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936
	0.900	2.396	2.379	2.364	2.351	2.340	2.330	2.320	2.312	2.305	2.298
10	0.990	4.772	4.706	4.650	4.601	4.558	4.520	4.487	4.457	4.430	4.405
	0.975	3.665	3.621	3.583	3.550	3.522	3.496	3.474	3.453	3.435	3.419
	0.950	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774
	0.900	2.302	2.284	2.269	2.255	2.244	2.233	2.224	2.215	2.208	2.201

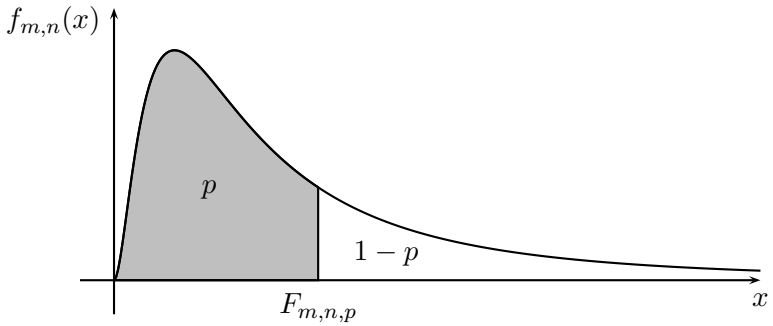


Table T.8: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	25	30	40	50	60	80	100	200	500	∞
1	0.990	6240	6260	6286	6302	6313	6326	6334	6350	6360	6366
	0.975	998.1	1001	1006	1008	1010	1012	1013	1016	1017	1018
	0.950	249.3	250.1	251.1	251.8	252.2	252.7	253.0	253.7	254.1	254.3
	0.900	62.05	62.26	62.53	62.69	62.79	62.93	63.01	63.17	63.26	63.33
2	0.990	99.46	99.47	99.48	99.48	99.48	99.48	99.49	99.49	99.50	99.50
	0.975	39.46	39.46	39.47	39.48	39.48	39.49	39.49	39.49	39.50	39.50
	0.950	19.46	19.46	19.47	19.48	19.48	19.48	19.49	19.49	19.49	19.50
	0.900	9.451	9.458	9.466	9.471	9.475	9.479	9.481	9.486	9.489	9.491
3	0.990	26.58	26.50	26.41	26.35	26.32	26.27	26.24	26.18	26.15	26.13
	0.975	14.12	14.08	14.04	14.01	13.99	13.97	13.96	13.93	13.91	13.90
	0.950	8.634	8.617	8.594	8.581	8.572	8.561	8.554	8.540	8.532	8.527
	0.900	5.175	5.168	5.160	5.155	5.151	5.147	5.144	5.139	5.136	5.134
4	0.990	13.91	13.84	13.75	13.69	13.65	13.61	13.58	13.52	13.49	13.46
	0.975	8.501	8.461	8.411	8.381	8.360	8.335	8.319	8.288	8.270	8.257
	0.950	5.769	5.746	5.717	5.699	5.688	5.673	5.664	5.646	5.635	5.628
	0.900	3.828	3.817	3.804	3.795	3.790	3.782	3.778	3.769	3.764	3.761
5	0.990	9.449	9.379	9.29	9.24	9.20	9.16	9.13	9.08	9.04	9.02
	0.975	6.268	6.227	6.175	6.144	6.123	6.096	6.080	6.048	6.028	6.015
	0.950	4.521	4.496	4.464	4.444	4.431	4.415	4.405	4.385	4.373	4.365
	0.900	3.187	3.174	3.157	3.147	3.140	3.132	3.126	3.116	3.109	3.105
6	0.990	7.296	7.229	7.143	7.091	7.057	7.013	6.987	6.934	6.901	6.880
	0.975	5.107	5.065	5.012	4.980	4.959	4.932	4.915	4.882	4.862	4.849
	0.950	3.835	3.808	3.774	3.754	3.740	3.722	3.712	3.690	3.678	3.669
	0.900	2.815	2.800	2.781	2.770	2.762	2.752	2.746	2.734	2.727	2.722
7	0.990	6.058	5.992	5.908	5.858	5.824	5.781	5.755	5.702	5.671	5.650
	0.975	4.405	4.362	4.309	4.276	4.254	4.227	4.210	4.176	4.156	4.142
	0.950	3.404	3.376	3.340	3.319	3.304	3.286	3.275	3.252	3.239	3.230
	0.900	2.571	2.555	2.535	2.523	2.514	2.504	2.497	2.484	2.476	2.471
8	0.990	5.263	5.198	5.116	5.065	5.032	4.989	4.963	4.911	4.880	4.859
	0.975	3.937	3.894	3.840	3.807	3.784	3.756	3.739	3.705	3.684	3.670
	0.950	3.108	3.079	3.043	3.020	3.005	2.986	2.975	2.951	2.937	2.928
	0.900	2.400	2.383	2.361	2.348	2.339	2.328	2.321	2.307	2.298	2.293
9	0.990	4.713	4.649	4.567	4.517	4.483	4.441	4.415	4.363	4.332	4.311
	0.975	3.604	3.560	3.505	3.472	3.449	3.421	3.403	3.368	3.347	3.333
	0.950	2.893	2.864	2.826	2.803	2.787	2.768	2.756	2.731	2.717	2.707
	0.900	2.272	2.255	2.232	2.218	2.208	2.196	2.189	2.174	2.165	2.159
10	0.990	4.311	4.247	4.165	4.115	4.082	4.039	4.014	3.962	3.930	3.909
	0.975	3.355	3.311	3.255	3.221	3.198	3.169	3.152	3.116	3.094	3.080
	0.950	2.730	2.700	2.661	2.637	2.621	2.601	2.588	2.563	2.548	2.538
	0.900	2.174	2.155	2.132	2.117	2.107	2.095	2.087	2.071	2.062	2.055

