Part I

Statistical Process and Quality Control

Introduction

Quality and Process Control

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

Statistical Process Control (SPC)
- Statistical process control is, first and foremost, a way of thinking which happens to have some tools attached.

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

Control Charts

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

Hypothesis test

Question:
is $|\bar{x} - \mu_0|$ significant?
- μ_0 target value
- $\bar{x}$ Arithmetic mean of the measurements

Hypothesis
Two-sided statistical test to check the two alternative hypotheses:

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

Statistical test:

Test statistic (z-test, since σ is known)

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

Critical value i.e. q-quantile of the normal distribution

$$P(|z| \leq z_q) = q \tag{3}$$

with $q = 1 - \frac{\alpha}{2}$, where $\alpha = 0.0027 \rightarrow z_q \approx 3$

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

Reverse it and determine acceptance and rejection limits of the test!

Control limits

$$UCL = \mu_0 + z_q \frac{\sigma}{\sqrt{n}} \qquad LCL = \mu_0 - z_q \frac{\sigma}{\sqrt{n}} \tag{4}$$

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.

Problem: In general, process standard deviation is unknown.
- Monitoring the mean and the variation of a process.
- First, monitoring the variation, then (if variation under control) monitoring the mean.

Solution: Control charts! :)

The Control Chart

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 1: Data Set with Mean, Standard Deviation and Range

Mean values

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{5}$$

Standard deviations

$$s_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2} \tag{6}$$

Ranges

$$R_i = max\{x_{ij} | j \in \{1,...,n\}\} - min\{x_{ij} | j \in \{1,...,n\}\} \tag{7}$$

for all $i \in \{1,...,k\}$

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

R Chart

Centerline

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{8}$$

Control limits

$$UCL = D_4\bar{R}; \qquad LCL = D_3\bar{R} \tag{9}$$

$\bar{x}$ Chart based on R chart

Control limits

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}}; \qquad LCL = \mu - 3\frac{\sigma}{\sqrt{n}} \tag{10}$$

Problem: μ and σ are in general unknown and must be estimated from the process data.
Two-stage process:
- 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.
- Use $\bar{R}$ to estimate the process standard deviation.

Centerline
We know that for an independent sample $x_1, ..., x_n$ from a normal distribution with parameters μ and σ the mean

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{11}$$

satisfies

$$E(\bar{x}) = \mu \quad and \quad Var(\bar{x}) = \frac{\sigma^2}{n} \tag{12}$$

The mean is an unbiased estimator with the standard error

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \tag{13}$$

Assumption: R chart is under statistical control.
- The value $\bar{R}$ is a reliable estimate for the mean range.
- The value $\bar{R}$ is a reliable estimate for the process standard deviation

$$\hat{\sigma} = \frac{\bar{R}}{d_2} \tag{14}$$

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

$$\bar{\bar{x}} = \frac{1}{k^\star}\sum_{i=1}^{k^\star} \bar{x}_i \tag{15}$$

Control limits

$$\begin{aligned} UCL =& \bar{\bar{x}} + 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_2\bar{R} \\ UCL =& \bar{\bar{x}} - 3\frac{\bar{R}}{d_2}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_2\bar{R} \end{aligned} \tag{16}$$

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

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

$$\bar{s} = \frac{1}{k}\sum_{i=1}^{k} s_i \tag{17}$$

Control limits

$$UCL = B_4\bar{s}; \quad LCL = B_3\bar{s} \tag{18}$$

$\bar{x}$ Chart based on s chart

Using an s chart of a process that is under control, the process standard deviation can be estimated by

$$\hat{\sigma} = \frac{\bar{s}}{c_4} \tag{19}$$

Any samples excluded for construction of the s 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

$$\hat{\mu} = \bar{\bar{x}} = \frac{1}{k^\star}\sum_{i=1}^{k^\star} \bar{x}_i \tag{20}$$

Control limits

$$\begin{aligned} UCL =& \bar{\bar{x}} + 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} + A_3\bar{s} \\ UCL =& \bar{\bar{x}} - 3\frac{\bar{s}}{c_4}\frac{1}{\sqrt{n}} \approx \bar{\bar{x}} - A_3\bar{s} \end{aligned} \tag{21}$$

Individual Control Charts

Individual control charts have exactly one measurement per sample.
Problem: You cannot estimate variability from a single measurement.
Idea: Use variation of two adjacent measurements.

