4. Inferential statistics

Objective of inferential statistics

- It is rarely possible to survey an entire population.
- Therefore, data is collected from a small subset of the population (sample) in order to draw conclusions about the entire population.
- Idea: The larger the sample, the more likely it is that the sample statistic will be close to the population value.
- Main assumption: Every element is independent from the other elements in the sample. The probability of drawing a certain element has no influence on which element to draw next.
- Each element in the population should have the same probability of being drawn into the sample.

The aim of inferential statistics is to draw conclusions about the underlying population from a sample. Since this conclusion goes beyond the observed, i.e. available data, it is subject to uncertainty.

Methods of inferential statistics

There are three methods or procedures for generalizing from a sample to a population:

- Point estimation.
- Interval estimation (confidence intervals).
- Statistical hypothesis tests.

A very important tool in inferential statistics is the standard error.

The standard error

Dispersion of the sample mean values (sample parameters):

- If several samples are taken from a population, the results will vary slightly. E.g., each sample will have a slightly different mean value.
- This dispersion of the sample means $\sigma_{\bar{x}}$ is also referred to as the **standard error (SE)** of the sample:

$$\sigma_{\bar{x}} = \frac{\sigma_{GG}}{\sqrt{n}}$$

- A small standard error means that the mean values of the different samples vary only slightly. The mean value of the sample is similar to the mean value of the population.
- The larger the standard error, the more the mean values of the different samples vary (and the less accurate the conclusion about the population value).

Relationship between standard error and sample size:

- The larger the sample, the smaller the dispersion of the sample mean will be.
- If the sample is always the same size as the population, then the mean value is identical every time. This means that the dispersion of the sample mean is zero.

4.1 Point estimation

Approach

- **Point estimation**: The unknown value of a parameter in the population is estimated by the value of a parameter in the random sample.
- **Example**: The average age of a representative sample of students at Emden/Leer University of Applied Sciences is $\overline{X} = 24$ years.
- In order to infer the population, in this case all students at Emden/Leer University of Applied Sciences, the mean value μ of the population is estimated by $\overline{X}$ of the sample:
- The estimated value for the unknown mean of the population μ is obtained by: $\hat{\mu} = \overline{X} = 24$.
- It is therefore assumed that the population of all students at Emden/Leer University of Applied Sciences has an average age of 24.

Notations for point estimator

μ	Expected value (theoretical mean value) of the population.
σ^2	Theoretical variance of the population.

Estimators are denoted with a hat (^) above the parameter:
$\hat{\mu}$ is an estimator for μ, and $\hat{\sigma}^2$ is an estimator for σ^2.

Common estimators:

- $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a common estimator for μ
- $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a common estimator for σ^2

Quality of point estimators

- As a rule, an estimated value will never match the true value of the population (unless the ppopulationis surveyed).
- It is not possible to make statements about the quality of the estimate.
- In the case of estimates that are unbiased, consistent, efficient and sufficient, it can be stated…
  - that the estimated value corresponds on average with the unknown parameter (unbiasedness),
  - that it will deviate less and less from the unknown parameter as the sample size increases (consistency),
  - that its variance is small (efficiency),
  - that all the information contained in the sample about the parameter to be estimated is used (sufficiency).

Bleymüller/Weißbach (2015), p. 118.

An inconsistent and biased estimator does not estimate what it is supposed to estimate. But a relatively inefficient or relatively insufficient estimator can still be good, for example if there is no relatively more efficient or relatively more sufficient estimator.

Exercise

Assume that the tips data set is a representative sample. Determine the following point estimates:

1. The average restaurant bill (total_bill).
2. The standard deviation of the tip (tip).
3. The average tip of males.

4.2 Interval estimation

The confidence interval

Limits of point estimators

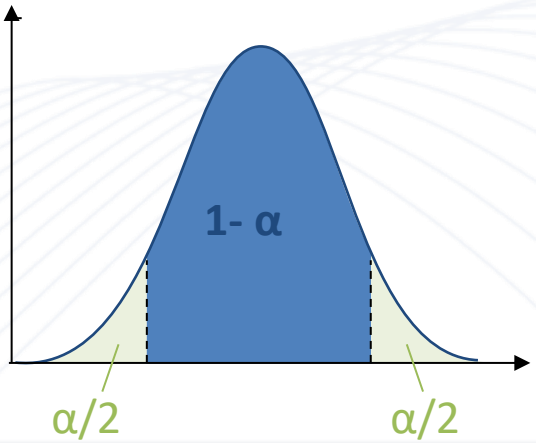
- The main disadvantage of point estimators is that it is not known how far the estimate deviates from the true value.
- However, the smaller the dispersion of the estimator (= standard error SE), the more accurate the estimator.

Basic idea of interval estimators

- Given a point estimator its and standard error, determine a range in which the true unknown parameter of the population is contained with an ex ante fixed probability.
- We trust that the unknown parameter is contained in this interval. This is why this interval is called the confidence interval.

The confidence interval

- The basic idea is to construct a range of values from which we think that the parameter of the population we are looking for falls within this interval.
- A **confidence interval** is constructed around the point estimate that covers the unknown parameter to be estimated with a probability (90% / 95% / 99%).
- The **confidence level 1-α** indicates the probability that the confidence interval contains the unknown parameter.
- Typically, confidence levels 1-α are high probabilities such as 90%, 95% or 99%.

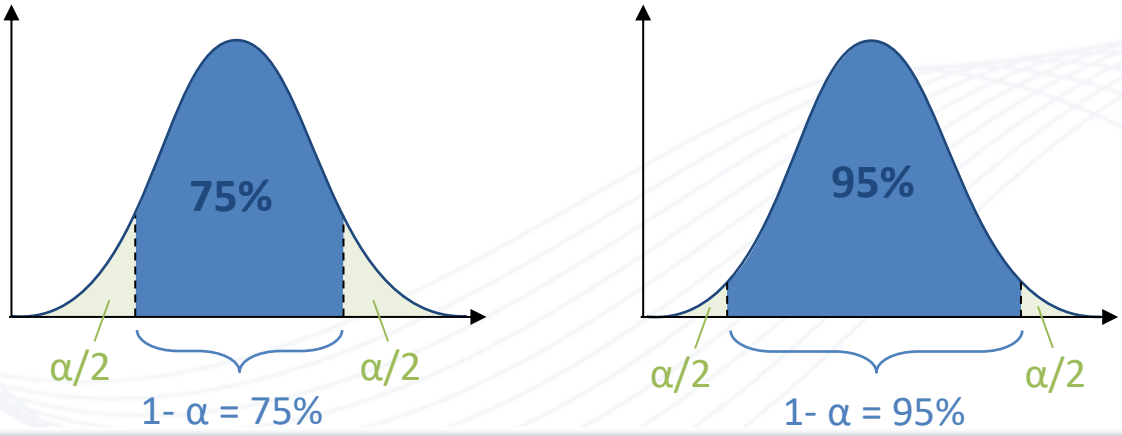


Interpretation of confidence intervals

Example: A point estimation for the average age of students from a sample is 24 years.

1. Assume for the **confidence level 1-α = 95%**, that the confidence interval is [16;32]. Interpretation: With 95% probability, the true unknown average age of the population is between 16 and 32 years.

2. Assume for the **confidence level 1-α = 75%**, that the confidence interval is [20;28]. Interpretation: With 75% probability, the true unknown average age of the population is between 20 and 28 years.



Accuracy, standard error, and confidence level

- The higher the confidence level 1-α (certainty), the larger (less precise) the confidence interval.
- The larger the standard error (dispersion of the estimate), the larger (less precise) the confidence interval.

In general:

A narrow confidence interval with a high confidence level indicates a good point estimate.

Exercise

Assume that the tip data set is a representative sample. Determine 90% and 95% confidence intervals for the following parameters:

1. The average restaurant bill.
2. The average tip of men.

R commands: t.test(x, conf.level=.95)

4.3 Hypothesis testing

Structure of a statistical hypothesis test

1. Formulate null hypothesis H_0 (and alternative hypothesis H_1).

The null hypothesis (H_0) typically reflects the current state of knowledge (status quo). The alternative hypothesis (H_1 or H_A) reflects the research question to be examined.

2. Defining the test level α.

The test level α represents the maximum probability up to which we want to be wrong if we reject H_0.

3. Derive the p-value p.

We carry out the hypothesis test under the assumption that the null hypothesis is true. To do this, we calculate the probability p, which measures the probability to observe the sample data under the assumption that H_0 is true. This probability p is called the p-value.

4. Making a test decision.

If the test results suggest that the data do not provide enough evidence for the alternative hypothesis (i.e. if $p > \alpha$), we do not reject the null hypothesis. Otherwise ($p \leq \alpha$), we reject the null hypothesis in favor of the alternative hypothesis.

More about H_0 and H_a

- The **null hypothesis H_0** always claims that there is **no effect/difference/ relationship**. Example: “There is nothing”, promotion and gender are independent.
- The **alternative hypothesis H_A** (or H_1) typically claims exactly the opposite of H_0. It claimes that there is an **effect/difference/relationship**. Example: “There is something”, promotion and gender are dependent.
- When testing hypotheses, an attempt is always made to reject the H_0.
- Only if the H_0 is rejected can a statement be made about the alternative H_A.
- A hypothesis (both H_0 and H_A) can never be *correct* or *proven*.
- Hypotheses are formulated for the population, but are rejected or not rejected on the basis of the present sample.
- Hypotheses should always be formulated precisely before data collection.

More about the test level α

- Whether a H_0 is rejected or not is decided on the basis of the α-level.
- The α-level is the **probability specified by the person carrying out the test** that the **H_0 is incorrectly rejected** (probability of error).
- Alternatively, the α-level can also be understood as the probability at which one is no longer ready to accept an observed effect in a sample as random.
- Typically, an α-level of **10%**, **5%**, **1%** or **0.1%** is used.

Types of errors in hypothesis testing

		Test decision	
		H_0	H_A
Truth	H_0	Right decision	Type I error, α-error, controlled by α
	H_A	Type II error, β-error	Right decision

Type I and type II error

- If H_0 is rejected, the **probability of a type I error** is limited by the test level α and is therefore **known**.
- If H_0 is not rejected, the **probability of a type II error** is **unknown**.
- The following applies: The higher the probability of a type I error, the lower the probability of a type II error and vice versa.
- But: Both probabilities do not add up to 1, i.e.:

probability of a type I error ≠ 1 - probability of a type II error.

→ Try to formulate H_0 in such a way that the alternative H_A reflects the research hypothesis. If H_0 is rejected, then the test result supports the research hypothesis with a known type I decision error.

Undirected (two-sided) and directed (one-sided) hypotheses

Undirected (two-sided) hypotheses

- Claim a difference/effect/relationship between groups/variables without making statements about the direction.
- Example research question: There is an influence of gender on the tip amount.
- The corresponding null hypothesis is: H_0: There is no effect of gender on the tip amount ($\mu_{TipFemales} = \mu_{TipMales}$).
- This null hypothesis is rejected if the average tips of males and females are significantly different, without taking into account whether men or women tip more on average.

Undirected (two-sided) and directed (one-sided) hypotheses

Directed (one-sided) hypotheses

- Contain statements about the direction of a difference/effect/relationship between groups/variables.
- Example research question: On average, men tip more than women.
- The associated null hypothesis is: H_0: On average, men tip at most as much as women ($\mu_{TipFemales} \geq \mu_{TtpMales}$).
- This null hypothesis is only rejected if the average tip for men is significantly higher than the average tip for women.

Underlying assumption for inference based on hypothesis testing

For estimation and testing to work, the observations in the sample should be independent and identically distributed: i.i.d.

This can be done by drawing a random sample from a population.

Alternatively, the data can also come from a randomized experiment.

5. Data preprocessing and data cleansing

5.1 Objectives and instruments

Overview

The main objectives of data pre-processing is to gain an initial overview of the data and to prepare the data for further analysis. To do so, the methods of descriptive statistics are used.

Methods descriptive statistics

- Graphical representations: Bar chart, histogram, boxplot, scatterplot, mosaicplot.
- Measures of position: mean, median, percentiles.
- Measures of dispersion: standard deviation, interquartile range, range.

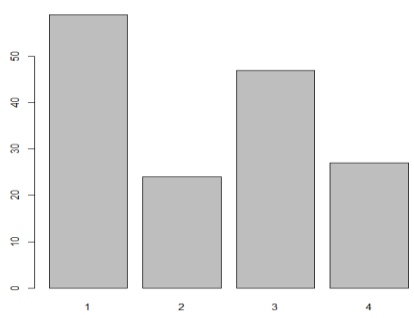
Application of descriptive statistics

- General: Description of the sample.
- Checking the data for:
  - Errors
  - Plausibility
  - Outliers
  - Missing values

Hussy; Schreier; Echterhoff (2013) Forschungsmethoden in Psychologie und Sozialwissenschaften fur Bachelor, pp. 169.

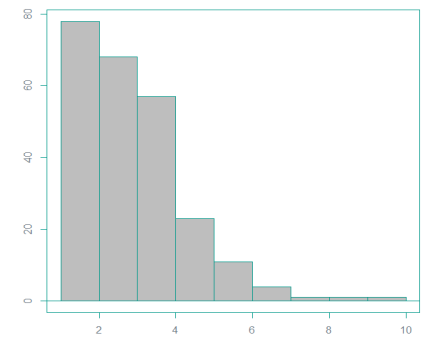
Graphics for the representation of the shape of the distribution

Bar chart



- Represents the absolute or relative frequency distribution of a categorical variable.
- Also suitable for displaying the shape of the distribution of a discrete at least ordinal variable.
- The height of the bars is proportional to the frequency.

Histogram



- Represents the distribution form of a continuous variable.
- The continuous variable is split into groups of equal size. The size of the areas is proportional to the frequency.
- The mean value is less meaningful for skewed distributions.

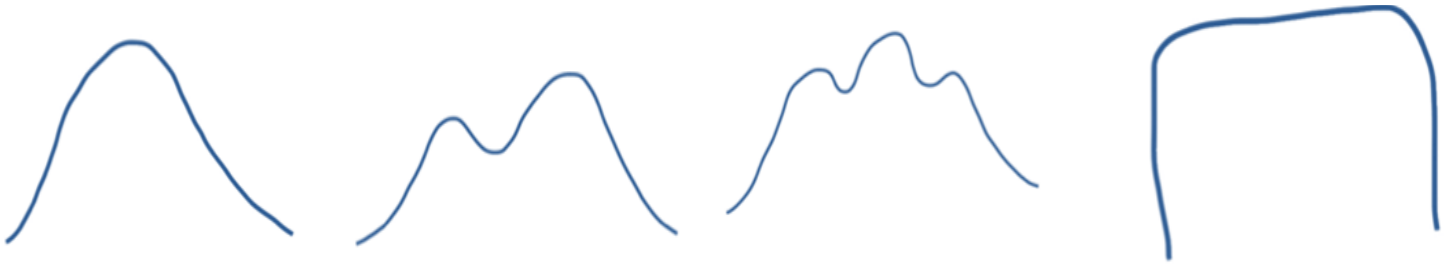
Lübke/Vogt (2014), pp. 19.

Common shapes of distributions

Histograms contain information about the…

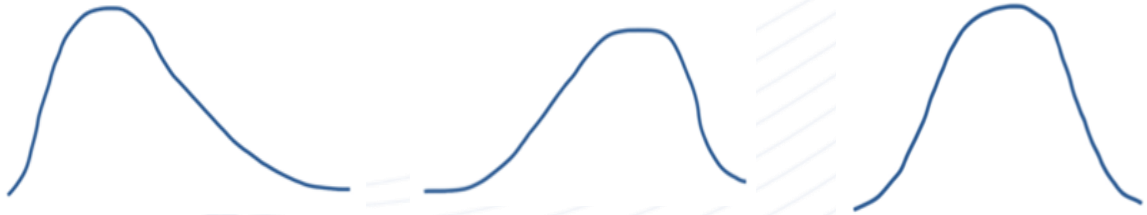
…modality of the distribution, and the

unimodal bimodal multimodal uniformly distributed



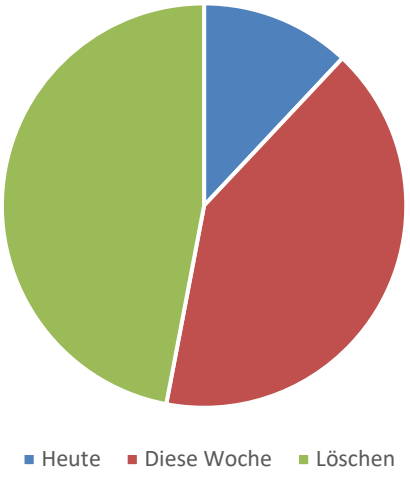
…skewness or symmetry of the distribution.

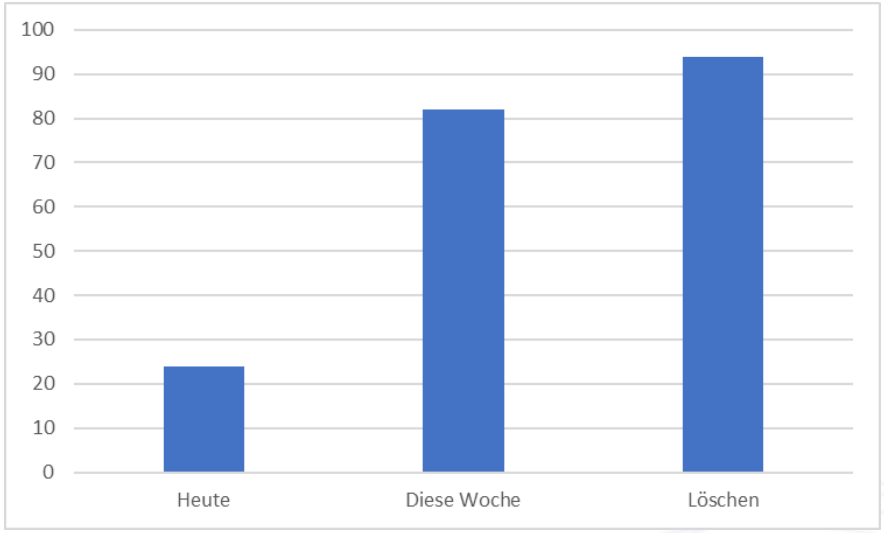
right skewed left skewed symmetrical



Bar charts versus pie charts

A bar chart displays much more information than a pie chart for the same data.