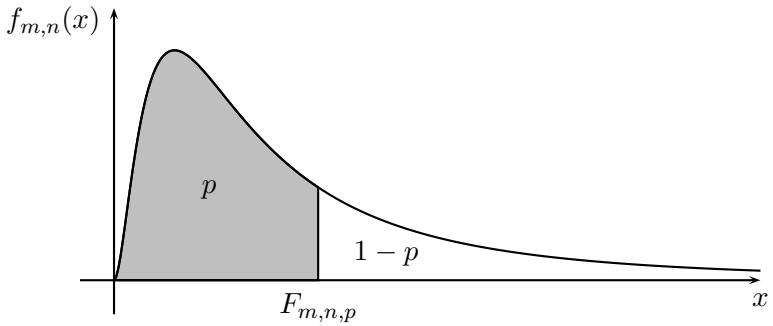


Table T.9: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	1	2	3	4	5	6	7	8	9	10
11	0.990	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
	0.975	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526
	0.950	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
	0.900	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
12	0.990	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
	0.975	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374
	0.950	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
	0.900	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
13	0.990	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
	0.975	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250
	0.950	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
	0.900	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
14	0.990	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
	0.975	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147
	0.950	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
	0.900	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
15	0.990	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
	0.975	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060
	0.950	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
	0.900	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
16	0.990	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
	0.975	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986
	0.950	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
	0.900	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
17	0.990	8.400	6.112	5.185	4.669	4.336	4.101	3.927	3.791	3.682	3.593
	0.975	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985	2.922
	0.950	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
	0.900	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
18	0.990	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
	0.975	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929	2.866
	0.950	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
	0.900	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
19	0.990	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
	0.975	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880	2.817
	0.950	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
	0.900	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
20	0.990	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
	0.975	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774
	0.950	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
	0.900	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937

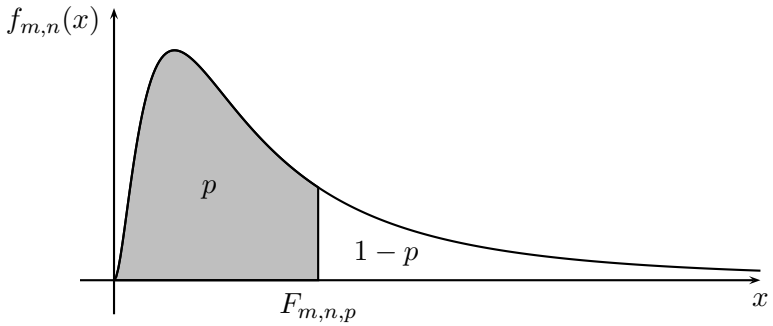


Table T.10: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	11	12	13	14	15	16	17	18	19	20
11	0.990	4.462	4.397	4.342	4.293	4.251	4.213	4.180	4.150	4.123	4.099
	0.975	3.474	3.430	3.392	3.359	3.330	3.304	3.282	3.261	3.243	3.226
	0.950	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646
	0.900	2.227	2.209	2.193	2.179	2.167	2.156	2.147	2.138	2.130	2.123
12	0.990	4.220	4.155	4.100	4.052	4.010	3.972	3.939	3.910	3.883	3.858
	0.975	3.321	3.277	3.239	3.206	3.177	3.152	3.129	3.108	3.090	3.073
	0.950	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544
	0.900	2.166	2.147	2.131	2.117	2.105	2.094	2.084	2.075	2.067	2.060
13	0.990	4.025	3.960	3.905	3.857	3.815	3.778	3.745	3.716	3.689	3.665
	0.975	3.197	3.153	3.115	3.082	3.053	3.027	3.004	2.983	2.965	2.948
	0.950	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459
	0.900	2.116	2.097	2.080	2.066	2.053	2.042	2.032	2.023	2.014	2.007
14	0.990	3.864	3.800	3.745	3.698	3.656	3.619	3.586	3.556	3.529	3.505
	0.975	3.095	3.050	3.012	2.979	2.949	2.923	2.900	2.879	2.861	2.844
	0.950	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388
	0.900	2.073	2.054	2.037	2.022	2.010	1.998	1.988	1.978	1.970	1.962
15	0.990	3.730	3.666	3.612	3.564	3.522	3.485	3.452	3.423	3.396	3.372
	0.975	3.008	2.963	2.925	2.891	2.862	2.836	2.813	2.792	2.773	2.756
	0.950	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328
	0.900	2.037	2.017	2.000	1.985	1.972	1.961	1.950	1.941	1.932	1.924
16	0.990	3.616	3.553	3.498	3.451	3.409	3.372	3.339	3.310	3.283	3.259
	0.975	2.934	2.889	2.851	2.817	2.788	2.761	2.738	2.717	2.698	2.681
	0.950	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276
	0.900	2.005	1.985	1.968	1.953	1.940	1.928	1.917	1.908	1.899	1.891
17	0.990	3.518	3.455	3.401	3.353	3.312	3.275	3.242	3.212	3.186	3.162
	0.975	2.870	2.825	2.786	2.753	2.723	2.697	2.673	2.652	2.633	2.616
	0.950	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230
	0.900	1.978	1.958	1.940	1.925	1.912	1.900	1.889	1.879	1.870	1.862
18	0.990	3.434	3.371	3.316	3.269	3.227	3.190	3.158	3.128	3.101	3.077
	0.975	2.814	2.769	2.730	2.696	2.667	2.640	2.617	2.596	2.576	2.559
	0.950	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191
	0.900	1.954	1.933	1.916	1.900	1.887	1.875	1.864	1.854	1.845	1.837
19	0.990	3.360	3.297	3.242	3.195	3.153	3.116	3.084	3.054	3.027	3.003
	0.975	2.765	2.720	2.681	2.647	2.617	2.591	2.567	2.546	2.526	2.509
	0.950	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155
	0.900	1.932	1.912	1.894	1.878	1.865	1.852	1.841	1.831	1.822	1.814
20	0.990	3.294	3.231	3.177	3.130	3.088	3.051	3.018	2.989	2.962	2.938
	0.975	2.721	2.676	2.637	2.603	2.573	2.547	2.523	2.501	2.482	2.464
	0.950	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124
	0.900	1.913	1.892	1.875	1.859	1.845	1.833	1.821	1.811	1.802	1.794