Moving ranges

$$MR_i = |x_{i+1} - x_i| \tag{22}$$

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

Arithmetic mean of the moving ranges

$$\overline{MR} = \frac{1}{n-1}\sum_{i=1}^{n-1} MR_i \tag{23}$$

Estimated process standard deviation

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{\overline{MR}}{1.128} \tag{24}$$

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.

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{25}$$

Control limits

$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128}; \quad LCL = \bar{x} - 3\frac{\overline{MR}}{1.128} \tag{26}$$

Control Charts for Attributes Data – p Chart

Number of defectives under number tested is a discrete random variable.

Given: Random sample of size n, of which D parts are defective We know: The number of defective D under n examined parts follows a binomial distribution with the unknown probability p of success.

Estimated probability

$$\hat{p} = \frac{D}{n} \tag{27}$$

Variance

$$Var(\hat{p}) = \frac{p(1-p)}{n} \tag{28}$$

Given:
- k random samples with $n_1, ..., n_k$ values.
- Each of these samples contains $d_1, ..., d_k$ defective products.

k relative frequencies

$$p_1 = \frac{d_1}{n_1}, \dots, p_k = \frac{d_k}{n_k} \tag{29}$$

Centerline
The centreline and the control limits of a p chart are again determined from a stable trial run with $k^\star$ valid samples.
Again $k^\star \leq k$ is the reduced number of samples.
Distinguish 2 cases:
1. The sample sizes $n_1, ..., n_k$ are all equal to n.
2. The sample sizes are not all equal.

Case 1

Centerline

$$\bar{p} = \frac{1}{k^\star}\sum_{i=1}^{k^\star} p_i \tag{30}$$

Control limits

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \quad LSL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{31}$$

Case 2

Centerline

$$\bar{p} = \frac{d_1 + \cdots + d_{k^\star}}{n_1 + \cdots + n_{k^\star}} \tag{32}$$

Control limits

$$UCL_i = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}; \quad LSL_i = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \tag{33}$$

The control limits now depend on the index i.

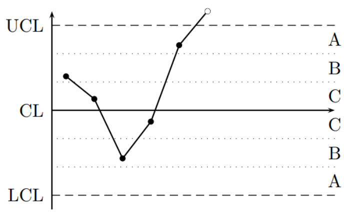
Statistical Properties of Control Charts

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.

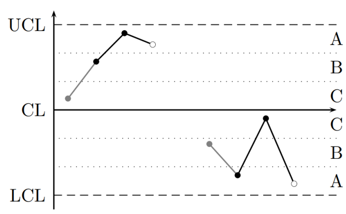
Interpretation of Control Charts

Western Electric Rules

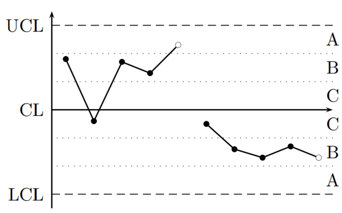
1. Any single data point falls outside the limit defined by UCL and LCL (beyond the 3σ-limit).
2. 2. Two out of three consecutive points fall beyond the limit defined by $\frac{2}{3}$ UCL and $\frac{2}{3}$ LCL on the same side of the centreline (beyond the 2σ-limit).
3. Four out of five consecutive points 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).



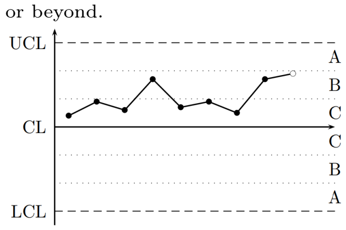
Rule 1: Any point beyond zone A.



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

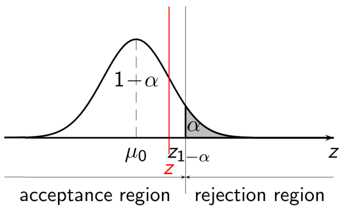
When monitoring a production process with a control chart, as with any statistical test, there are two wrong decisions possible.

$$\begin{aligned} H_0 :& \mu_0 = \mu \text{ i.e. process is not disturbed} \\ H_1 :& \mu_0 \neq \mu \text{ i.e. process is disturbed, } \mu_1 \text{ is true} \end{aligned} \tag{34}$$

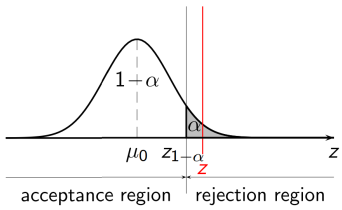
Denoted by:
- μ_0 the target value of the process
- μ the considered statistic, eg. $\mu = \bar{x}$ or $\mu = R$
- μ_1 the true value of the considered statistic.