- Size ratios are easier to recognize.
- Axis scaling allows more precise representation of absolute or relative frequencies.

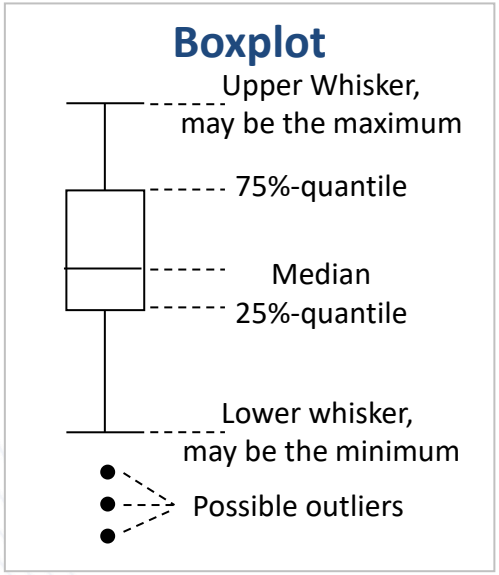
Avoid the pie! If necessary, use pies for data with few categories that can be easily distinguished visually.

Graphics for identifying potential outliers

Definition:

A single value that differs significantly from all other values is called an **outlier**.

According to (2009), p. 41.



- Represents the most important percentiles of a numerical variable.
- The box contains the middle 50% of the data, the height of the box corresponds to the interquartile range.
- The length of the whiskers have a maximum of 1.5 x the height of the box.
- Individual points outside the whiskers represent possible outliers.
- Outliers distort the mean value.

Bleymüller/Weißbach (2015), p. 29.

Measures of position

Arithmetic mean	• Sum of the individual values divided by the total number of individual values. • Suitable for numerical variables. • Can be distorted by outliers and skewed distributions.
Median (50% quantile)	• Lies exactly in the middle: One half of the individual values is at most this large, the other half of the individual values is at least this large. • Requires at least ordinal scaled variables. • Robust against outliers and skewed distributions.
p% quantile	• p% of the individual values is at most this large, (1-p)% of the individual values is at least this large. • Important quantiles: 0% (minimum), 25% (lower quartile), 75% (upper quartile), 100% (maximum).

Bleymüller/Weißbach (2015), p. 29.

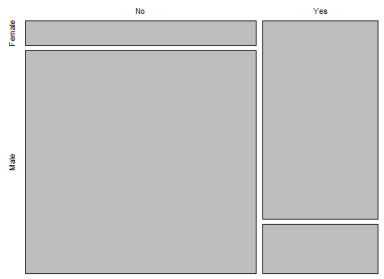
Measures of dispersion

Standard deviation	• Square root of the mean sum of squared deviations from the mean value. • Suitable for numerical variables. • Can be distorted by outliers and skewed distributions.
Inter quartile range	• Upper quartile minus lower quartile. Contains the middle 50% of the data. • Suitable for numerical variables. • Robust against outliers and skewed distributions.
Range	• Maximum minus minimum. • Suitable for numerical variables. • Can be extremely distorted by outliers and skewed distributions.

Lübke/Vogt (2014), p. 34.

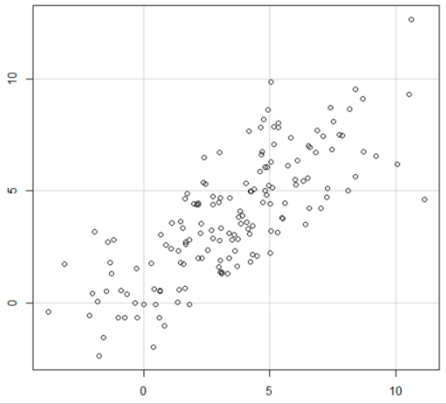
Graphics to illustrate the relationships between two variables

Mosaicplot



- Represents the absolute or relative joint frequency distribution of two categorical variables.
- The size of the area is proportional to the frequency of the combination of variables.

Scatterplot



- Represents the relationship between two numerical variables as a point cloud in the coordinate system.
- Each individual point represents a single observation in the data set.
- The position of each individual point is determined by the values of the two numerical variables for this observation.

Friendly (1994); Lübke/Vogt (2014), pp. 80 & 83.

Measures of correlation

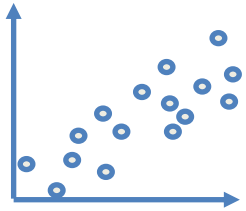
Correlation coefficient r	• The Pearson correlation coefficient examines the linear relationship between two numerical variables. • The Spearman correlation coefficient examines the monotonic relationship between two variables that are at least ordinally scaled. • The value range is between -1 and +1. • The Spearman correlation is robust against outliers, whereas the Pearson correlation is not.
Contingency coefficient C	• The contingency coefficient examines the relationship between two nominally scaled variables. • The value range is between 0 and +1.

Lübke/Vogt (2014), p. 79.

Correlation analysis

A scatterplot and the correlation coefficient r can be used to analyze the relationship between two at least ordinally scaled variables:

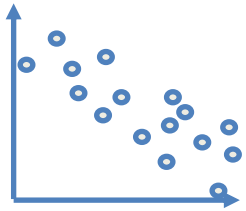
Positive correlation



Upward oriented point cloud

$0 < r \leq 1$

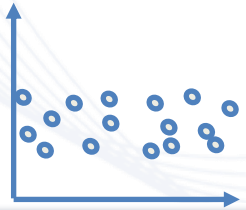
Negative correlation



Downward oriented point cloud

$-1 \leq r < 0$

No correlation



Point cloud is without orientation

$r \approx 0$

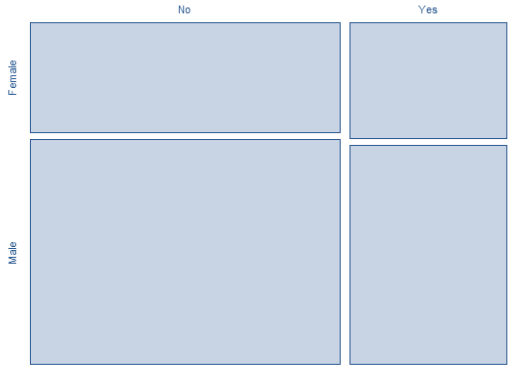
Lübke/Vogt (2014), p. 80.

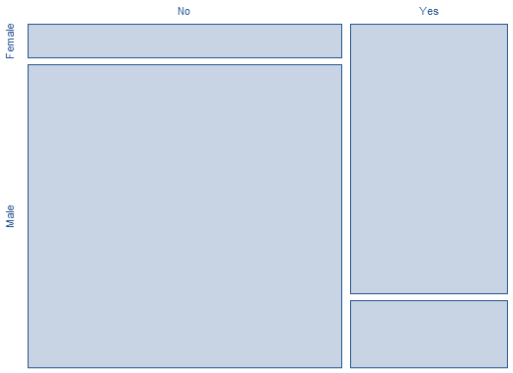
Contingency analysis

The mosaic plot and the contingency coefficient C can be used to examine the relationship between two nominal (categorical) variables:

No correlation: The ratios of the areas in the rows and columns are the same and C is close to 0.

The mosaic plot shows a “scandinavian cross”.

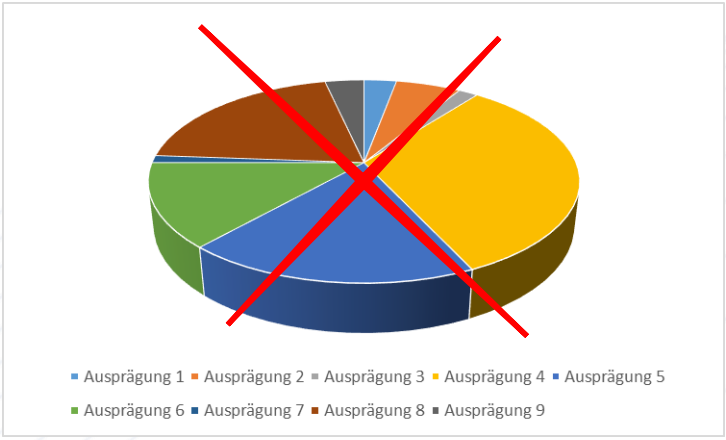




Correlation: The ratios of the areas in the rows and columns are different, and C is (significantly) greater than 0.

Tips for graphics

- Display lots of numbers, otherwise you don't need a graphic.
- Avoid distraction from the main message.
- Encourage visual comparison.
- Different colors only if it supports comparison.
- Avoid 3d.
- Pay attention to axis scaling.
- Visualize first, then calculate summary statistics.



Possible objectives for descriptive data analysis

- What is the most frequent values?
- Which values are rare or not present?
- Are the values plausible, or do they contain errors?
- Are there outliers?
- Is the distribution symmetrical or skewed?
- Are the answers widely scattered or are they close together?
- Are there missing values?

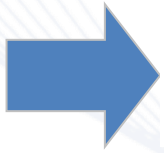
The result is reliable information on the quality and usability of the data.

"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

Picture: https://lovestats.wordpress.com/dman/, accessed 22.07.2020.

Plausibility assessment

- An **error** occurs when an answer does not match the question, i.e. is obviously wrong.
- Example: “On a scale of 1 to 10, indicate your agreement with the current government policy” - answer: 11.
- A **contradiction** occurs when answers to different questions do not match in terms of content.
- Example: “In which year did you graduate from high school?” - Answer: 1993; “When were you born?” - Answer: January 12, 1991.
- There is an **anomaly** if, for example, answers are very extreme or if answers are implausible, but not necessarily so.
- Example: “In what year were you born?” - Answer: 1921.



Application of the methods of descriptive statistics, followed by detailed analysis of the individual data records. Individual data that is not plausible is deleted.

Missing values

Definition:
If an observation has no value for a variable, it is called **missing value**.

Missing values are categorized into three categories according to their causes:

- **Missing Completely At Random (MCAR)**: Data are MCAR when the probability of missing data on a variable X is unrelated to other measured variables and to the values of X itself. In other words, missingness is completely unsystematic.
- **Example**: A participant in a panel study moves to another city for unknown reasons, i.e. it is unrelated to other variables in the data set.

Baraldi/Enders (2010), pp. 6-8.

Missing values (continued)

- **Missing At Random (MAR)**: Data are MAR if missingness is related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable (i.e., the hypothetical values that would have resulted had the data been complete).
- **Example**: A participant in a panel study on purchasing behavior does not provide complete purchase data due to alcohol consumption after previous alcohol purchases.
- **Missing Not At Random (MNAR)**: Data are MNAR if the probability of missing data is systematically related to the hypothetical values that are missing. In other words, the MNAR mechanism describes data that are missing based on the would-be values of the missing scores.
- **Example**: Consider a reading test where poor readers fail to respond to certain test items because they do not understand the accompanying vignette. The probability of a missing reading score is directly related to reading ability.

Baraldi/Enders (2010), pp. 6-8.

Handling missing values: Deletion

Listwise deletion (complete-case analysis, casewise deletion):

- Cases with missing values are discarded, so the analyses are restricted to cases that have complete data.
- Major advantage: It produces a complete data set.
- Disadvantages:
  - Deleting incomplete records can dramatically reduce the total sample size.
  - Listwise deletion assumes that the data are MCAR. When this assumption is violated – as it often is in real research settings – the analyses will produce biased estimates.

Pairwise deletion (available-case analysis):

- Incomplete cases are removed on an analysis-by-analysis basis, such that any given case may contribute to some analyses but not to others.
- Advantage: It minimizes the number of cases discarded in any given analysis.
- Disadvantages: Assumes that the data are MCAR. Otherwise, pairwise deletion can produce biased estimates.

Baraldi/Enders (2010), pp. 10-11.

Handling missing values: Single imputation and mean imputation

Single imputation: The researcher imputes (i.e., “fills in”) the missing data with seemingly suitable replacement values. There exist three common approaches: mean imputation, regression imputation, and stochastic regression imputation.

Mean imputation:

- Replaces missing values with the arithmetic mean of the available data, either from other variables in the data set or from other values of the same variable.
- Advantage: Easy to calculate.
- Disadvantages:
  - Reduces correlation with other variables.
  - Produces biased estimates for MAR data.

Baraldi/Enders (2010), pp. 11-14.

Handling missing values: Regression imputation

Regression imputation:

- Replaces missing values with predicted scores from a regression equation.
- Complete cases are used to estimate a regression equation where the incomplete variable serves as the outcome and the complete variables are the predictor.
- Advantage: Produces unbiased estimates of the mean when the data are MCAR or MAR.
- Disadvantage: Increases correlation with other variables, i.e. biases measures of association and variability.

Baraldi/Enders (2010), pp. 11-14.

Handling missing values: Stochastic regression imputation

Stochastic regression imputation:

- Replaces missing values with predicted scores from a regression equation and adds a random error term to each predicted score.
- The error term is a random number generated from a normal distribution with a mean of zero and a variance equal to the residual variance from the preceding regression analysis.
- Advantage: Produces parameter estimates that are unbiased under both the MCAR and MAR assumptions.
- Disadvantage: Provides no mechanism for adjusting the standard errors to compensate for the fact that the imputed values are just guesses about the true score values. Consequently, the standard errors are inappropriately small, and significance tests will have excessive Type I error rates.

Baraldi/Enders (2010), pp. 11-14.

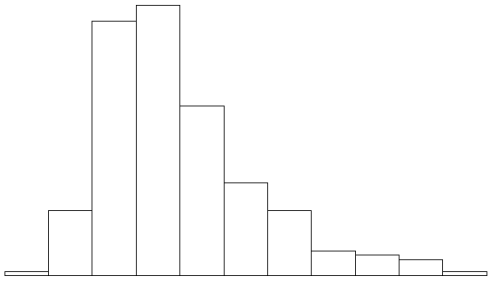
Exercise

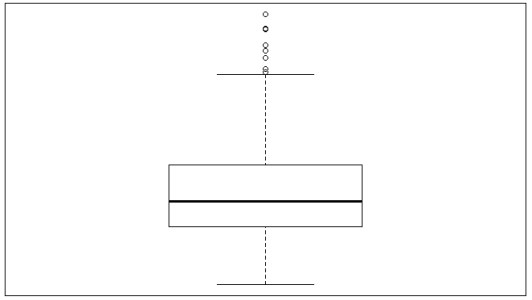
Practise data preprocessing and data cleansing.

5.2 Variable transformations

Repetition: Symmetry and outliers

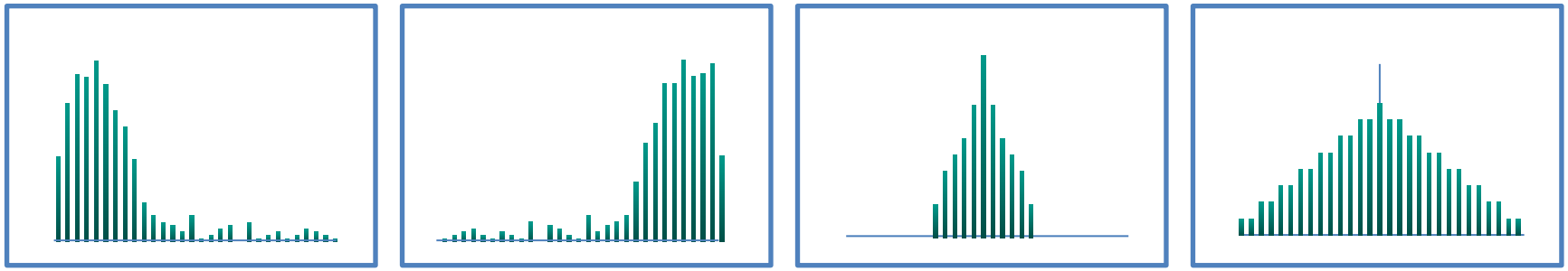
- A symmetrical distribution, in particular a normal distribution without outliers, is an important precondition for the applicability of numerous statistical methods.
- The normal distribution is a distribution in which most values are grouped around the mean and a few values are further away from the mean.
- A histogram can be used to graphically represent the distribution shape.
- Outliers can be identified using a box plot.





Deviations from the normal distribution

- There are often deviations from the normal distribution, for example:



right skewed or left skewed

thin tailed (platykurtic) or fat tailed (leptokurtic)

- Many statistical methods lead to “good” results even with deviating forms of distribution.
- Alternatively, a data transformation can be useful, but this makes it more difficult to interpret the results.
- Frequently observed: Transformation with the natural logarithm due to right skewed and fat tailed distributions.

Skewness and excess kurtosis

- The **skewness** S denotes how symmetrical a distribution is.

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{s^3}$$

s: standard deviation

- For a symmetrical distribution: S = 0, for a right skewed (left skewed) distribution: S > 0 (S < 0).
- The **excess kurtosis** E denotes the kurtosis, i.e. whether there is more or less probability mass in the tails of the distribution compared to a normal distribution.

$$E = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3$$

s: standard deviation

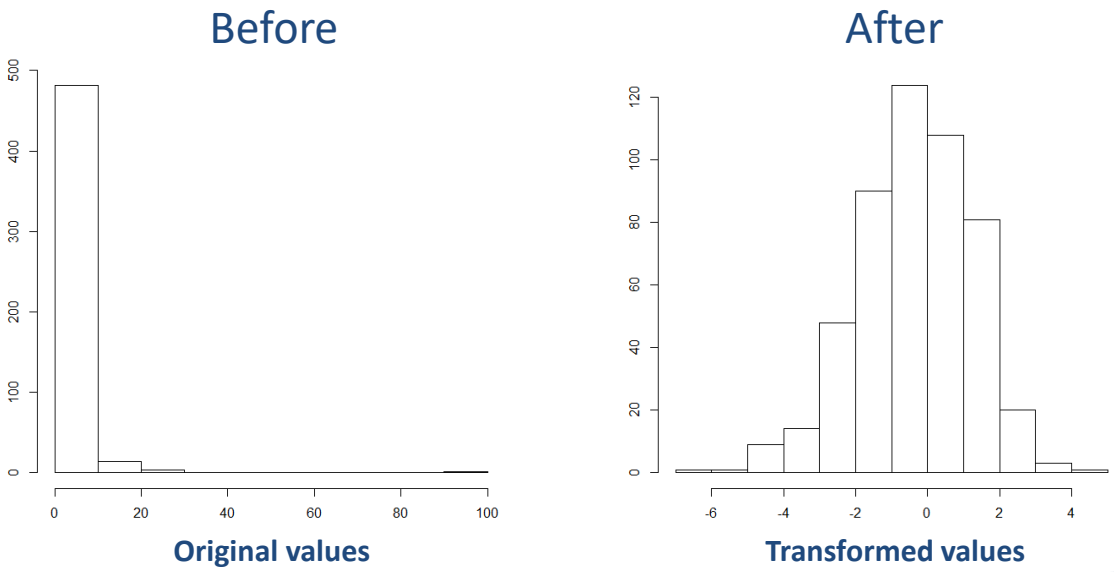
- For a normal distribution: E = 0, for fat tails (thin tails): E > 0 (E < 0).