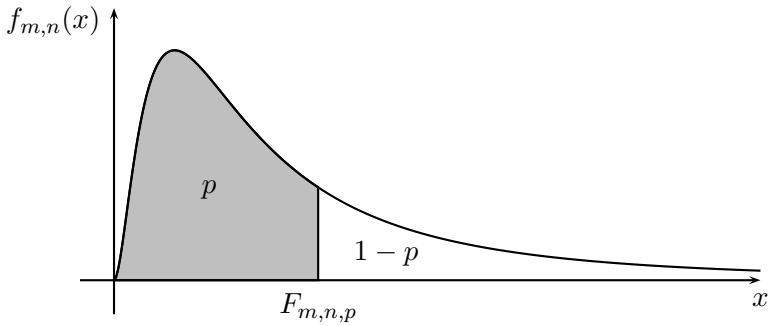


Table T.11: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	25	30	40	50	60	80	100	200	500	∞
11	0.990	4.005	3.941	3.860	3.810	3.776	3.734	3.708	3.656	3.624	3.603
	0.975	3.162	3.118	3.061	3.027	3.004	2.974	2.956	2.920	2.898	2.883
	0.950	2.601	2.570	2.531	2.507	2.490	2.469	2.457	2.431	2.415	2.405
	0.900	2.095	2.076	2.052	2.036	2.026	2.013	2.005	1.989	1.979	1.972
12	0.990	3.765	3.701	3.619	3.569	3.535	3.493	3.467	3.414	3.382	3.361
	0.975	3.008	2.963	2.906	2.871	2.848	2.818	2.800	2.763	2.740	2.725
	0.950	2.498	2.466	2.426	2.401	2.384	2.363	2.350	2.323	2.307	2.296
	0.900	2.031	2.011	1.986	1.970	1.960	1.946	1.938	1.921	1.911	1.904
13	0.990	3.571	3.507	3.425	3.375	3.341	3.298	3.272	3.219	3.187	3.165
	0.975	2.882	2.837	2.780	2.744	2.720	2.690	2.671	2.634	2.611	2.596
	0.950	2.412	2.380	2.339	2.314	2.297	2.275	2.261	2.234	2.218	2.206
	0.900	1.978	1.958	1.931	1.915	1.904	1.890	1.882	1.864	1.853	1.846
14	0.990	3.412	3.348	3.266	3.215	3.181	3.138	3.112	3.059	3.026	3.004
	0.975	2.778	2.732	2.674	2.638	2.614	2.583	2.565	2.526	2.503	2.487
	0.950	2.341	2.308	2.266	2.241	2.223	2.201	2.187	2.159	2.142	2.131
	0.900	1.933	1.912	1.885	1.869	1.857	1.843	1.834	1.816	1.805	1.797
15	0.990	3.278	3.214	3.132	3.081	3.047	3.004	2.977	2.923	2.891	2.869
	0.975	2.689	2.644	2.585	2.549	2.524	2.493	2.474	2.435	2.411	2.395
	0.950	2.280	2.247	2.204	2.178	2.160	2.137	2.123	2.095	2.078	2.066
	0.900	1.894	1.873	1.845	1.828	1.817	1.802	1.793	1.774	1.763	1.755
16	0.990	3.165	3.101	3.018	2.967	2.933	2.889	2.863	2.808	2.775	2.753
	0.975	2.614	2.568	2.509	2.472	2.447	2.415	2.396	2.357	2.333	2.316
	0.950	2.227	2.194	2.151	2.124	2.106	2.083	2.068	2.039	2.022	2.010
	0.900	1.860	1.839	1.811	1.793	1.782	1.766	1.757	1.738	1.726	1.718
17	0.990	3.068	3.003	2.920	2.869	2.835	2.791	2.764	2.709	2.676	2.653
	0.975	2.548	2.502	2.442	2.405	2.380	2.348	2.329	2.289	2.264	2.248
	0.950	2.181	2.148	2.104	2.077	2.058	2.035	2.020	1.991	1.973	1.960
	0.900	1.831	1.809	1.781	1.763	1.751	1.735	1.726	1.706	1.694	1.686
18	0.990	2.983	2.919	2.835	2.784	2.749	2.705	2.678	2.623	2.589	2.566
	0.975	2.491	2.445	2.384	2.347	2.321	2.289	2.269	2.229	2.204	2.187
	0.950	2.141	2.107	2.063	2.035	2.017	1.993	1.978	1.948	1.929	1.917
	0.900	1.805	1.783	1.754	1.736	1.723	1.707	1.698	1.678	1.665	1.657
19	0.990	2.909	2.844	2.761	2.709	2.674	2.630	2.602	2.547	2.512	2.489
	0.975	2.441	2.394	2.333	2.295	2.270	2.237	2.217	2.176	2.150	2.133
	0.950	2.106	2.071	2.026	1.999	1.980	1.955	1.940	1.910	1.891	1.878
	0.900	1.782	1.759	1.730	1.711	1.699	1.683	1.673	1.652	1.639	1.631
20	0.990	2.843	2.778	2.695	2.643	2.608	2.563	2.535	2.479	2.445	2.421
	0.975	2.396	2.349	2.287	2.249	2.223	2.190	2.170	2.128	2.103	2.085
	0.950	2.074	2.039	1.994	1.966	1.946	1.922	1.907	1.875	1.856	1.843
	0.900	1.761	1.738	1.708	1.690	1.677	1.660	1.650	1.629	1.616	1.607