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

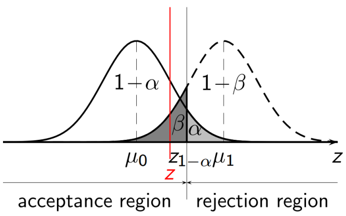


Let h_0 be true: Since $z < z_{1-\alpha}$ the null hypothesis is accepted. This is the right decision, which is made with probability $1-\alpha$.

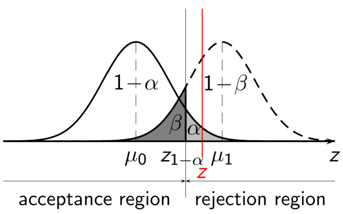


Let h_0 be true: Since $z \geq z_{1-\alpha}$ the null hypothesis is rejected. This is the wrong decision (type I error), which is made with probability α.

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 $z \leq z_1 - \alpha$ 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.

$$\text{power} = P(\text{reject } H_0 | H_1 \text{ is true}) = 1 - \beta \tag{35}$$

Power function
Probability to reject the null hypothesis H_0 if μ_1 is true.

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}, \tag{36}$$

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

$$g(\mu_1) = g(\delta\sigma + \mu_0) = \tilde{g}(\delta). \tag{37}$$

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

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

$$g(\mu_1) = \tilde{g}(0) = \alpha. \tag{38}$$

Disturbed Process
For a disturbed process we have

$$\tilde{g}(\delta) = \Phi(\delta\sqrt{n} - z_q, 0, 1) + \Phi(-\delta\sqrt{n} - z_q, 0, 1) \tag{39}$$

with Φ being

$$\Phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \tag{40}$$

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.

$$ARL(\delta) = \frac{1}{p(\mu)} = \frac{1}{g(\mu_1)} = \frac{1}{\tilde{g}(\delta)} \tag{41}$$

Undisturbed Process
If we have an undisturbed process with $\mu_1 = \mu_0$, i.e. with $\delta = 0$, then it follows from equation 38 that

$$ARL(0) = \frac{1}{\alpha}. \tag{42}$$

Disturbed Process
To determine the average run length of a disturbed process with $\mu = \mu_1$, i.e. $\sigma = \frac{\mu_1 - \mu_0}{\sigma}$ we use the power function $\tilde{g}(\delta)$ from equation 39.

Process Capability

The specification limit (SL) is defined by

$$SL = \frac{USL + LSL}{2} \tag{43}$$

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

$$C_p = \frac{USL - LSL}{6\sigma} \tag{44}$$

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
- Decision to interfere with the manufacturing process is based on the result of the current sample.
- No consideration of the development of the manufacturing process in the past (except with western electric rules).

Modern control charts
- have a memory.

Idea
Linear combination of mean values $\bar{x}_j$ of samples from the past

$$y_i = \alpha_i + \sum_{j=1}^{i} \beta_j \bar{x}_j, \tag{45}$$

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

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

$$\begin{aligned} C_i^+ &= max\{0, \bar{x}_i - (\mu_0 + K) + C_{i-1}^+\}, \\ C_i^- &= max\{0, (\mu_0 - K) - \bar{x}_i + C_{i-1}^+\}. \end{aligned} \tag{46}$$

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

$$K = \frac{\Delta}{2}. \tag{47}$$

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 intercal 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 $H = 5\hat{\sigma}$

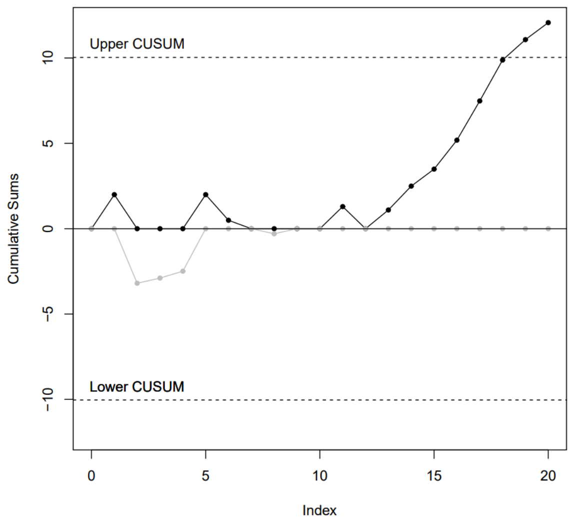


Figure 2: CUSUM-Chart

EWMA - Exponentially Weighted Moving Average

Idea
- Monitoring means.
- Wheights β_j decay exponentially.

