Statistics II │ 1

Introduction to statistics, Linear regression model

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

12 September 2022 (updated 16 September 2022)

▶ Probability theory *(teorie pravděpodobnosti)*:

branch of mathematics, deals with the description of random phenomena and experiments; provides theoretical ideas, definitions, derivations, assertions and proofs for describing and working with random phenomena and experiments.

▶ (Mathematical) Statistics *(matematická statistika)*:

deals with the collection, organization, analysis, interpretation and presentation of data; uses the tools of probability theory.

▶ Statistical / machine learning *(statistické / strojové učení)*

Descriptive statistics *(popisná statistika)*:

- ▶ summarizes the sample using summary statistics and indices;
- ▶ frequency tables, sample moments (mean, variance, standard deviation, skewness, kurtosis) and quantiles (median, quartiles, IQR), contingency tables, sample correlation.

Exploratory Data Analysis *(exploratorní analýza dat)*:

- ▶ analysis of the data, usually using visualization methods;
- ▶ frequency plot, boxplot, histogram, scatter plot, QQ plot.

Statistical inference *(statistická inference)*:

- ▶ derives probabilistic properties (parameters or probability distribution) of the population based on the analysis of the data sample;
- ▶ requires the so-called model – assumptions about the population and the sample;
- ▶ estimates of parameters – point and interval (confidence intervals), testing statistical hypothesis, prediction, classification, clustering.

Parametric methods:

- ▶ the model assumes a probability distribution or some class of them, parameters of these distributions are estimated;
- ▶ most of *classical* methods, e. g., *t*-test, linear regression model, multiple regression model, generalized linear models, analysis of variance, correlation analysis.

Nonparametric methods:

- ▶ millimalistic assumptions for the model are specified, no specific probability distribution is required;
- ▶ e. g., rank statistics and test and corresponding variants of ANOVA, correlation analysis; Functional Data Analysis.

Semiparametric methods:

- ▶ a combination of both approaches;
- ▶ e. g., Cox model of proportional hazards in survival analysis.

Stevens, Stanley Smith (1946). On the Theory of Scales of Measurement. Science 103 (2684), 677–680.

Nominal data *(nominální data)*:

- ▶ defined operations: $=$, $\neq$, classification, set membership;
- ▶ categorial data, discrete, in *R*: `factor`;
- ▶ values cannot be compared and ordered;
- ▶ e. g., blood type;
- ▶ for two categories: dichotomous data, often TRUE/FALSE or 1/0;
- ▶ dummy variable $1/0$ encodes membership of a specific category.

Ordinal data *(ordinální data)*:

- ▶ additionally defined: order, rank;
- ▶ additional operations: $<$, $>$, comparison, sorting;
- ▶ categorial data, discrete, in *R*: `ordered factor`;
- ▶ distance between values cannot be quantified;
- ▶ e. g., the highest education attained, achieved grade in a course.

Interval data *(intervalová data)*:

- ▶ additionally defined: distance;
- ▶ additional operations: $+$, $-$;
- ▶ typically continuous numerical data, in *R*: `numeric`;
- ▶ ratio of values cannot be quantified, zero is not correctly defined;
- ▶ e. g., temperature in $^\circ$C.

Ratio data *(poměrová data)*:

- ▶ additionally defined: ratio;
- ▶ additional operations: $*$, $/$;
- ▶ typically continuous numerical data, in *R*: `numeric`;
- ▶ all physical variables in accordance with SI, e. g., temperature in K.

Continuous data can be treated (with a certain loss of information) as well as discrete data: we divide the data into intervals, which further play the role of categories, so-called interval data.

- ▶ H_0: Null hypothesis *(nulová hypotéza)*, the statement being tested,
- ▶ H_1: Alternative hypothesis *(alternativní hypotéza)*, the statement being tested against H_0,
- ▶ statistical test **rejects** *(zamítne)*, or **does not reject** *(nezamítne)*, H_0 in favor of H_1.

	H_1 is true	H_0 is true
test rejects H_0 in favor of H_1	*True Positive* right decision $P = 1 - \beta$	*False Positive* Type I error $P = \alpha$
test does not reject H_0 in favor of H_1	*False Negative* Type II error $P = \beta$	*True Negative* right decision $P = 1 - \alpha$

- ▶ Level of significance of the test $= \alpha = P(H_0 \text{ rejected} \mid H_0 \text{ true})$
- ▶ Power of the test $= 1 - \beta = P(H_0 \text{ rejected} \mid H_1 \text{ true})$
- ▶ Cannot be ensured both $\alpha = 0$ and $\beta = 0$, even $\alpha, \beta \to \min$.
- ▶ General methodology – Neyman-Pearson lemma: Good criterion for the selection of hypotheses is a likelihood ratio.

▶ *Classically*, using critical region *(kritický obor)*:

1. Set up H_0 a H_1,

2. choose a model corresponding to data annd hypotheses,

3. choose suitable test and test statistic T with known probability distribution under H_0,

4. decide about α, β and sample size, in our case $\alpha = 0.05$,

5. calculate the observed value t of the test statistic T,

6. calculate critical region W corresponding to t and H_1,

7. H_0 is rejected in favor of H_1, if and only if $t \in W$.

▶ Using p-value *(p-hodnota)*, a common method nowadays, especially in software:

6. calculate p-value p of the test,

7. H_0 is rejected in favor of H_1, if and only if $p < \alpha$.

▶ Using $100(1 - \alpha)\%$ confidence interval *(interval spolehlivosti)* for suitable parameter.

Definition (p-value)

P-value p is the (highest) probability that, under validity of H_0, the test statistic T exhibits an equally or more extreme value than the value t observed on the test sample.

$$p = 2 \min\{P(T \geq t), P(T \leq t)\}; \text{ or } p = P(T \geq t); \text{ or } p = P(T \leq t);$$

according to the variant of H_1 (two-sided or one-sided).

▶ *P*-value is just a tool for deciding whether or not to reject H_0 in favor of H_1; it does not quantify the significance of the observed effect.

▶ If the test is performed correctly, it is ensured that $P(\text{Type I error}) \leq \alpha$.

▶ Neither $p = P(H_0)$ nor $p = P(\overline{H_1})$.

▶ The power of a test can usually be increased using larger sample.

▶ Non-rejection of H_0 does not imply H_0 is true.
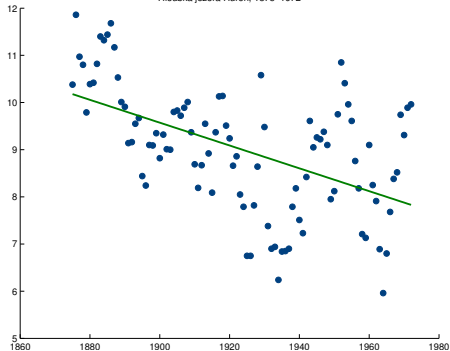
Linear regression model

- ▶ Examination of the relationship between two numerical quantities, non-random independent variable x and observed random variable Y.
- ▶ Data: pairs (x_i, Y_i), $i = 1, \ldots, n$.
- ▶ Regression model *(regresní model)*:
$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$
- ▶ x_i = known points (vectors) of fixed plan *(pevný plán)*,
- ▶ Y_i = observed (measured) values,
- ▶ $m(x)$ = regression function *(regresní funkce)* in the form of a function linear in parameters,
- ▶ ε_i = measurement errors, $\mathsf{E}(\varepsilon_i) = 0$, $\mathsf{Var}(\varepsilon_i) = \sigma^2$, $\mathsf{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
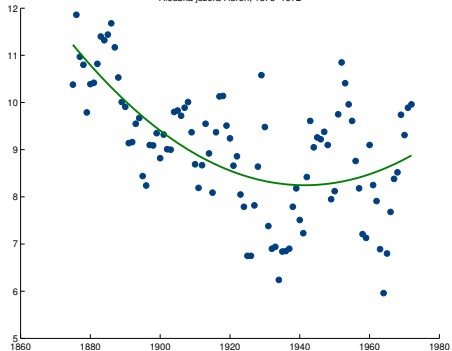- ▶ Task: given data (x_i, Y_i), find the *suitable* regression function $m(x)$.

