Part I

Statistical Process and Quality Control

Introduction

What is Statistical Quality and Process Control

Definition of Quality
- Quality means fitness for use. and
- Quality is inversely proportional to variability.
- My definition: Something is of quality if *Is* and *should* are the same

Statistical Process Control (SPC)
Statistical process control is commonly understood as a way to optimise production and manufacturing processes.

The Magnificent Seven
The magnificent seven 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

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.

| causes | interruptions/quarter | | | | total |
	1.	2.	3.	4.	
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

Pareto Chart
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.
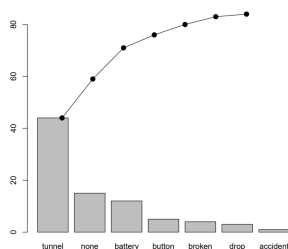

Figure 1: Pareto chart.

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.
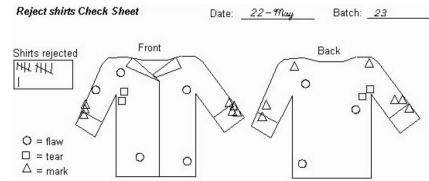

Figure 2: Defect concentration diagram or location plot: The locations of the various defects in the quality control of shirts are marked.

Cause-and-Effect Diagram
The cause-and-effect diagram is a tool in the form of a fishbone for the systematic identification of causes which make problems.
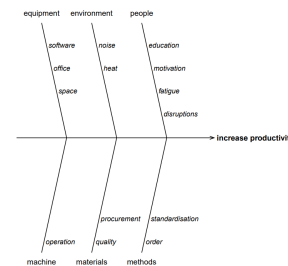

Figure 3: Cause-and-effect diagram.

Control Charts

The basis of a control chart is a statistical hypothesis test.

Hypothesis test

With a hypothesis test we want to check if an assumption (hypothesis) correct. In the case of control charts we usually want to check wheter the process is under controll. We assume that the process is under controll, meaning that the estimated mean $\bar{x}$ is equal to the target mean μ_0. The corresponding Hypothesis look as follows:

$$H_0 : \mu_0 = \bar{x} \text{ i.e. process is not disturbed}$$
$$H_1 : \mu_0 \neq \bar{x} \text{ i.e. process is disturbed} \tag{1}$$

To check the H_0 hypothesis we use a test statistic. Depending on the circumstances we use a diferent test statistic. If the standard deviation σ is known, we use a z-test. If the standard deviation is unknown, a t-test is used.

Test statistic (z-test, since σ is known)
The difference between the estimated mean and the target mean $|\bar{x} - \mu_0|$ in units of standard deviations is given by

$$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}. \tag{2}$$

Given a confidence level (for example $\alpha = 0.0027$) we can calculate the critical difference of the estimated mean from the target mean (again in units of standard deviations), for which H_0 stays true with the given confidence. For $\alpha = 0.0027$ the critical distance is 3 standard deviations, ($z_q \approx 3$). i.e. the probability of getting a more extream result (more than 3 standard deviations difference) is 0.27%

Statistical conclusion
If $|z| \leq z_q \rightarrow$ accept null hypothesis, i.e. process is not disturbed.
If $|z| < z_q \rightarrow$ reject null hypothesis, i.e. process is disturbed.

Control limits With the critical distance z_q we can calculate the control limits

$$UCL = \mu_0 + z_q \frac{\sigma}{\sqrt{n}} \qquad LCL = \mu_0 - z_q \frac{\sigma}{\sqrt{n}} \tag{3}$$

Statistical conclusion
If $LCL \leq \bar{x} \leq UCL \rightarrow$ process is not disturbed.
If $\bar{x} < LCL$ or $UCL < \bar{x} \rightarrow$ process is disturbed.

The Control Chart

We introduce the most commonly used Shewhart control chart for monitoring the mean and the variation of a process. 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.

No.	sample values						mean	sd	range
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

Figure 4: Data Set with Mean, Standard Deviation and Range

Mean values

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \tag{4}$$

Standard deviations

$$s_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2} \tag{5}$$

Ranges

$$R_i = max\{x_{ij}|j \in \{1, ..., n\}\} - min\{x_{ij}|j \in \{1, ..., n\}\} \tag{6}$$

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

Firs we monitor the variation of the process with the R- Chart. If the variation is under control, we will consider the chart for the mean $\bar{x}$.

R-Chart
Centerline
The centreline (CL) of the R chart is denoted by $\bar{R}$ and is calculated from the arithmetic mean of the ranges of the k random samples, i.e.

$$\bar{R} = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{7}$$

Control limits

$$UCL = D_4 \bar{R}; \qquad LCL = D_3 \bar{R} \tag{8}$$

The constants D_3 and D_4 depend only on the sample size n and can be found in tables.

$\bar{x}$ Chart based on R chart
Control limits
The control limits of the $\bar{x}$ chart can be calculated using the mean μ, the standard error and the significance level of $z_q = 3$.

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}}; \qquad LCL = \mu - 3\frac{\sigma}{\sqrt{n}} \tag{9}$$

The problem with that is that μ and σ are in general unknown and must be estimated from the process data.
This is a two-stage process. First make sure that the process standard deviation (R chart) is under statistical control. That is, if some samples are out of bounds, it is recommended to omit these measurements and recalculate the limits. Next we can use $\bar{R}$ to estimate the process standard deviation.

$$\hat{\sigma} = \frac{\bar{R}}{d_2} \tag{10}$$

Centerline
Any samples excluded for construction of the R chart should also be disregarded for construction of the $\bar{x}$ chart. This results in a sample of $k^\star$ valid samples, (where $k^\star$ denotes the reduced number of samples). Mean values of $\bar{x}_1, ..., \bar{x}_{k^\star}$ provide an estimate of μ, i.e

$$\bar{\bar{x}} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} \bar{x}_i \tag{11}$$

Control limits
The control limits then are

$$UCL = \bar{\bar{x}} + 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_2\bar{R}$$
$$UCL = \bar{\bar{x}} - 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_2\bar{R} \tag{12}$$

Control Chart with $\bar{x}$ and s

We again monitor the variation (this time with an s chart) and then consider the chart for the mean $\bar{x}$.

s Chart
Centerline
The centreline of the s chart is denoted by $\bar{s}$ and is calculated from the arithmetic mean of the standard deviations

$$\bar{s} = \frac{1}{k} \sum_{i=1}^{k} s_i \tag{13}$$

Control limits
Analogously to above, the control limits are given by

$$UCL = B_4\bar{s}; \qquad LCL = B_3\bar{s} \tag{14}$$

$\bar{x}$ Chart based on s chart
Using an s chart of a process that is under control, the process standard deviation can be estimated by

$$\hat{\sigma} = \frac{\bar{s}}{c_4} \tag{15}$$

Centerline
Any samples excluded for construction of the s chart should again also be disregarded for construction of the $\bar{x}$ chart. This results in a sample of $k^\star$ valid samples, (where $k^\star$ denotes the reduced number of samples). Mean values of $\bar{x}_1, ..., \bar{x}_{k^\star}$ provide an estimate of μ, i.e

$$\hat{\mu} = \bar{\bar{x}} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} \bar{x}_i \tag{16}$$

Control limits

$$UCL = \bar{\bar{x}} + 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_3\bar{s}$$
$$UCL = \bar{\bar{x}} - 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_3\bar{s} \tag{17}$$

Individual Control Charts

Individual control charts have exactly one measurement per sample. The problem with that is that one cannot estimate variability from a single measurement. The solution to that is to use variation of two adjacent measurements to measure the process variability.

The moving range is defined as

$$MR_i = |x_{i+1} - x_i| \tag{18}$$

for all $i \in \{1, ..., n-1\}$.

The arithmetic mean of all moving ranges

$$\overline{MR} = \frac{1}{n-1} \sum_{i=1}^{n-1} MR_i \tag{19}$$

is a reasonable estimate for the process standard deviation

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{\overline{MR}}{1.128} \tag{20}$$

Since two neighboring measurements were used to calculate the moving ranges we have $d_2 = 1.128$.

Centerline
The centerline for the individuals control chart is the arithmetic mean of the measured values.

$$\bar{x} = \frac{1}{k} \sum_{i=1}^{k} x_i \tag{21}$$

Control limits

$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128}; \qquad LCL = \bar{x} - 3\frac{\overline{MR}}{1.128} \tag{22}$$

Andre Gut

Control Charts for Attributes Data – p Chart

Sometimes it is useful to mark a product as defective or useable. In such a case we need a control chart for attributes data.

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. The relative frequency

$$\hat{p} = \frac{D}{n} \tag{23}$$

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

$$Var(\hat{p}) = \frac{p(1-p)}{n} \tag{24}$$

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

$$p_1 = \frac{d_1}{n_1}, ..., p_k = \frac{d_k}{n_k} \tag{25}$$

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 $k^\star$ valid samples.

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

$$\bar{p} = \frac{1}{k^\star} \sum_{i=1}^{k^\star} p_i \tag{26}$$

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

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \qquad LSL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{27}$$

- 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

$$\bar{p} = \frac{d_1 + \cdots + d_{k^\star}}{n_1 + \cdots + n_{k^\star}} \tag{28}$$

  and the control limits by

$$UCL_i = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}; \qquad LSL_i = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \tag{29}$$

Statistical Properties of Control Charts

Aim of process control using control charts: Keep the process under statistical control. Or, if it is not at the beginning, to put it into statistical control by improving production conditions.

Interpretation of Control Charts

Western Electric Rules

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

2. 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).

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).

4. Nine consecutive points fall on the same side of the centreline (so-called run).

Rule 1: Any point beyond zone A.

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

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

Rule 4: Nine consecutive points fall on the same side of the centreline.
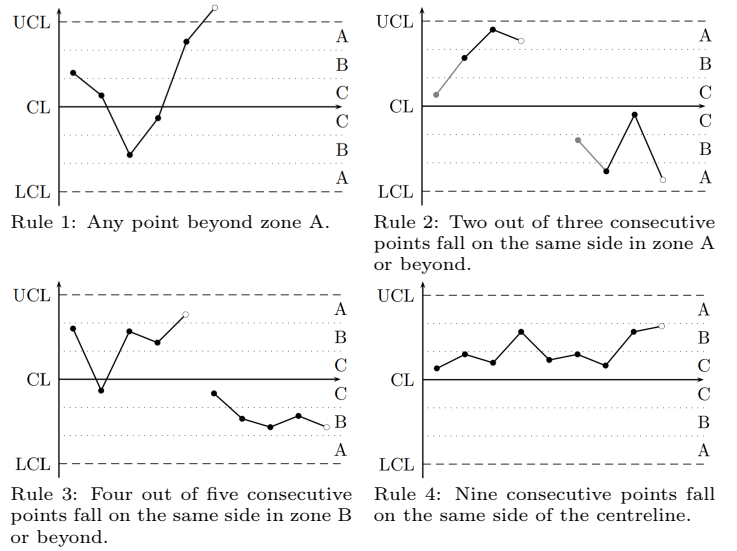
Type I Error and Type II Error

When monitoring a production process with a control chart, as with any statistical test, there are two wrong decisions possible.

$$
\begin{aligned}
H_0 &: \mu_0 = \mu \text{ i.e. process is not disturbed}\\
H_1 &: \mu_0 \neq \mu \text{ i.e. process is disturbed, } \mu_1 \text{ is true}
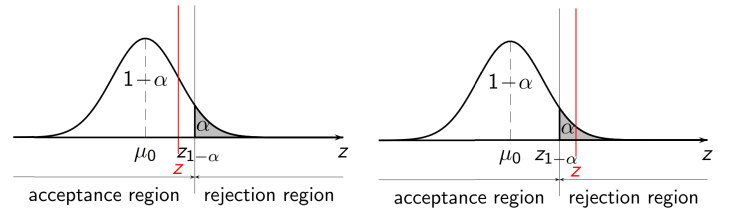\end{aligned} \tag{30}
$$

Denoted by:
- μ_0 the target value of the process
- μ the considered statistic, eg. $\mu = \bar{x}$ or $\mu = R$
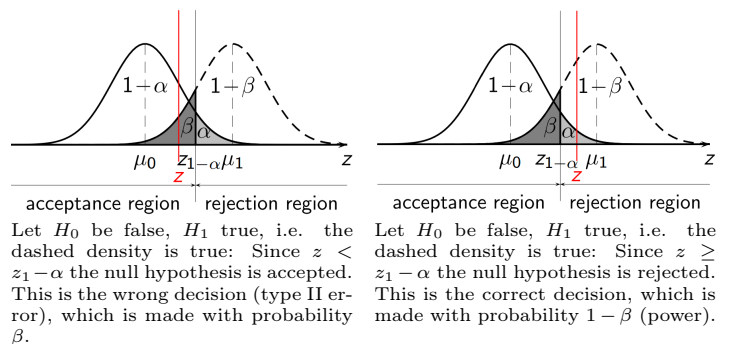- μ_1 the true value of the considered statistic.