Overview over possible data transformation methods

- Statistics knows a whole range of data transformations for smoothing extreme skewness or extreme kurtosis.
- A very important transformation is the logarithmic transformation ln(x), for percentages or growth rates also ln(x + 1). However, it has the disadvantage that it can only be used for variables with positive values, which should also be at least 1.
- Other possible transformations are (see e.g. Hartung (2009) “Statistik”, p. 349.):
  - The reciprocal transformation x^{-1},
  - the root transformation $\sqrt{x}$,
  - the power transformation x^n,
  - the Box Cox transformation,
  - the arcsine-sine-transformation,
  - the Fisher‘s z-transformation $\operatorname{arctanh}(x)$.
- It may be necessary to apply several different transformations in succession.

Examples for transformations - Logarithm

- The transformation with the natural logarithm can be used for extremely right-skewed distributions.

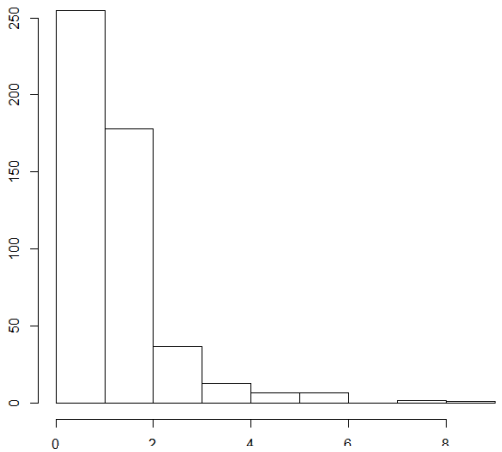


- As the logarithm is only defined for positive values, the values may have to be transformed beforehand, e.g. by adding the amount of the minimum +1: x_new = x + abs(min(x)) + 1.

Examples for transformations - Root

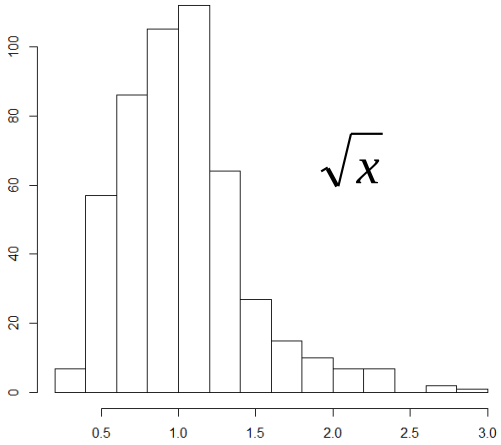
- If the skewness to the right is not so extreme, the root transformation can be used, which may be repeated several times:

Before



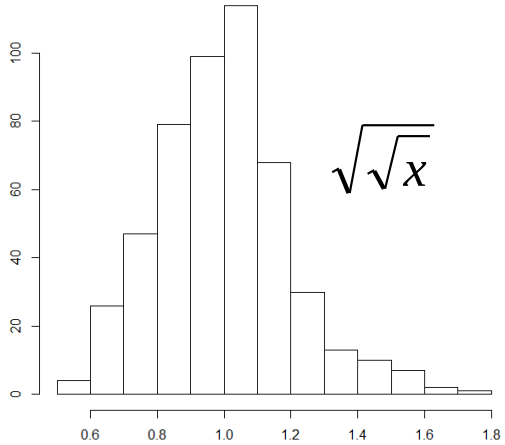
Original values

After



Transformed values

After

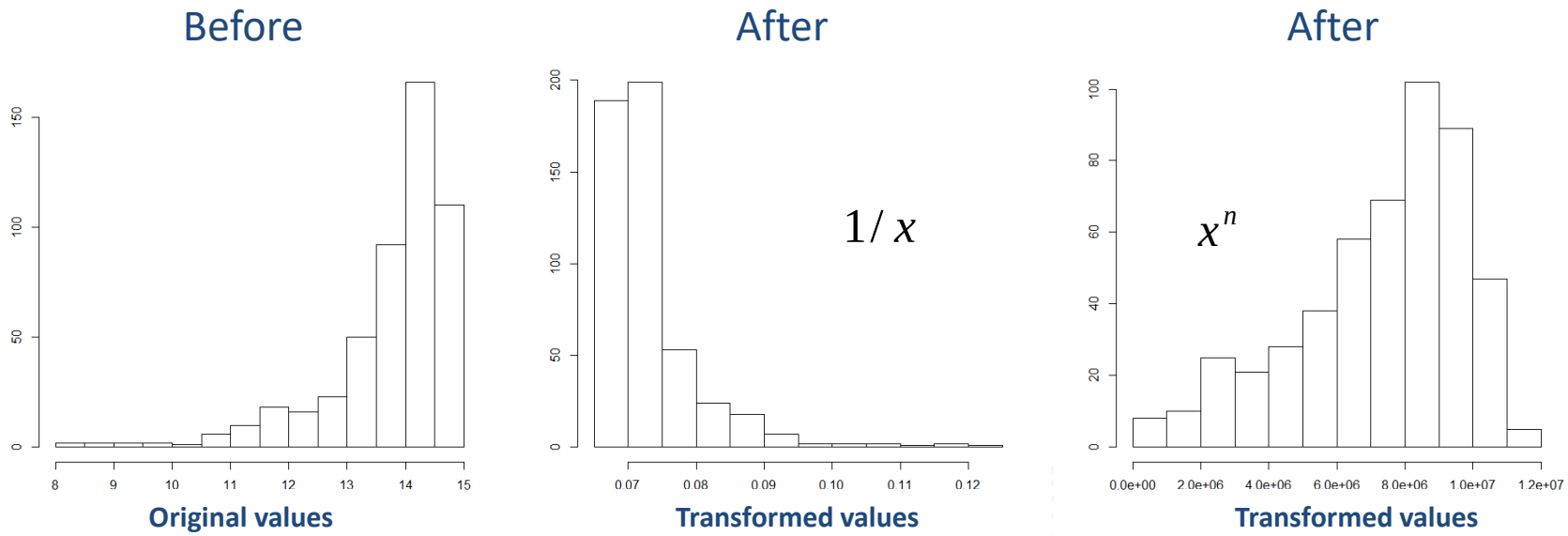


Transformed values

- Here, too, it may be necessary to transform beforehand, as the root can only be taken from positive numbers.

Examples for transformations – reciprocal and power transformation

- Reciprocal and power transformations can be applied to left-skewed distributions:



- After the reciprocal transformation, a further transformation may be required for right-skewed distributions.

Exercise

Calculate the skewness and excess kurtosis for the tip from the tips data set and try to reduce the skewness and kurtosis using various transformations.

R commands: describe() from the psych package

6. Linear Regression

Introduction

- Linear regression is a method that can be used to model the relationship between several variables.
- Linear regression is suitable for the following situation:
  - A variable (dependent variable, target variable) depends on one or more other variables (independent variables, influencing variables, determinants, success factors). There is therefore a **known dependency structure**.
  - The relationship between independent and dependent variables is linear or at least linearized.
  - The dependent variable and (at least) one independent variable are numerically scaled.

Basic terms and notations

Y	Dependent or target variable. Must be numerical.
$X_1,\ldots,X_K$	Independent variables. At least one of them must be numerical.

Simple linear regression: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{für } i = 1, \ldots, n$

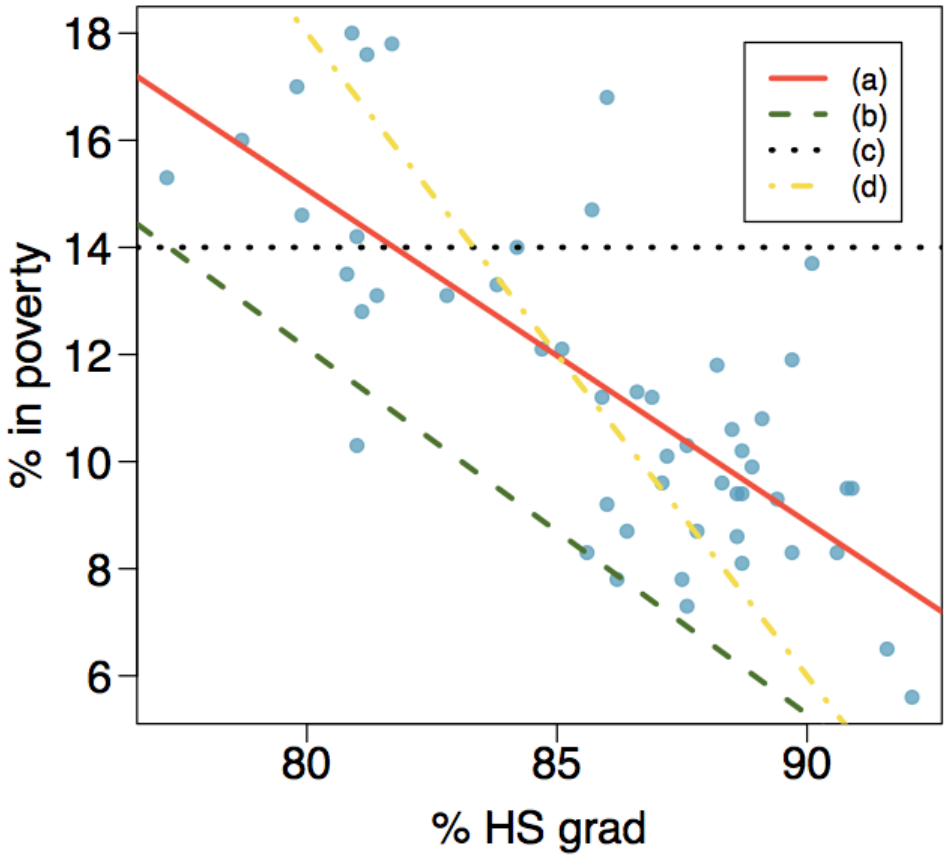
Multiple linear regression: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \; (i = 1, \ldots, n)$

y_i	Observed values for the target variable Y
x_{ki}	Observed values for the independent variables X_k
β_0	Intercept, constant (intersection with the Y-axis)
β_k	Regression coefficients (slopes)
i	Index for the observations ($i=1,\ldots,n$)
k	Index for the independent variables X ($k=1,\ldots,K$)
ε_i	Residual values, error values.

6.1 Ordinary least squares and the coefficient of determination

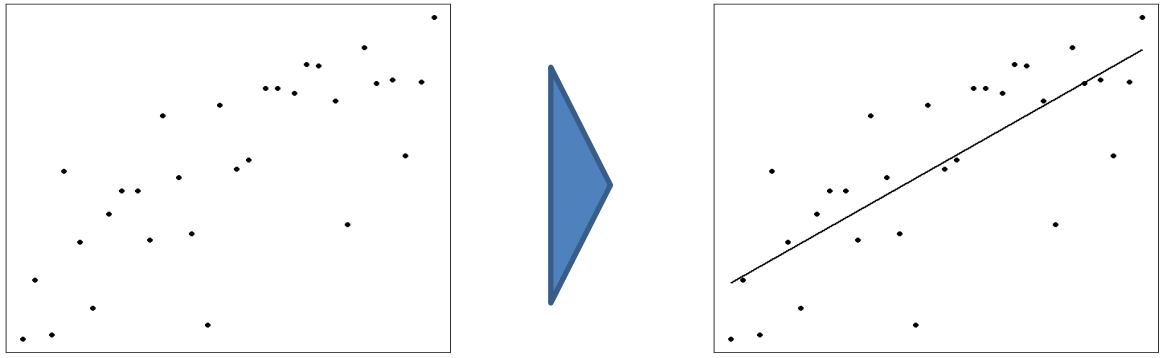
Exercise

Which of the following lines best describes the linear relationship between the percentage of people living in poverty and the percentage of high school graduates?



Basic idea of a linear regression

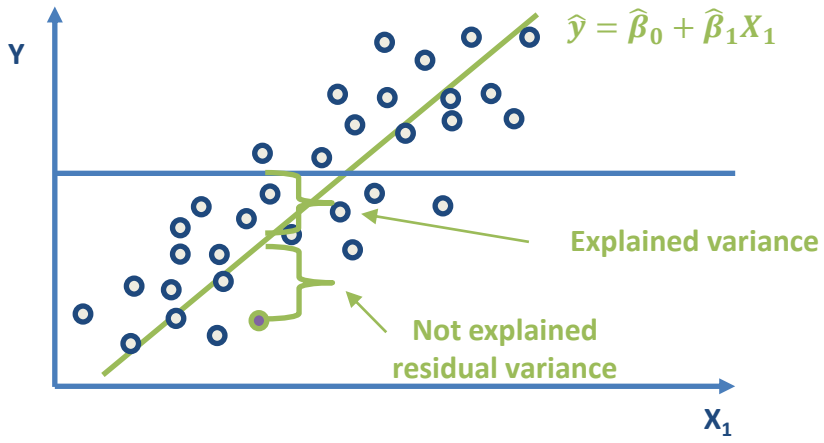
- The basic idea of linear regression is to describe a point cloud as well as possible using a linear function.
- In the simplest case, one independent variable, this can be represented well in a scatter diagram:



- The Ordinary Least Squares (OLS) method fits a straight line to the point cloud in such a way that the sum of the squared distances from each individual point to the straight line (residuals) is minimized. This is why linear regression is also called OLS regression.

Fitting the regression line

Consider a simple linear regression:



SST = SSR + SSE

SST: Total sum of squares (total variance)
SSR: Regression sum of squares (explained variance by the regression line).
SSE: Error sum of squares (not explained residual variance).

The total variance SST is split up in an explained variance SSR and a not explained residual variance SSE.

Ordinary least squares

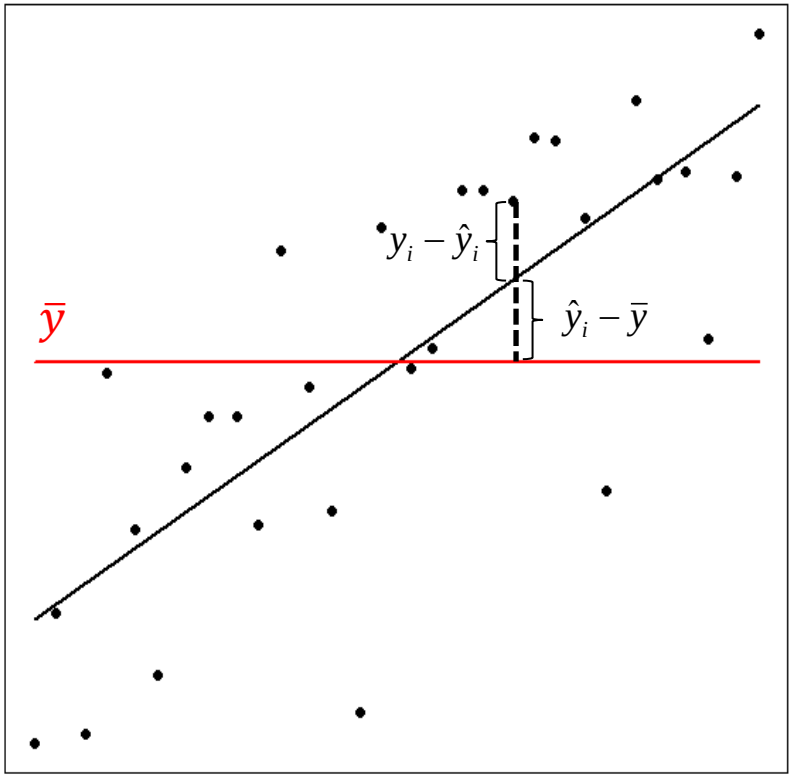
- Ordinary least squares identify the regression line that best represents the empirical values, i.e. around which the points in the scatter plot deviate/scatter minimally.
- This means that the best regression line is the one with the lowest error in the prediction of the Y values based on the X values.
- The deviation of each point from the regression line is called the residual or error and is denoted by ε_i or u_i.
- Meaning of the residual:
  - The variance of the y-values (total variance) is the sum of the variance of the residuals (unexplained variance) and the variance of the predicted (estimated) ŷ-values (explained variance).
  - Residuals thus contain the proportions of the dependent variable Y that are not captured by the independent variables X.
  - These proportions contain measurement errors of Y as well as components that can be explained by other variables that are not directly related to the independent variables.

Derivation of the multiple coefficient of determination R^2

- The starting point for assessing the global model quality is the mean value ȳ for the target variable y. The distance to the mean value ȳ can now be determined for each value y_i.
- The deviation of each point y_i from the mean value ȳ can be divided into a deviation explained by the regression line and a deviation not explained by the regression line:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Total deviation	Explained deviation	Not explained deviation



$\hat{y}_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}$ is the values estimated by the regression line for y_i.

Derivation of the multiple coefficient of determination R^2

- It can be shown that this relationship is maintained if the deviations are first squared and then added up:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SST	SSR	SSE
Total sum of squares	Regression sum of squares	Error sum of squares

- SSE is just the sum of the squared residuals. As a reminder: The sum of the squared residuals is minimized by the best possible regression line.
- The model quality is good if the share of the variance explained by the regression line in the total variance is as large as possible or if the share of the unexplained variance in the total variance is as small as possible.