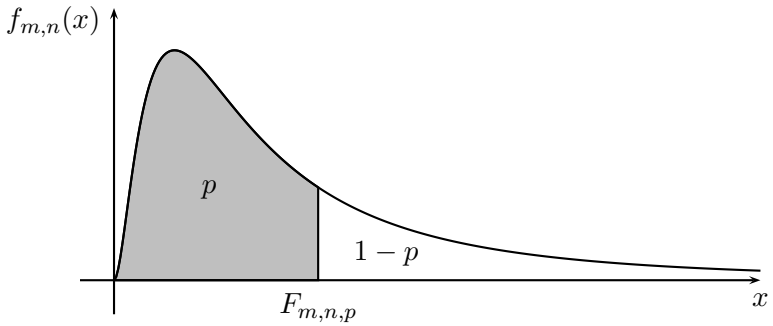


Table T.12: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	1	2	3	4	5	6	7	8	9	10
25	0.990	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
	0.975	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613
	0.950	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
	0.900	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
30	0.990	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979
	0.975	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511
	0.950	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
	0.900	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
40	0.990	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
	0.975	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388
	0.950	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
	0.900	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763
50	0.990	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
	0.975	5.340	3.975	3.390	3.054	2.833	2.674	2.553	2.458	2.381	2.317
	0.950	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
	0.900	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
60	0.990	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
	0.975	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270
	0.950	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
	0.900	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707
80	0.990	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551
	0.975	5.218	3.864	3.284	2.950	2.730	2.571	2.450	2.355	2.277	2.213
	0.950	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
	0.900	2.769	2.370	2.154	2.016	1.921	1.849	1.793	1.748	1.711	1.680
100	0.990	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
	0.975	5.179	3.828	3.250	2.917	2.696	2.537	2.417	2.321	2.244	2.179
	0.950	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
	0.900	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663
200	0.990	6.763	4.713	3.881	3.414	3.110	2.893	2.730	2.601	2.497	2.411
	0.975	5.100	3.758	3.182	2.850	2.630	2.472	2.351	2.256	2.178	2.113
	0.950	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
	0.900	2.731	2.329	2.111	1.973	1.876	1.804	1.747	1.701	1.663	1.631
500	0.990	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356
	0.975	5.054	3.716	3.142	2.811	2.592	2.434	2.313	2.217	2.139	2.074
	0.950	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
	0.900	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644	1.612
∞	0.990	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321
	0.975	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192	2.114	2.048
	0.950	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831
	0.900	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599

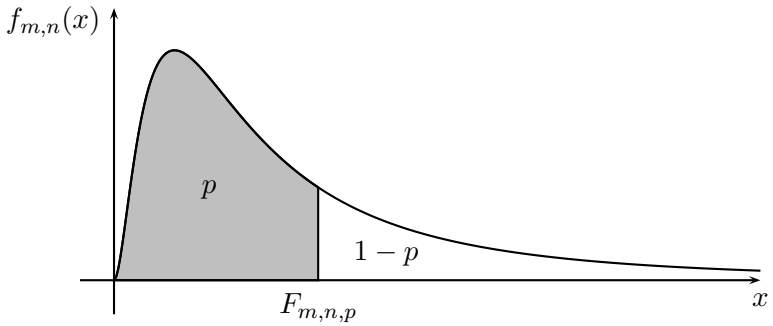


Table T.13: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	11	12	13	14	15	16	17	18	19	20
25	0.990	3.056	2.993	2.939	2.892	2.850	2.813	2.780	2.751	2.724	2.699
	0.975	2.560	2.515	2.476	2.441	2.411	2.384	2.360	2.338	2.318	2.300
	0.950	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007
	0.900	1.841	1.820	1.802	1.785	1.771	1.758	1.746	1.736	1.726	1.718
30	0.990	2.906	2.843	2.789	2.742	2.700	2.663	2.630	2.600	2.573	2.549
	0.975	2.458	2.412	2.372	2.338	2.307	2.280	2.255	2.233	2.213	2.195
	0.950	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932
	0.900	1.794	1.773	1.754	1.737	1.722	1.709	1.697	1.686	1.676	1.667
40	0.990	2.727	2.665	2.611	2.563	2.522	2.484	2.451	2.421	2.394	2.369
	0.975	2.334	2.288	2.248	2.213	2.182	2.154	2.129	2.107	2.086	2.068
	0.950	2.038	2.003	1.974	1.948	1.924	1.904	1.885	1.868	1.853	1.839
	0.900	1.737	1.715	1.695	1.678	1.662	1.649	1.636	1.625	1.615	1.605
50	0.990	2.625	2.563	2.508	2.461	2.419	2.382	2.348	2.318	2.290	2.265
	0.975	2.263	2.216	2.176	2.140	2.109	2.081	2.056	2.033	2.012	1.993
	0.950	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784
	0.900	1.703	1.680	1.660	1.643	1.627	1.613	1.600	1.588	1.578	1.568
60	0.990	2.559	2.496	2.442	2.394	2.352	2.315	2.281	2.251	2.223	2.198
	0.975	2.216	2.169	2.129	2.093	2.061	2.033	2.008	1.985	1.964	1.944
	0.950	1.952	1.917	1.887	1.860	1.836	1.815	1.796	1.778	1.763	1.748
	0.900	1.680	1.657	1.637	1.619	1.603	1.589	1.576	1.564	1.553	1.543
80	0.990	2.478	2.415	2.361	2.313	2.271	2.233	2.199	2.169	2.141	2.115
	0.975	2.158	2.111	2.071	2.035	2.003	1.974	1.948	1.925	1.904	1.884
	0.950	1.910	1.875	1.845	1.817	1.793	1.772	1.752	1.734	1.718	1.703
	0.900	1.653	1.629	1.609	1.590	1.574	1.559	1.546	1.534	1.523	1.513
100	0.990	2.430	2.368	2.313	2.265	2.223	2.185	2.151	2.120	2.092	2.067
	0.975	2.124	2.077	2.036	2.000	1.968	1.939	1.913	1.890	1.868	1.849
	0.950	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676
	0.900	1.636	1.612	1.592	1.573	1.557	1.542	1.528	1.516	1.505	1.494
200	0.990	2.338	2.275	2.220	2.172	2.129	2.091	2.057	2.026	1.997	1.971
	0.975	2.058	2.010	1.969	1.932	1.900	1.870	1.844	1.820	1.798	1.778
	0.950	1.837	1.801	1.769	1.742	1.717	1.694	1.674	1.656	1.639	1.623
	0.900	1.603	1.579	1.558	1.539	1.522	1.507	1.493	1.480	1.468	1.458
500	0.990	2.283	2.220	2.166	2.117	2.075	2.036	2.002	1.970	1.942	1.915
	0.975	2.019	1.971	1.929	1.892	1.859	1.830	1.803	1.779	1.757	1.736
	0.950	1.808	1.772	1.740	1.712	1.686	1.664	1.643	1.625	1.607	1.592
	0.900	1.583	1.559	1.537	1.518	1.501	1.485	1.471	1.458	1.446	1.435
∞	0.990	2.248	2.185	2.130	2.082	2.039	2.000	1.965	1.934	1.905	1.878
	0.975	1.993	1.945	1.903	1.866	1.833	1.803	1.776	1.751	1.729	1.708
	0.950	1.789	1.752	1.720	1.692	1.666	1.644	1.623	1.604	1.587	1.571
	0.900	1.570	1.546	1.524	1.505	1.487	1.471	1.457	1.444	1.432	1.421