Two wrong decisions possible:
- 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.
- 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

Type I Error

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 $1 - \alpha$.

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.

Type II Error

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.

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).

Power Function and Operating Characteristic

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, i.e.

$$\text{power} = P(\text{reject } H_0 | H_1 \text{ is true}) = 1 - \beta \qquad (31)$$

Power function
Probability to reject the null hypothesis H_0 if H_1 is true.

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}, \qquad (32)$$

or $\mu_1 = \delta\sigma + \mu_0$.
The variable δ is a normalized 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

$$g(\mu_1) = g(\delta\sigma + \mu_0) = \tilde{g}(\delta). \qquad (33)$$

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

In the following figure four power function curves for different sample sizes $n = 2$, $n = 5$, $n = 10$ and $n = 20$ are plotted. For a given δ (deviation of the actual value of the process to the target value measured in units of σ) we can read off the plot the probability of a type II error, i.e. an omitted alarm.
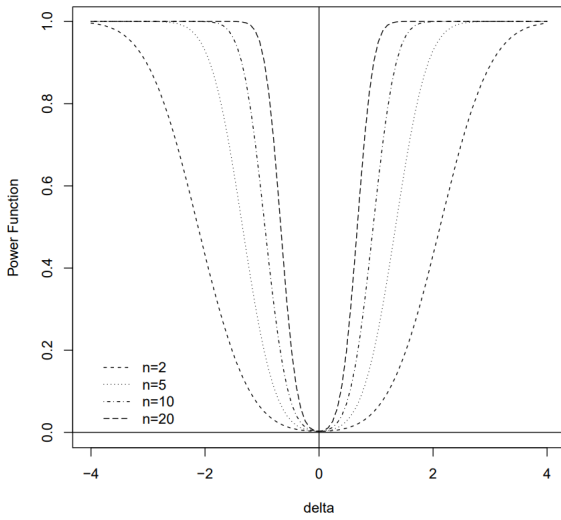

Figure 5: Power Function Curve

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

$$g(\mu_1) = \tilde{g}(0) = \alpha. \qquad (34)$$

Disturbed Process
For a disturbed process we have

$$\tilde{g}(\delta) = \Phi(\delta\sqrt{n} - z_q, 0, 1) + \Phi(-\delta\sqrt{n} - z_q, 0, 1) \qquad (35)$$

with Φ being

$$\Phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \qquad (36)$$

Average Run Length

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 ARL is the expected value of the probability i.e. the likelihood of exceeding the control limits when performing a test.

$$ARL(\delta) = \frac{1}{p(\mu)} = \frac{1}{g(\mu_1)} = \frac{1}{\tilde{g}(\delta)} \qquad (37)$$

Undisturbed Process
If we have an undisturbed process with $\mu_1 = \mu_0$, i.e. with $\delta = 0$, then it follows from equation 34 that

$$ARL(0) = \frac{1}{\alpha}. \qquad (38)$$

Disturbed Process
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 equation 35.

Process Capability

The specification limit (SL) is defined by

$$SL = \frac{USL + LSL}{2} \qquad (39)$$

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

$$C_p = \frac{USL - LSL}{6\sigma} \qquad (40)$$

The capability process ratio C_p expresses the ratio of the width of the tolerance range to the width of the process range.

- $C_p = 1$ 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.

Control Charts with Memory

Classical Shewhart control charts
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.

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.

$$y_i = \alpha_i + \sum_{j=1}^{i} \beta_j \bar{x}_j, \qquad (41)$$

where α_i and the weights $\beta_1, ..., \beta_i$ can be arbitrary real numbers where the sum of all $\beta s = 1$.
Depending on how the weights are chosen, we get another type of control chart.

CUSUM - Cumulative Sum Control Chart

Idea
The CUSUM chart plots the cumulative sums of deviations of measurement values from the target value.

Recursive procedure
Using two statistics C^+, resp. C^- the CUSUM chart sums up deviations above, resp. below the target value

$$\begin{aligned} C_i^+ &= max\{0, \bar{x}_i - (\mu_0 + K) + C_{i-1}^+\}, \\ C_i^- &= max\{0, (\mu_0 - K) - \bar{x}_i + C_{i-1}^+\}. \end{aligned} \qquad (42)$$

C^+ and C^- only sum up deviations from the target value, which are greater than the reference value K. The starting values of the recursion are $C^+ = 0$ and $C^- = 0$.

If a shift of Δ is to be detected then set

$$K = \frac{\Delta}{2}. \qquad (43)$$

The constant K is called reference value.

If the process is under controll the expected values of the statistic are both 0.
If the process is not under control, then the statistic sums up the deviations. If the sum of de deviations (C^+ and C^-) exceed the decision interval H, then we should stop the process and look for the cause.

Rule of thumb for choosing the constants K and H: Let $\hat{\sigma}$ be an estimate for the process standard deviation.

- reference value $= K\frac{\hat{\sigma}}{2}$
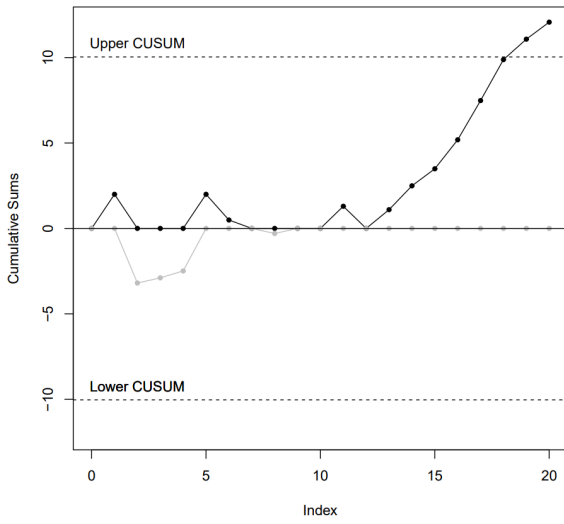- decision interval $H = 5\hat{\sigma}$

Figure 6: CUSUM-Chart
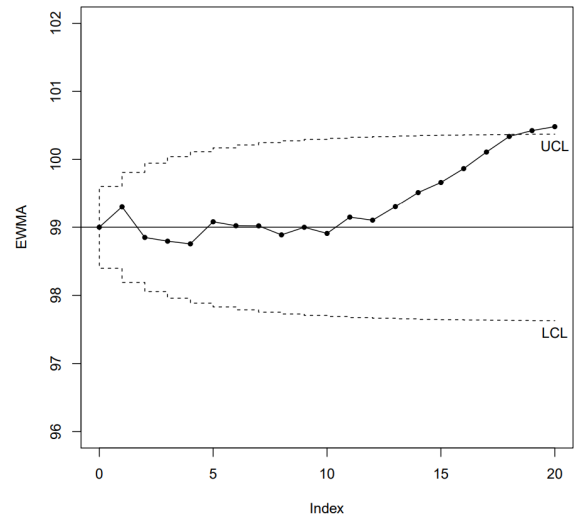

Figure 7: EWMA-Chart

EWMA - Exponentially Weighted Moving Average

Idea
Monitor the mean and weight its influence on the statistic with the exponentially decaying factor β_j.

Smoothing parameter λ
Lambda lays between 0 and 1. The smoothing parameter λ determines the influence of the previous sample mean $\bar{x}_j$ on the statistic. The smaller λ, the more values $\bar{x}_j$ are used for the decision. For $\lambda = 1$ only one sample is used and we get the well-known Shewhart $\bar{x}$ chart.

Weights

$$
\begin{aligned}
\alpha_i &= (1-\lambda)^1 \mu_0 \\
\beta_j &= \lambda(1-\lambda)^{i-j} \text{with } j \in \{1, 2, ..., i\}.
\end{aligned}
\tag{44}
$$

Statistics

$$
y_i = (1-\lambda)^i \mu_0 + \lambda \sum_{j=1}^{i} (1-\lambda)^{i-j} \bar{x}_j
\tag{45}
$$

Start: $y_0 = \mu_0$

The same with recursion

$$
y_i = (1-\lambda)y_{i-1} + \lambda \bar{x}_j
\tag{46}
$$

Assumptions
If the process is under control, then $\bar{x}_i$ comes from a normal distribution with the expected value μ_0 and the standard deviation $\frac{\sigma}{\sqrt{n}}$.
The standard deviation is either known or can be estimated from data. The statistic y_i is then also normally distributed with $E(y_i) = \mu_0$ and

$$
Var(y_i) = \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\frac{\sigma^2}{n}
\tag{47}
$$

Control limits
These assumptions lead to the 3σ control limits:

$$
\begin{aligned}
UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}} \\
LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}}
\end{aligned}
\tag{48}
$$

The asymtotic control limits are:

$$
\begin{aligned}
UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \\
LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}}
\end{aligned}
\tag{49}
$$

Estimate of process standard error

$$
\hat{\sigma} = \frac{\bar{s}}{c_4}
\tag{50}
$$

Acceptance Sampling

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.

Idea
Draw random samples from a lot and make a decision on the quality of the lot based on this information.

Acceptance Sampling Plans for Attributes

Percent defective
We have a lot of N parts (known). The number of defective parts m (unknown) with $0 \leq m \leq N$ is observed. Thus, the percent defective is:

$$
p = \frac{m}{N}
\tag{51}
$$

Acceptance sampling plan consitst of:

- Sample size n,
- the number of defective parts x in the sample, where $0 \leq x < n$,
- the acceptance number c, where $0 \leq c < n$ and
- the rule: if $x \leq c$, then the lot is accepted, if $x > c$, then the lot is rejected.

Hypothesis test
The rule above can also be formulated as a hypothesis test:

$$
\begin{aligned}
H_0 &: x \leq c, \text{ i.e. the lot is not rejected.} \\
H_1 &: x > c, \text{ i.e. the lot is rejected.}
\end{aligned}
\tag{52}
$$

As with any hypothesis test the following two errors can be made.

Type I Error
H_0 true, but rejected (Lot is good but is rejected)
Probability for this to happen (false negative): α. (Producer's risk)
The producer wants to avoid type I error, i.e. does not want good lots to be returned.