$m(x) = \beta_0 + \beta_1 x$

Hloubka jezera Huron, 1875–1972

$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

Hloubka jezera Huron, 1875–1972

Assume the regression function

$$m(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{il} = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij}, \quad i = 1, \ldots, n,$$

which is a linear function of unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

Linear regression model

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1l} + \varepsilon_1, \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nl} + \varepsilon_n,
\end{aligned}
$$

written in matrix form as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1l} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nl} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}, \quad \text{i. e.,} \quad Y = X\beta + \varepsilon.$$

$$Y = X\beta + \varepsilon$$

- ▶ $\beta = (\beta_0, \beta_1, \ldots, \beta_l)'$ = vector of $k = l + 1$ regression coefficients *(regresní koeficienty)*,
- ▶ X = regression / design matrix *(matice plánu)* consists of $(n \times k)$ nonrandom numbers x_{ij}, regressors / predictors *(regresory / prediktory)*,
- ▶ $n > k$,
- ▶ $r(X) = k = l + 1$, i. e., the design matrix has full rank *(plná hodnost)*, its columns are linearly independent.

Random errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$:

- ▶ are nonsystematic: $E(\varepsilon_i) = 0$, i.e., $E(\varepsilon) = 0$ and $E(Y) = X\beta$,
- ▶ have homogeneous variance: $\text{Var}(\varepsilon_i) = \sigma^2 > 0$,
- ▶ are mutually uncorrelated: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$;
- ▶ variance-covariace matrix *(kovarianční matice)* of the vector of observations is $\text{Var}(Y) = \text{Var}(\varepsilon) = \sigma^2 I_n$.
- ▶ Hence, observations are uncorrelated and have homogeneous variance.

Optimization: find such $\boldsymbol{\beta}$ which minimizes the sum of quadratic deviations,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Big[Y_i - \beta_0 - \sum_{j=1}^{l}\beta_j x_{ij}\Big]^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \longrightarrow \min.$$

► Ordinary Least Squares (OLS) estimate *(odhad metodou nejmenších čtverců)*

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_l) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

► predicted / fitted values

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \quad \text{i. e.,} \quad \widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{l}\widehat{\beta}_j x_{ij},$$

► residuals *(rezidua)* $r_i = Y_i - \widehat{Y}_i$,

► residual sum of squares *(reziduální součet čtverců)*

$$S_e = S(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sum_{i=1}^{n}\Big[Y_i - \widehat{\beta}_0 - \sum_{j=1}^{l}\widehat{\beta}_j X_{ij}\Big]^2 = \sum_{i=1}^{n}r_i^2.$$

Theorem (Gaussov-Markovov)

OLS estimate $\widehat{\beta}_{\mathsf{OLS}}$ is BLUE = Best Linear Unbiased Estimate *(nejlepší nestranný lineární odhad)* of vector β and its variance-covariance matrix *(kovarianční matice)* is $\mathsf{Var}(\widehat{\beta}_{\mathsf{OLS}}) = \sigma^2 \, (X'X)^{-1}$.

Theorem

Fitted values $\widehat{Y} = HY$,   residual sum of squares $S_e = Y'(I_n - H)Y$,

where $H = X\,(X'X)^{-1}\,X'$ is so-called hat matrix.
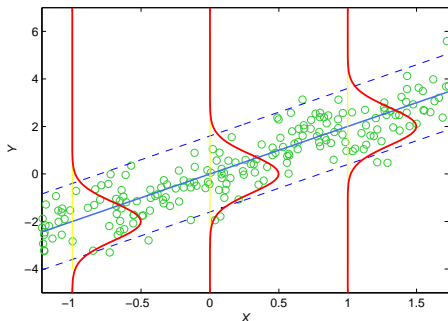

Theorem

$$\widehat{\sigma^2}_{\mathsf{OLS}} = \frac{S_e}{n-l-1} = \frac{S_e}{n-k}$$

is an unbiased estimate of the variance σ^2 of random errors.

Additionally, let us assume that the observations have n-dimensional gaussian (normal) disribution

$$\boldsymbol{Y} \sim \mathsf{N}_n\Big(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\Big).$$

Theorem

▶ OLS estimate has gaussian distribution, $\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} \sim \mathsf{N}_k\Big(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\Big)$,

▶ statistic $K = (n-k)\dfrac{\widehat{\sigma^2}_{\mathsf{OLS}}}{\sigma^2} \sim \chi^2(n-k)$ has chi-square distribution,

▶ OLS estimate $\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}$ and statistic K are independent.

H_0: $\beta_j = 0$, i.e., regression coefficient β_j is not significant,
H_1: $\beta_j \neq 0$, i.e., regression coefficient β_j is significant

Under H_0, test statistic

$$T_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma^2}_{\text{OLS}} \, (\boldsymbol{X'X})_{jj}^{-1}}}$$

has Student $t_{1-\alpha/2}(n-k)$ probability distribution.

H_0 is rejected at the level of significance α, if $|T_j| \geq t_{1-\alpha/2}(n-k)$.

$100(1-\alpha)\%$ confidence interval for regression coefficient β_j is

$$\left(-\sqrt{\widehat{\sigma^2}_{\text{OLS}} \, (\boldsymbol{X'X})_{jj}^{-1}} \, t_{1-\alpha/2}(n-k), \quad \sqrt{\widehat{\sigma^2}_{\text{OLS}} \, (\boldsymbol{X'X})_{jj}^{-1}} \, t_{1-\alpha/2}(n-k) \right).$$

H_0: $\beta_1 = \cdots = \beta_l$,
H_1: $\exists j \in \{1, \ldots, l\} : \beta_j \neq 0$, i.e., at least one β_j is significant

Note that the intercept β_0 is not included in these hypotheses.

Under H_0, test statistic

$$F = \frac{1}{k-1} \cdot \frac{S_{\widehat{Y}}}{\widehat{\sigma^2}_{\mathsf{OLS}}} = \frac{n-k}{k-1} \cdot \frac{S_{\widehat{Y}}}{S_e}$$

where $S_{\widehat{Y}} = \sum\limits_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$ is regression sum of squares,

has Fisher-Snedecor $F(k-1, n-k)$ probability distribution.

H_0 is rejected at the level of significance α, if $F \geq F_{1-\alpha}(k-1, n-k)$.

Definition

Coefficient of determination *(index determinace)* R squared:

$$R^2 = \frac{S_{\widehat{Y}}}{S_T} = 1 - \frac{S_e}{S_T},$$

where $S_{\widehat{Y}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$ and $S_T = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ is total sum of squares.

Adjusted *(korigovaný)* coefficient of determination R bar squared:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2).$$

▶ If the predicted values exactly match the observed values, then $R^2 = 1$.

▶ Linear regression model with only the intercept β_0 has $R^2 = 0$.