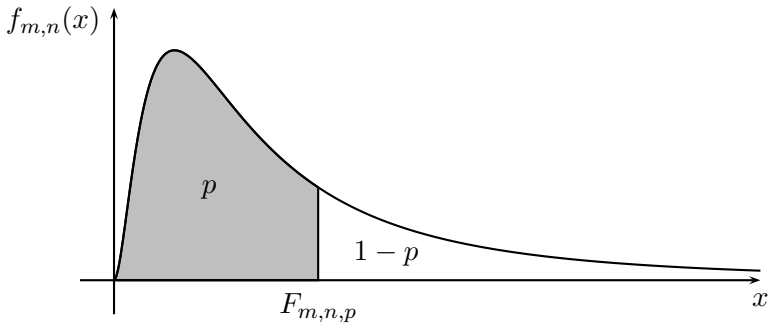


Table T.14: p-quantiles $F_{m,n,p}$ for F-**distribution** with (m, n) degrees of freedom. It is $F_{n,m,1-p} = \frac{1}{F_{m,n,p}}$.

		m									
n	q	25	30	40	50	60	80	100	200	500	∞
25	0.990	2.604	2.538	2.453	2.400	2.364	2.317	2.289	2.230	2.194	2.170
	0.975	2.230	2.182	2.118	2.079	2.052	2.017	1.996	1.952	1.924	1.906
	0.950	1.955	1.919	1.872	1.842	1.822	1.796	1.779	1.746	1.725	1.711
	0.900	1.683	1.659	1.627	1.607	1.593	1.576	1.565	1.542	1.527	1.518
30	0.990	2.453	2.386	2.299	2.245	2.208	2.160	2.131	2.070	2.032	2.006
	0.975	2.124	2.074	2.009	1.968	1.940	1.904	1.882	1.835	1.806	1.787
	0.950	1.878	1.841	1.792	1.761	1.740	1.712	1.695	1.660	1.637	1.622
	0.900	1.632	1.606	1.573	1.552	1.538	1.519	1.507	1.482	1.467	1.456
40	0.990	2.271	2.203	2.114	2.058	2.019	1.969	1.938	1.874	1.833	1.805
	0.975	1.994	1.943	1.875	1.832	1.803	1.764	1.741	1.691	1.659	1.637
	0.950	1.783	1.744	1.693	1.660	1.637	1.608	1.589	1.551	1.526	1.509
	0.900	1.568	1.541	1.506	1.483	1.467	1.447	1.434	1.406	1.389	1.377
50	0.990	2.167	2.098	2.007	1.949	1.909	1.857	1.825	1.757	1.713	1.683
	0.975	1.919	1.866	1.796	1.752	1.721	1.681	1.656	1.603	1.569	1.545
	0.950	1.727	1.687	1.634	1.599	1.576	1.544	1.525	1.484	1.457	1.438
	0.900	1.529	1.502	1.465	1.441	1.424	1.402	1.388	1.359	1.340	1.327
60	0.990	2.098	2.028	1.936	1.877	1.836	1.783	1.749	1.678	1.633	1.601
	0.975	1.869	1.815	1.744	1.699	1.667	1.625	1.599	1.543	1.507	1.482
	0.950	1.690	1.649	1.594	1.559	1.534	1.502	1.481	1.438	1.409	1.389
	0.900	1.504	1.476	1.437	1.413	1.395	1.372	1.358	1.326	1.306	1.292
80	0.990	2.015	1.944	1.849	1.788	1.746	1.690	1.655	1.579	1.530	1.494
	0.975	1.807	1.752	1.679	1.632	1.599	1.555	1.527	1.467	1.428	1.400
	0.950	1.644	1.602	1.545	1.508	1.482	1.448	1.426	1.379	1.347	1.325
	0.900	1.472	1.443	1.403	1.377	1.358	1.334	1.318	1.284	1.261	1.245
100	0.990	1.965	1.893	1.797	1.735	1.692	1.634	1.598	1.518	1.466	1.427
	0.975	1.770	1.715	1.640	1.592	1.558	1.512	1.483	1.420	1.378	1.347
	0.950	1.616	1.573	1.515	1.477	1.450	1.415	1.392	1.342	1.308	1.283
	0.900	1.453	1.423	1.382	1.355	1.336	1.310	1.293	1.257	1.232	1.214
200	0.990	1.868	1.794	1.694	1.629	1.583	1.521	1.481	1.391	1.328	1.279
	0.975	1.698	1.640	1.562	1.511	1.474	1.425	1.393	1.320	1.269	1.229
	0.950	1.561	1.516	1.455	1.415	1.386	1.346	1.321	1.263	1.221	1.189
	0.900	1.414	1.383	1.339	1.310	1.289	1.261	1.242	1.199	1.168	1.144
500	0.990	1.810	1.735	1.633	1.566	1.517	1.452	1.408	1.308	1.232	1.165
	0.975	1.655	1.596	1.515	1.462	1.423	1.370	1.336	1.254	1.192	1.137
	0.950	1.528	1.482	1.419	1.376	1.345	1.303	1.275	1.210	1.159	1.113
	0.900	1.391	1.358	1.313	1.282	1.260	1.229	1.209	1.160	1.122	1.087
∞	0.990	1.773	1.696	1.592	1.523	1.473	1.404	1.358	1.247	1.153	1.000
	0.975	1.626	1.566	1.484	1.428	1.388	1.333	1.296	1.205	1.128	1.000
	0.950	1.506	1.459	1.394	1.350	1.318	1.274	1.243	1.170	1.106	1.000
	0.900	1.375	1.342	1.295	1.263	1.240	1.207	1.185	1.130	1.082	1.000