Type II Error
H_0 false, but accepted (Lot is bad but is accepted)
Probability for this to happen (false positive): β. (Consumer's risk)
The consumer wants to avoid type II error, i.e. does not want to accept bad lots.

In an agreement between a producer and a consumer the following parameter must de defined:

- α: The producer's acceptable probability of falsly rejecting a good lot,
- β: The consumer's acceptable probability of falsly accepting a bad lot,

- p_α the producer's minimal percentage defective needed for a lot to be returned (The producer does not want to take back lots with $p < p_\alpha$.),

- p_β the consumer's maximal percentage defective needed for a lot to be accepted (The consumer wants to reject lots with $p_\beta < p$ whenever possible).

Operating Characteristic

The operating characteristic, short OC, is the probability of accepting a lot as a function of the percent defective p.

The number of defective units X is a hypergeometric random variable. There are N parts in total, among which m defective and $N - m$ are good parts. 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

$$OC(p) = P(X \leq c) = \sum_{k=0}^{c} P(X = k) = \sum_{k=0}^{c} \frac{\binom{pN}{k}\binom{N-pN}{n-k}}{\binom{N}{n}} \qquad (53)$$

For a given lot size N, the parameters n and c of the acceptance sampling plan determine the form of the OC curve.

Ideal OC Curve
If all parts of a lot get checked we get an ideal acceptance sampling plan:
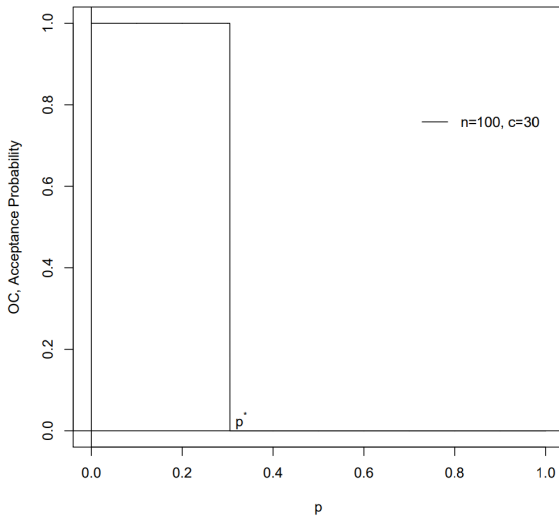

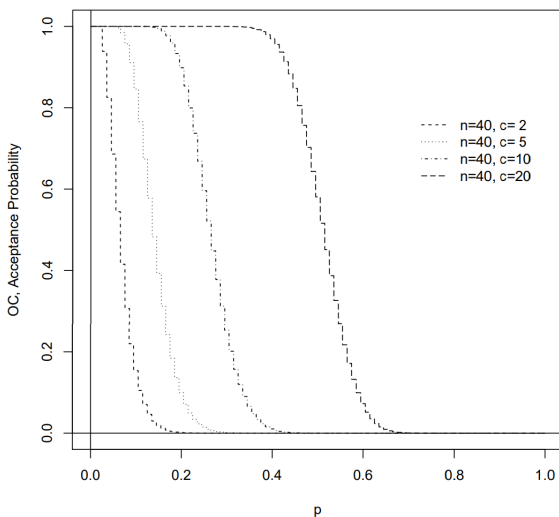Figure 8: OC curve of an ideal acceptance sampling plan

OC Curves - Examples
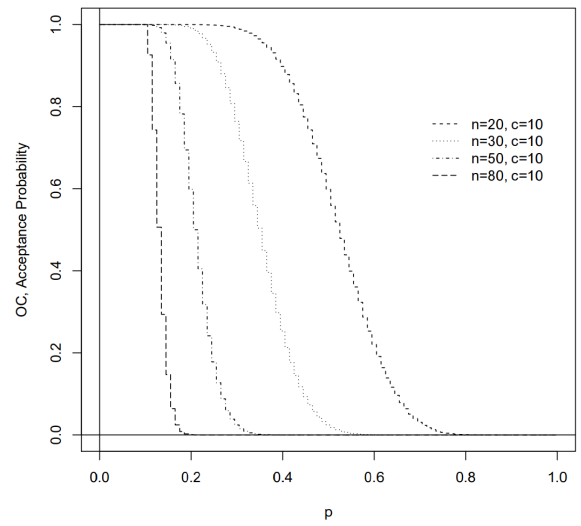

Figure 9: OC curve with fixed n and increasing c

Figure 10: OC curve with increasing n and fixed c

Parameters of an Acceptance Sampling Plan

n and c are to be chosen such that:

- n is as small as possible,

- the producer risk is at most equal to α, i.e. $OC(p_\alpha) \leq 1 - \alpha$, and

- the consumer risk is at most equal to β, i.e. $OC(p_\beta) \geq \beta$

The pair $(p_\alpha, 1 - \alpha)$ is called producer risk point and the pair (p_β, β) is called consumer risk point.

n and c are calculated by brute force. The resulting OC curve is not a perfect fit since only integer values can be choosen.
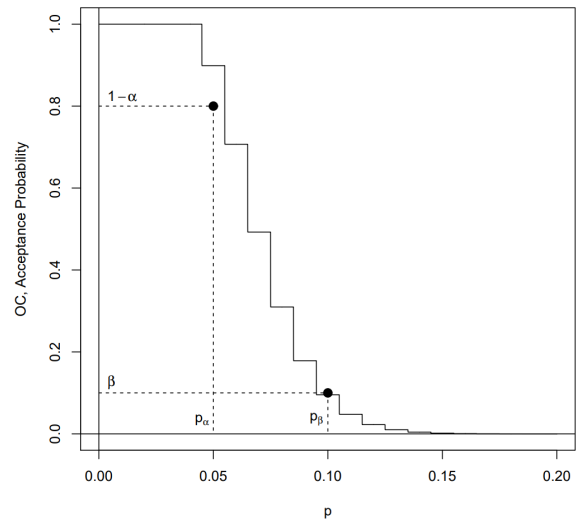
Real OC Curve

Figure 11: OC curve of a real acceptance sampling plan.

Acceptance Sampling Plans for Variables

It is always possible to reduce an acceptance sampling plan for variables to an acceptance sampling plan for attributes by saying:

- If $LSL \leq x \leq USL$, then the part is fit for use.

- If $x < LSL$ or $USL < x$, then the part is rejected

By counting the number of rejected parts, we again have an acceptance sampling plan for attributes.

Part II
Multiple Regression

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.

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.

The model for a simple linear regression is
$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{54}$$
where β_0 and β_1 are unknown and fixed parameters.

The perdictor/input/explanatory variable x is deterministic
The response/output variable y is a random variable
ε is the uncorrelated random error with
$$E(\varepsilon) = 0 \tag{55}$$
and unknown variance
$$\text{Var}(\varepsilon) = \sigma^2. \tag{56}$$
Since the error ε is a random value, the response y is also a random value.

The aim is to find the parameters $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$ such that model fits the data as well as possible.

It is assumend, that the error ε is normaly distributed.
$$E \approx \mathcal{N}(0, \sigma^2) \Rightarrow E(y) = \beta_0 + \beta_1 x \tag{57}$$

Estimation of the Parameters

The parameters β_0 and β_1 are estimated using the method of least square, i.e. by minimising:
$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2, \tag{58}$$
with
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \ \text{ with } \ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{59}$$
where
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \ \text{ and } \ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{60}$$
with
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ \text{ and } \ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{61}$$

Residual
The difference between the ith observed value y_i and its fitted value $\hat{y}_i$ is the residual
$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i) \tag{62}$$
where the mean of the residual is given by
$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0 \tag{63}$$

Unbiased estimator
The unbiased estimator $\hat{\sigma}^2$ is obtained from the error sum of squares
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \tag{64}$$
Notice that we have to divide by the degrees of freedom $n-2$ to make the estimator unbiased. (2 because we have two parameters to find).

Distribution of the estimators
To be able to answer how well the model fits the data, whether the model can be used as a reliable predictor and whether the the assumptions of constant variance and uncorrelated errors are met, we need to know the distributions of the estimators β_0 and β_1.
Since the estimators β_0 and β_1 are calculated using the random variable y, the estimators themselves are random variables which follow the distributions
$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \tag{65}$$
and
$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{66}$$

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.

Test of a Slope
We use the following hypothesis
$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \\ H_1 &: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{67}$$

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. Where $\beta_{1.0}$ is the slope we want to test.

We can estimate the standard error with the error sum of squares.
$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \ \text{ with } \ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{68}$$

The test statistic is the following
$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{se}(\hat{\beta}_1)} \tag{69}$$

Under the null hypothesis the test statistic T follows a Student's t-distribution with $n-2$ degrees of freedom. Values of t are usually found in a table.
If T is smaller than t (from the table), we accept the null hypothesis and conclude that the data agrees also with a model with the slope $\beta_{1,0}$.

In many applications we test the two alternative hypotheses
$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{70}$$

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. These hypotheses are used in the R-output!

P-Value
The P-Value is the probability, under the null hypothesis, of obtaining a result equal to or more extreme than what was actually estimated. If the value T of the test statistic is larger than the critival value t (from the table) the null hypothesis is not rejected i.e. it is plausible to assume that the data fits the model.

The test statistic 69 is accepted on the significance level α if
$$t_{\frac{\alpha}{2}, n-2} \leq T \leq t_{1-\frac{\alpha}{2}, n-2}, \tag{71}$$

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.

Confidence Interval
which leads us to the following confidence interval on the slope
$$\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} \cdot \text{se}(\beta_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2}, n-2} \cdot \text{se}(\beta_1) \tag{72}$$

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} \cdot \text{se}(\beta_1), \hat{\beta}_1 - t_{1-\frac{\alpha}{2}, n-2} \cdot \text{se}(\beta_1)\right] \tag{73}$$

Interpretation: The true value of β_1 lies in the confidence interval with high probability.

Andre Gut

Confidence Interval of the Response

The estimated response of a regression model is as follows:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{74}$$

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

$$H_0 : \hat{y}_0 = \mu_0 \tag{75}$$

The estimator $\hat{y}_0$ is normally distributed, unbiased and has the variance

$$\text{Var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \tag{76}$$

As usual σ^2 is in general unknown and has to be estimated from the data (equation 64). The test statistic about the response is

$$T = \frac{\hat{y}_0 - \mu_0}{\text{se}(\hat{y}_0)}, \tag{77}$$

with the standard error

$$\text{se}(\hat{y}_0) = \hat{\sigma} \sqrt{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} }. \tag{78}$$

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

Confidence Interval of the Response
The confidence interval on the response at the point x_0 is

$$\hat{y}_0 - t_{\frac{\alpha}{2}, n-2} \cdot \text{se}(\hat{y}_0) \leq \mu_0 \leq \hat{y}_0 + t_{1-\frac{\alpha}{2}, n-2} \cdot \text{se}(\hat{y}_0) \tag{79}$$

Prediction Interval

The point estimate of a new value of the response y_0 is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{80}$$

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 the random variable $y_0 - \hat{y}_0$ has to be considered.