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

$$\begin{aligned} \alpha_i &= (1-\lambda)^1\mu_0 \\ \beta_j &= \lambda(1-\lambda)^{i-j} \text{with } j \in \{1, 2, ..., i\}. \end{aligned} \tag{48}$$

Statistics

$$y_i = (1-\lambda)^i\mu_0 + \lambda\sum_{j=1}^{i}(1-\lambda)^{i-j}\bar{x}_j \tag{49}$$

Start: $y_0 = \mu_0$

The same with recursion

$$y_i = (1-\lambda)y_{i-1} + \lambda\bar{x}_j \tag{50}$$

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

$$Var(y_i) = \frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})\frac{\sigma^2}{n} \tag{51}$$

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

$$\begin{aligned} UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}} \\ LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{52}$$

The asymtotic control limits are:

$$\begin{aligned} UCL_i &= \mu_0 + 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \\ LCL_i &= \mu_0 - 3\sqrt{\frac{\lambda}{2-\lambda}}\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{53}$$

Estimate of process standard error

$$\hat{\sigma} = \frac{\bar{s}}{c_4} \tag{54}$$

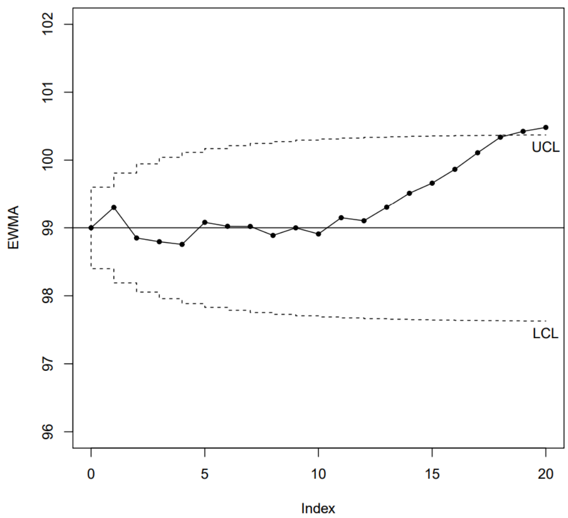


Figure 3: EWMA-Chart

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.

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:

$$p = \frac{m}{N} \tag{55}$$

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:

$$\begin{aligned} H_0 &: x \le c, \text{ i.e. the lot is not rejected.} \\ H_1 &: x > c, \text{ i.e. the lot is rejected.} \end{aligned} \tag{56}$$

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. So a sample of size n is drawn at once. The probability that among these n parts are at most c parts defective is given by

$$OC(p) = P(X \le c) = \sum_{k=0}^{c} P(X = k) = \sum_{k=0}^{c} \frac{\binom{pN}{k}\binom{N-pN}{n-k}}{\binom{N}{n}} \tag{57}$$

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:

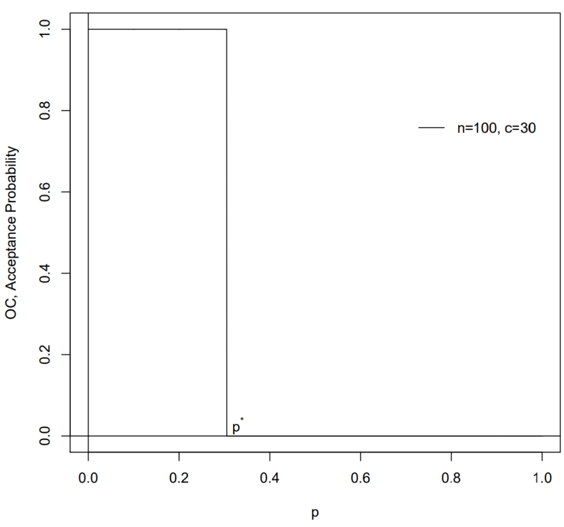


Figure 4: OC curve of an ideal acceptance sampling plan

OC Curves - Examples

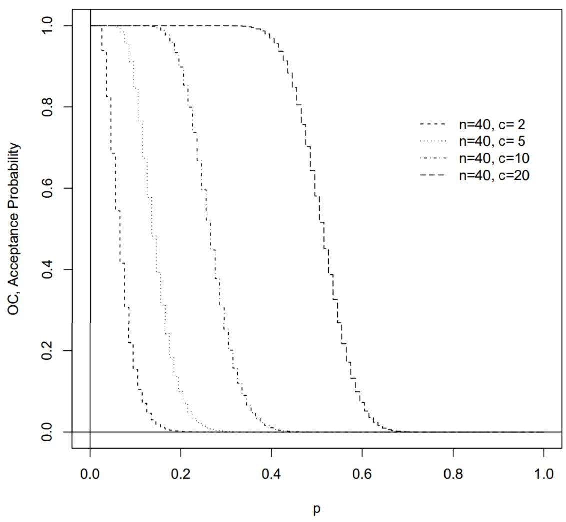


Figure 5: OC curve with fixed n and increasing c

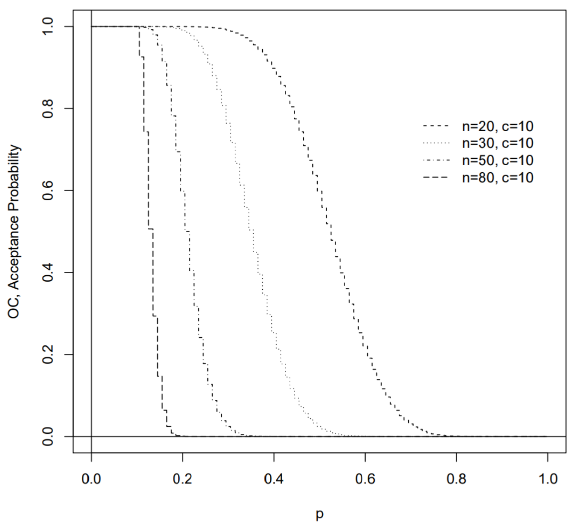


Figure 6: OC curve with increasing n and fixed c

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. $OC(p_\alpha) \le 1 - \alpha$, and
- the consumer risk is at most equal to β, i.e. $OC(p_\beta) \ge \beta$

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

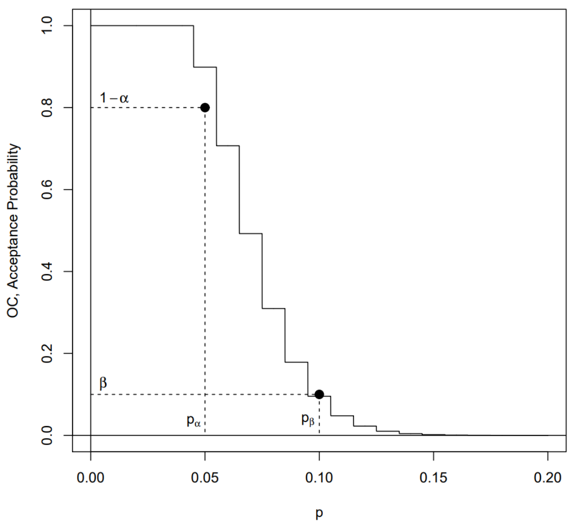


Figure 7: OC curve of a real acceptance sampling plan.

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

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{58}$$

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

$$E(\varepsilon) = 0 \tag{59}$$

and unknown variance

$$\mathrm{Var}(\varepsilon) = \sigma^2. \tag{60}$$

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.

$$E \approx \mathcal{N}(0, \sigma^2) \Rightarrow E(y) = \beta_0 + \beta_1 x \tag{61}$$

Estimation of the Parameters

The parameters β_0 and β_1 are estimated using the method of least square, i.e. by minimising:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2, \tag{62}$$

with

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ with } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{63}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ and } S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{64}$$

with

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{65}$$

Residual
The difference between the ith observed value y_i and its fitted value $\hat{y}_i$ is the residual

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i) \tag{66}$$

where the mean of the residual is given by

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0 \tag{67}$$

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

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \tag{68}$$

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

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \tag{69}$$

and

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{70}$$

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

$$\begin{aligned} H_0 :& \beta_1 = \beta_{1,0} \\ H_1 :& \beta_1 \neq \beta_{1,0} \end{aligned} \tag{71}$$

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.