Table T.15: Factors for constructing control charts. Source: [23], p. 762.

	$\bar{x}$ chart			R chart			s chart			
n	A_1	A_2	A_3	D_3	D_4	d_2	B_3	B_4	c_4	n
2	3.760	1.880	2.659	0.000	3.267	1.128	0.000	3.267	0.7979	2
3	2.394	1.023	1.954	0.000	2.575	1.693	0.000	2.568	0.8862	3
4	1.880	0.729	1.628	0.000	2.282	2.059	0.000	2.266	0.9213	4
5	1.596	0.577	1.427	0.000	2.115	2.326	0.000	2.089	0.9400	5
6	1.410	0.483	1.287	0.000	2.004	2.534	0.030	1.970	0.9515	6
7	1.277	0.419	1.182	0.076	1.924	2.704	0.118	1.882	0.9594	7
8	1.175	0.373	1.099	0.136	1.864	2.847	0.185	1.815	0.9650	8
9	1.094	0.337	1.032	0.184	1.816	2.970	0.239	1.761	0.9693	9
10	1.028	0.308	0.975	0.223	1.777	3.078	0.284	1.716	0.9727	10
11	0.973	0.285	0.927	0.256	1.744	3.173	0.321	1.679	0.9754	11
12	0.925	0.266	0.886	0.284	1.716	3.258	0.354	1.646	0.9776	12
13	0.884	0.249	0.850	0.308	1.692	3.336	0.382	1.618	0.9794	13
14	0.848	0.235	0.817	0.329	1.671	3.407	0.406	1.594	0.9810	14
15	0.816	0.223	0.789	0.348	1.652	3.472	0.428	1.572	0.9823	15
16	0.788	0.212	0.763	0.364	1.636	3.532	0.448	1.552	0.9835	16
17	0.762	0.203	0.739	0.379	1.621	3.588	0.466	1.534	0.9845	17
18	0.738	0.194	0.718	0.392	1.608	3.640	0.482	1.518	0.9854	18
19	0.717	0.187	0.698	0.404	1.596	3.689	0.497	1.503	0.9862	19
20	0.697	0.180	0.680	0.414	1.586	3.735	0.510	1.490	0.9869	20
21	0.679	0.173	0.663	0.425	1.575	3.778	0.523	1.477	0.9876	21
22	0.662	0.167	0.647	0.434	1.566	3.819	0.534	1.466	0.9882	22
23	0.647	0.162	0.633	0.443	1.557	3.858	0.545	1.455	0.9887	23
24	0.632	0.157	0.619	0.452	1.548	3.895	0.555	1.445	0.9892	24
25	0.619	0.153	0.606	0.459	1.541	3.931	0.565	1.435	0.9896	25

Bibliography

[1] F. J. Anscombe, *Graphs in statistical analysis*, American Statistician, 27, 17-21, 1973.

[2] D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis & Its Applications*, John Wiley & Sons, 1988.

[3] G. E. P. Box, J. S. Hunter and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition, John Wiley & Sons, 2005.

[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th Edition, John Wiley & Sons, 2015.

[5] K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, 1965.

[6] D. R. Cox and E. J. Snell, *Applied Statistics Principles and Examples*, Chapman & Hall, 1981.

[7] P. L. Davies, *Data features*, Statistica Neerlandica, 49, 185–245, 1995.

[8] A. C. Davison, *Statistical Models*, Cambridge University Press, 2008.

[9] J. Devore and N. Farnum, *Applied Statistics for Engineers and Scientists*, 3rd Edition, Duxbury, Thomson, 2013.

[10] W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis*, McGraw-Hill, 1983.

[11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman & Hall, 1993.

[12] P. D. Haaland, *Experimental Design in Biotechnology*, Vol. 105 of Statistics, Textbooks and Monographs, Marcel Dekker, New York, 1989.