To calculate its variance we use the formula for the difference of two uncorrelated random variables

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) \tag{81}$$

Therefore

$$\begin{aligned} \text{Var}(y_0 - \hat{y}_0) &= \text{Var}(y_0) + \text{Var}(\hat{y}_0) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned} \tag{82}$$

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.
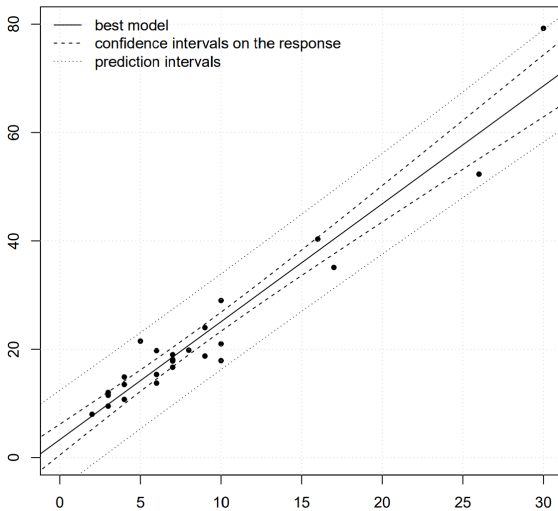
Prediction Interval
The prediction interval at the point x_0 is

$$\hat{y}_0 - t_{\frac{\alpha}{2}, n-2} \cdot \text{se}(y_0 - \hat{y}_0) \leq \hat{y}_0 \leq \hat{y}_0 + t_{1-\frac{\alpha}{2}, n-2} \cdot \text{se}(y_0 - \hat{y}_0) \tag{83}$$

with

$$\text{se}(y_0 - \hat{y}_0) = \hat{\sigma} \sqrt{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} } \tag{84}$$

Residual Analysis

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.

Introduction

In our simple linear regression models we made the following assumptions:

1. The relationship between response and the regressors is linear.

2. The error ε has mean zero.

3. The error ε has constant variance σ^2

4. The errors are uncorrelated.

5. The errors are normally distributed

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

$$e_i = y_i - \hat{y}_i \quad \text{for all} \quad i \in 1, ..., n \tag{85}$$

instead. The residuals are estimators of the random errors.

Scaled Residuals
If the errors are normally distributed, so are the residuals of a least-squares estimate. Since the variance of the residuals depends on x_0 it is not equal to the variance of the errors. Therefore we use scaled residuals

$$\tilde{e}_i = \frac{e_i}{\sqrt{ 1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) }} \quad \text{for all} \quad i \in 1, ..., n \tag{86}$$

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.

Coefficient of Determination
The coefficient of determination is a measure of the linear relationship between the response variable and the fit.

In simple linear regression with only one explanatory variable the coefficient of determination is defined by

$$R^2 = \frac{SS_{\text{fit}}}{S_{yy}} \tag{87}$$

where

$$\begin{aligned} SS_{\text{fit}} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}}_i)^2 = \hat{\beta}_1^2 S_{xx} \\ SS_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \end{aligned} \tag{88}$$

Therefore the coefficient of determination is

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \text{Cor}(x, y)^2 \tag{89}$$

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 $\hat{y}$

$$R^2 = \text{Cor}(y, \hat{y})^2 = \frac{S_{y\hat{y}}^2}{S_{yy} S_{\hat{y}\hat{y}}} \tag{90}$$

The coefficient of determination is a global measure for the goodness of fit and it says nothing about the suitability of the regression model.

Diagnostic Tools

What follows is a description of 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.

In an exam question on the topic of residuals analysis, the following four questions must be answered in order to obtain the full score:

- Which model assumption is being checked?
- What should the plot look like if the assumption was true?
- What is observed?
- What is the conclusion?

Tukey-Anscombe Plot (Residuals vs Fitted)

The idea of this plot is to plot residuals e_i versus fitted values $\hat{y}_i$. With the Tukey-Anscombe Plot we can check whether the second model assumption holds true.

Residuals in any interval of the Tukey-Anscombe plot should vary randomly around the horizontal line at zero.
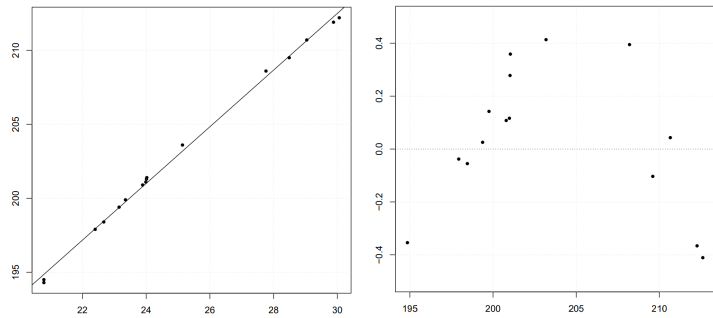

Best linear model

Tukey-Anscombe plot of the model. The points are not spread uniformly around the horizontal line indicating a bad model.

Model assumption number two states that the errors should have mean zero. To check this assumption it is best to smooth the data in the Tukey-Anscombe plot. We are still left with the question: Is such a curved curve due to chance possible? An informal method to find answers is based on bootstrap simulations

Bootstrap Simulation

What follows is a brief explanation of the bootstrap simulation:

- Calculate the best linear model with Var $= \hat{\sigma}^2$ and calculate the smooth curve of the residuals.
- Simulate new observations y_i^* which correspond to the fitted model. Draw n random numbers e_i^* from a standard normal distribution and create the simulated observations.
- 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.
- Repeat 19 times. $(1 + 19$ observations)
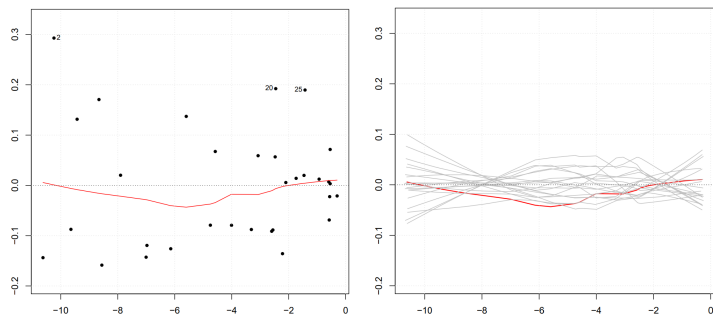
Tukey-Anscombe plot with smoothed residuals (red). The smoothed curve shows a clear curvature (which is bad).

Tukey-Anscombe plot with 19 simulated smoothed residuals (gray). The curved smooth curve is not an extreme curve among all 20. We conclude that the model fits the data.

Scale-Location Plot

The idea of this plot is to plot the residuals of square-root of absolute standardised residuals $\sqrt{|\tilde{e}_{std,i}|}$ versus fitted values $\hat{y}_i$. With the scale-location plot we can check whether the third model assumption holds true.

If the smoothed curve of square-root of the absolute standardised residuals is approximately horizontal, then the errors have equal variance.
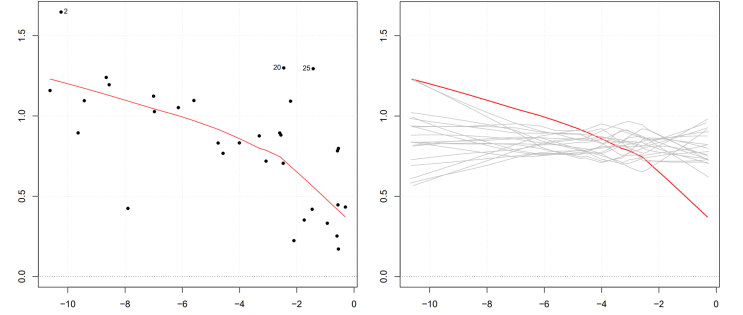
Scale-location plot with smoothed residuals (red). We observe a clear downward trend in residual scattering.

Scale-location plot with 19 simulated smoothed residuals (gray). The simulation confirms the extraordinary behavior.

Normal q-q Plot (and Histogram)

The idea of the histogram is to compare histogram of residuals e_i with normal density function with parameters 0 and $\hat{\sigma}^2$. It is often difficult to compare a histogram with a bell shaped curve and the histogram is very sensitive to the number of histogram cells and the breakpoints between cells.

The idea of the normal q-q plot is to plot quantiles of the empirical distribution of the of the standardised residuals $\tilde{e}_{std,i}$ versus 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.
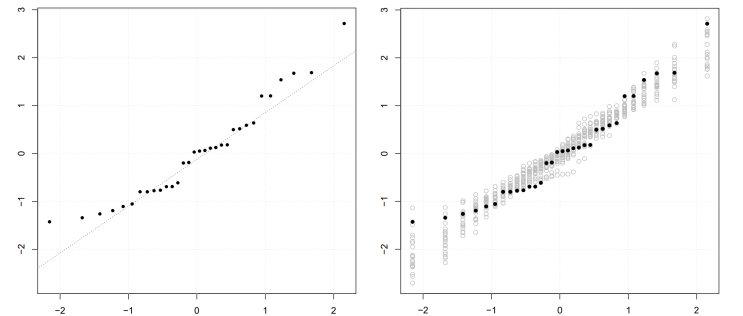
q-q plot with standardised residuals. We can observe a discrepancy to normality

q-q plot with 19 simulated standardised residuals (gray). The simulation shows us that this discrepancy might be due to random fluctuations

Treatment of Model Violations

Non-Constant Variance of Random Errors

Often the assumption is violated that the variance of random errors is constant.

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 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.

- **Logarithmic** for continuous positive variables:

$$z \mapsto \log(z) \tag{91}$$

  Logarithmic rules:

$$\log(xy) = \log(x) + \log(y) \qquad \log\left(\frac{x}{y}\right) = \log(x) - \log(y)$$

$$\log\left(x^y\right) = y \log(x) \qquad \log\left(\frac{1}{x}\right) = -\log(x)$$

- **Square-root** for continuous and discrete positive variables:

$$z \mapsto \sqrt{z} \tag{92}$$

- **Arcsine** for proportions:
$$z \mapsto \arcsin(\sqrt{z}) \tag{93}$$

- **Logit** for proportions:
$$z \mapsto \log\left(\frac{z + \varepsilon_1}{1 + \varepsilon_2 - z}\right) \tag{94}$$

Outliers
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.

Heavy-Tailed Distributions
If the q-q plot shows a symmetric distribution but with heavy tails, then transformations are useless. One possibility is to omit the most extreme observations such that the heavy tails disappear. 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.

Independence

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 subsequent residuals in a scatter plot.
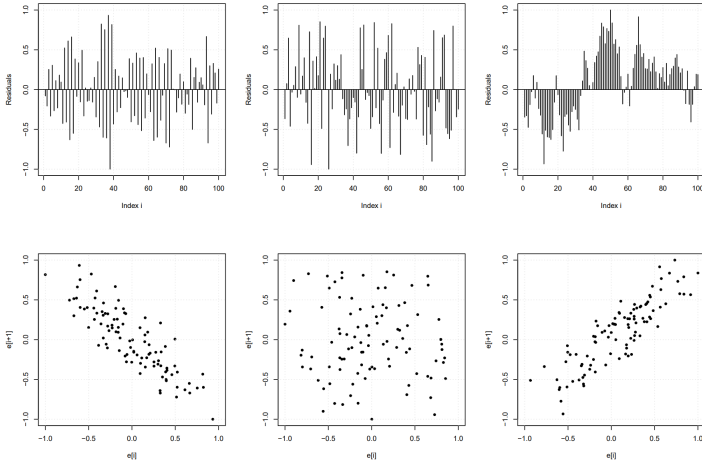
- **Negative correlation**
  After a positive (negative) residual the chance of observing a negative (positive) residual is high, i.e. very strong alternation of the sign of the residuals.
  In the scatter plot the points lie in an ellipse with negative slope of the first principal axis.

- **No correlation**
  The residuals are completely random, no obvious pattern present.
  In the scatter plot the points show no pattern.