Multiple and adjusted coefficient of determination R^2

- The multiple coefficient of determination R^2 is defined as:

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Proportion of the variance explained by the regression line (SSR) in the total variance (SST). R^2 lies (as a proportion value) between 0 (poor model quality) and 1 (good model quality).
- Problem: Each additional independent variable added increases R^2.
- Therefore, the corrected R^2 (adjusted R^2) is usually used:

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$$

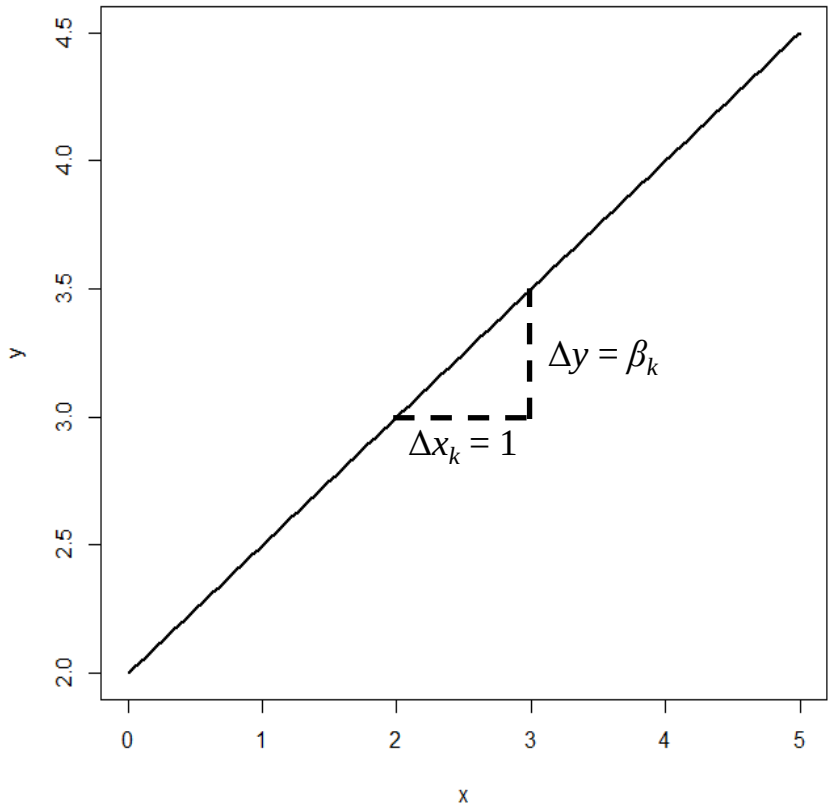
n – Number of observations
K – Number of independent variables

- The explanatory power of the model only improves if the adjusted R^2 increases by adding another independent variable.

6.2 Regression coefficients

Interpretation of the regression coefficients

The regression coefficients can be interpreted as slopes, as in any linear function.

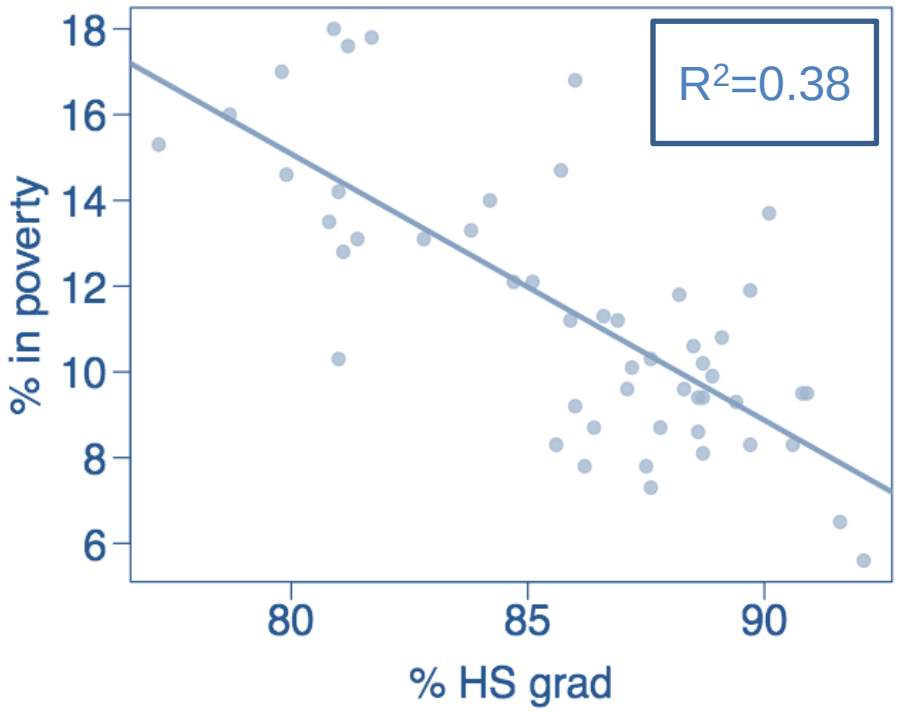


Let $\hat{y} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_k$ be the regression equation:

- $\hat{\beta}_0$ is the intersection with the Y-axis (intercept). This often cannot be interpreted in an economically meaningful way.
- $\hat{\beta}_k > 0$: If x_k increases (decreases) by one unit, y increases (decreases) by $\hat{\beta}_k$ units.
- $\hat{\beta}_k < 0$: If x_k increases (decreases) by one unit, y decreases (increases) by $\hat{\beta}_k$ units.
- The larger the value of $\hat{\beta}_k$, the stronger the influence from x_k on y.

Example for a regression line

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretation slope

If the high school graduation rate increases by one percentage point, the proportion of people living in poverty can be expected to fall by 0.62 percentage points on average.

Interpretation intercept

In states with no high school graduates at all, an average of 64.68% of residents are expected to live in poverty.

Interpretation coefficient of determination R^2

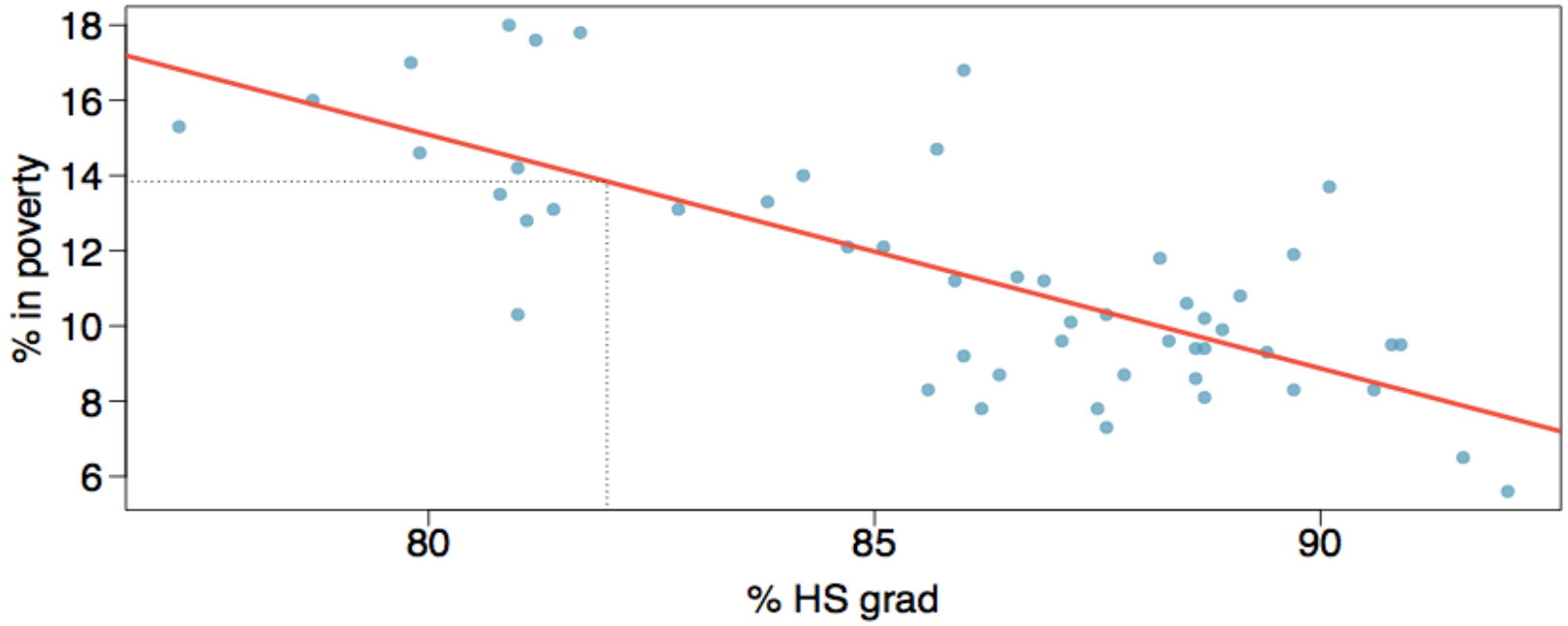
The differences (variance) in the poverty rate can be explained to 38% by the differences (variance) in the high school graduation rate.

Prediction

The regression line can be used for prediction.

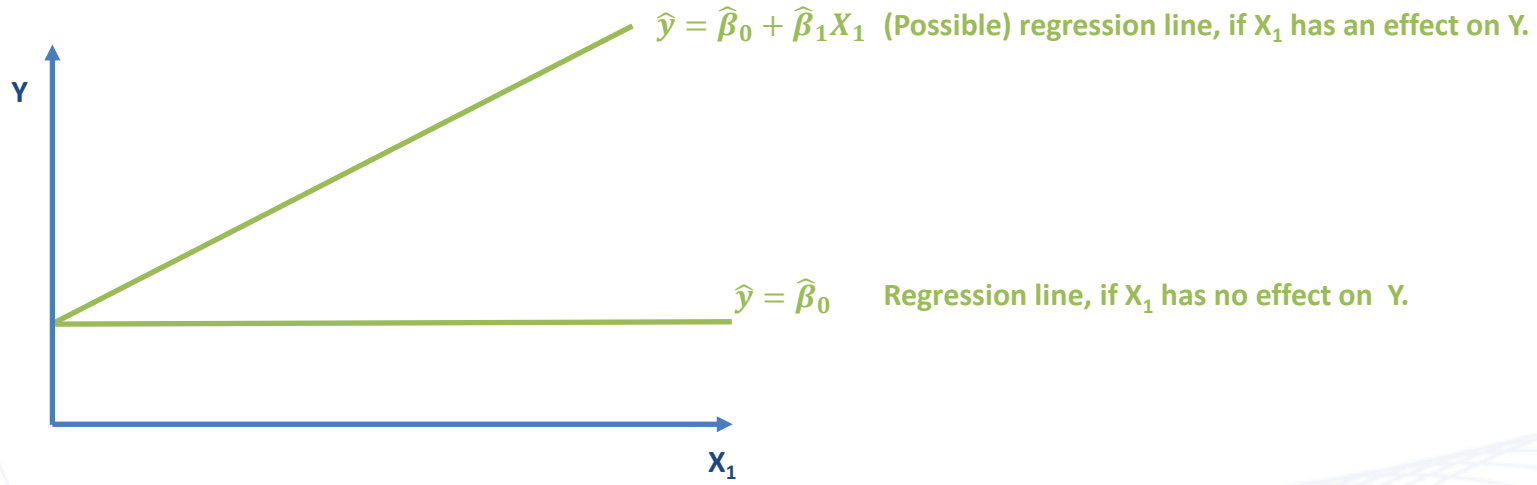
Example for the regression line $\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$

The proportion of people living in poverty for an 82% high school graduation rate is to be predicted: 64.68 - 0.62 x 82 = 13.84, i.e. an average poverty rate of 13.84% is expected. However, as for every point estimation, there is some uncertainty in the predicted value.



Hypothesis testing

An independent variable X_1 has an effect on the target variable Y, if the corresponding regression coefficient β_1 does not equal 0.



- Null hypothesis H_0: X_1 has no effect on Y, i.e. H_0: $\beta_1 = 0$.
- H_0 is rejected, if $\hat{\beta}_1$ differs significantly from 0, i.e. $p = P(\hat{\beta}_1 \mid H_0) \leq \alpha$.

t-test to assess the significance of the regression coefficients

- The t-test tests the null hypothesis H_0: $\beta_k = 0$ against the alternative H_1: $\beta_k \neq 0$, $0 \leq k \leq K$ for each regression coefficient β_k.
- The t-test checks whether the independent variable X_k has an effect on the target variable.
- If β_k is significantly different from 0, X_k has an effect on the target variable. Otherwise, X_k has no effect on the target variable.
- The ratio of the estimated regression coefficient $\hat{\beta}_k$ to its estimated standard error follows a t-distribution with n-K-1 degrees of freedom:

$$T = \frac{\hat{\beta}_k}{\widehat{sd}(\hat{\beta}_k)} \sim t_{n-K-1}$$

- It is common practice to carry out a two-sided test for the test level $\alpha = 0.05$. Simple rule of thumb (depending on sample size): If |T| is greater than 2, the null hypothesis is rejected.

6.3 Linear Regression using 

Preparation

- Linear regressions are best performed in R using the lm() function.
- Scatter plots can be generated with the plot() function. However, the scatterplot() function from the car package provides some additional useful options. For example, adding a linear trend line is easier than with the plot() function.
- The scatterplot command requires the car package, which must be activated and installed if necessary.

R Script

```
# Activate the car package
library(car)    # scatterplot()
```

- The B3 data set is used as an example data set. It can be activated as described below. The klaR package may need to be installed. For a description of the B3 data set, see the following slide.

R Script

```
# Activate the B3 data set
data(B3, package="klaR")
```

B3 data set

- Data set B3 from the klaR package contains data on the business cycle in western Germany, collected quarterly for the years 1955 (starting with the fourth quarter) to 1994 (fourth quarter).
- The row numbering indicates first the year, then the quarter. Thus 1977.4 stands for the fourth quarter of 1977.
- The data set contains a total of 157 quarters and 14 variables. The variables with JW refer to the annual changes.

R Script

```
# Data structure of the B3 data set
str(B3)
```

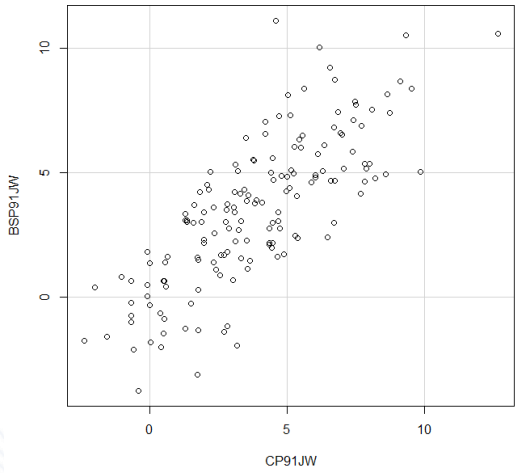
- Documentation on the klaR package, including a description of the B3 data set, can be found here: https://CRAN.R-project.org/package=klaR.

Scatter plot

- The relationship between the growth in gross national product (dependent variable) and the growth in private consumption (independent variable) is to be shown.

R Script

```
# Scatter plot for growth in gross national product (BSP91JW) depending
# on the growth in private consumption (CP91JW).
plot(BSP91JW~CP91JW, data=B3)
```



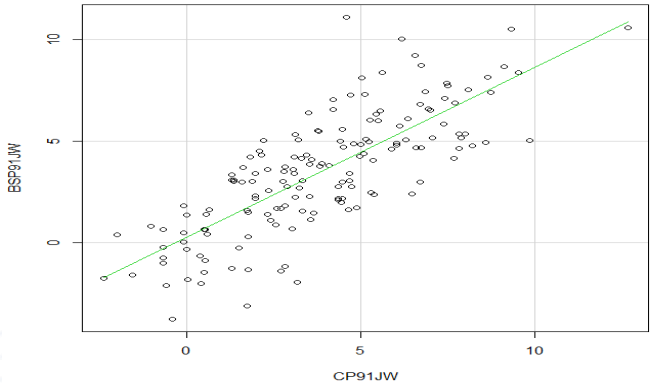
Interpretation:
- Growth in gross national product depends on growth in private consumption.
- The correlation is positive.

Scatter plot

- The abline(lm(BSP91JW~CP91JW, data=B3)) command adds a linear trend line:

R Script

```
# Scatter plot with linear trend for growth in gross national product
# BSP91JW) depending on the growth in private consumption (CP91JW).
plot(BSP91JW~CP91JW, data=B3)
abline(lm(BSP91JW~CP91JW, data=B3))

# Alternative: Use the scatterplot() function
scatterplot(BSP91JW~CP91JW, data=B3, regLine=TRUE, smooth=FALSE,
  boxplots="", id=FALSE)
```



Linear regression with numerical variables

Consider the following regression model:

$$BSP91JW = \beta_0 + \beta_1 \cdot CP91JW + \beta_2 \cdot ZINSK + u$$

R Script

```
# Carrying out the linear regression with numerical variables
linreg <- lm(BSP91JW~CP91JW+ZINSK, data=B3)
summary(linreg)
```

Output

```
lm(formula = BSP91JW ~ CP91JW + ZINSK, data = B3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8871 -1.2223  0.0711  1.1815  6.9945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.306215   0.525633   0.583    0.561
CP91JW       0.835542   0.057587  14.509   <2e-16 ***
ZINSK       -0.005384   0.063917  -0.084    0.933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 154 degrees of freedom
Multiple R-squared:  0.6017, Adjusted R-squared:  0.5966
F-statistic: 116.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

The results are interpreted separately for the global model fit and the regression coefficients on the following two slides.

Results of the linear regression – model fit

Output

```
lm(formula = BSP91JW ~ CP91JW + ZINSK, data = B3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8871 -1.2223  0.0711  1.1815  6.9945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.306215   0.525633   0.583    0.561
CP91JW       0.835542   0.057587  14.509   <2e-16 ***
ZINSK       -0.005384   0.063917  -0.084    0.933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 154 degrees of freedom
Multiple R-squared:  0.6017,    Adjusted R-squared:  0.5966
F-statistic: 116.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

Coefficient of determination

Adjusted coefficient of determination

p-value: F test for model fit

- The adjusted coefficient of determination takes on a value of just under 0.6. This is very good for only two variables.
- The global F-test rejects the null hypothesis that all independent variables taken together have no influence on the dependent variable.