[13] A. Kiermeier, *Visualising and Assessing Acceptance Sampling Plans: The R Package AcceptanceSampling*, Journal of Statistical Software, Vol. 26/6, 2008.

[14] A. Kiermeier, *The AcceptanceSampling Package*, `https://cran.r-project.org/web/packages/AcceptanceSampling/`.

[15] M. Kühlmeyer, *Statistische Auswertungsmethoden für Ingenieure*, Springer-Verlag, 2001.

[16] J. Lawson, *Design and Analysis of Experiments with R*, Chapman & Hall, 2013.

[17] R. V. Lenth, *Response-Surface Methods in R, Using* `rsm`, 2012.

[18] R. V. Lenth, *Surface Plots in the* `rsm` *Package*, 2012.

[19] R. A. Maronna, D. R. Martin and V. J. Yohai, *Robust Statistics: Theory and Methods*, Wiley Series in Probability and Statistics, 2006.

[20] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd Edition, Chapman & Hall, 1989.

[21] D. C. Montgomery, *Design and Analysis of Experiments*, 8th Edition, John Wiley & Sons, Inc., 2012.

[22] D. C. Montgomery, *Introduction to Statistical Quality Control*, 6th Edition, John Wiley & Sons, Inc., 2014.

[23] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 6th Edition, John Wiley & Sons, Inc., 2013.

[24] D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th Edition, John Wiley & Sons, Inc., 2012.

[25] R. H. Myers and D. C. Montgomery, *Response Surface Methodology*, 3rd Edition, John Wiley & Sons, Inc., 2009.

[26] J. S. Oakland, *Statistical Process Control*, Butterworth-Heinemann, 2007.

[27] S. L. Pfleeger, *Experimental Design and Analysis in Software Engineering*, ACM SIGSOFT Software Engineering Notes, Vol. 20, No. 2, 14–16, 1995.

[28] J. A. Rice, *Mathematical Statistics and Data Analysis*, 3rd Edition, Duxbury, Thomson, 2007.

[29] H. Rinne and H.-J. Mittag, *Statistische Methoden der Qualitätssicherung*, 3. Auflage, Hanser-Verlag, München Wien, 1995.

[30] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, 1987.

[31] T. P. Ryan, *Statistical Methods for Quality Improvement*, 3rd Edition, John Wiley & Sons, Inc., 2011.

[32] A. Ruckstuhl, *MSE-Modul Angewandte Statistik und Datenanalyse: Regression und Design of Experiment*, IDP Institut für Datenanalyse und Prozessdesign, School of Engineering, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2018.

[33] L. Scrucca, *qcc: An R package for quality control charting and statistical process control*, R-News, Vol. 4/1, 2004, p.11–17.

[34] L. Scrucca, *The qcc Package*, `https://cran.r-project.org/web/packages/qcc/`.

[35] J. S. Simonoff, *Smoothing Methods in Statistics*, Springer Series in Statistics, Springer-Verlag New York, 1996.

[36] M. Steiner-Curtis, *Datenanalyse*, Skriptum FHNW, 2020.

[37] M. Steiner-Curtis, *Wahrscheinlichkeitstheorie und Statistik*, Skriptum FHNW, 2020.

[38] R. Storm, *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle*, 12. Auflage, Carl Hanser Verlag, 2007.

[39] M. Tableman and J. S. Kim, *Survival Analysis Using S*, 1st Edition, Chapman & Hall, 2003.

[40] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th Edition, Springer, 2002.

[41] D. J. Wheeler and D. S. Chambers, *Understanding Statistical Process Control*, SPC-Press, Inc., 1992.

[42] R. Wehrens, *Chemometrics with R*, 1st Edition, Springer, 2011.

[43] S. Weisberg, *Applied Linear Regression*, 3rd Edition, Wiley-Interscience, 2005.

Index

acceptance
  number, 52
  probability, 53
  sampling, 51
  sampling plan, 51, 52
    attributes, 52
    variables, 59
adjusted R^2_{adj}, 154
AIC, 154
Akaike information criterion, 154
alarm
  false, 29
  omitted, 29
alias structure, 203
aliased, 203
all subset selection, 157
analysis of variance
  two factor, 180
  two way, 186
analysis of variance table, 179
analysis of variance test, 179
ANOVA-test, 179
ARL, 34
autocorrelation, 102

backward elimination, 155
block, 186
Bonferroni, 180
bootstrap simulation, 92
breakdown point, 143

capability process ratio, 37
cause-and-effect diagram, 8, 169
centreline, 14
chart
  R, 13, 14
  $\bar{x}$, 13, 14, 18
  p, 23
  s, 18
  attributes data, 23
  individuals, 20
  mean, 13
  Pareto, 5, 201
  range, 13
  Shewhart, 13
check sheet, 5
chronological order, 169
CL, 14
coefficient of determination, 88, 115
collinearity, 157
  measure, 158
comparison of two straight lines, 124
confidence interval, 76, 112
  frequentist interpretation, 77
  response, 77, 114
  slope, 76
confounded, 202
control
  area, 13
  chart, 11
  limit, 12
    lower, 14
    upper, 14
Cook's distances measure, 135
Mallow's C_p statistic, 154
CUSUM, 42

decision interval, 43
defect concentration diagram, 7
defining relation, 203
degrees of freedom, 70
design
  balanced, 169
  block, 169
  complete block, 170
  complete factorial, 170
  completely randomised, 170
  defining relation, 203

experimental, 170
factorial, 170
factorial 2^k, 171, 193
first-order, 212
fractional factorial, 171
fractional factorial 2^k, 171, 202
generators, 203
incomplete, 171
matrix, 194, 196
matrix regression, 108
orthogonal, 169
rotatable central composite, 217
second-order central composite, 217
distribution
binomial, 25, 26
χ, 233, 234
F, 235–243
geometric, 34, 38
hypergeometric, 59
normal, 9, 31
standard normal, 230
Student's t, 232
DoE, 167
dummy-variable, 119

effect
aliased, 203
confounded, 202
error, 66
type I, 29, 52
type II, 29, 52
error sum of squares, 70, 179
EWMA, 45
experiment
design of, 167
randomised, 168
experimental design, 170
extrapolation, 81