- **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.
  In the scatter plot the points lie in an ellipse with positive slope of the first principal axis.

Multiple Linear Regression

A multiple regression model is a model which involves more than one regressor variable.

Model and Estimation

We generalise the concept of a simple linear regression model with only one explanatory variable to a multiple linear regression model
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon, \tag{95}$$

β_0 is the intercept
β_k is the slope in the x_k direction for all $k \in 1, 2, ..., m$.

Least-Squares Estimation of the Regression Coefficient
The general idea and philosophy of the multiple linear regression model are the same as in the simple linear regression model. We will again estimate the coefficients β_0 and $\beta_1, \beta_2, ..., \beta_m$ with the method of least-squares. To do this we will use matrix notation.
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{96}$$

Now we want to find the vector of least-squares estimators $\boldsymbol{\beta}$, which minimise
$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{97}$$

The solution of the normal equations is the least-squares estimator
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} \tag{98}$$

It is now straight forward to calculate the fitted values
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} = \mathbf{H}\mathbf{y} \tag{99}$$

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 e can be written as
$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{100}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.
Aufg. 8.5.1 is important!!

Tests on the Significance of any Individual Regression Coefficient
It is possible to show that $\hat{\boldsymbol{\beta}}$ follows an $(m+1)$-dimensional multivariate normal distribution with expected value
$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{101}$$

and $(m+1) \times (m+1)$ covariance matrix
$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1} \tag{102}$$

Using the above we can derive a test procedure and corresponding confidence intervals for each coefficient β_0 and β_k for $k \in \{1, 2, ..., m\}$. The hypothesis for testing the significance of any individual regression coefficient, such as β_k, are
$$\begin{aligned} H_0 &: \beta_k = \beta_{k,0} \\ H_1 &: \beta_k \neq \beta_{k,0}. \end{aligned} \tag{103}$$

The test statistic for this hypothesis is
$$T_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\mathrm{se}(\hat{\beta}_k)} \quad \text{with} \quad \mathrm{se}(\hat{\beta}_k) = \hat{\sigma}\sqrt{\left((\mathbf{X}^t \mathbf{X})^{-1}\right)_{kk}} \tag{104}$$

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. The degrees of freedom equals the denominator of
$$\hat{\sigma}^2 = \frac{1}{n - (m+1)} \sum_{i=1}^{n} e_i^2 = \frac{\mathbf{e}^t \mathbf{e}}{n - (m+1)} \tag{105}$$

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.

Confidence Intervals on the Regression Coefficients
The corresponding $100(1 - \alpha)$ percent confidence interval on β_k is
$$\hat{\beta}_k - t_{\frac{\alpha}{2}, n-(m+1)} \cdot \mathrm{se}(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{1-\frac{\alpha}{2}, n-(m+1)} \cdot \mathrm{se}(\hat{\beta}_k), \tag{106}$$

where $\mathrm{se}(\hat{\beta}_k)$ can be found in above 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

Confidence Interval of the Response
We want to construct a confidence interval on the response at a particular point. Define the $(m+1) \times 1$ vector
$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0m} \end{pmatrix} \tag{107}$$

and calculate the estimated value
$$\hat{y}_0 = \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \tag{108}$$

This is an unbiased, normally distributed estimator with variance

$$\text{Var}(\hat{y}_0) = \sigma^2 \mathbf{x}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_0 \qquad (109)$$

Therefore a $100(1-\alpha)$ percent confidence interval on the response at the point $(x_{01}, x_{02}, ..., x_{0m})$ is

$$\hat{y}_0 - t_{\frac{\alpha}{2}, n-(m+1)} \cdot \text{se}(\hat{y}_0) \le \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \le \hat{y}_0 + t_{1-\frac{\alpha}{2}, n-(m+1)} \cdot \text{se}(\hat{y}_0), \quad (110)$$

with

$$\text{se}(\hat{y}_0) = \hat{\sigma} \sqrt{\mathbf{x}_0^t \left(\mathbf{X}_0^t \mathbf{X}\right)^{-1} \mathbf{x}_0} \qquad (111)$$

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.

Prediction Interval
The regression model can be used to predict future observations on y corresponding to particular values of regressor variables, for example, $(x_{01}, x_{02}, ..., x_{0m})$. A point estimate of the future observation is

$$\hat{y}_0 = \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \qquad (112)$$

where $\mathbf{x}_0$ was defined above.
A $100(1-\alpha)$ percent prediction interval for the future observation $(x_{01}, x_{02}, ..., x_{0m})$ is

$$\hat{y}_0 - t_{\frac{\alpha}{2}, n-(m+1)} \cdot \sqrt{\sigma^2 + \text{se}(\hat{y}_0)^2} \le \mathbf{x}_0^t \hat{\boldsymbol{\beta}} \le \hat{y}_0 + t_{1-\frac{\alpha}{2}, n-(m+1)} \cdot \sqrt{\sigma^2 + \text{se}(\hat{y}_0)^2}, \qquad (113)$$

where $\text{se}(\hat{y}_0)$ was defined above 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.

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 $\hat{y}$

$$R^2 = \text{Cor}(y, \hat{y})^2. \qquad (114)$$

The coefficient of determination is a measure of the linear relationship between the response variable and the fit.

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.
Explanatory variables can also be strongly (linearly) correlated. It is however difficult to interpret the results of the regression analysis and should be avoided.

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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad (115)$$

If we define the new variables `lin` $= x$ and `quad` $= x^2$, then we obtain mathematically the same model

$$y = \beta_0 + \beta_1 \cdot \texttt{lin} + \beta_2 \cdot \texttt{quad} + \varepsilon \qquad (116)$$

which is now a multiple linear regression model.

Nonlinear Functions and Linear Regression
Experience shows that in most real-world regression problems, the variables must be transformed to get the best results. Four examples:

- Inverting the nonlinear equations

$$y = \frac{1}{a + be^{-x}} \qquad \text{gives} \qquad \frac{1}{y} = a + be^{-x}$$

$$y = \frac{ax}{b + x} \qquad \text{gives} \qquad \frac{1}{y} = \frac{1}{a} + \frac{a}{b}\frac{1}{x}$$

- Taking logarithms of the nonlinear equations

$$y = ax^b \qquad \text{gives} \qquad \ln(y) = \ln(a) + b\ln(x)$$

$$y = ax^{bg(x)} \qquad \text{gives} \qquad \ln(y) = \ln(a) + bg(x)$$

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.

Binary Explanatory Variables
In all our examples, the explanatory variables have been continuous so far. This is not necessary.
If the binary variable x_b is the only variable in the model $y = \beta_0 + \beta_1 x_b + \varepsilon$ then

$$\begin{aligned} y &= \beta_0 + \varepsilon & \text{if} & & x_b &= 0 \\ y &= \beta_0 + \beta_1 + \varepsilon & \text{if} & & x_b &= 1 \end{aligned}$$

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.

Factor Variables
If a discrete variable has more than two levels it is called a factor.
To be able to use such a factor in a regression model we have to introduce for every level a so-called dummy-variable

$$x_k^{(j)} = \left(x_{k,1}^{(j)}, \ldots, x_{k,i}^{(j)}, \ldots, x_{k,n}^{(j)}\right). \qquad (117)$$

Now we can add these $s - 1$ dummy variables to the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + (\beta_{k,2} x_k^{(2)} + \cdots + \beta_{k,s} x_k^{(s)}) + \cdots + \beta_m x_m + \varepsilon. \qquad (118)$$

Realise that for a factor with s levels we only added $s - 1$ dummy-variables. This is done to make the estimation unique by setting artificially $\beta_{k,1} = 0$. Example:

```
Position:       int   1 1 1 2 2 1 ...   3 3 4 ...   4 3 ...   3
...
Position1.dum:  num   1 1 1 0 0 1 ...   0 0 0 ...   0 0 ...   0
Position2.dum:  num   0 0 0 1 1 0 ...   0 0 0 ...   0 0 ...   0
Position3.dum:  num   0 0 0 0 0 0 ...   1 1 0 ...   0 1 ...   1
Position4.dum:  num   0 0 0 0 0 0 ...   0 0 1 ...   1 0 ...   0
```

Comparison of Regression Models with F-Test

We need 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 with $p = m+1$ coefficients and a smaller model with q coefficients $\beta_{j1}, ..., \beta_{jq}$ 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

$$\begin{aligned} H_0 &: \beta_{j1} = \beta_{j2} = \ldots \beta_{jq} = 0 \\ H_1 &: \beta_{j_k} \ne \text{ for at least one } k \in \{1, \ldots, q\} \end{aligned} \qquad (119)$$

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

$$F = \frac{n - p}{q} \frac{\text{SS}_E^* - \text{SS}_E}{\text{SS}_E} \qquad (120)$$

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.

Sensitivity and Robustness

Residual Analysis

The graphical tools Tukey-Anscombe plot, scale-location plot and q-q plot which check goodness of fit for the simple linear regression model can also be used for multiple linear regression models.
In addition to the Tukey-Anscombe plots mentioned above the *Residuals vs. Leverage plot* is introduced which will be discussed in the next section.

In multiple regression, it is no longer clear which explanatory variable might cause a 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 $\hat{y}$.

Andre Gut

Influential Observations

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.

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.

$$d_i = \frac{(\hat{\mathbf{y}}_{(-1)} - \hat{\mathbf{y}})^t (\hat{\mathbf{y}}_{(-1)} - \hat{\mathbf{y}})}{p\hat{\sigma}^2} \tag{121}$$

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 leads to the equivalent form

$$d_i = \frac{\tilde{e}_{std,i}^2}{p} \frac{h_{ii}}{1 - h_{ii}} \tag{122}$$

Cook's distance measure is a function of the ith standardised residual $\tilde{e}_{std,i}$ and the so-called leverages h_{ii} of the ith observation. The leverages

$$h_{ii} = \mathbf{H}_{ii} = (\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)_{ii} \tag{123}$$

are the diagonal entries of the hat matrix $\mathbf{H}$. Leverages h_{ii} satisfy $0 \geq h_{ii} \geq 1$ and the mean of all leverages is always $\frac{p}{n}$.

Interpretation of leverages:

- A large leverage h_{ii} has a big influence on the fitted values.
- A large leverage h_{ii} reduces the variance of the ith residual, and therefore the ith observation is close to regression line (hyperplane).
- 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?
Rule of thumb: Observations are dangerous if

$$h_{ii} > 2\frac{p}{n} \tag{124}$$

Huber's classification:
Observations with

- $h_{ii} \leq 0.2$ are harmless.
- $0.2 < h_{ii} < 0.5$ are potentially dangerous.
- $0.5 < h_{ii}$ should be avoided.

Rule of thumb for Cook's distance measure:

- If $d_i > 1$ then the ith observation is dangerous.
- If $d_i \leq 1$ then the ith observation is harmless.

Artificial data with two outliers. Original model (black) and estimated best fit (red).

Standardised residuals versus leverages with two outliers. Contours of Cook's distance measure (red).

Weighted linear regression

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.

we want to find the vector of least-squares estimators β, which minimise

$$S(\beta) = \sum_{i=1}^{n} w_i \varepsilon_i^2 = \varepsilon^t \mathbf{W} \varepsilon = (\mathbf{y} - \mathbf{X}\beta)^t \mathbf{W}(\mathbf{y} - \mathbf{X}\beta) \tag{125}$$

which simplifies to the so-called weighted least-squares normal equations

$$\mathbf{X}^t\mathbf{W}\mathbf{y} = \mathbf{X}^t\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} \tag{126}$$

The solution of the normal equations is the least-squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y} \tag{127}$$

Robust Methods

The goal of robust statistics is to find parameter estimates as if the outliers did not exist

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.