Results of the linear regression – regression coefficients

Output

```
lm(formula = BSP91JW ~ CP91JW + ZINSK, data = B3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8871 -1.2223  0.0711  1.1815  6.9945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.306215   0.525633   0.583    0.561
CP91JW      0.835542   0.057587  14.509   <2e-16 ***
ZINSK      -0.005384   0.063917  -0.084    0.933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 154 degrees of freedom
Multiple R-squared:  0.6017, Adjusted R-squared:  0.5966
F-statistic: 116.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

Parameter estimators

t values

p values

- t values greater than 2 or p values less than 5% mean that the corresponding variable (here: CP91JW) has a significant effect on BSP91JW.
- The effect of the parameter ZINSK and the intercept, on the other hand, are not significant at the 5% level.
- Thus, the estimated regression equation is: $\widehat{BSP91JW} = 0{,}835542 \cdot CP91JW$.

Results of the linear regression – inconsistency

Output

```
lm(formula = BSP91JW ~ CP91JW + ZINSK, data = B3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8871 -1.2223  0.0711  1.1815  6.9945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.306215   0.525633   0.583    0.561
CP91JW       0.835542   0.057587  14.509   <2e-16 ***
ZINSK       -0.005384   0.063917  -0.084    0.933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 154 degrees of freedom
Multiple R-squared:  0.6017, Adjusted R-squared:  0.5966
F-statistic: 116.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

p values for the regression coefficients

Adjusted coefficient of determination

- The coefficient of determination is very good at approx. 0.6 with only two independent variables, but only one of the two independent variables has a significant influence on the dependent variable. This must be checked.
- More on this in the section on regression diagnostics.

Standardized regression coefficients

- To make the regression coefficients comparable, the independent variable and the dependent variable must first be standardized using the scale() function.
- A linear regression is then carried out with the standardized variables. The intercept must be omitted here, as a regression line with standardized variables runs through the origin. In R, the intercept is eliminated by the parameter -1 in the model equation.
- This results in the following regression equation, the addition z stands for standardized variables: BSP91JWz = $\beta_1 \cdot$ CP91JWz + $\beta_2 \cdot$ ZINSKz + u

R Script

```
# Standardization (z transformation)
B3[,"BSP91JWz"] <- scale(B3[,"BSP91JW"])
B3[,"CP91JWz"] <- scale(B3[,"CP91JW"])
B3[,"ZINSKz"] <- scale(B3[,"ZINSK"])

# Linear regression with standardized regression coefficients
linreg <- lm(BSP91JWz~CP91JWz+ZINSKz-1, data=B3)
summary(linreg)
```

Results – standardized regression coefficients

Output

```
Call:
lm(formula = BSP91JWz ~ CP91JWz + ZINSKz - 1, data = B3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64313 -0.41097  0.02391  0.39723  2.35167

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
CP91JWz 0.774336   0.053196  14.556   <2e-16 ***
ZINSKz -0.004495   0.053196  -0.084    0.933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6331 on 155 degrees of freedom
Multiple R-squared:  0.6017, Adjusted R-squared:  0.5966
F-statistic: 117.1 on 2 and 155 DF,  p-value: < 2.2e-16
```

Parameter estimators for the standardized regression coefficients.

Interpretation: CP91JW is a more important variable than ZINSK.

N.B. - The following interpretation **does not apply**: If CP91JW increases by 1 percentage point, BSP91JW increases by 0.77 percentage points.

Linear regression with categorial variables

- The dummy variable is created automatically in R if the variable has the data type factor. Otherwise, the variable must be converted using the factor() function.
- The smallest value in alphanumeric order, is selected as the reference category (i.e.: 1 before 2, a before b, ...).
- The contrasts() function can be used to display the contrast matrix and thus the current reference category.
- The reference category is changed with the relevel() function.
- Example: BSP91JW = $\beta_1 \cdot$ CP91JW + $\beta_2 \cdot$ PHASEN + u

R Script

```
# Carry out the linear regression with categorial variables
linreg <- lm(BSP91JW~CP91JW+factor(PHASEN), data=B3)
summary(linreg)

# Display the contrast matrix and change the reference category to
# economic phase 2
contrasts(B3$PHASEN)
B3$PHASEN <- relevel(B3$PHASEN, ref="2")
```

Results of the linear regression – categorial variables

Output

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.5487 -1.1171 -0.0629  1.0859  6.2843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34859    0.29035   4.645 7.32e-06 ***
CP91JW       0.75806    0.05379  14.094  < 2e-16 ***
PHASEN[T.2]  0.09107    0.43569   0.209    0.835
PHASEN[T.3] -1.35136    0.32887  -4.109 6.47e-05 ***
PHASEN[T.4] -2.24137    0.39392  -5.690 6.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 152 degrees of freedom
Multiple R-squared:  0.6885, Adjusted R-squared:  0.6803
F-statistic: 83.98 on 4 and 152 DF,  p-value: < 2.2e-16
```

Parameter estimators for the dummy variables' regression coefficients.

P values for the dummy variables' regression coefficients.

Interpretation:
PHASEN[T.2]: Since the regression coefficient at the 5% level is not significant, the following holds: The growth of GNP91JW is just as high in phase 1 as in phase 2.

PHASEN[T.3]: The regression coefficient is significant at the 5% level. That means: The growth of BSP91JW in phase 3 is 1.35 percent points lower than in phase 1.

Linearized regression

- Sometimes independent variables must be transformed before they can be used in a linear regression, for example if extreme skewness of the data is to be adjusted by a logarithm transformation.
- Transformations can either be carried out in advance (as in standardization with scale()) or in the regression statement itself. In the latter case, the operators + - * : / ^ must be masked by the I() function.
- If the regression coefficients are also to be standardized, the scale() function must be applied after the transformation.
- Example: $BSP91JW = \beta_0 + \beta_1 \cdot CP91JW + \beta_2 \cdot \ln(CP91JW) + \beta_3 \cdot CP91JW^2 + u$

R Script

```
# Linearized regression without standardized regression coefficients
linreg <- lm(BSP91JW~CP91JW+log(CP91JW+abs(min(CP91JW))+1)+I(CP91JW^2),
  data=B3)
summary(linreg)