F-statistic, 123
F-test to compare two models, 123
factor, 119, 181
factorial design, 170
first aid transformations, 98
first-order
design, 212
model, 212
fishbone approach, 8
fitted value, 70, 183
forward and backward selection, 155
forward selection, 154

generalised linear model, 161
goodness of fit tests, 91
gross error sensitivity, 143

half-normal plot, 199
hat matrix, 109
heteroscedasticity, 85
high, 171, 193
histogram, 9, 17
homoscedasticity, 85
Huber's ψ-function, 144

identity, 203
influence
diagnostics, 135
function, 143
interaction, 122, 187

LCL, 14
least-squares
generalised, 104
normal equations, 69, 109
level, 181
leverage
Huber's classification, 136
leverages, 135
linear model, 178
loss function
Huber, 144
least squares, 69, 144
Tukey's bisquare, 148
lot, 51
low, 171, 193
LSL, 11, 37

M-estimator, 145
MM-estimator, 148
magnificent seven, 4
Mallow's C_p statistic, 154
mean
arithmetic, 143
mean squares, 179
measure for collinearity, 158

median, 143
median of absolute values, 145
method
    least squares, 68
    steepest ascent, 214
model
    accuracy, 153
    complexity, 153
    empirical, 67, 151
    first-order, 212
    linear, 175
    mechanistic, 67, 151
    second-order, 217
    selection methods, 153
moving range, 21
multiple
    R^2, 115
    comparisons, 180
    contrasts, 180

nonlinear
    least-squares regression, 117
    regression, 161
nonparametric regression, 161
normal equations, 69, 109
notched box plot, 177

OC curve, 32, 53
    ideal, 53
operating characteristic, 30, 32, 53
orthogonal, 169
outliers, 100
overall mean, 178

P-value, 75
Pareto, Vilfredo F. D., 1848-1923, 6
partial least squares regression, 162
pattern recognition, 28
PCR, 37
percent defective, 52
plot
    box-, 9
    half-normal, 199
    location, 7
    notched box-, 177
    q-q, 16, 95
    quantile-quantile, 16, 95
    scale-location, 94
    Tukey-Anscombe, 91
power function, 30
prediction interval, 79, 115
principal component regression, 162
principle
    Pareto, 5
process
    capability, 36
    control, 3
    out of control, 27
    standard deviation, 15
    under control, 27
ψ-function, 145
    Huber's, 144
    redescending, 148

q-q norm, 9, 10
quality, 3

R-squared, 88
    multiple, 115
R_{adj}-squared
    adjusted, 154
R-package
    `AcceptanceSampling`, 62
    `MASS`, 146, 148
    `RGL`, 148
    `car`, 159
    `daewr`, 199
    `openxlsx`, 213
    `qcc`, 61
    `robustbase`, 146
    `rsm`, 224
randomisation, 169
redescending ψ-function, 148
reference value, 42
regression
    generalised least-squares, 104
    linear, 65
    linear model, 67
    M-estimator, 145
    MM-estimator, 148
    multiple, 107
    multiple linear, 108
    nonlinear, 161
    nonlinear functions, 117
    nonlinear least-squares, 117

nonparametric, 161
partial least squares, 162
polynomial, 116
principal component, 162
quadratic, 116
residual analysis, 86
weighted multiple, 140
replicate, 170, 187
residual, 70, 183
analysis, 86
scaled, 85
standardised, 86, 136
residual analysis, 188
response surface, 212
response surface methodology, 211
risk
consumer, 52
point
consumer, 56
producer, 56
producer, 52
robust statistics, 143
rotatable central composite design, 217
RSM, 211
run, 28, 193
average length, 34
length, 34

sample size, 52
scale-location plot, 94
second-order central composite design, 217
second-order model, 217
selection
all subset, 157
sensitivity, 143
Shewhart, Walter A., 1891-1967, 4
significance of regression, 74
simulation, 92
SL, 37
smoothing parameter, 46
SPC, 4
specification limit, 37
lower, 11, 37
upper, 11, 37
standard error, 73
statistical
process control, 4
quality control, 3
Student's t-test, 117
sum of squares, 179
total corrected, 179
treatment, 179
survival analysis, 162

table, 229
analysis of variance, 179
F-distribution, 235–243
factors for control charts, 244
quantile
χ-distribution, 233, 234
standard normal distribution, 231
Student's t-distribution, 232
standard normal distribution, 230
target value, 11, 42, 45
test
analysis of variance, 179
ANOVA, 179
curvature in 2^k design, 215
regression coefficient, 111
slope, 73
two-sample, 176
tilde, 66
time series analysis, 162
tolerance limit, 11, 37
total corrected sum of squares, 179
transformation
arcsine, 98
logarithmic, 98
logit, 98
square-root, 98
treatment, 186
treatment effect, 178
treatment sum of squares, 179
trimming, 101
Tukey's bisquare, 148
Tukey-Anscombe plot, 91
two factor analysis of variance, 180, 186
two way analysis of variance, 186
two-sample test, 176

UCL, 14
USL, 11, 37

variable
binary, 117

dummy, 119
explanatory, 67
factor, 119
predictor, 67
primary, 168
response, 67
secondary, 168
variance inflation factor, 158
VIF, 158

weight, 140
weighted normal equations, 140
western electric rules, 27
world
model, 68
real, 68

Yates standard order, 194