▶ R^2 quantifies the fraction of the variance in the data which is explained by the linear regression model.
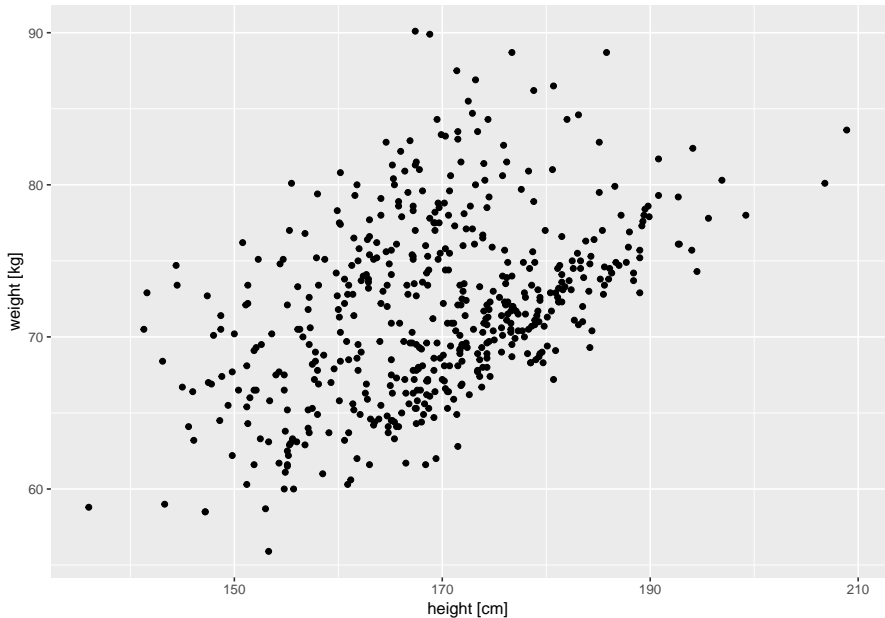
Data

```
'data.frame': 528 obs. of  21 variables:
 $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
 $ Height  : num  173 160 165 164 161 ...
 $ Weight  : num  67.7 71.3 66.3 65.5 60.3 69 75 69.6 74.7 76.1 ...
```

```
m1 <- lm(Weight ~ Height, data = dt) # or
m1 <- lm(Weight ~ 1 + Height, data = dt)
summary(m1)
```

```
Call:
lm(formula = Weight ~ Height, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-12.277  -3.832  -1.121   3.487  18.685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.9708     3.4759    9.485   <2e-16 ***
Height       0.2296     0.0205   11.201   <2e-16 ***

Residual standard error: 5.26 on 526 degrees of freedom
Multiple R-squared:  0.1926,	Adjusted R-squared:  0.1911
F-statistic: 125.5 on 1 and 526 DF,  p-value: < 2.2e-16
```

OLS estimate with 95% confidence and prediction intervals
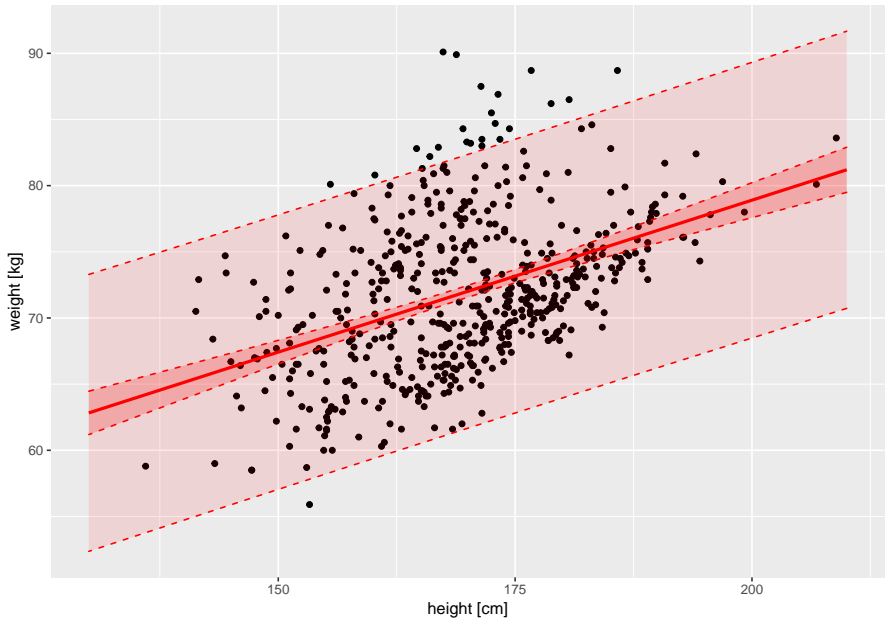
▶ value of the regression function – point estimate of the value of the regression function $m(x)$ for given x;
  predict(..., type = "none")

▶ confidence interval *(interval spolehlivost pro hodnoty regresní funkce)* = interval estimate of the value of the regression function $m(x)$ for given x;
  predict(..., type = "confidence")

▶ confidence band *(pás spolehlivosti kolem regresní funkce)* – band estimate for the whole regression function $m(x)$

▶ prediction *(predikce pozorování)* – point estimate $\widehat{Y}(x)$ for given x;
  predict(..., type = "none")

▶ prediction interval *(predikční interval)* – for the predicted observation $\widehat{Y}(x)$ for given x;
  predict(..., type = "prediction")

Example: weight vs. height of people
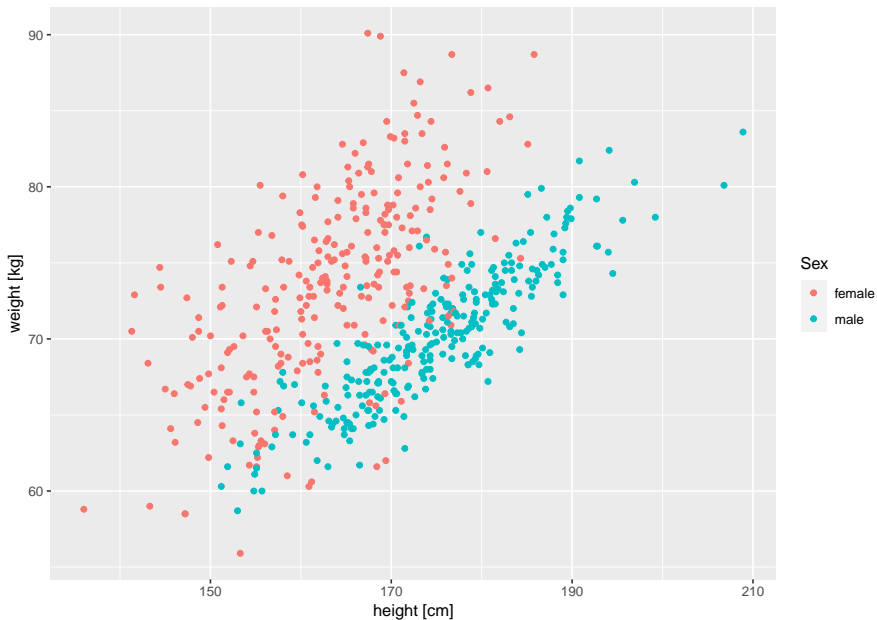
```
m3 <- lm(Weight ~ 1 + Height + Sex, data = dt)
```

```
Call:
lm(formula = Weight ~ 1 + Height + Sex, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-14.1307 -2.0173  0.0736  1.9999 14.8114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.79944    2.88700   1.662    0.097 .
Height       0.42108    0.01761  23.905   <2e-16 ***
Sexmale     -8.29905    0.39340 -21.096   <2e-16 ***

Residual standard error: 3.873 on 525 degrees of freedom
Multiple R-squared:  0.563,  Adjusted R-squared:  0.5614
F-statistic: 338.2 on 2 and 525 DF,  p-value: < 2.2e-16
```
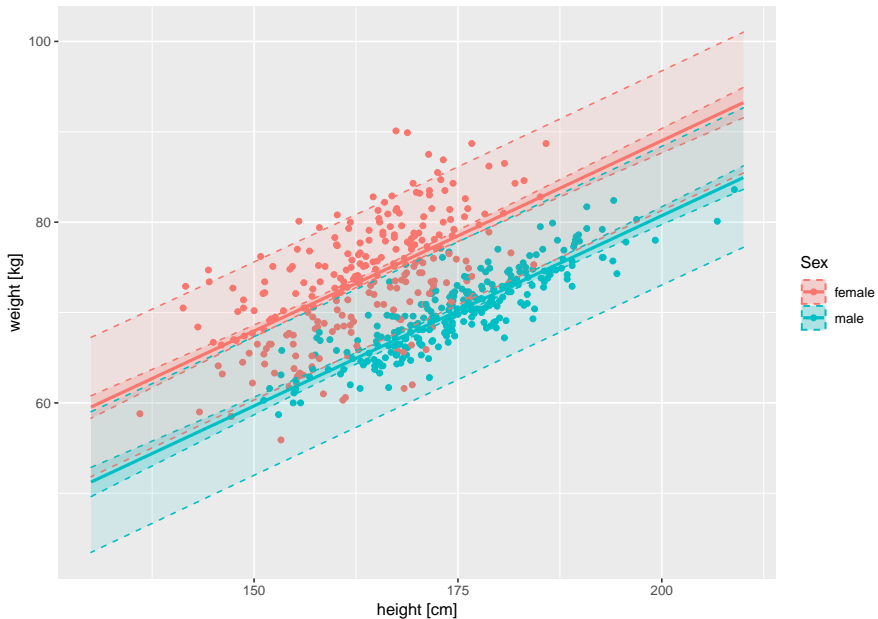
OLS estimate with 95% confidence and prediction intervals