Breakdown Point
The breakdown point is the minimal proportion of incorrect observations which cause completely unrealistic estimates.
Gross Error Sensitivity
The gross error sensitivity is based on the influence function and measures the maximum effect of a single observation on the estimated value.

Robust Regression M-Estimator
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

$$\Psi(e) = w_i e_i \tag{128}$$

must be limited. This can be achieved, for instance by the Hubert's ψ -function.

Figure 12: Loss function ρ of least-squares estimate (a), ψ-function (b) and corresponding weight function w (c).

Figure 13: Loss function ρ (a) of Huber's ψ-function (b) and corresponding weight function w (c).

The corresponding weights are then simply

$$w_i = \frac{\psi(e_i)}{e_i} \tag{129}$$

The threshold c in Huber's ψ-function is set to be $c = 1.345$ which is based on theoretical efficiency considerations.

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. To be able to handle bad leverage points we need a redescending ψ-function such that observations with large absolute residuals are gradually ignored.

Modified Robust Regression M-Estimator
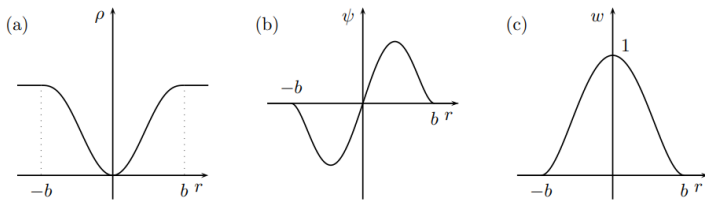
Figure 14: Loss function ρ (a) of Tukey's bisquare ψ-function (b) and corresponding weight function w (c).

The threshold b of Tukey's bisquare is set to be $b = 4.685$ which is based on theoretical efficiency considerations. The regression MM-estimator has a breakdown point of $\frac{1}{2}$ and an efficiency and an asymptotic distribution like the regression M-estimator.

Variable Selection and Modeling

In practice, regression is used in many situations. The approach depends on the previous knowledge and subject-specific questions.

- Ideally, the relationships between the response y and the explanatory variables $x_1, ..., x_m$ is clear in advance. The functional relationships are often based on physical, chemical or other engineering or scientific theory. We call these **mechanistic models**. In this case we are most often interested in good estimates of the parameters, confidence and prediction intervals.
- In the other 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. The functional relationships are unknown and so we have to find **empirical models**.
- 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.

Model Selection Methods

We will proceed the approach with model selection methods which are based on statistical measures of goodness of fit. In the following we will define three statistical measures of model accuracy which also take the model complexity into consideration.

- Adjusted R-squared
- Akaike information criterion (AIC)
- Mallow's C_p statistic is

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.
  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
  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \qquad (130)$$
  with all possible explanatory variables. Use `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.

Collinearity

Collinearity means that an explanatory variable x_j can almost be represented by the others as a linear combination.

$$x_k \approx \gamma_0 + \sum_{j, j \neq k}^{m} \gamma_j x_j \qquad (131)$$

If this relation is exakt then the least-squares estimator is not uniquely defined and software implementations may have problems.

Measure for Collinearity

If the equation above is only approximately satisfied we can consider it as a linear multiple regression model. For each explanatory variable x_k define the coefficient of determination R_k^2 of the model. The **variance inflation factor** of variable x_k is then defined by

$$\text{VIF}_k = \frac{1}{1 - R_k^2} \qquad (132)$$

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, that the standard error of the estimate of the coefficient β_k is increased by a factor of

$$\sqrt{\text{VIF}_k} \qquad (133)$$

In an ideal situation with no collinearity we would have $R_k^2 = 0$ and therefore $\text{VIF}_k = 1$ and hence no increase in the standard error of the estimated coefficient $\hat{\beta}_k$.

Modelling Strategies

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.

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**. More complicated functions can be fitted.

- **Nonparametric regression**. Less assumptions on the models are made (smoothing).

- **Generalised linear model**. The response variable is allowed to have error distribution models other than the normal distribution (logistic, Poisson and gammaregression).

- **Survival analysis**. Censored data (reliability theory, survival and hazard function).

- **Time series analysis**. The errors can be correlated.

- **Principal component** and **partial least squares regression**. (Many) more variables than observations are available: large p, small n (chemometrics).

Part III
Design of Experiment

Design of Experiment

Many scientific studies are concerned with the understanding of causal relationships. 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 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.

Terminology and Concepts

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.

Basic idea of an experiment: Primary and secondary variables can be varied in a prescribed controlled manner.

Random error: Sum of all unknown influences.

Balanced design: Design with factor variables

Orthogonal design: Design with continuous variables.

Block designs: Record additional variables, i.e. batches, production lot or origin of product.

- Precision can be increased (if variable is significant).

- Interpretability of the results can be simplified.

Randomisation: Perform experiments in random order

- Chronological order as independent as possible from all recorded and unrecorded variables.

- Prevent influences due to the system of the experimental conditions.

- Avoid learning effects or aging of a device.

Replicates: Multiple measurements with same experimental conditions.

- Improved accuracy of the statements.

- Estimate of the variation of the random errors available.

- Existence of interactions in analysis of variance can be tested.

- Replicates are not allowed to be measured consecutively, otherwise they are not independent.

- Randomise replicates despite the increased experimental effort.

Experimental Design

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

Experimental design: : List of experimental conditions that determine how each factor is to be varied and at which levels.

Completely randomised design:

- One multi-level primary factor variable.

- One or more (but always the same number) measurements are made per level.

- The design is processed in random order to avoid systematic effects.

Complete block design:

- A secondary factor influences the target variable in addition to the primary factor.

- Every level of the primary factor (eg. treatment) is used at least once in each block.

Complete factorial design:

- Examine k factors with L levels each.

- Design with L^k measurements is needed.

- Often too many measurements.

2^k factorial designs:

- In screening experiments, many factors are investigated for their influence on the response variable.

- Examine only 2 levels (high and low) per factor.

2^k fractional factorial designs:

- To save even more resources.

- Only a balanced part of the complete 2^k design is realised.

Planning and Conducting a Study

1. **Scientific question:** Avoid situations without precise questions and hypotheses.

2. **Response and explanatory variables:** Find all influential factors, eg. with a cause-and-effect diagram.

3. **Preliminary experiments:** Experience with the topic and the measurement equipment is important.

4. **Planning:** It is worth going through the question, eg. with simulated data. The evaluation of the data in relation to the main questions is clear.

5. **Sample size:** The magnitude of the random fluctuations and the size of the expected effects has to be known. Often, the question must be downsized.

6. **Data:** 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.

7. **Data cleanup:** Checking the plausibility of the data and visualising it saves a lot of effort.

8. **Evaluation, interpretation:** The methodology for evaluating the data in relation to the main question is clear.

9. **Report, presentation:** For whom is the report mainly intended?

10. **Completion:** It is worthwhile to collect and discuss the experiences made with all participants.

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.

One Factor Analysis of Variance

With the one factor analysis of variance we want to check whether a factor (for example the hardwood concentration in paper) has a significant effect.

We check this by asking if there is a difference between one group and another group. The corresponding hypothesis are:

$$H_0 = \mu_1 = \mu_2 = \cdots = \mu_g \text{ (i.e. all groups obey the same model)}$$
$$H_1 = \mu_i \neq \mu_j \text{ j for at least one pair } i,j \text{ with } i \neq j. \tag{134}$$

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

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ with } i \in \{1, 2, ..., g\} \text{ and } j \in \{1, 2, ..., m\} \tag{135}$$

Notice that the model can also be written as

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \text{ with } i \in \{1, 2, ..., g\} \text{ and } j \in \{1, 2, ..., m\} \tag{136}$$

with corresponding hypothesis

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$
$$H_1 : \tau_i \neq 0 \text{ for at least one i.} \tag{137}$$

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.

We are looking for a test statistic that takes extreme values if the means ($\bar{y}.$ with the ".") of the g groups differ. If the variance of the group means S_b^2 compared to the variance of the observations within the groups S_w^2 is large, then we reject the null hypothesis. Therefore the test statistics for the analysis of variance test (ANOVA) is

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{S_b^2}{S_w^2}$$
$$= \frac{\frac{1}{g-1} \sum_{i=1}^{g} m \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2}{\frac{1}{n-g} \sum_{i=1}^{g} \sum_{j=1}^{m} \left(\bar{y}_{ij} - \bar{y}_{i.}\right)^2} \tag{138}$$

where $\bar{y}..$ 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
Tratments (groups)	SS_G	$df_G = g - 1$	$MS_G = \frac{SS_G}{df_G}$
Error (residuals)	SS_E	$df_E = n - g$	$MS_E = \frac{SS_E}{df_E}$
Total	SS_T	$df_T = n - 1$	

The sums of squares are defined by

$$SS_G = \sum_{i=1}^{g} m \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$
$$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{m} \left(y_{ij} - \bar{y}_{i.}\right)^2$$
$$SS_T = \sum_{i=1}^{g} \sum_{j=1}^{m} \left(y_{ij} - \bar{y}_{..}\right)^2$$

It is possible to show that

$$SS_T = SS_G + SS_E \tag{139}$$

The test statistic of the analysis of variance test can also be written as

$$F = \frac{MS_G}{MS_E} \tag{140}$$

Under the null hypothesis the F-statistic follows an F-distribution with $(df_G, df_E) = (g - 1, n - g)$ degrees of freedom. Critical values of the F-distribution can be found in tables.

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.

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

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \text{ with } i \in \{1, 2, ..., a\} \text{ and } j \in \{1, 2, ..., b\} \tag{141}$$

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

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{142}$$

is based on an overall value μ, an effect α_i which describes the factor A and an effect β_j which describes the factor B.

To be able to estimate the parameters uniquely we need to assume two constraints

$$\sum_{i=1}^{a} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0 \tag{143}$$

The parameters $\mu, \alpha_1, \alpha_2, ..., \alpha_a$ and $\beta_1, \beta_2, ..., \beta_b$ can be estimated from the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \text{ with } i \in \{1, 2, ..., a\} \text{ and } j \in \{1, 2, ..., b\} \tag{144}$$

For every level i and j we obtain

$$\bar{y}_{i.} = \mu + \alpha_i$$
$$\bar{y}_{.j} = \mu + \beta_j$$
$$\bar{y}_{..} = \mu \tag{145}$$

All together we obtain the estimates

$$\hat{\mu} = \bar{y}_{..}$$
$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \text{ for all } i \in \{1, 2, ..., a\}$$
$$\hat{\beta}_i = \bar{y}_{.j} - \bar{y}_{..} \text{ for all } j \in \{1, 2, ..., b\} \tag{146}$$

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	$df_A = a - 1$	MS_A	MS_A/MS_E
Factor B	SS_B	$df_B = b - 1$	MS_B	MS_B/MS_E
Error	SS_E	$df_E = n - a - b + 1$	MS_E	
Total	SS_T	$df_T = n - 1$		

The sums of squares are defined by

$$SS_A = \sum_{i=1}^{a} \sum_{j=1}^{b} \hat{\alpha}_i^2 = b \sum_{i=1}^{a} \hat{\alpha}_i^2$$
$$SS_B = \sum_{i=1}^{a} \sum_{j=1}^{b} \hat{\beta}_j^2 = b \sum_{j=1}^{b} \hat{\beta}_j^2$$
$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \hat{e}_{ij}^2$$
$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \left(y_{ij} - \bar{y}_{..}\right)^2 \tag{147}$$

We test the alternative hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$
$$H_1 : \alpha_i \neq 0 \text{ for at least one i.} \tag{148}$$

i.e. all effects of the factor A are zero, with the test statistic

$$F_A = \frac{MS_A}{MS_E} = \frac{\frac{SS_A}{df_A}}{\frac{SS_E}{df_E}} \tag{149}$$

Under the null hypothesis the statistic follows an F-distribution with

$$(df_A, df_E) = (a - 1, n - a - b + 1) = (a - 1, (a - 1)(b - 1)) \tag{150}$$

degrees of freedom. Critical values of the F-distribution can be found in tables. Analogously, we can test the factor B.