# Linearized regression with standardized regression coefficients
linreg <- lm(scale(BSP91JW)~scale(CP91JW)+scale(log(CP91JW
  +abs(min(CP91JW)) +1))+scale(CP91JW^2)-1, data=B3)
summary(linreg)
```

Kleiber/Zeileis (2008), S. 65 ff.

Exercise

Exercise the linear regression using the tips data set.

R commands: lm(), plot()

6.4 Regression diagnostics

Application requirements and robustness of linear regression

- Linear regression is subject to a number of application requirements. The results of linear regression are only valid if these are fulfilled.
- Furthermore, linear regression is not robust against outliers. Outliers can bias the estimates for the regression coefficients $\hat{\beta}_k$ and thus also the results for the t tests for these coefficients.
- This section deals with the various application requirements and methods for checking the requirements and robustness.
- These methods are generally referred to as regression diagnostics or residual analysis, as the majority of these methods are based on the analysis of residuals.

Application requirements and robustness of linear regression

Application requirements

AR1: The relationship between the dependent and the independent variables must be linear. Important: Non-linearity in the variables is possible!

AR2: The expected value of the residuals is zero.

AR3: The residuals must not correlate with each other (no autocorrelation).

AR4: The variance of the residuals is constant and finite (homoscedasticity or no heteroscedasticity).

AR5: There must not be a very strong linear relationship between the individual independent variables (no or at most low multicollinearity).

AR6: The residuals are normally distributed.

AR7: There is no correlation between the residuals and the independent variable(s) (independent variables must not be endogenous).

Robustness

R1: There are no outliers in the dependent or the independent variable(s) or both.

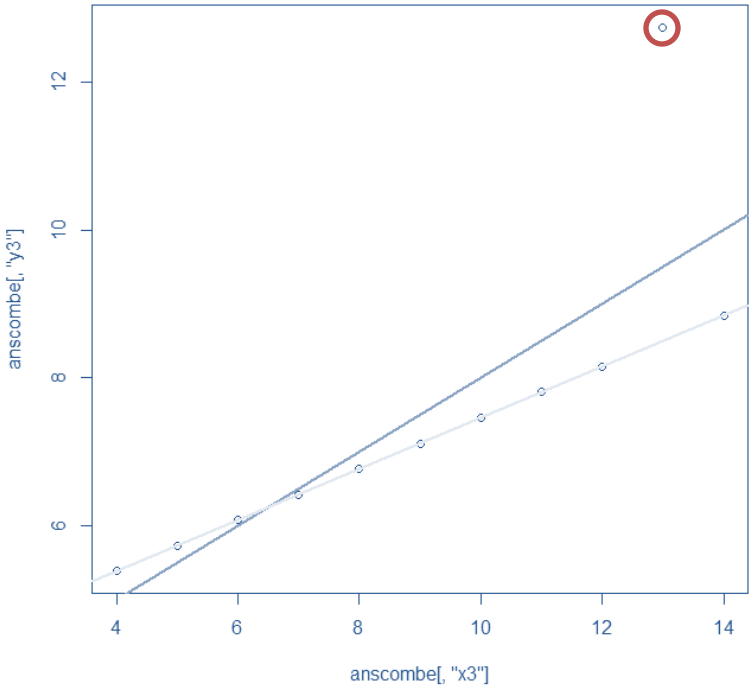
Relevance of the application requirements

- If requirements AR2, AR3 and AR4 are fulfilled, the least squares estimate of the regression coefficients (according to the Gauss-Markov theorem) is unbiased and efficient (see the section on the quality criteria for estimators) and even fulfills the so-called BLUE property (BLUE - Best Linear Unbiased Estimator).
- If requirement AR5 is fulfilled, the t tests for significance of the regression coefficients are reliable.
- The fulfillment of requirement AR6 is generally not considered to be particularly strict, as the central limit theorem applies for larger sample sizes. Consequently, in practice, the various tests for linear regressions such as the t tests for the regression coefficients, the F test for the coefficient of determination and tests of regression diagnostics such as the Durbin-Watson test and Breusch-Pagan test are also carried out without a prior test of the residuals for normal distribution.

Robustness of the regression line against outliers

Influence of individual observations on the regression line

- Individual observations that do not follow the trend of the vast majority of observations are referred to as outliers. Outliers are characterized by a large value for their residual.
- Changing or even omitting these outliers would have a significant influence on the course of the regression line.



If the highlighted point were omitted when fitting the regression line to the data, the line would run exactly through the other points (light blue line).

Omitting other points would not have such a significant effect on the course of the regression line.

Methods for outlier detection

- The example clearly shows that outliers distort the regression line. The estimates for the regression coefficients $\hat{\beta}_k$ are biased and therefore also the results for the t tests.
- It may make sense to remove outliers before the actual regression diagnostics. To do this, outliers must first be detected.
- The basic idea of outlier detection is to examine the influence of individual observations on the regression line. If this is large, the data point is a potential outlier.
- The following graphics and methods for outlier detection are presented:
  - Cook‘s Distance
  - Leverage values

Leverage values

- **Leverage values** (hat values) h_{ii} measure the deviation of the values of the independent variables for observation i from the average values of the independent variables $x_1, \ldots x_k$.
- The higher h_{ii} the larger the distance of observation i from the average values of the independent variables. **Large leverage values indicate possible outliers**.
- Leverage values range between 0 and 1 and are considered large if one of the following conditions is met:
  - If they are twice as large as the quotient of the number of independent variables *k* and the sample size *n*.
  - If they are greater than 0.5.
  - If individual values show a clear distance to the other leverage values.

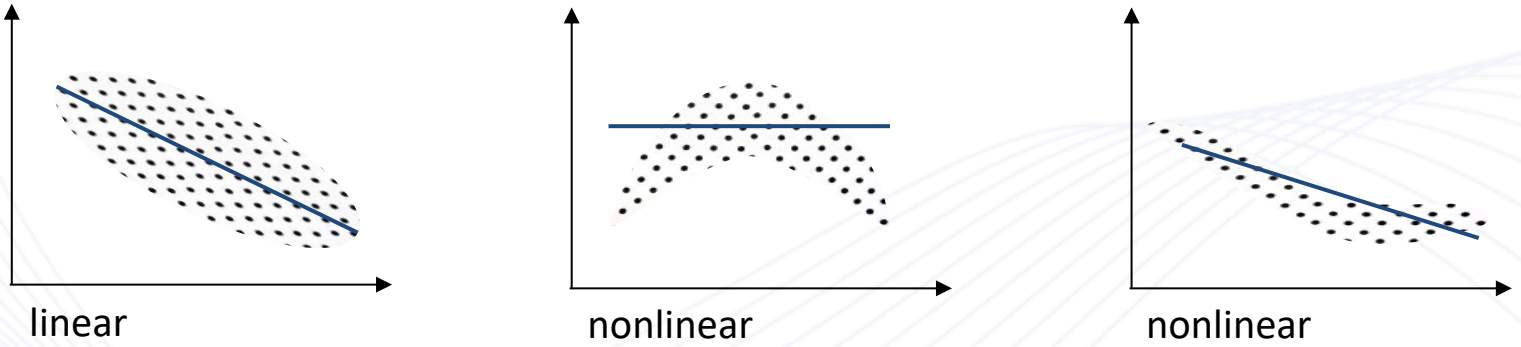
Cook‘s Distance

- **Cook's Distance** can be calculated for each individual observation *i*.
- To do so, a new regression is carried out omitting the data of observation *i.* The new regression therefore uses one data point less.
- The newly determined regression coefficients differ more or less from the original regression coefficients $\hat{\beta}_k$.
- Cook's Distance summarizes this difference between the new regression coefficients and the original ones in one figure.
- **Large values for Cook's Distance** mean a large difference in the regression coefficients and thus a strong influence of the corresponding individual observation and thus **indicate potential outliers**.
- A value for Cook's Distance is large if it deviates significantly from the other values for Cook's Distance.

Linearity of the relationship between dependent and independent variables

Consequences of the violation of the linearity requirement

- In the case of non-linearity, the regression line no longer provides the best estimator (i.e. they no longer minimize the distance between actual and estimated values).
- The result is a bias of the estimated values of the parameters $\hat{\beta}_k$ resulting in inconsistent estimators, i.e. the estimated values no longer converge towards the true parameters β_k as the sample size increases.
- Examples:

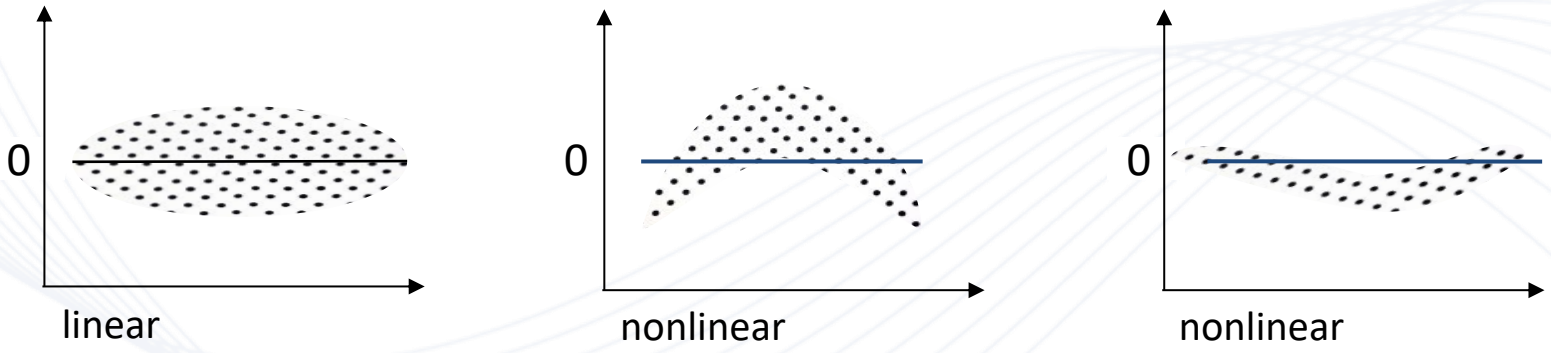


Residuals against fitted values plot

- One graphical option is a **scatter plot of the residuals** (y-axis) **against the estimated values** of the dependent variable $\hat{y}$, the so-called fitted values (x-axis). These are calculated from the regression equation using the parameter estimators:

$$\hat{y} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_k$$

- A linear relationship can be assumed if the residuals scatter randomly and without recognizable patterns around 0. Otherwise it indicates a nonlinear relationship between the independent and dependent variable.
- Examples:



RESET test for misspecification

- **Ramsey's REgression Specification Error Test (RESET test)** generally tests for misspecification of a regression model. Causes for a misspecification can be either missing important explanatory variables or a nonlinear relationship between dependent and independent variables.
- Ramsey was able to show that the fitted values $\hat{y}$ in exponentiated form ($\hat{y}^2, \hat{y}^3, \ldots$) are a suitable approximation for a misspecification, whereby an exponentiation up to the fourth power is considered sufficient.
- Let $y = \beta_0 + \sum \beta_k x_k + u$ be a regression model and let $\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_k x_k$ be the corresponding fitted values. Let $y = \beta_0 + \sum \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \gamma_3 \hat{y}^4 + u$ be the regression model extended by the first four powers of the fitted values.
- If the additionally included independent variables $\hat{y}^2, \hat{y}^3, \hat{y}^4$ taken together make a significant explanatory contribution, a misspecification can be assumed. In this case, the sum of the squared residuals SSE (error sum of squares) for the extended model e is smaller than for the original model u, i.e. SSE(e) < SSE(u).

RESET test for misspecification

- This allows the construction of an F test:

$$F = \frac{\left(SSE(u) - SSE(e)\right)/L}{SSE(e)/\left(n-K-L-1\right)}$$

F follows an F distribution with L and $n-K-L-1$ degrees of freedom

- Notation:

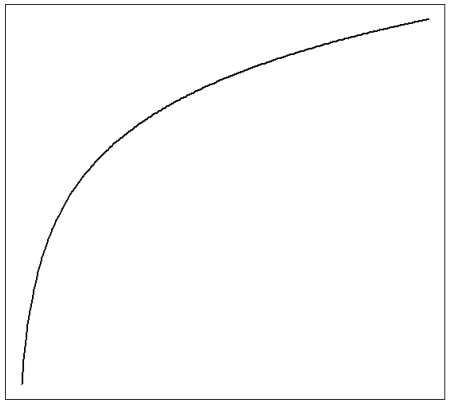
L	Number of powers of $\hat{y}$, usually L=3.
n	Sample size
K	Number of independent variables

- The null hypothesis is H_0: No misspecification.
- H_0 is rejected for large values of F and alternative H_1: Misspecification holds.

Strategies to deal with non-linearity

Nonlinear relationships between the independent and dependent variables can be eliminated by transforming the independent variables, especially for simple linear regressions.

Examples:

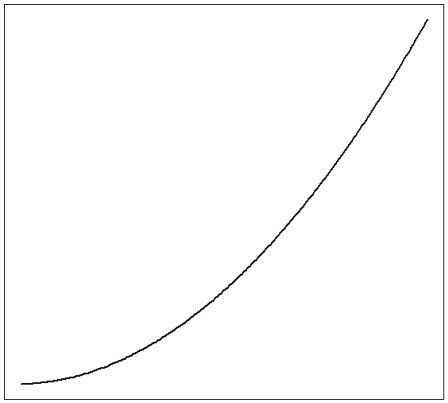


Feasible functions:

- Logarithm $\ln(x)$
- Root function $\sqrt{x}$

Example:

$y = \beta_0 + \beta_1 \cdot \ln(x) + u$

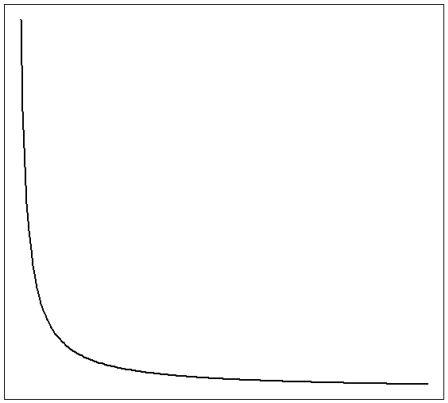


Feasible functions:

- Exponential function $\exp(x)$
- Parabola x^2

Example:

$y = \beta_0 + \beta_1 \cdot \exp(x) + u$



Feasible functions:

- Exponential function $\exp(-x)$
- Hyperbolic function $1/x$

Example:

$y = \beta_0 + \beta_1 \cdot 1/x + u$

Expected value for residuals

Consequences of an expected value for residuals unequal zero

- If the expected value of the residuals is 0, the residual only includes random effects that cause negative and positive deviations between the observed and estimated values for the target variable. These random fluctuations then balance each other out on average.
- A violation of this assumption occurs, for example, if the values for the dependent variable are measured too high or too low by a constant term. In this case, the residual contains a systematic effect.
- If a constant term, the intercept β_0, is included in the model, this systematic effect is captured by the least squares estimation in this constant term. In this case, the expected value of the residuals is always 0.
- However, if the intercept is omitted from the model, the systematic effect described above affects the estimates of the regression coefficients, which are then biased.
- Consequently, the intercept may only be omitted in justified exceptional cases, namely when it is certain that there is no systematic constant effect (e.g. for standardized values).

Autocorrelation

Autocorrelation

- **Autocorrelation** refers to the correlation between two typically adjacent residual terms e_i and e_j.
- Autocorrelation typically occurs if
  - a relevant regressor is not considered in the model (**misspecification**),
  - the functional form of a regressor is incorrectly specified (**non-linearity**),
  - there is a **spurious regression**, i.e. there is a statistically significant relationship that cannot be justified logically, but which is due to an underlying common trend (analogous to spurious correlation). An indication of a spurious regression is a high coefficient of determination R^2 with a simultaneous very strong positive autocorrelation.
- Autocorrelation is often observed in longitudinal regressions (**time series**) with capital market data, but spatial autocorrelation is also possible in cross-sectional data.
- Autocorrelation can occur either with adjacent residuals (**first-order autocorrelation**) or with “more distant” residuals (**higher-order autocorrelation**).

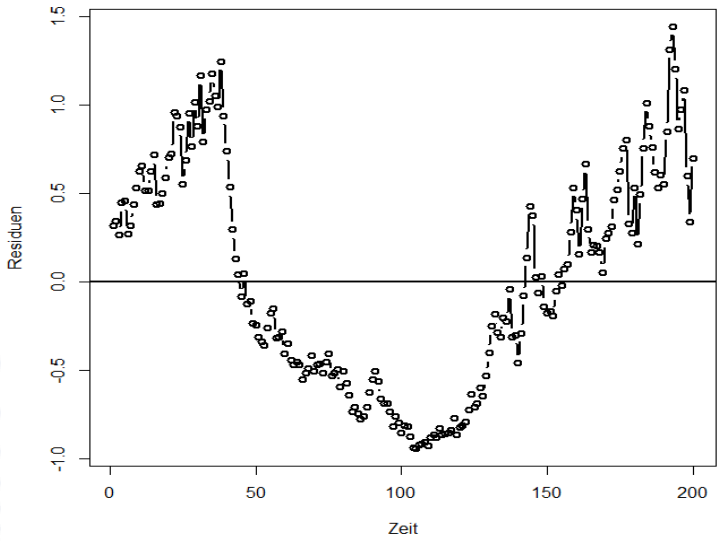
Consequences of autocorrelation

- Residuals do not scatter randomly around the regression line, but depend on the residuals of the preceding values.
- The least squares estimators $\hat{\beta}_k$ are unbiased and consistent but not efficient (no BLUE property). The variance of the estimators is no longer valid.
- That means, t test and F test provide misleading results. A least squares estimation for the linear regression is not feasible!

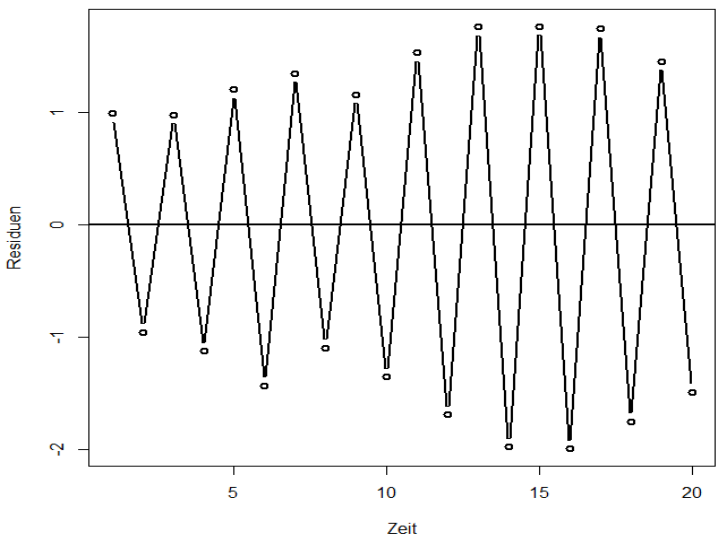
Graphical representation of autocorrelation

- Autocorrelations are best represented graphically with a scatter plot for the residuals or the target variable (y-axis) against the time variable (x-axis).
- The following typical phenomena arise for positive and negative autocorrelations (first order, i.e. adjacent residuals):

Positive autocorrelation



Negative autocorrelation



- The graphical representation of higher-order autocorrelations can be done using the autocorrelation function, which is a method of time series analysis.

Durbin-Watson test for first order autocorrelation

- The Durbin-Watson test compares the residuals of two adjacent observations u_t und u_{t-1} (r denotes the correlation of two adjacent residuals):

$$DW = \frac{\sum_{t=2}^{T} (u_t - u_{t-1})^2}{\sum_{t=1}^{T} u_t^2} \approx 2(1-r) \qquad \text{Value range: } 0 \leq DW \leq 4$$

- The null hypothesis is that there is no first order autocorrelation.
- If two adjacent residuals are almost equal, i.e. the target variable is subject to a trend, then DW is also small. Low values of DW therefore indicate positive autocorrelation.
- Conversely, large jumps in the residuals lead to a large DW, so high values of DW indicate negative autocorrelation.
- In general, the following applies:

  $DW = 0 \rightarrow r = 1 \rightarrow$ Perfect positive autocorrelation

  $DW = 2 \rightarrow r = 0 \rightarrow$ No autocorrelation

  $DW = 4 \rightarrow r = -1 \rightarrow$ Perfect negative autocorrelation

Criticism of the Durbin-Watson test

Using the Durbin-Watson test requires the consideration of some issues:

- If the null hypothesis is rejected, the DW test gives no indication of the reasons for rejecting the null hypothesis and how the model should be modified.
- The test is only suitable for first order autocorrelation.
- As a rule of thumb, higher order autocorrelation is rarely observed if there is no first order autocorrelation.
- The distribution of the DW value cannot be determined exactly, so that critical values have been tabulated using simulation studies. Furthermore, these critical values depend on the specific data matrix, which means that there is an undecidability range. If the value for DW lies within this range, it is not possible to decide whether the null hypothesis shall be rejected or not. This makes the DW test “unhandy”.

Breusch-Godfrey test for m-th order autocorrelation

- The Breusch-Godfrey test for autocorrelation assumes a linear regression $Y_t = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u_t$. If an *m*-th order autocorrelation is present, the coefficients of the auxiliary regression $u_t = \alpha_0 + \alpha_1 u_{t-1} + \ldots + \alpha_m u_{t-m} + \varepsilon_t$ are not all equal to 0.
- The corresponding null hypothesis is H_0: All $\alpha_i = 0$ or H_0: There is no *m*-th order autocorrelation.
- The corresponding test statistic nR^2 is asymptotically chi-square distributed with *m* degrees of freedom, where R^2 is the coefficient of determination of the auxiliary regression. For large sample sizes *n*, the null hypothesis can therefore be tested with a chi-square test.
- If R^2 is sufficiently large, the null hypothesis is rejected and an *m*-th order autocorrelation is assumed.
- Before applying this test, a decision must be made as to the maximum order of the autocorrelation. If *m* is set too low, a higher order autocorrelation may not be recognized; if *m* is set too high, the power of the test is poor.

Strategies to deal with autocorrelation

- If there is autocorrelation, the first step is to check whether the model is misspecified.
- It must be checked whether important independent variables are missing in the model or whether there is a nonlinear relationship between the independent and target variables.
- If the autocorrelation cannot be eliminated in this way, time series analysis methods (ARIMA / GARCH models) may help.
- It is also possible to replace the incorrect variances for the least squares estimators with the correct variances.
- Robust estimators yielding corrected p values can also be used, see e.g. Kleiber/Zeileis (2008), p. 106.
- Finally, suitable variable transformations such as the Cochrane-Orcutt estimators can also lead to the elimination of autocorrelation.

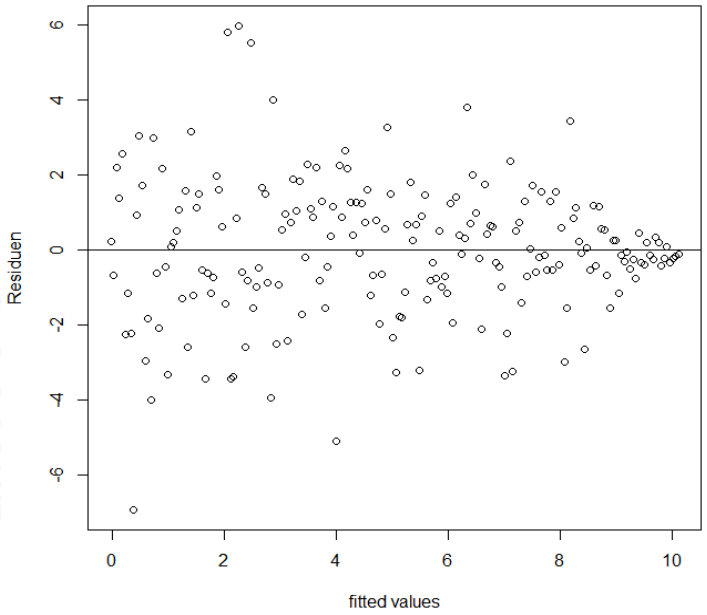
Heteroscedasticity

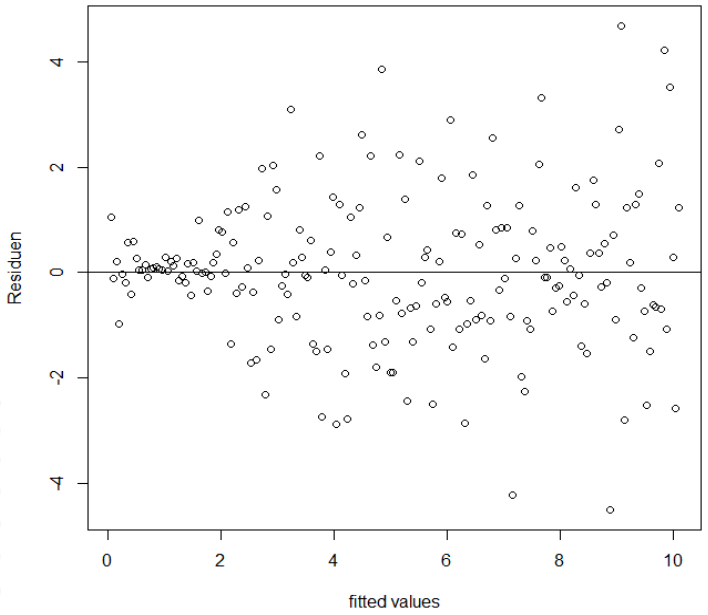
Heteroscedasticity and its effects

- If the dispersion of the residuals in a series of values of the predicted dependent variable is not constant, i.e. if the error terms scatter systematically, then heteroscedasticity is present.
- Heteroscedasticity typically occurs with
  - cross-sectional surveys,
  - data with a measurement error, where the measurement error shows a trend,
  - data from the financial markets, such as exchange rates or returns on securities,
  - non-consideration of a relevant regressor in the model (misspecification) and
  - incorrect specification of the functional form of a regressor (non-linearity).
- The effects of heteroscedasticity are the same as for autocorrelation.
- The least squares estimators $\hat{\beta}_k$ are unbiased and consistent but not efficient (no BLUE property). The variance of the estimators is no longer valid.
- t test and F test provide misleading results. A least squares estimation for the linear regression is not feasible.

Graphical representation of heteroscedasticity

- Heteroscedasticity is best represented graphically with a residual plot (plot of the residuals (y-axis) against the fitted values (x-axis)).
- The following typical patterns result for the residuals against the fitted values in the presence of heteroscedasticity:





Tests for heteroscedasticity

- A large number of tests for heteroscedasticity are listed in the literature. Of these, the **White test** and the **Breusch-Pagan test** are presented here.
- These two tests have in common that homoscedasticity is assumed as the null hypothesis and that the residuals of the regression model are examined to test the null hypothesis.
- The differences between the two tests lie on the one hand in the normal distribution assumption of the residuals and on the other hand in the general validity of the alternative hypothesis.

Breusch-Pagan test

- The Breusch-Pagan test uses an auxiliary regression to check whether the residual variance $Var(u_i) = \sigma_i^2$ depends on more than one variable:

$$\sigma_i^2 = \alpha_0 + \alpha_1 Z_{1i} + \ldots + \alpha_K Z_{Ki}$$

- The independent variables Z_{ki} can be the independent variables of the original regression model or other variables.
- The null hypothesis H_0: $\alpha_1 = \ldots = \alpha_K = 0$ is tested against the alternative that at least one $\alpha_k \neq 0$. If the null hypothesis is true, then $\sigma_i^2 = \alpha_0$, i.e. constant, and homoscedasticity is present.
- The test statistic for the Breusch-Pagan test is nR_e^2, where R_e^2 is the coefficient of determination of the auxiliary regression.
- Under the null hypothesis, the test statistic is asymptotically chi-square distributed, i.e. the associated test is a chi-square test.

White test

- The main disadvantage of the Breusch-Pagan test is that the residuals of the original regression must be normally distributed. Otherwise, the test statistic is not asymptotically chi-squared distributed and the test cannot be used.
- The White test, on the other hand, can also be used with non-normally distributed residuals. Similar to the Breusch-Pagan test, the residual variance is modeled using an auxiliary regression.
- The independent variables of the original regression model, their squares and cross products form the independent variables of the auxiliary regression. The auxiliary regression also has an intercept. In the case of two independent variables x_1 and x_2, this results in

$$\sigma_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x^2_{1i} + \alpha_4 x^2_{2i} + \alpha_5 x_{1i} x_{2i}$$

- The null hypothesis is H_0: $\sigma_i^2 = \sigma^2$ for all i=1,…,n, meaning homoscedasticity.
- The test statistic is nR_e^2, where R_e^2 is the coefficient of determination of the auxiliary regression.
- Under null hypothesis, this is asymptotically chi-square distributed.

Strategies to deal with heteroscedasticity

- If heteroscedasticity is present, the first step is to check whether the model is misspecified, analogous to autocorrelation.
- For example, it must be checked whether important independent variables are missing from the model or whether there is a non-linear relationship between the influencing and target variables.
- If there is a so-called volatility cluster, time series analysis methods (ARCH / GARCH models) may help.
- It is also possible to replace the incorrect variances for the least squares estimators with the correct variances, for example by using the so-called white standard errors.
- Robust estimators yielding corrected p values can also be used, see e.g. Kleiber/Zeileis (2008), p. 106.
- If the residuals' variance is known, a transformation of the dependent and all independent variables can help. The original dependent and independent variables are divided by the variances of the confounding variables and then a new regression is carried out with the transformed data.

Multicollinearity

Multicollinearity and its effects

- **Multicollinearity** refers to a high correlation between the independent variables. This always occurs when several independent variables measure more or less the same.
- Perfect multicollinearity means, an independent variable can be mapped linearly by one or more other independent variables, e.g. $X_1 = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + \ldots \alpha_{K\text{-}1} X_K$. In this case, it is not possible to estimate the regression coefficients.
- With increasing multicollinearity, it is possible to estimate the regression coefficients and the overall influence of all independent variables on the dependent variable is correctly modeled. But it is no longer possible to attribute the overall influence to the individual independent variables.
- The estimated regression coefficients become unstable; small changes in the underlying data result in significant changes in the estimated regression coefficients. The standard errors of the regression coefficients increase, i.e. the t-values decrease.

Detecting multicollinearity

- An initial indication of the potential presence of multicollinearity can be obtained by considering the coefficient of determination and the significance of the regression coefficients.
- A high coefficient of determination with simultaneous non-significance of the majority of the regression coefficients is a clear indication of multicollinearity.
- A simple correlation analysis, either using a correlation matrix of all independent variables or using pairwise scatter plots with two independent variables, also provides an indication of multicollinearity.
- However, only pairwise dependencies can be detected in this way. Correlations greater than 0.8 are regarded as critical in the literature.
- The standard strategy is the calculation of variance inflation factors (VIF). The variance inflation factors are the factors by which the variances of the regression coefficients increase with increasing multicollinearity.

Derivation of the variance inflation factors

- If an independent variable X_k can be expressed as a linear combination of the other *K*-1 independent variables, this relationship can be modeled using multiple regression. The fit of this relationship is reflected in the coefficient of determination R^2_k of this regression.
- The variance inflation factor is then calculated as

$$\mathrm{VIF_k} = \frac{1}{1 - R_k^2}$$

- The better an independent variable can be expressed by a linear combination of all other independent variables, the higher R^2_k and the higher VIF_k.
- A value of 10 or higher for VIF_k is considered critical, which corresponds to a coefficient of determination of at least 0.9. Other sources already describe a value of 5 as critical, which corresponds to a coefficient of determination of at least 0.8.

Strategies to deal with multicollinearity

- If the aim of the modeling is to uncover an overall effect of all variables, so that the consideration of individual variables is not relevant, multicollinearity can be neglected.
- If individual independent variables show a very high correlation with the other independent variables, these can be removed from the model. The resulting loss of information is small. However, this can lead to a misspecification of the model.
- The highly correlated variable is regressed on the other explanatory variables. The resulting residuals remain in the original model as independent variables (orthogonalization). However, if there are several highly correlated variables, the order of orthogonalization of the individual independent variables influences the results of the regression.

Strategies to deal with multicollinearity

- A principal component analysis can also be performed to replace the correlated variables with principal components and perform the regression with the principal components.
- Occasionally, it is also possible to increase the amount of data. This can lead to a reduction in multicollinearity.
- Two variables that are correlated with each other can possibly be combined into one. E.g.: The variable Y is regressed on X_1 (height) and X_2 (body weight). X_1 and X_2 are highly correlated. thus, these variables can be combined into one, the “body mass index”, i.e. the weight divided by the square of the height.
- Sometimes it is useful to use “biased estimators”, for example the ridge regression (RR). This calculates biased coefficient estimators with much smaller standard errors than with OLS. RR is included in most statistical software including R.

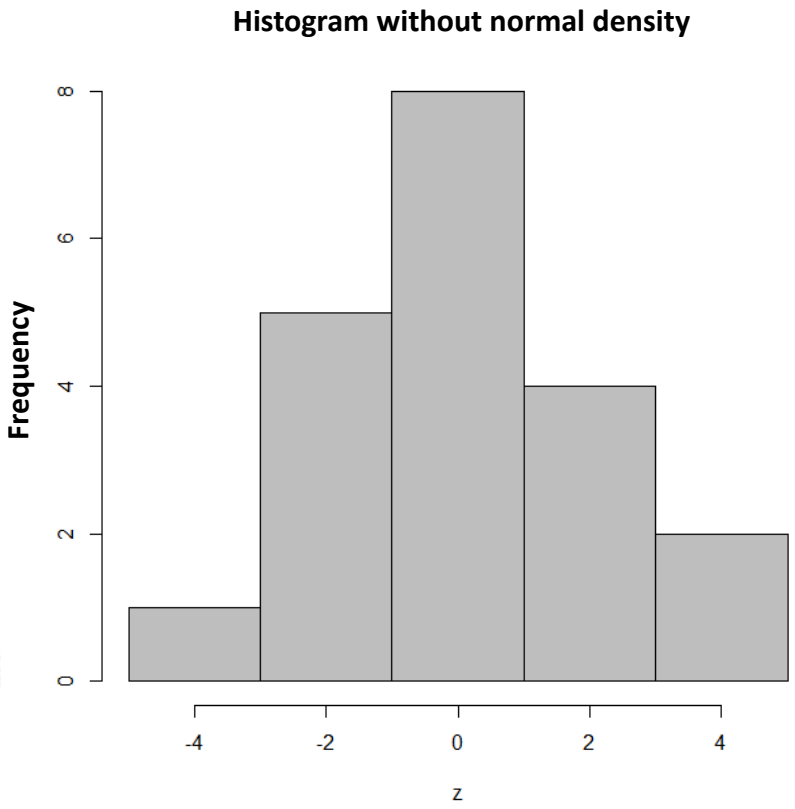
Normal distribution of the residuals

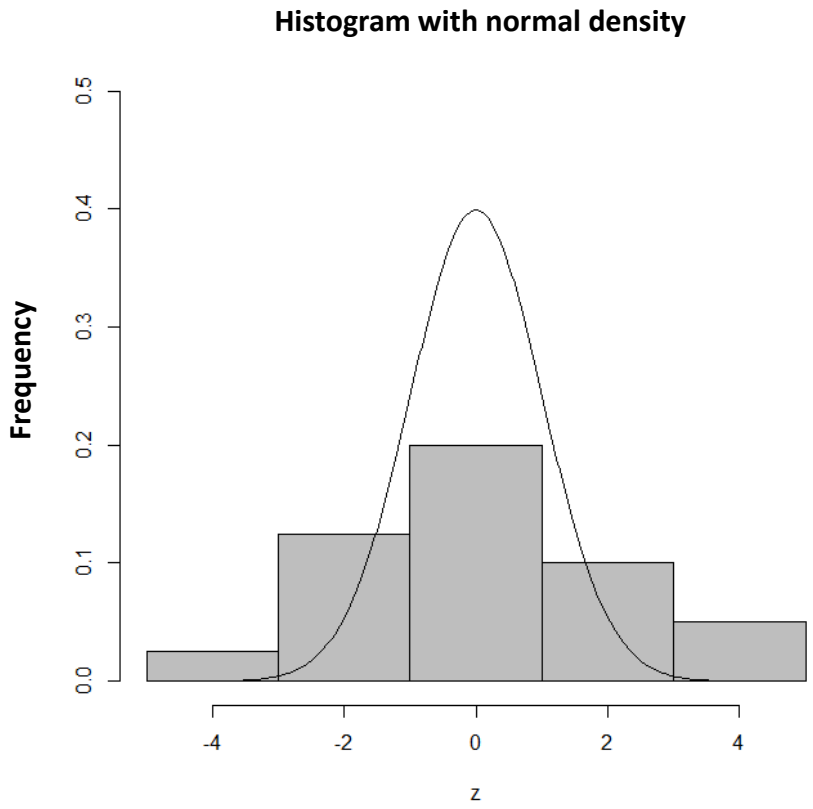
Normal distribution of the residuals

- It has already been pointed out that the normal distribution assumption is often not tested, as the central limit theorem applies and therefore a normal distribution of the residuals can be assumed for sample sizes that are not too small.
- If the normal distribution assumption is nevertheless to be examined, various graphical and analytical methods are available.
- Graphical methods include the histogram or the stem and leaf diagram for displaying the distribution of the residuals and the normal probability plot.
- Tests include the tests from Shapiro-Wilk, Jarque-Bera, Kolmogoroff-Smirnov and others.

Histogram

- **Histograms** allow the visual inspection of the distribution of the residuals. A symmetrical distribution indicates normality.





Stem and leaf diagram

- A **stem and leaf diagram** consists of two columns. The first column contains the equivalence classes in relation to which the individual values are summarized. This is often the number before the decimal point. The second column then contains the individual values, sorted by size. This is the first number after the decimal point.
- Example: A stem and leaf diagram is to be created for the values 0.2, 0.5, 0.7, 1.1, 1.3, 1.8, 1.8, 2.6, 2.8, 3.1:

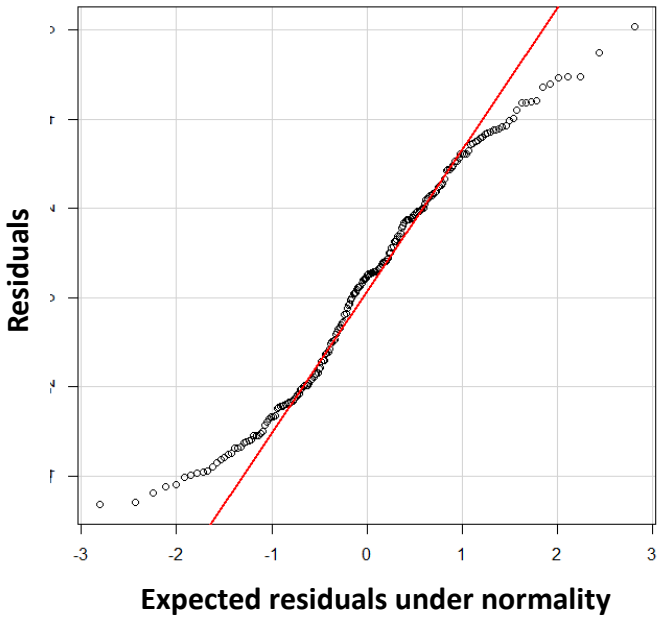
```
0 | 2 5 7
1 | 1 3 8 8
2 | 6 8
3 | 1
```

- If the data follow a normal distribution, then the stem and leaf diagram corresponds to the density of the normal distribution rotated by 90°.

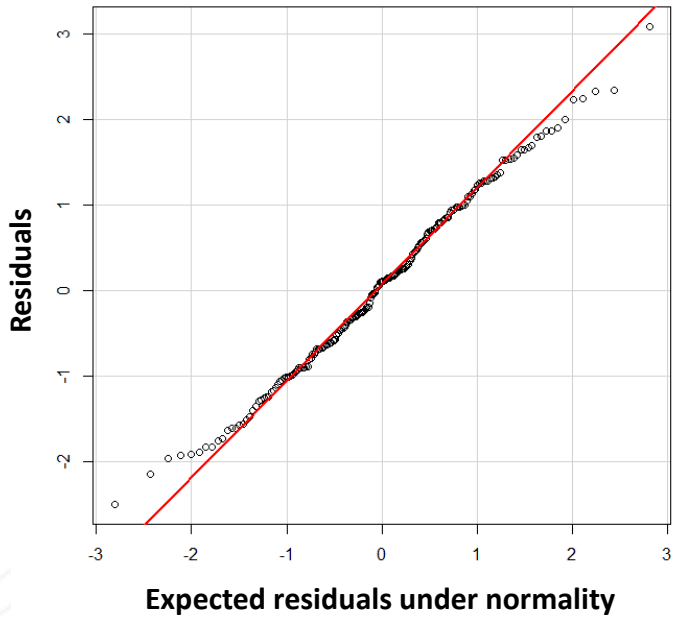
Normal probability plot

- In the normal probability plot (quantile-comparison plot, quantile-quantile plot, qq plot), each value for the residuals is plotted against its expected value under normal distribution in a scatter plot. If the distribution is normal, all points are close to or on the diagonal. However, if there are systematic deviations of the individual points from the diagonal, this is an indication of a non-normal distribution.

Not normal distributed



Normal distributed



Advantages of linear regression over correlation

- More than one independent variable can be taken into account.
- The model quality (explanatory power) can be specified.
- The strength and direction of an influence (leverage effect) can be determined.
- Forecasts are possible.

But: The distinction between dependent and independent variables requires a theoretical basis!

6.5 Regression diagnostics using 

Preparations

- The lmtest package must be activated to perform the regression diagnostics. The package car is also required for the scatter plot matrix.

R Script