random variable	*náhodná veličina*	X
random sample	*máhodný výběr*	$\mathbf{X}$
mean	*střední hodnota*	$\mathsf{E}(X);\ \mu_X$
variance	*rozptyl*	$\mathsf{Var}(X);\ \sigma_X^2$
standard deviation	*směrodatná odchylka*	σ_X
variance-covariance matrix	*kovarianční matice*	$\mathsf{Var}(\mathbf{X})$
sample mean	*výběrový průměr*	$\overline{X}$
sample variance	*výběrový rozptyl*	S_X^2
sample standard deviation	*výběrová sm. odchylka*	S_X
statistic	*statistika*	$T(\mathbf{X})$
probability distribution	*rozdělení pravděpodobnosti*	$\mathsf{N};\mathsf{t};\mathsf{F};\chi^2;\dots$
quantile	*kvantil*	
median	*medián*	
estimate	*odhad*	$\widehat{\mu};\widehat{\sigma}^2;\dots$
confidence interval	*interval spolehlivosti*	
prediction interval	*predikční interval*	

Statistics II │ 2

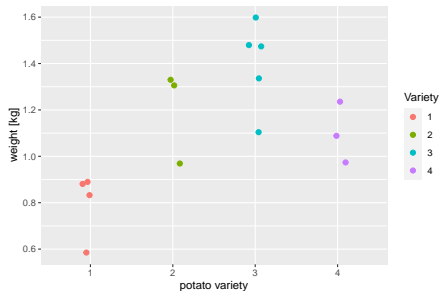Analysis of variance (ANOVA)

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

19 September 2022

One-way (single-factor) ANOVA

Analysis of 4 varieties of potatoes based on the weights of the clusters of potato tubers.

variety	weight [kg]
1	0.9, 0.8, 0.6, 0.9
2	1.3, 1.0, 1.3
3	1.3, 1.5, 1.6, 1.1, 1.5
4	1.1, 1.2, 1.0

At a 5% significance level, test the null hypothesis that the mean weight of a cluster of potatoe tubers does not depend on the variety. If you reject the null hypothesis, find which pairs of varieties significantly differ.

▶ variety: grouping factor – categorical, nominal or ordinal type
▶ weight: observed random variable – numerical, interval or ratio type

- The factor A has $a \geq 3$ levels.
- The ith level has n_i observations $Y_{i1}, \ldots, Y_{i n_i}$, which form a random sample from $N(\mu_i, \sigma^2)$ probability distribution, $i = 1, \ldots, a$.
- Y_{ij}: first index – group by the level of the factor, second index – order in the group.
- The particular random samples are stochastically independent.
- Model of one-way ANOVA *(jednofaktorová analýza rozptylu / jednoduché třídění)*:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where ε_{ij} are stochastically independent random variables with $N(0, \sigma^2)$ probability distribution, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$.

level	count	observations	sum	average	distribution
1	n_1	$\boldsymbol{Y}_1 = (Y_{11}, \ldots, Y_{1n_1})'$	$Y_{1\cdot} = \sum\limits_{j=1}^{n_1} Y_{1j}$	$\overline{Y}_{1\cdot} = \frac{1}{n_1} Y_{1\cdot}$	$Y_{1j} \sim N(\mu_1, \sigma^2)$
2	n_2	$\boldsymbol{Y}_2 = (Y_{21}, \ldots, Y_{2n_2})'$	$Y_{2\cdot} = \sum\limits_{j=1}^{n_2} Y_{2j}$	$\overline{Y}_{2\cdot} = \frac{1}{n_2} Y_{2\cdot}$	$Y_{2j} \sim N(\mu_2, \sigma^2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	n_a	$\boldsymbol{Y}_a = (Y_{a1}, \ldots, Y_{an_a})'$	$Y_{a\cdot} = \sum\limits_{j=1}^{n_a} Y_{aj}$	$\overline{Y}_{a\cdot} = \frac{1}{n_a} Y_{a\cdot}$	$Y_{aj} \sim N(\mu_a, \sigma^2)$
	n		$Y_{\cdot\cdot} = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} Y_{ij}$	$\overline{Y}_{\cdot\cdot} = \frac{1}{n} Y_{\cdot\cdot}$	

Dot – summing over the index, overline – averaging.

At a significance level of α, we test H_0 against H_1,

H_0: all levels of the factor have equal means,

H_1: at least one pair of levels has different means.

- ▶ It is a generalization of the two-sample t-test.
- ▶ Note that it is not the same as to apply the two-sample t-test to each of $a(a-1)/2$ pairs of levels. This so-called multiple testing problem) does not guarantee that $P(\text{Type I error}) \leq \alpha$.
- ▶ Significance level corrections (e.g., Bonferroni correction) are not feasible for a large number of levels.
- ▶ In the 1930s, R. A. Fisher introduced the analysis of variance (ANOVA) *(analýza rozptylu)*, which guarantees $P(\text{Type I error}) = \alpha$.

If the null hypothesis H_0 is rejected, we are further interested in finding which pairs of levels have significantly different means. The so-called multiple comparison *(mnohonásobné porovnávání)* methods are used for this:

- ▶ Tukey's method – preferred if all groups have similar sample sizes;
- ▶ Scheffé's method – preferred if sample sizes are considerably different.