Andre Gut

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.

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{151}$$

with $i \in \{1, 2, ..., a\}, j \in \{1, 2, ..., b\}$ and $k \in \{1, 2, ..., c\}$. The observations on the same combination (i, j) are called **replicates**.

If the polygonal lines in the interaction plot are not almost parallel, then the additive model has to be modified too. The model with two factors and an **interaction** is

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{152}$$

with $i \in \{1, 2, ..., a\}, j \in \{1, 2, ..., b\}$ and $k \in \{1, 2, ..., c\}$. To be able to properly estimate the coefficients we have to add further constraints on the interaction term γ_{ij}

$$\sum_{i=1}^{a} \gamma_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{b} \gamma_{ij} = 0 \tag{153}$$

It is very important to know that the parameter estimate of the model with interaction, needs more than one observation per combination (i, j). The estimates for the parameters can be found analogously as in the chapter above, for the model with two factors. We obtain the estimates

$$\hat{\mu} = \bar{y}_{...}$$
$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} \text{ for all } i \in \{1, 2, ..., a\}$$
$$\hat{\beta}_i = \bar{y}_{.j.} - \bar{y}_{...} \text{ for all } j \in \{1, 2, ..., b\}$$
$$\hat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} \text{ for all } i \in \{1, 2, ..., a\}, \ j \in \{1, 2, ..., b\} \tag{154}$$

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	$df_A = a - 1$	MS_A	MS_A / MS_E
Factor B	SS_B	$df_B = b - 1$	MS_B	MS_B / MS_E
Interaction I	SS_I	$df_I = (a-1)(b-1)$	MS_I	MS_I / MS_E
Error	SS_E	$df_E = ab(c-1)$	MS_E	
Total	SS_T	$df_T = n - 1$		

The sums of squares are defined by

$$SS_A = bc \sum_{i=1}^{a} \hat{\alpha}_i^2$$
$$SS_B = ac \sum_{i=1}^{b} \hat{\beta}_j^2$$
$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \hat{e}_{ijk}^2 \tag{155}$$
$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(y_{ijk} - \bar{y}_{...} \right)^2$$

and

$$SS_I = SS_T - SS_A - SS_B - SS_E \tag{156}$$

We test the alternative hypothesis

$$H_0 : \gamma_{11} = \gamma_{12} = \gamma_{13} \cdots = \gamma_{ab} = 0$$
$$H_1 : \gamma_{ij} \neq 0 \text{ for at least one combinationi } (i, j). \tag{157}$$

i.e. all effects of the interaction I are zero, with the test statistic

$$F_I = \frac{MS_I}{MS_E} = \frac{\frac{SS_I}{df_I}}{\frac{SS_E}{df_E}} \tag{158}$$

Under the null hypothesis the statistic follows an F-distribution with

$$(df_I, df_E) = (ab(c-1), (a-1)(b-1)) \tag{159}$$

degrees of freedom.
Since a block cannot interact with the treatment, no interactions with block designs are allowed.

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 previously mentioned 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.

Screening Experiments

A production process is usually influenced by many factors. We are now concentrating on an essential factor of the product that should become the target of our investigations. 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.

2^k Factorial Designs

The number of experiments 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 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. 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	Example Data
	A	B	C		
1	-1	-1	-1	(1)	34
2	$+1$	-1	-1	a	26
3	-1	$+1$	-1	b	33
4	$+1$	$+1$	-1	ab	21
5	-1	-1	$+1$	c	24
6	$+1$	-1	$+1$	ac	23
7	-1	$+1$	$+1$	bc	19
8	$+1$	$+1$	$+1$	abc	18

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.

Example of main effect

$$\text{main effect for } A = \frac{1}{4}(26 + 21 + 23 + 18) - (34 + 33 + 24 + 19) \tag{160}$$
$$= 22.0 - 27.5 = -5.5$$

Example of interaction effect

mean if $A = -1$ and $C = -1$: 33.5   mean if $A = +1$ and $C = -1$: 23.5
mean if $A = -1$ and $C = +1$: 21.5   mean if $A = +1$ and $C = +1$: 20.5

$$\text{interaction effect } A : C = \frac{1}{2}((33.5 - 21.5) - (23.5 - 20.5))$$
$$= \frac{1}{2}(12 - 3) = 4.5 \tag{161}$$

The calculation of the interaction effect can be made simpler by expanding the matrix with additional columns by adding all possible products between the columns of the matrix.

Run	Factors			Interactions			
	A	B	C	$A:B$	$A:C$	$B:C$	$A:B:C$
1	-1	-1	-1	$+1$	$+1$	$+1$	-1
2	$+1$	-1	-1	-1	-1	$+1$	$+1$
3	-1	$+1$	-1	-1	$+1$	-1	$+1$
4	$+1$	$+1$	-1	$+1$	-1	-1	-1
5	-1	-1	$+1$	$+1$	-1	-1	$+1$
6	$+1$	-1	$+1$	-1	$+1$	-1	-1
7	-1	$+1$	$+1$	-1	-1	$+1$	-1
8	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

For instance, the column of the interaction $A : B$ is obtained by multiplying the columns of the factors A and B. The interaction effect can then alternatively be calculated as follows:

$$A : C = \frac{1}{2^{3-1}}(+34 - 26 + 33 - 21 - 24 + 23 - 19 + 18) = 4.5 \quad (162)$$

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. More than one measurement in each test situation is needed (Most often too expensive) in order to find out which effects are significantly different from zero. In R this analysis can then be done with the `aov` command.

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. The quantile of the ith ordered observation can be calculated according to the formula

$$q_i = \frac{1}{2} + \frac{1}{2}\frac{i - \frac{3}{8}}{e + \frac{1}{4}}, \quad (163)$$

where e denotes the number of estimated effects.
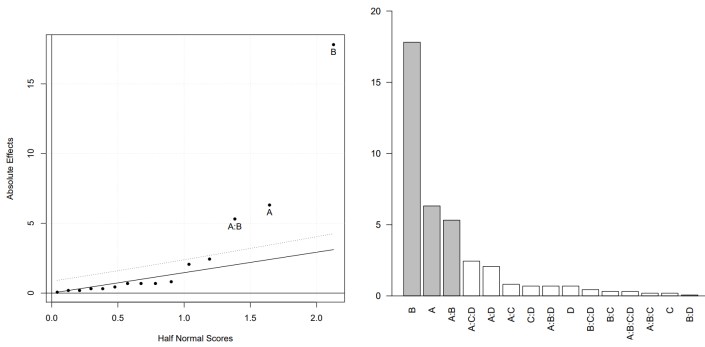
Half normal plot of the estimated main and interaction effects. The non-significant effects vary around the diagonal solid line and are below the dotted line. The significant effects are labeled with the name of the factor.

Pareto Chart. The significant main effects A and B and the interaction effect $A : B$ (gray) clearly dominate the non-significant effects (white).

2^k Fractional Factorial Designs

Few major effects and few interaction effects are significant. 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. Because the interaction effect of the highest order is most likely to be insignificant, the corresponding column in the following table is now assigned to a new main effect D. 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.

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. The entire set of aliases in a fractional factorial design is called alias structure of the design.

Effect		Alias
I	=	$ABCD$
A	=	BCD
B	=	ACD
C	=	ABD
D	=	ABC
AB	=	CD
AC	=	BD
AD	=	BC

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.

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.
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 suitabledesign and additional experiments.

This kind of procedure is a response surface exploration which is based on a sequence of experimental design.

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

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu + \alpha + \beta_j + \varepsilon_{ij} \quad (164)$$

If both explanatory variables x_1 and x_2 are continuous then the model is

$$y = f(x_1, x_2) + \varepsilon \quad (165)$$

where ε represents the noise or error observed in the response y. If we denote the expected response by $E(y) = f(x_1, x_2) = \eta$, then the surface represented by

$$\eta = f(x_1, x_2) + \varepsilon \quad (166)$$

is called a response surface.

We start with a so-called **first-order design**, a 2^k design with additional measurements in the center of the test conditions.

If the response is well modeled by a linear function of the k explanatory variables, the approximating function is the **first-order model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (167)$$

Example
A first-order test design and the results of the measurement are reported in the following table.

Run	Variables in original units		Variables in coded untis		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	63.6

The design is given in coded form and in original units.
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.
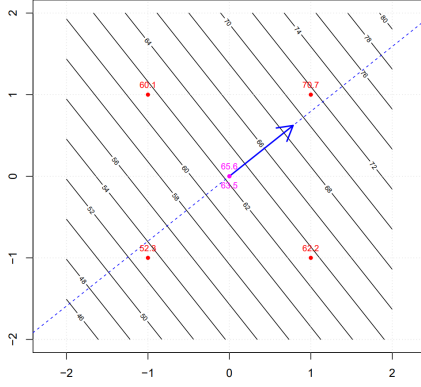The gradient of the response function

$$f(\texttt{Temp}, \texttt{Time}) = \beta_0 + \beta_1 \cdot \texttt{Temp} + \beta_2 \cdot \texttt{Time} \quad (168)$$

is

$$\text{grad}\,(f(\texttt{Temp}, \texttt{Time})) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 5.125 \\ 4.075 \end{pmatrix} \quad (169)$$

The estimated gradient points into the direction of the largest change in the coded `Temp-Time`-domain. Notice that this direction leads us to a maximal increase of the response.

There are two reasons why we extended the 2^2 design with two additional measurements in the centre of the test conditions.

- 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.

- 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. 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

$$H_0 : \text{ no curvature in the data,}$$
$$H_1 : \text{ curvature in the data.} \tag{170}$$

The test statistic for the curvature in a 2^k factorial design is

$$t_{\text{curve}} = \frac{\bar{y}_c - \bar{y}_f}{\sqrt{s_c^2 \left(\frac{1}{n_c} + \frac{1}{2^k} \right)}} \tag{171}$$

Where s_c^2 is the empirical variance. Under the null hypothesis the test statistic t_{curv} follows a t-distribution with $n_c - 1$ degrees of freedom.

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.

Second-Order Response Surface with Interactions

There are two reasons to use a second-order polynomial:

- If we are close to the optimum and the estimated response plane is almost horizontal, i.e. the estimated coefficients are almost zero.

- If there is curvature in the system, then a polynomial of higher degree must be used, such as the second-order model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon \tag{172}$$

In the case of two explanatory variables the model simplifies to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{173}$$

The 2^k design is not enough to estimate the parameters in the second-order model. To keep the effort within limits, so-called *rotatable central composite design* or *second-order central composite design* are used. 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. $(\pm 1, \pm\sqrt{2}, 0)$

Example (continuation)
The second-order test plan and the results of the measurements are reported in the following table.