```
# Required Packages
library(lmtest)
library(car) # scatterplotMatrix()
```

- The B3 dataset is again used as an example data set. For a description of the data set, see https://CRAN.R-project.org/package=klaR.
- The starting point is a linear regression with two independent variables.

R Script

```
# Activate B3 data set
data(B3, package="klaR")

# Linear regression: BSP91JW ~ CP91JW + ZINSK
linreg <- lm(BSP91JW~CP91JW+ZINSK, data=B3)
summary(linreg)
```

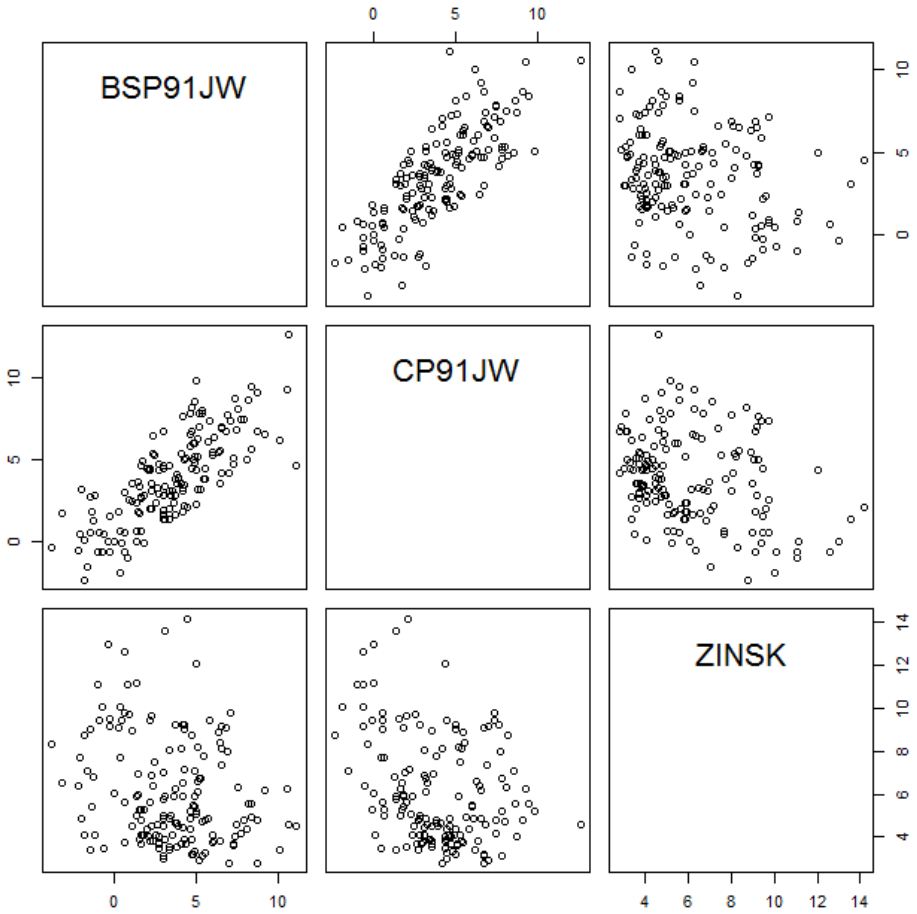
Robustness

- The pairwise scatter plots are obtained with the scatterplotMatrix() function.
- Cook's distance is obtained with cooks.distance(linreg), and the leverage values are obtained with hat(model.matrix(linreg)).

R Script

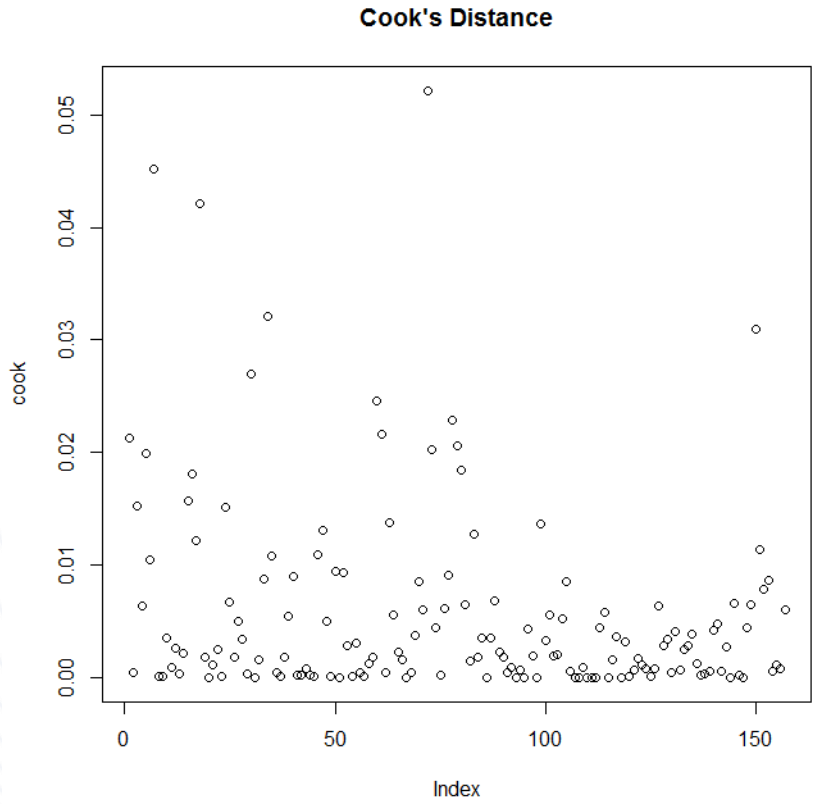
```
# Scatter plot matrix
scatterplotMatrix(~BSP91JW+CP91JW+ZINSK, regLine=FALSE,
                  smooth=FALSE, diagonal = FALSE, data=B3)

# Cook's distance
cook <- cooks.distance(linreg)
plot(cook, main="Cook's distance")
B3[cook > 0.5,] # displays data rows for values > .5