Definition (M_A, one-way ANOVA model)

Observations Y_{ij} follow model M_A, if

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \varepsilon_{ij},$$

for $i = 1, \ldots, a, j = 1, \ldots, n_i$,

where ε_{ij} are *i.i.d.* random variables with $N(0, \sigma^2)$ probability distribution,

- μ = overall / grand mean *(střední hodnota)* of random variable Y,
- α_i = the effect *(efekt)* of the ith level of factor A,
- $\mu_i = \mu + \alpha_i$ = mean of Y by the ith level of factor A,
- ε_{ij} = random errors.

Equivalent expressions of the hypotheses

$H_0: \alpha_1 = \cdots = \alpha_a = 0,$

$H_1: \exists i \in \{1, \ldots, a\} : \alpha_i \neq 0$

$H_0: \mu_1 = \cdots = \mu_a,$

$H_1: \exists i, j \in \{1, \ldots, a\} : \mu_i \neq \mu_j$

Definition (M_0, minimal / null model)

Under H_0, observations Y_{ij} follow model M_0, a submodel of M_A,

$$Y_{ij} = \mu + \varepsilon_{ij}$$

Model M:

$$Y = X\beta + \varepsilon = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \vdots & & \ddots & \mathbf{1}_{n_{a-1}} & \mathbf{0} \\ \mathbf{1}_{n_a} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}_{n_a} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Solving the system of *normal equations*: $X'X\beta = X'Y$:

$$X'X = \begin{pmatrix} n & n_1 & n_2 & \cdots & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & \cdots & 0 \\ n_2 & 0 & n_2 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ n_{a-1} & \vdots & & \ddots & n_{a-1} & 0 \\ n_a & 0 & \cdots & \cdots & 0 & n_a \end{pmatrix}, X'Y = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \cdots & \mathbf{1}'_{n_{a-1}} & \mathbf{1}'_{n_a} \\ \mathbf{1}'_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{n_2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{1}'_{n_{a-1}} & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}'_{n_a} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_{a-1} \\ Y_a \end{pmatrix} = \begin{pmatrix} Y_{..} \\ Y_{1.} \\ \vdots \\ \vdots \\ Y_{a-1.} \\ Y_{a.} \end{pmatrix}.$$

The design matrix X is not of full rank *(plné hodnosti)*. (Calculate its rank.)

Least-squares estimate of the vector of parameters in linear regression model:
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$
But the design matrix $\boldsymbol{X}$ is not of full rank, thus $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist.

Any pseudoinverse matrix *(psudoinverzní matice)* $(\boldsymbol{X}'\boldsymbol{X})^-$ can be used instead, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$, e.g., $(\boldsymbol{X}'\boldsymbol{X})^- = \text{diag}\left(0, \frac{1}{n_1}, \ldots, \frac{1}{n_a}\right)$; or one additional equation is necessary.

Usually, the additional equation is
$$\sum_{i=1}^{a} n_i \alpha_i = 0,$$
leading to following estimators:

overall (grand) mean:	$\widehat{\mu} = \overline{Y}_{\bullet\bullet}$
effects (A):	$\widehat{\alpha}_i = \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
mean of group ($A = i$):	$\widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i = \overline{Y}_{i\bullet}$
- in model M_0:	$\widehat{\mu}_i = \widehat{\mu} = \overline{Y}_{\bullet\bullet}$

▶ Total sum of squares *(celkový součet čtverců)*
= variability of observations around the overall mean:

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{..}\right)^2 \qquad \sim \chi^2(df_T = n - 1),$$

▶ Between-groups sum of squares *(regresní součet čtverců)*
= variability of group means around the overall mean, i.e., explained by factor A:

$$S_A = \sum_{i=1}^{a} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 \qquad \sim \chi^2(df_A = a - 1),$$

▶ Within-groups / error / residual sum of squares *(reziduální součet čtverců)*
= variability of observations within each group around the group mean, i.e., unexplained by factor A:

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i.}\right)^2 \qquad \sim \chi^2(df_e = n - a).$$

The *df* quantities are degrees of freedom *(stupně volnosti)* of the statistics.

Theorem

$$S_T = S_A + S_e$$

The testing in one-way ANOVA relies on comparison of models M and M_0.

Theorem (Omnibus ANOVA F-test)

$$F_A = \frac{MS_A}{MS_e} = \frac{\dfrac{S_A}{df_A}}{\dfrac{S_e}{df_e}} = \frac{\dfrac{S_A}{a-1}}{\dfrac{S_e}{n-a}} = \frac{\dfrac{S_T - S_e}{df_T - df_e}}{\dfrac{S_e}{df_e}} = \left(\frac{S_T}{S_e} - 1\right)\frac{n-a}{a-1}.$$

Under the null hypothesis H_0, i.e, if M_0 is correct,
the statistic F_A has Fisher-Snedecor F$(a-1, n-a)$ probability distribution
with $(a-1)$ and $(n-a)$ degrees of freedom.

The null hypothesis H_0 is rejected, i.e., factor A is not significant, if

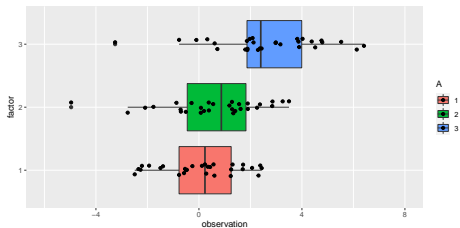$$F_A \geq F_{1-\alpha}(a-1, n-a).$$

Why the test statistic F_A has Fisher-Snedecor probability distribution?
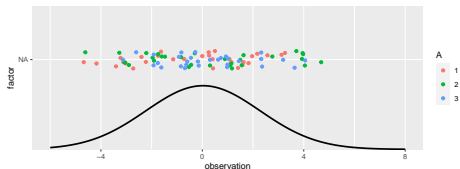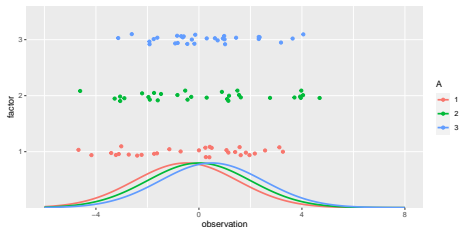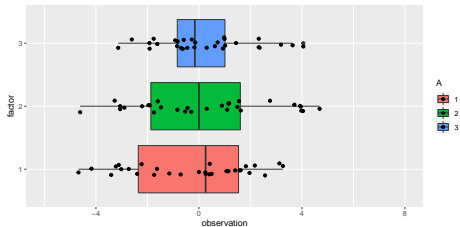
$$n_1 = n_2 = n_3 = 30,$$

$$\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 0.5, \sigma^2 = 4$$

```
     S.T         S.A        S.e
450.6129  6.443252  444.1697

           Df  Sum Sq  Mean Sq  F value
A           2     6.4    3.222    0.631
Residuals  87   444.2    5.105
```
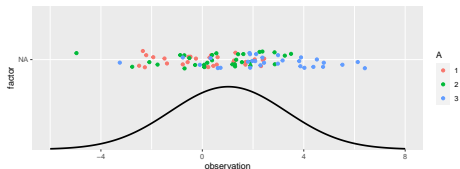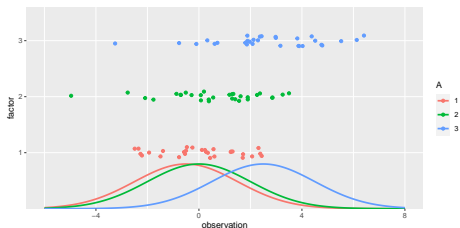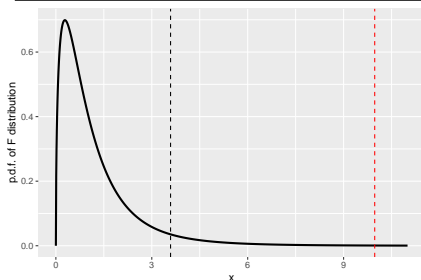
$$n_1 = n_2 = n_3 = 30,$$

$$\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 2.5, \sigma^2 = 4$$

```
       S.T        S.A        S.e
560.0406 218.7568 341.2838
```