Run	Variables in original units		Variables in coded untis		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 second-order linear model in original variables is

$$y = \beta_0 + \beta_1 T + \beta_2 t + \beta_{11} T^2 + \beta_{22} t^2 + \beta_{12} T \cdot t + \varepsilon \tag{174}$$

Now we fit this model with the interaction term to the data in original units of the rotatable central composite design and get the estimated second-order response function.
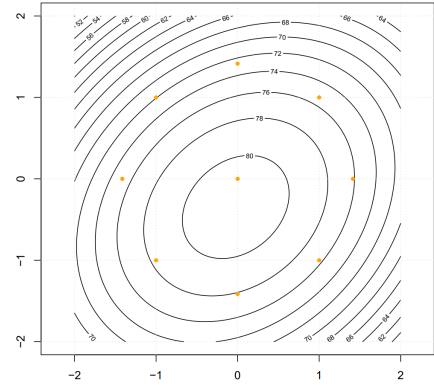
A second-order response surface can describe different types of surfaces depending on its parameters.

- 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.

The optimum of the response function can be calculated analytically. The stationary point can be found by setting all k partial derivatives equal to zero. In the case of two explanatory variables we obtain

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = \beta_1 + 2\beta_{11} x_1 + \beta_{12} x_2 = 0$$
$$\frac{\partial}{\partial x_2} f(x_1, x_2) = \beta_2 + \beta_{12} x_1 + 2\beta_{22} x_2 = 0 \tag{175}$$

and the estimated stationary point is

$$\hat{x}_1 = \frac{\hat{\beta}_{12}\hat{\beta}_2 - 2\hat{\beta}_{22}\hat{\beta}_1}{4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2} \quad \text{and} \quad \hat{x}_2 = \frac{\hat{\beta}_{12}\hat{\beta}_1 - 2\hat{\beta}_{11}\hat{\beta}_2}{4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2} \tag{176}$$

If the estimated determinant of the Hessian matrix is

$$4\hat{\beta}_{11}\hat{\beta}_{22} - \hat{\beta}_{12}^2 > 0 \tag{177}$$

then we have a maximum (or a minimum) otherwise a useless saddle point.

Summary Response Surface Methodology

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 design in the vicinity of the optimum found in Step 3. Estimate the second-order response surface. Find analytically an estimate of the stationary point.

All experiments in one plot
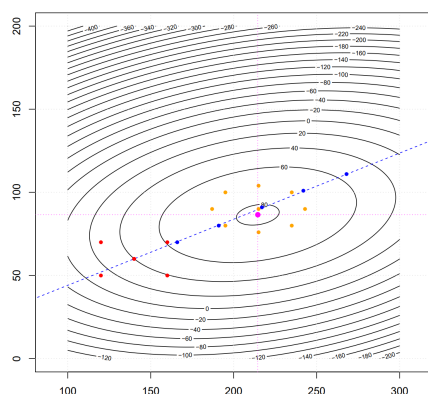
Contour plot in original units of the estimated second-order response surface with measurements of the three experiments: Experiment 1 is a 2^2 factorial design with additional doubled measurement in the centre (red), experiment 2 is an optimisation along the gradient (blue) and experiment 3 is a rotatable central composite design (orange). Optimal settings (magenta).

Part IV

R

Basics

Normal Distribution

Normal Distribution Function `pnorm`
Probability of X being less than 8.10 with `mean = 8.00` and `sd = 0.05`

$$P(X \leq 8.10) = \texttt{pnorm(q=8.10, mean=8.00, sd=0.05)} = 0.97725$$

Quantile Function `qnorm`
Value of x for which $P(X \leq x) = 0.97725$.

$$\texttt{qnorm(p=0.97725, mean=8.00, sd=0.05)} = 8.10$$

Binomial Distribution

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.
Density Function `dbinom`
What is the probability of throwing exactly two times head?

$$P(X = 2) = \texttt{dbinom(x=2, size=5, prob=0.53)} = 0.2916388$$

Distribution Function `pbinom`
What is the probability of throwing a maximum of three times head?

$$P(X \leq 3) = \texttt{pbinom(q=3, size=5, prob=0.53)} = 0.7727541$$

Geometric Distribution

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.
Density Function `dgeom`
What is the probability of throwing exactly two times?

$$P(X = 2) = \texttt{dgeom(x=2-1, prob=0.3)} = 0.21$$

Distribution Function `pgeom`
What is the probability of at most three throws?

$$P(X \leq 3) = \texttt{pgeom(q=3-1, prob=0.3)} = 0.657$$

Note that `x` and `q` in the functions `dgeom` and `pgeom` are the number of failed attempts until the first success.

Hypergeometric Distribution

Four out of ten lots win.
Density Function `dhyper`
What is the probability that out of five randomly selected lots there will be exactly two winning lots.

$$P(X = 2) = \texttt{dhyper(x=2, m=4, n=10-4, k=5)} = 0.4761905$$

Distribution Function `phyper`
What is the probability that out of five randomly selected lots there will be a maximum of three winning lots.

$$P(X \leq 3) = \texttt{phyper(q=3, m=4, n=10-4, k=5)} = 0.9761905$$

Utility function for finding sampling plans

Description
Find the sampling plan with smallest sample size (single sample only) such that specified Producer Risk Point (PRP) and Consumer Risk Point (CRP) are both met.
Arguments

PRP	The Producer Risk Point in the form of a two element numeric vector of the form `c(pd, pa)`. The first element, `pd`, specifies the quality level at which to evaluate the plan. The second element, `pa`, indicates the minimum probability of acceptance to be achieved by the plan.
CRP	The Consumer Risk Point in the form of a two element numeric vector of the form `c(pd, pa)`. The first element, `pd`, specifies the quality level at which to evaluate the plan. The second element, `pa`, indicates the maximum probability of acceptance to be achieved by the plan.
type	The distribution which the sampling plan is based on. Possible values are `binomial`, `hypergeom`, `poisson` and `normal`.
N	The size of the population from which samples are drawn. Only applicable for `type="hypergeom"`.
s.type	The type of the standard deviation. A value of `known` results in a sampling plan based on the population standard deviation, while a value of `unknown` results in the use of the sample standard deviation.

Usage
```
find.plan(PRP, CRP, type="binomial")
find.plan(PRP, CRP, type="hypergeom", N)
find.plan(PRP, CRP, type="normal", s.type="unknown")
```

Example
We want to find the smallest sample size n and the acceptence number c. The line
```
find.plan(PRP=c(p.alpha, 1-alpha), CRP=c(p.beta, beta),
type="hypergeom", N=100)
```
yields values for n and c.
`p.alpha` and `p.beta` are the probabilities of acceptance and rejection and `alpha` and `beta` are the producer and consumer risk.

The real producers risk can then be found with
`1-func.OC(p=p.alpha,N,n=n,c=c)`
and the real consumers risk with
`func.OC(p=p.beta,N,n=n,c=c)`.
Where `func.OC` is the function of the operating characteristic (eq. 53).

Linear Regression

Description
`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Usage
```
lm(formula, data, ...)
```

Arguments

formula	an object of class `"formula"`: a symbolic description of the model to be fitted. The details of model specification are given under *Details*.
data	an optional data frame, list or environment containing the variables in the model.

Details

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`. A terms specification of the form `first + second` indicates all the terms in `first` together with all the terms in `second` with duplicates removed. A specification of the form `first:second` indicates the set of terms obtained by taking the interactions of all terms in `first` with all terms in `second`. The specification `first*second` indicates the cross of `first` and `second`. This is the same as `first + second + first:second`.

A formula has an implied intercept term. To remove this use either `y ∼ x - 1` or `y ∼ 0 + x`.

Value

`lm` returns an object of class `lm` or for multiple responses of class `c("mlm", "lm")`.

The functions `summary` and `anova` are used to obtain and print a summary and analysis of variance table of the results.

Simple Linear Regression

We want to fit a linear model where the model has the form `Time ∼ Volume`. In R we do this with the following line
`mod <- lm(formula = Time ~ Volume, data = data)`

With the following line we get a summary of the fited model:

```
> 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
```

Confidence Interval on the Slope

```
# 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
```

Confidence Interval on the Response

We construct 95% confidence intervals on the response with the following line:
```
> 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
```

Where `data.new` is an array with equally spaced x-values covering at least the range of the Volume-data.

Prediction Interval on the Response

We construct 95% prediction intervals with the following lines:
```
> 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
```

Correlation

Multiple Linear Regression

We want to fit a multiple linear regression model where the model has the form `Vibration.log ∼ Distance.log + Charge.log`. In R we do this with the following line:
`mod <- lm(Vibration.log ~ Distance.log + Charge.log, data)`
The summary of that model is the following:
```
> 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
```

Factor Variables

If we want to use factor variables in a linear regression model we have do define them as such. The following lines tell R to use the position as a factor variable
```
# define Position as a factor
> data$Position.fac <- as.factor(data$Position)
```
The summary of the model is the following
```
> #   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
```

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.

F-test to compare two models

We come back to the explosion example and want to ask the question whether the factor `Position.fac` has a significant influence on the vibrations.

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.

F-test with `drop1` command

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.

```
> 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`. 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.

Model Selection

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")
```

According to the stepwise variable selection in both directions based on the AIC criterion the following model results.

```
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
```

Collinearity

We calculate the variance inflation factor for all explanatory variables used in an optimal model.

```
# 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

Analysis of Variance

Description
Fit an analysis of variance model by a call to lm for each stratum.

Usage
```
aov(formula, data, ...)
```

Arguments

formula	A formula specifying the model. (Analogus to the `lm` function).
data	A data frame in which the variables specified in the formula will be found.

Example

```
# 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
```

A low P-value means that the effect of, say the `Batch`-effect, is significant.

For the result to be valid, the following must be fulfilled: The errors must be independent and normally distributed with expected value 0 and constant variance σ^2. Furthermore, the two factors must overlap additively. With the interaction plot (`interaction.plot()`) it can be checked graphically whether the factor effects overlap additively. If they overlap additively, the lines in the interaction plot must be parallel except for random deviations.
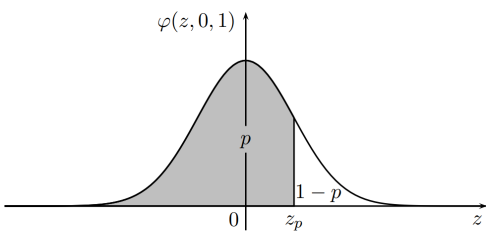
Donations

Thank you for your contribution:) I hope this summary was of help to you.

If you have any ideas for improvement, spot some typos or other mistakes - don't hesitate to contact me: andre.gut@hslu.ch

Appendix

## Standard Normal Distribution

$\varphi(z,0,1)$

p ... $1-p$ ... 0 ... z_p ... z

p-quantile 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

Students t-distribution

$f_n(t)$

p ... $1-p$ ... 0 ... $t_{n,p}$ ... t
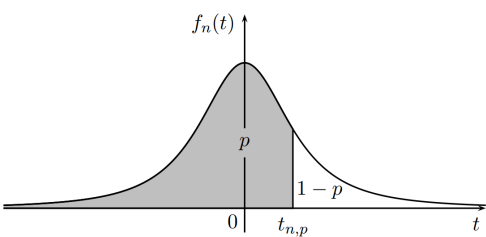
p-quantile $t_{n,p}$ for Student's t-distribution with n degrees of freedom. The density function is symmetric, therefore $t_{n,1-p} = -t_{n,p}$.

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	∞

Andre Gut

Factors for constructing control charts

n	$\bar{x}$ chart			R chart			s chart			n
	A_1	A_2	A_3	D_3	D_4	d_2	B_3	B_4	c_4	
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

Andre Gut