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \text{ with } S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{72}$$

The test statistic is the following

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{se}(\hat{\beta}_1)} \tag{73}$$

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 $\beta_{1,0}$.

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 73 is accepted on the significance level α if

$$t_{\frac{\alpha}{2},n-2} \leq T \leq t_{-\frac{\alpha}{2},n-2}, \tag{74}$$

Confidence Interval
which leads us to the following confidence interval on the slope

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\beta_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\beta_1) \tag{75}$$

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\beta_1), \hat{\beta}_1 - t_{-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\beta_1)\right] \tag{76}$$

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

Confidence Interval of the Response

The estimated response of a regression model is as follows:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{77}$$

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

$$H_0 : \hat{y}_0 = \mu_0 \tag{78}$$

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

$$\mathrm{Var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \tag{79}$$

As usual σ^2 is in general unknown and has to be estimated from the data (equation 68). The test statistic about the response is

$$T = \frac{\hat{y}_0 - \mu_0}{\mathrm{se}(\hat{y}_0)}, \tag{80}$$

with the standard error

$$\mathrm{se}(\hat{y}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \tag{81}$$

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

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{y}_0) \leq \mu_0 \leq \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(\hat{y}_0) \tag{82}$$

Prediction Interval

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

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{83}$$

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

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{84}$$

Therefore

$$\begin{aligned}\mathrm{Var}(y_0 - \hat{y}_0) &= \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\end{aligned} \tag{85}$$

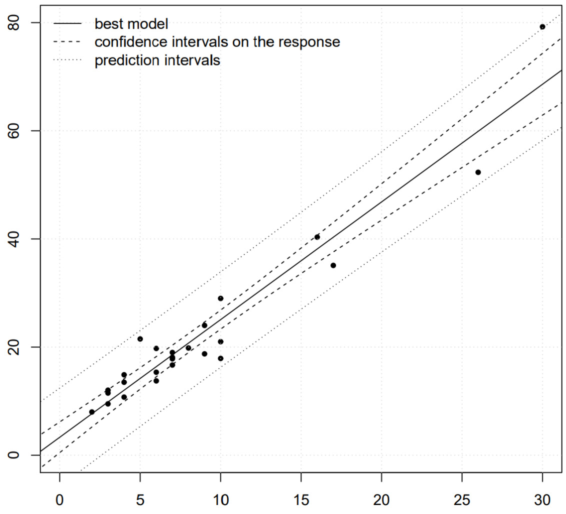
The prediction interval is wider than the confidence interval on the response, since the variance of the future observation has to be considered too. As usual σ^2 is in general unknown and has to be estimated from the data.

Prediction Interval
The prediction interval at the point x_0 is

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2} \cdot \mathrm{se}(y_0 - \hat{y}_0) \leq \hat{y}_0 \leq \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot \mathrm{se}(y_0 - \hat{y}_0) \tag{86}$$

with

$$\mathrm{se}(y_0 - \hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{87}$$



Residual Analysis

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

$$e_i = y_i - \hat{y}_i \quad \text{for all} \quad i \in 1, ..., n \tag{88}$$

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

$$\tilde{e}_i = \frac{e_i}{\sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}} \quad \text{for all} \quad i \in 1, ..., n \tag{89}$$

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

$$R^2 = \frac{SS_{\text{fit}}}{S_{yy}} \tag{90}$$

where

$$\begin{aligned}SS_{\text{fit}} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2 = \hat{\beta}_1^2 S_{xx} \\ SS_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y}_i)^2\end{aligned} \tag{91}$$

Therefore the coefficient of determination is

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \mathrm{Cor}(x, y)^2 \tag{92}$$

In multiple regression and simple linear regression with at least one explanatory variable the coefficient of determination is identical to the squared correlation between the response variable y and the fitted values $\hat{y}$

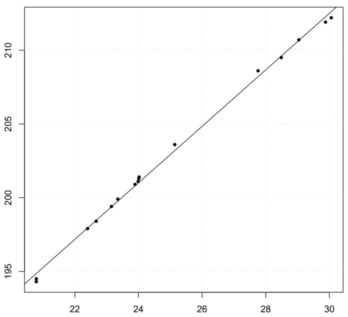
$$R^2 = \mathrm{Cor}(y, \hat{y})^2 = \frac{S_{y\hat{y}}^2}{S_{yy}S_{\hat{y}\hat{y}}} \tag{93}$$

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

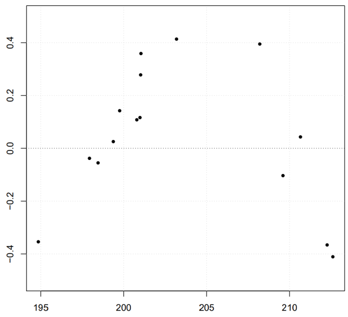
Diagnostic Tools

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.