	Df	Sum Sq	Mean Sq	F value
A	2	218.8	109.38	27.88
Residuals	87	341.3	3.92	

source of variability	degrees of freedom	sum of squares	mean squares	value of the test statistic	p-value
group	$df_A = a - 1$	S_A	$MS_A = \dfrac{S_A}{df_A}$	$F_A = \dfrac{MS_A}{MS_e}$	p_A
residual	$df_e = n - a$	S_e	$MS_e = \dfrac{S_e}{df_e}$		
total	$df_T = n - 1$	S_T			

Example – potatoes:

```
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Variety     3   0.816  0.27200    9.973  0.0018
Residuals  11   0.300  0.02727
```

For chosen $k \neq l$, we test the equality of means in the kth and lth level:

H_0: $\mu_k = \mu_l$,
H_1: $\mu_k \neq \mu_l$

Theorem (Tukey's method)

H_0 is rejected at the level of significance α, if

$$\left| \overline{Y}_{k\cdot} - \overline{Y}_{l\cdot} \right| \geq \sqrt{\frac{S_e}{(n-a)n_k}} \, q_{1-\alpha}(a, n-a),$$

where $q_{1-\alpha}(a, n-a)$ are quantiles (numerically computed) of the studentized range.

Theorem (Scheffé's method)

H_0 is rejected at the level of significance α, if

$$\left| \overline{Y}_{k\cdot} - \overline{Y}_{l\cdot} \right| \geq \sqrt{S_e \frac{a-1}{n-a} \left(\frac{1}{n_k} + \frac{1}{n_l} \right) F_{1-\alpha}(a-1, n-a)}.$$

Different parametrization of the one-way ANOVA is used by most of the statistical software (including *R*).

Definition

Random variables Y_{ij} follow the model

$$Y_{ij} = \mu^* + \alpha_i^* + \varepsilon_{ij},$$

for $i = 1, \ldots, a, j = 1, \ldots, n_i$,
where ε_{ij} are *i.i.d.* random variables with $N(0, \sigma^2)$ probability distribution,

▶ $\mu^* = \mu_1$ = mean of the first level of factor A, i.e., $\alpha_1^* = 0$,

▶ α_i^* = the effect of the ith level of factor A, $\alpha_1^* = 0$ is fixed,

▶ $\mu_i = \mu^* + \alpha_i^*$ = mean of Y by the ith level of factor A.

Equivalent expressions of the hypotheses

$$H_0: \alpha_2^* = \cdots = \alpha_a^* = 0,$$
$$H_1: \exists i \in \{2, \ldots, a\} : \alpha_i^* \neq 0$$

To verify the homogeneity of variances, i.e. to verify the consistency of variances in individual levels of the factor, we use

▶ Levene's test,
▶ Bartlett's test.

To verify the normality of the observations in each group, we use

▶ normal QQ-plot,
▶ Lilliefors test,
▶ Shapiro–Wilk test.

Theorem (Levene's test)

Denote $Z_{ij} = |Y_{ij} - \overline{Y}_{i\cdot}^*|$, where $\overline{Y}_{i\cdot}^*$ is sample mean / median / 10% trimmed mean. Under the hypothesis of homogeneity of variances, test statistic

$$L = \frac{\frac{1}{a-1} \sum\limits_{i=1}^{a} n_i \left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)^2}{\frac{1}{n-a} \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i\cdot}\right)^2} \text{ has } \mathsf{F}(a-1, n-a) \text{ probability distribution.}$$

The homogeneity of variances is rejected at the level of significance α, if $L \geq F_{1-\alpha}(a-1, n-a)$.

Theorem (Bartlett's test)

Under the hypothesis of homogeneity of variances, test statistic

$$B = \frac{1}{C}\left[(n-a)\ln\frac{S_e}{n-a} - \sum_{i=1}^{a}(n_i-1)\ln S_i^2\right] \text{ has asymptotically } \chi^2(a-1)$$
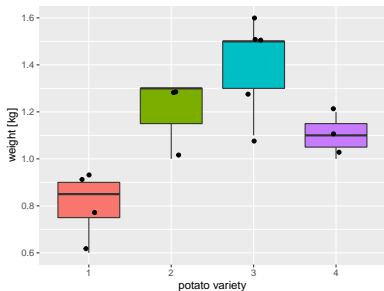
probability distribution. Here, S_j^2 denotes the sample variance in the ith level of the factor and $C = 1 + \frac{1}{3(a-1)}\left(\sum\limits_{j=1}^{a}\frac{1}{n_j-1} - \frac{1}{n-a}\right)$. The homogeneity of variances is rejected at the level of significance α, if $B \geq \chi^2_{1-\alpha}(a-1)$.

```
'data.frame': 15 obs. of  2 variables:
 $ Weight : num  0.9 0.8 0.6 0.9 1.3 1 1.3 1.3 1.5 1.6 ...
 $ Variety: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 2 2 3 3 3 ...
```
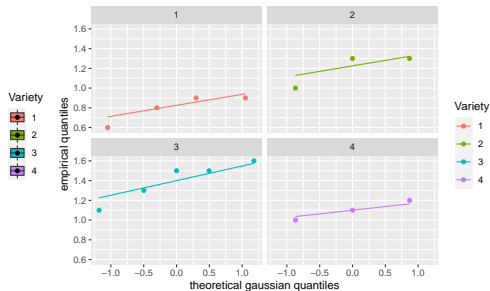


```
   Bartlett test of homogeneity of variances
data:  Weight by Variety
Bartlett's K-squared=1.0417, df=3, p-value=0.7912

Levene's Test for Homogeneity of Variance (center=median)
      Df F value Pr(>F)
group  3  0.1874  0.9027

   Variety Shapiro.p.value
   <fct>            <dbl>
1 1                0.161
2 2                0
3 3                0.440
```

```
# using "aov"
aov.model <- aov(Weight~Variety, data=dt)
# or using "lm" and "anova"
M.A <- lm(Weight~Variety, data=dt)
anova.model <- anova(M.A)
```

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Variety	3	0.816	0.27200	9.973	0.0018	**
Residuals	11	0.300	0.02727			

Estimates of coefficients of the linear regression model

```
(Intercept)    Variety2    Variety3    Variety4
      0.8         0.4         0.6         0.3
```

Estimates of effects and means

```
                                              Tables of means
                                              Grand mean
Tables of effects                             1.14
 Variety                                       Variety
      1    2    3    4                              1   2   3   4
   -0.34 0.06 0.26 -0.04                         0.8 1.2 1.4 1.1
rep  4.00 3.00 5.00 3.00                       rep 4.0 3.0 5.0 3.0
```

Tukey's method:

```
     diff          lwr         upr      p adj
2-1  0.4   0.02040199  0.77959801  0.0381806
3-1  0.6   0.26659524  0.93340476  0.0010299
4-1  0.3  -0.07959801  0.67959801  0.1391459
3-2  0.2  -0.16296512  0.56296512  0.3885221
4-2 -0.1  -0.50580735  0.30580735  0.8783019
4-3 -0.3  -0.66296512  0.06296512  0.1172041
```