# Leverage values
lev <- hat(model.matrix(linreg))
plot(lev, main="Leverage values")
B3[lev > 0.5,] # displays data rows for values > .5
```

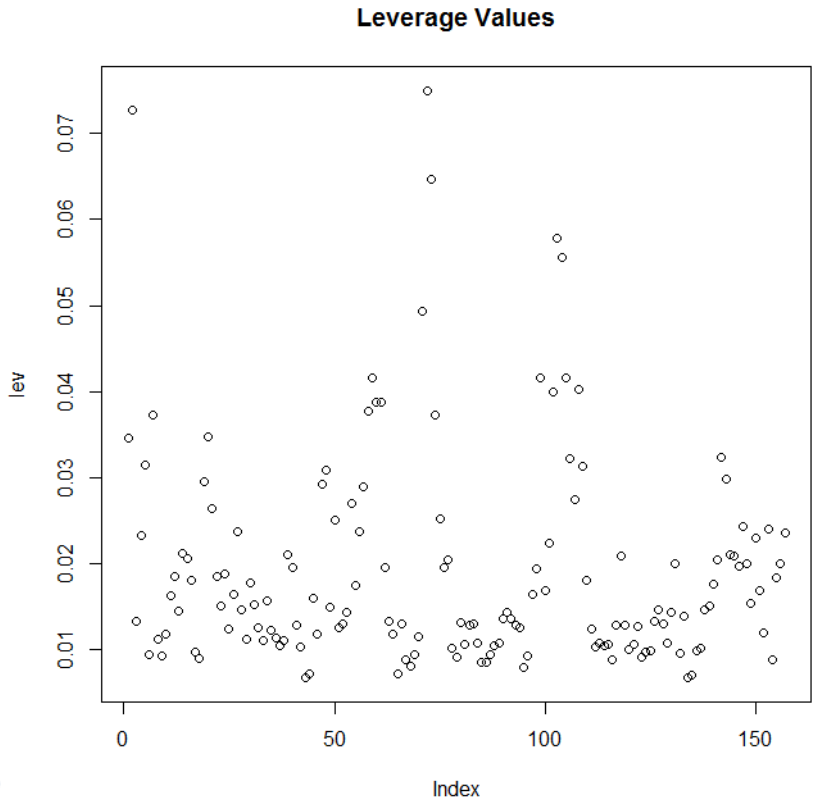
Robustness – Results



Interpretation:
Potential outliers can be identified in isolated cases.

Robustness – Results





Interpretation:
Even if individual values stand out from the mass of other points, there is no single value that stands out very clearly from the other values. Overall, there do not appear to be any outliers.

Linearity – Residuals against fitted values plot and RESET test

- This graphic can be generated with the plot() function. It is recommended to add a horizontal line for 0 (abline(h=0)).
- The fitted(linreg) command calculates the fitted values, and the resid(linreg) command calculates the residuals for the linreg regression model.

R Script

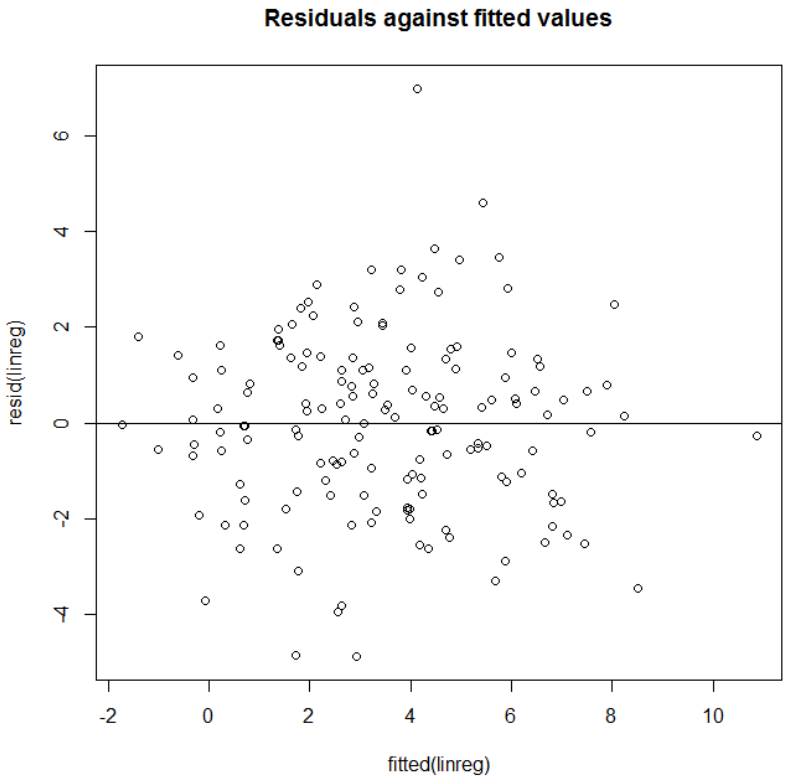
```
# Residuals against fitted values plot
plot(fitted(linreg), resid(linreg))
abline(h=0)
```

- The second option of the resettest() function can be used to set which powers are to be taken into account. 2:4 takes into account the second, third and fourth power. By default, only the second and third powers are taken into account.
- The option type="fitted" is used to determine the powers for the fitted values.

R Script

```
# RESET test for misspecification
resettest(linreg, 2:4, type="fitted")
```

Linearity – Results



Output

```
> resettest(linreg, 2:4, type="fitted")

        RESET test

data:  linreg
RESET = 1.3078, df1 = 3, df2 = 151, p-value = 0.274
```

P value for the RESET test

Interpretation:
The residuals scatter randomly around 0, there is no recognizable pattern, which does not indicate specification errors.

The RESET test does not reject the null hypothesis of “no specification error” (p = 0.274).

Overall, a linear relationship can be assumed.

Autocorrelation – Graphical representation and hypothesis testing

- The already known functions scatterplot() and plot() generate this graphic relatively easily. If the residuals are not available, the dependent variable can be used instead.
- The option type =“b” generates both points and lines connecting the points.

R Script

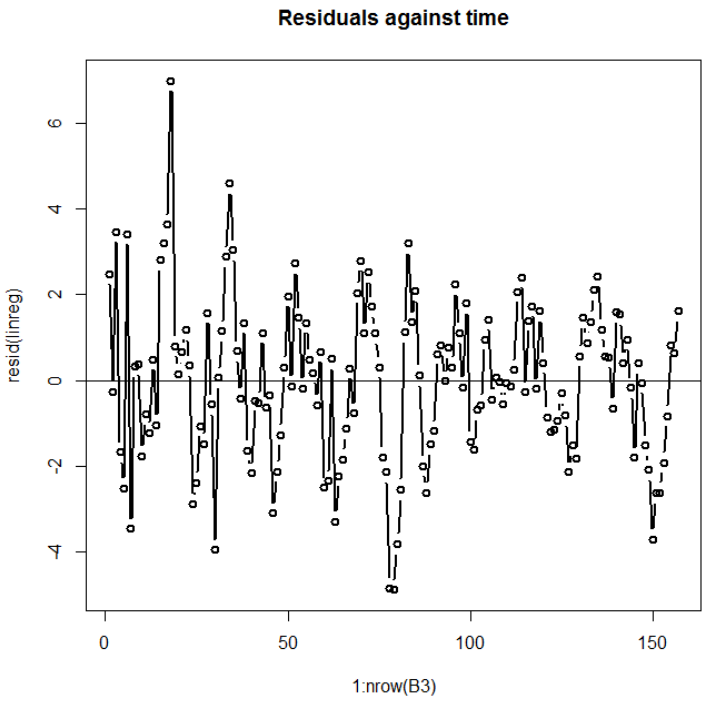
```
# Residual plot
plot(1:nrow(B3), resid(linreg), type="b", lwd=2)
abline(h=0)
```

- The DW test requires the specification of the alternative (option alternative=), i.e. whether the test is two-sided or one-sided.
- The BG test requires the specification of the order (option order=) up to which an autocorrelation is to be tested.

R Script

```
# Durbin-Watson test for autocorrelation
dwtest(linreg, alternative="two.sided")
# Breusch-Godfrey test for autocorrelation
bgtest(linreg, order=1)
```

Autocorrelation – Results



Output

```
        Durbin-Watson test

data:  linreg
DW = 0.9991, p-value = 1.085e-10
alternative hypothesis: true autocorrelation is not 0


> bgtest(linreg, order=1)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  linreg
LM test = 40.003, df = 1, p-value = 2.536e-10
```

p values for the two tests

Interpretation:
Waves can be recognized in the residual graph, there could be positive autocorrelation.

Both tests reject the null hypothesis of “no first-order autocorrelation” ($p < 0.0001$).

Overall, autocorrelation can be assumed.

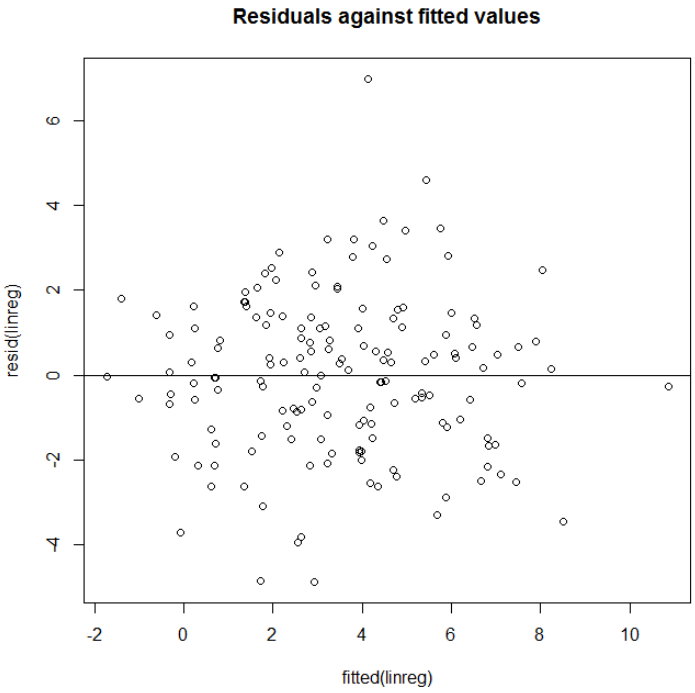
Heteroscedasticity – Breusch-Pagan test and White test

- Both tests can be carried out using the bptest() function. They both use the same independent variables as the original regression model as the default for modeling the residual variance.
- For the White test, the squares of the independent variables and their interactions must also be specified.
  - The I() function is required for the squares, see also the explanations on the implementation of linearized regression in R.
  - Interactions are specified as V1 * V2; if there are more than two variables, it is sufficient to specify the interaction between all variables involved, i.e. V1 * V2 * V3, the respective pairwise etc. interactions do not have to be specified separately.

R Script

```
# Breusch-Pagan test for heteroscedasticity
bptest(linreg)

# White test for heteroscedasticity
bptest(linreg, ~ I(CP91JW^2) + I(ZINSK^2) + CP91JW * ZINSK, data = B3)
```

Heteroscedasticity – Results



Output

```
        Goldfeld-Quandt test

data:  linreg
GQ = 1.1218, df1 = 72, df2 = 71, p-value = 0.6289

        studentized Breusch-Pagan test

data:  linreg
BP = 2.5355, df = 2, p-value = 0.2815

        studentized Breusch-Pagan test

data:  linreg
BP = 3.7547, df = 5, p-value = 0.5852
```

The additional 3 regression parameters required for the white test can be recognized on the basis of the degrees of freedom - in contrast to the BP test, which only shows the two original regression parameters.

Interpretation:

The scatter plot does not show a “funnel”, which indicates homoscedasticity.

The tests do not reject the null hypothesis of “homoscedasticity” (p >= 0.2815).

Overall, homoscedasticity appears to be present.

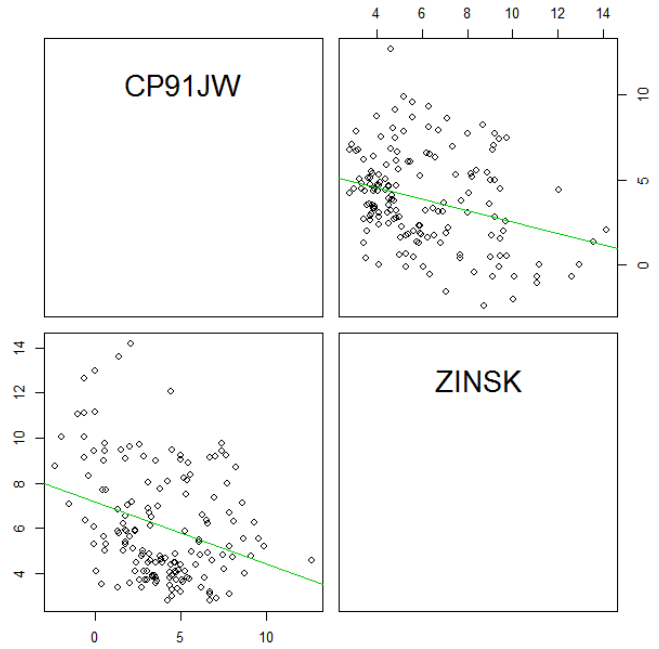
Multicollinearity – Pairwise correlations and VIF

- Pairwise scatter plots can be generated with the scatterplotMatrix() command.
- The cor() command generates the correlation matrix for the pairwise correlations.
- vif() generates the variance inflation factors.

R Script

```
# Pairwise scatter plots
scatterplotMatrix(~CP91JW+ZINSK, regLine=TRUE, smooth=FALSE,
     diagonal = FALSE, data=B3)

# Pairwise correlations
cor(B3[,c("CP91JW","ZINSK")], use="complete")

# Variance inflation factors
vif(linreg)
```

Multicollinearity – Results



Output

```
> cor(B3[,c("CP91JW","ZINSK")], use="complete")
           CP91JW      ZINSK
CP91JW  1.0000000 -0.3033194
ZINSK  -0.3033194  1.0000000

> vif(linreg)
  CP91JW    ZINSK
1.101325 1.101325
```

VIFs for both independent variables.

Correlation.

Interpretation:
The scatter plot matrix shows a negative correlation between the two independent variables. However, the corresponding correlation of -0.3 is not particularly high.

At around 1.1, the variance inflation factors are well below the threshold values of 5 and 10.

Overall, there does not appear to be any multicollinearity.

Normal distribution of the residuals – Histogram

- The hist() command generates the histogram.
- The freq= option is used to set whether absolute (freq=TRUE) or relative (freq=FALSE) frequencies are to be displayed. If the density of a normal distribution is also to be plotted in the histogram, relative frequencies are required.
- The density of the normal distribution can be added using the lines() function.

R Script

```
# Histogram without normal probability density
hist(resid(linreg), col="gray", main="Histogram for the residuals",
  ylab="Frequency", xlab="Residuals")

# Histogram with normal probability density. Generate the density
  first, then draw the histogram.
x=seq(min(resid(linreg)),max(resid(linreg)),length=200)
y=dnorm(x,mean=0,sd=1)
hist(resid(linreg), freq=FALSE, col="gray", ylim=c(0,0.5),
  main="Histogram for the residuals", ylab="Relative frequency",
  xlab="Residuals")
lines(x,y)
```

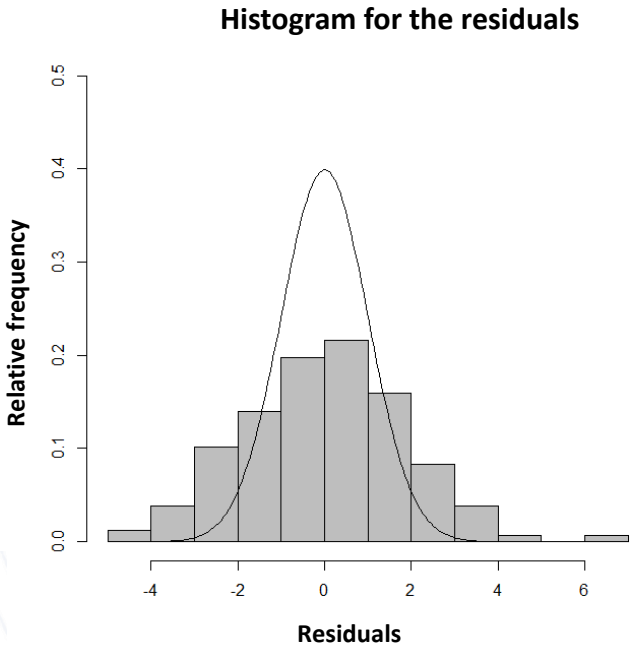
Normal distribution of the residuals – Further visualizations

- A stem and leaf diagram is generated using the stem() function.
- A normal probability plot can be generated using the qqnorm() or qqPlot() functions.
- The option dist="norm" is used to compare the observed residuals with the expected residuals of a normal distribution; the option line="quartiles" is used to add the diagonal.

R Script

```
# Stem and leaf diagram
stem(resid(linreg))

# Normal probability plot
qqPlot(resid(linreg), dist="norm", line="quartiles", envelope=FALSE)
```

Normal distribution of the residuals – Results (1/2)



Output

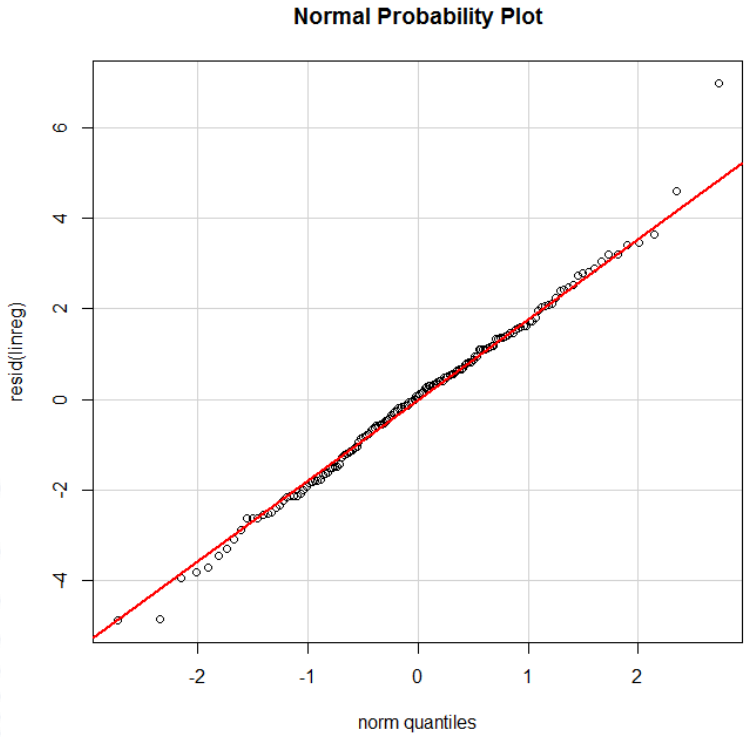
```
The decimal point is at the |
-4 | 980
-3 | 87531
-2 | 9666555442211110
-1 | 9988887665555432221110
-0 | 998888776666655543333222211100
 0 | 11122233333344445555566677788889
 1 | 00111112223344445556666778
 2 | 00111244557889
 3 | 022456
 4 | 6
 5 |
 6 |
 7 | 0
```

The vertical line marks the decimal point.
Reading example: The three residuals with the largest negative deviation from the regression line have the values -4.9, -4.8 and -4.0.

Interpretation:
The histogram indicates a right skewed distribution and thus not a normal distribution.

The stem and leaf diagram reveals that the skewness to the right can be attributed to an outlier (value for the residual of 7.0).

Normal distribution of the residuals – Results (2/2)



Interpretation:
The QQ plot shows only minor systematic deviations from the diagonal.

Overall, a normal distribution of the residuals can be assumed.

Exercise

Perform regression diagnostics with the tips data set.

Note: Many of the following commands come from the lmtest package.

R commands: lm(), plot(), resid(), cooks.distance(), hat(model.matrix(), resettest(), dwtest(), bgtest(), bptest(), hist(), stem(), qqnorm(), qqline()

Exercise

Perform multiple regression including regression diagnostics with the tips data set.

Note: Many of the following commands come from the lmtest package.

R commands: lm(), plot(), resid(), cooks.distance(), hat(model.matrix(), resettest(), dwtest(), bgtest(), bptest(), scatterplotMatrix() aus dem Paket car, cor(), vif(), hist(), stem(), qqnorm(), qqline()