Tukey-Anscombe Plot
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.



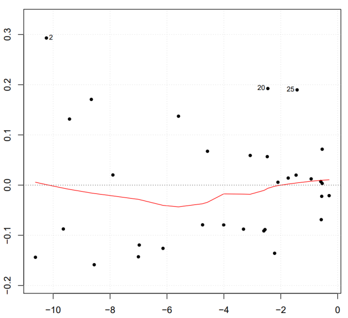
Best linear model



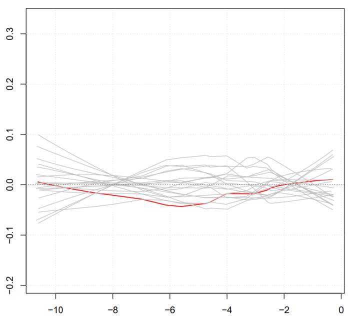
Tukey-Anscombe plot of the model. The points are not spread uniformly around the horizontal line indicating a bad model.

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 briev 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 based off the fitted model. - Calculate the smooth residuals of the simulated obervation and add it to the diagram.
- Repeat 19 times. (1 + 19 observations)

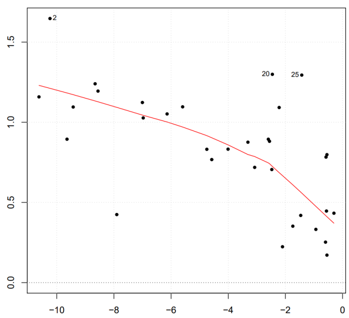


Tukey-Anscombe plot with smoothed residuals (red). The smoothed curve shows a clear curvature (which is bad).

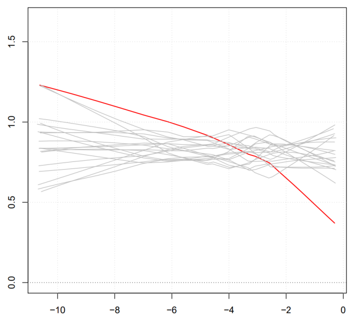


Tukey-Anscombe plot with 19 simulated smoothed residuals (gray). The curved smooth curve is not an extreme curve among all 20. We conclude that the model fits the data.

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.

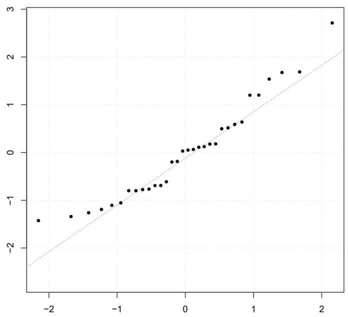


Scale-location plot with smoothed residuals (red). We observe a clear downward trend in residual scattering.

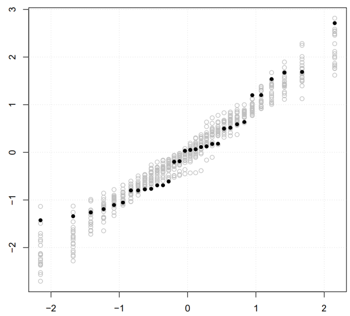


Scale-location plot with 19 simulated smoothed residuals (gray). The simulation confirms the extraordinary behavior.

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

Todo: Understand left and right-skewed plots

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.