Significantly different: 1−2 and 1−3
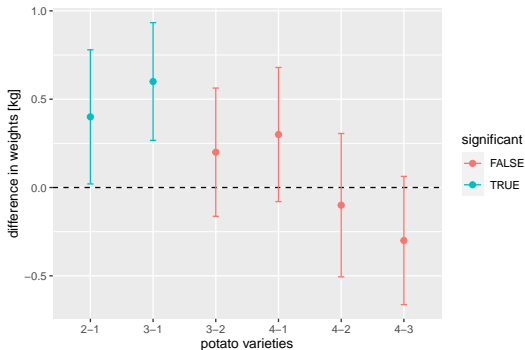
Scheffé's method:

```
  Weight groups
3   1.4      a
2   1.2      ab
4   1.1      ab
1   0.8      b
```

Significantly different: 1−3

Using ANOVA to compare nested linear regression models

Assume two linear regression models for the **same data** of size n:

1. model M_1 with design matrix X_1 with rank $r_1 = r(X_1)$
   and residual sum of squares S_{e1};
2. **submodel** M_2 of model M_1 with design matrix X_2 with rank $r_2 = r(X_2)$
   which is formed by omitting some columns of X_1,
   and with residual sum of squares S_{e2}.

— **Asumming the validity of model M_1, we further test** —

H_0: model M_2 is valid, too, i.e., M_1 can be simplified to M_2,
H_1: model M_2 is not valid.

Under the null hypothesis H_0, test statistic

$$F = \frac{\frac{S_{e2}-S_{e1}}{r_1 - r_2}}{\frac{S_{e1}}{n - r_1}} = \frac{S_{e2} - S_{e1}}{S_{e1}} \cdot \frac{n - r_1}{r_1 - r_2} = \left(\frac{S_{e2}}{S_{e1}} - 1 \right) \frac{n - r_1}{r_1 - r_2}$$

has Fisher-Snedecor $F(r_1 - r_2, n - r_1)$ probability distribution.

H_0 is rejected at the level of significance α, if $F \geq F_{1-\alpha}(r_1 - r_2, n - r_1)$.

Obviously $r_2 < r_1 < n$, $S_{e2} \geq S_{e1}$. Compare F statistic with one-way ANOVA.
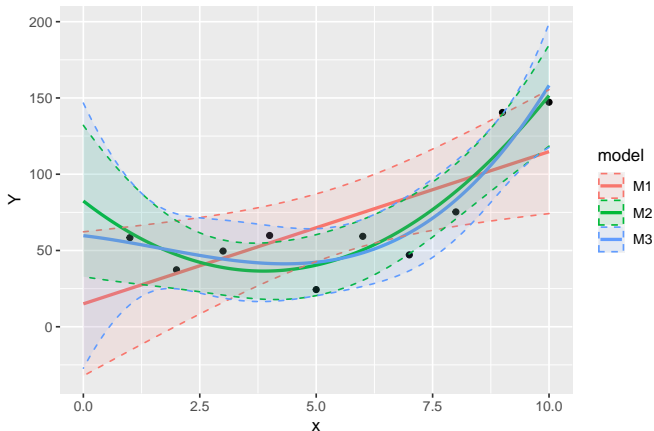
```
'data.frame': 10 obs. of  2 variables:
$ x: int   1 2 3 4 5 6 7 8 9 10
$ Y: num   58.4 37.3 49.6 59.9 24.4 ...
```

▶ $M_3 : \widehat{Y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
▶ $M_2 : \widehat{Y} = \beta_0 + \beta_1 x + \beta_2 x^2$
▶ $M_1 : \widehat{Y} = \beta_0 + \beta_1 x$

$$M_3$$

```
lm(formula=Y~1 + x + I(x^2) + I(x^3), data=dt)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.7207    35.5280   1.681   0.144
x           -3.5743    26.6299  -0.134   0.898
I(x^2)      -1.3063     5.4929  -0.238   0.820
I(x^3)       0.2649     0.3294   0.804   0.452
Residual standard error: 18.31 on 6 degrees of freedom
Multiple R-squared: 0.8694, Adjusted R-squared: 0.8042
F-statistic: 13.32 on 3 and 6 DF, p-value: 0.004624
```

$$M_2$$

```
lm(formula=Y~1 + x + I(x^2), data=dt)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.4500    20.9805   3.930   0.00568 **
x           -23.7340    8.7623  -2.709   0.03026 *
I(x^2)        3.0647     0.7763   3.948   0.00555 **
Residual standard error: 17.84 on 7 degrees of freedom
Multiple R-squared: 0.8554, Adjusted R-squared: 0.814
F-statistic: 20.7 on 2 and 7 DF, p-value: 0.001151
```

$$M_1$$

```
lm(formula=Y~1 + x, data=dt)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.027     20.475    0.734   0.4840
x            9.978       3.300    3.024   0.0165 *
Residual standard error: 29.97 on 8 degrees of freedom
Multiple R-squared: 0.5333, Adjusted R-squared: 0.475
F-statistic: 9.143 on 1 and 8 DF, p-value: 0.01647
```

Check residuals…
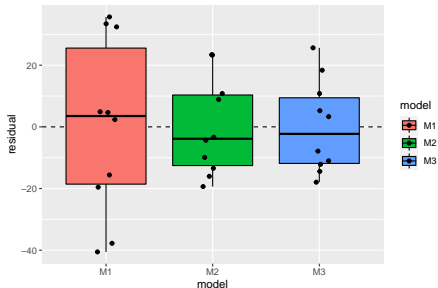
Compare M_3 and M_2

```
anova(M3, M2)

Analysis of Variance Table
Model 1: Y~1 + x + I(x^2) + I(x^3)
Model 2: Y~1 + x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 2010.7
2      7 2227.4 -1   -216.76 0.6469 0.4519
```
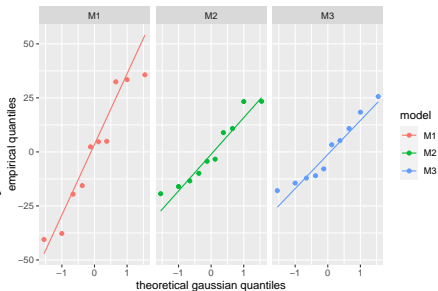
Compare M_2 and M_1

```
anova(M2, M1)

Analysis of Variance Table
Model 1: Y~1 + x + I(x^2)
Model 2: Y~1 + x
  Res.Df    RSS Df Sum of Sq       F Pr(>F)
1      7 2227.4
2      8 7186.6 -1   -4959.2 15.585 0.0055
```

And the winner is… M_2

Two-way ANOVA

Examination of hay yields (tons per hectare) based on the type of soil (normal; sour) and fertilizer (none; dung; calcium).

| soil type (A)	fertilizer (B) none	dung	calcium
normal	2.8, 3.2, 3.0, 3.0	3.7, 3.6, 3.9, 3.6	3.4, 3.8, 3.7, 3.6
sour	3.1, 2.7, 3.0, 2.9	3.4, 3.4, 3.0, 3.8	4.2, 4.0, 4.1, 3.9

At a 5% significance level, test following hypotheses:

► Soil type does not significantly affect hay yields.

► Method of fertilization does not significantly affect hay yields.

► Soil type and method of fertilization do not interact with respect to hay yields.