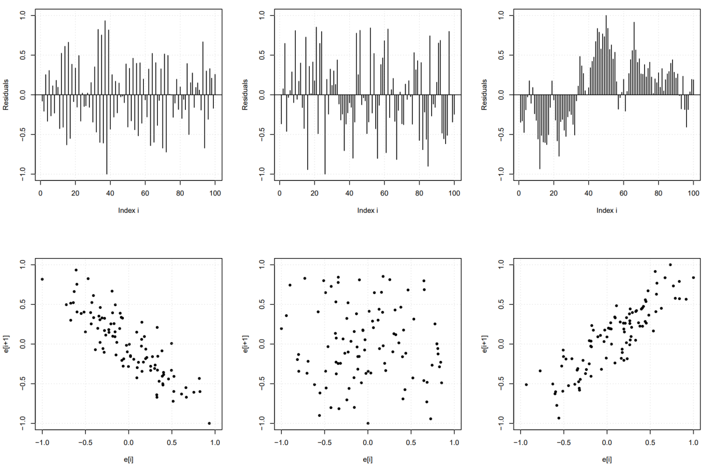
- **Logarithmic** for continuous positive variables:
$$z \mapsto \log(z) \tag{94}$$
- **Square-root** for continuous and discrete positive variables:
$$z \mapsto \sqrt{z} \tag{95}$$
- **Arcsine** for proportions:
$$z \mapsto \arcsin(\sqrt{z}) \tag{96}$$
- **Logit** for proportions:
$$z \mapsto \log\left(\frac{z+\varepsilon_1}{1+\varepsilon_2-z}\right) \tag{97}$$

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon, \tag{98}$$

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 teh multiple linear regression modle are the same as in the simple linear regression model. To calculate the according factors the matrix notation is used.

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{99}$$

Aufg. 8.5.1 is important!!

Diversity of Modelling Possibilities

Polynomial Regression
If we want to use a polynomial of degree two to describe a relationship, then the model has the following form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{100}$$

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

$$y = \beta_0 + \beta_1 \cdot \texttt{lin} + \beta_2 \cdot \texttt{quad} + \varepsilon \tag{101}$$

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

$$\begin{aligned} y =& \frac{1}{a + be^{-x}} && \text{gives} && \frac{1}{y} = a + be^{-x} \\ y =& \frac{ax}{b + x} && \text{gives} && \frac{1}{y} = \frac{1}{a} + \frac{a}{b}\frac{1}{x} \end{aligned} \tag{102}$$

- Taking logarithms of the nonlinear equations

$$\begin{aligned} y &= ax^b && \text{gives} && \ln(y) = \ln(a) + b\ln(x) \\ y &= ax^{bg(x)} && \text{gives} && \ln(y) = \ln(a) + bg(x) \end{aligned} \tag{103}$$

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

$$y = \beta_0 + \beta_1 x_b + \varepsilon \tag{104}$$

then

$$\begin{aligned} y =& \beta_0 + \varepsilon && \text{if } x_b = 0 \\ y =& \beta_0 + \beta_1 + \varepsilon && \text{if } x_b = 1 \end{aligned} \tag{105}$$

Factor Variables
Explanatory variables can be strongly (linearly) correlated. ⇒difficult to interpret the results of the regression analysis and should be avoided.
Binary Explanatory Variables
Variable is discrete with values 0 and 1.

Factor Variables

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 $\hat{y}$.

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.

$$d_i = \frac{(\hat{\mathbf{y}}_{(-1)} - \hat{\mathbf{y}})^t(\hat{\mathbf{y}}_{(-1)} - \hat{\mathbf{y}})}{p\hat{\sigma}^2} \tag{106}$$

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

$$d_i = \frac{\tilde{e}^2_{std,i}}{p}\frac{h_{ii}}{1 - h_{ii}} \tag{107}$$

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

$$h_{ii} = \mathbf{H}_{ii} = (\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)_{ii} \tag{108}$$

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

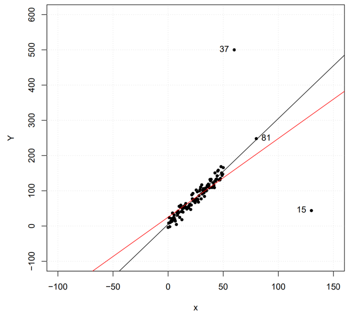
$$h_{ii} > 2\frac{p}{n} \tag{109}$$

Huber's classification:
Observations with

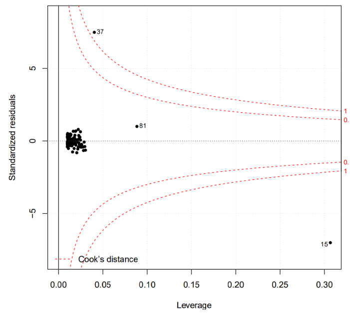
- $h_{ii} \leq 0.2$ are harmless.
- $0.2 < h_{ii} < 0.5$ are potentially dangerous.
- $0.5 < h_{ii}$ should be avoided.

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).

Part III

Design of Experiment