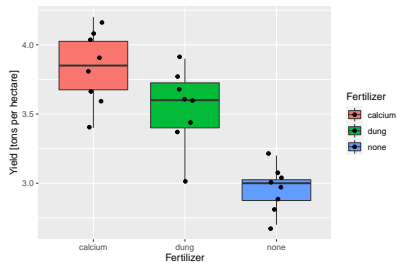If you reject the null hypothesis, find which pairs differ.
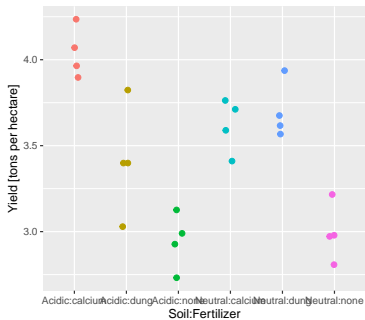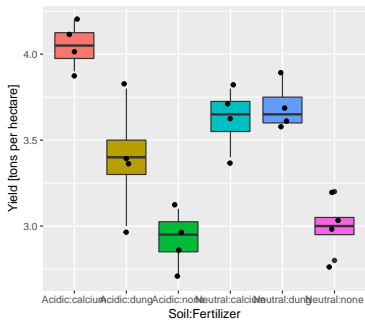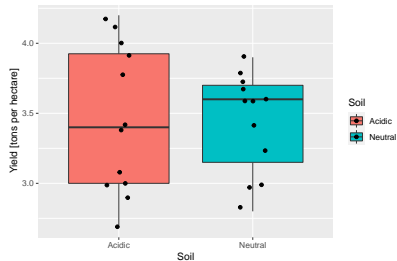
- Two factors, A has $a \geq 2$ levels, B has $b \geq 2$ levels
- The combination of ith level of factor A and jth level of factor B has n_{ij} observations $Y_{ij1}, \ldots, Y_{ijn_{ij}}$, which form a random sample from $N(\mu_{ij}, \sigma^2)$ probability distribution.
- Y_{ijk}: first index – group by the level of factor A, second index – group by the level of factor B, third index – order in the group.
- The particular random samples are stochastically independent.
- Model of two-way ANOVA *(dvoufaktorová analýza rozptylu / dvojné třídění)*:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

where $\varepsilon_{ij}k$ are stochastically independent random variables with $N(0, \sigma^2)$ probability distribution, $i = 1, \ldots, a, \ j = 1, \ldots, b, \ j = k, \ldots, n_{ij}$.

$$
\begin{aligned}
M_+: \quad & Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk} \\
M_2: \quad & Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
M_B: \quad & Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\
M_A: \quad & Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\
M_0: \quad & Y_{ijk} = \mu + \varepsilon_{ijk}
\end{aligned}
$$

for $i = 1, \ldots, a$, for $i = j, \ldots, b$, $k = 1, \ldots, n_{ij}$,
where ε_{ijk} are *i.i.d.* random variables with $N(0, \sigma^2)$ probability distribution,

- ▶ μ = overall mean (grand mean) of the random variable Y,
- ▶ α_i = the (row) effect of the ith level of factor A,
- ▶ β_j = the (column) effect of the jth level of factor B,
- ▶ λ_{ij} = the interaction of the ith level of factor A and jth level of factor B,
- ▶ $\mu_{ij} = \mu + \alpha_i + \beta_j + \lambda_{ij}$ = mean of Y by the ith level of factor A and jth level of factor B,
- ▶ ε_{ij} = random errors.

$$\begin{aligned}
M_+: \quad & Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk} \\
M_2: \quad & Y_{ijk} = \mu + \alpha_i + \beta_j \phantom{{}+ \lambda_{ij}} + \varepsilon_{ijk} \\
M_B: \quad & Y_{ijk} = \mu + \alpha_i \phantom{{}+ \beta_j + \lambda_{ij}} + \varepsilon_{ijk} \\
M_A: \quad & Y_{ijk} = \mu \phantom{{}+ \alpha_i} + \beta_j \phantom{{}+ \lambda_{ij}} + \varepsilon_{ijk} \\
M_0: \quad & Y_{ijk} = \mu \phantom{{}+ \alpha_i + \beta_j + \lambda_{ij}} + \varepsilon_{ijk}
\end{aligned}$$

Interactions

H_{0AB}: all $\lambda_{ij} = 0$, i.e., the interaction is not significant,

H_{1AB}: $\exists\, i, j : \lambda_{ij} \neq 0$, i.e., the interaction is significant

---

Factor B

H_{0B}: all $\beta_j = 0$, i.e., factor B is not significant,

H_{1B}: $\exists\, j : \beta_j \neq 0$, i.e., factor B is significant

---

Factor A

H_{0A}: all $\alpha_i = 0$, i.e., factor A is not significant,

H_{1A}: $\exists\, i : \alpha_i \neq 0$, i.e., factor A is significant

---

Possible sequences of submodels

$M_+ \xrightarrow{H_{0AB}} M_2 \xrightarrow{H_{0B}} M_B \xrightarrow{H_{0A}} M_0,$ or $M_+ \xrightarrow{H_{0AB}} M_2 \xrightarrow{H_{0A}} M_A \xrightarrow{H_{0B}} M_0$

Model M_+ (and its submodels similarly) is written as linear regression model

$$Y = X\,(\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \lambda_{11}, \ldots, \lambda_{ab})'$$

with design matrix X of size $n \times (1 + a + b + a\,b)$ and rank $r(X) = a\,b$.

Additional equations $\quad \displaystyle\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \lambda_{ij} = 0, \quad \sum_{j=1}^{b} \lambda_{ij} = 0$

leads to model of full rank with following estimators:

overall (grand) mean:	$\widehat{\mu} = \overline{Y}_{\text{...}}$
interactions:	$\widehat{\lambda}_{ij} = \overline{Y}_{ij\text{.}} - \overline{Y}_{i\text{..}} - \overline{Y}_{\text{.}j\text{.}} + \overline{Y}_{\text{...}}$
row (A) effects:	$\widehat{\alpha}_i = \overline{Y}_{i\text{..}} - \overline{Y}_{\text{...}}$
column (B) effects:	$\widehat{\beta}_j = \overline{Y}_{\text{.}j\text{.}} - \overline{Y}_{\text{...}}$

mean of group $(A = i, B = j)$:	$\widehat{\mu}_{ij} = \overline{Y}_{ij\text{.}}$
- in model M_2:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j = \overline{Y}_{i\text{..}} + \overline{Y}_{\text{.}j\text{.}} - \overline{Y}_{\text{...}}$
- in model M_B:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i = \overline{Y}_{i\text{..}}$
- in model M_A:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\beta}_j = \overline{Y}_{\text{.}j\text{.}}$
- in model M_0:	$\widehat{\mu}_{ij} = \widehat{\mu} = \overline{Y}_{\text{..}}$

▶ $S_T = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} \sum\limits_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}...)^2,$ $\sim \chi^2(df_T = n - 1),$

▶ $S_{AB} = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} n_{ij} (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}...)^2 \sim \chi^2(df_{AB} = (a-1)(b-1)),$

▶ $S_B = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} n_{ij} (\overline{Y}_{.j.} - \overline{Y}...)^2,$ $\sim \chi^2(df_B = b - 1),$

▶ $S_A = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} n_{ij} (\overline{Y}_{i..} - \overline{Y}...)^2,$ $\sim \chi^2(df_A = a - 1),$

▶ $S_e = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} \sum\limits_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij})^2,$ $\sim \chi^2(df_e = n - ab).$

Theorem

$$S_T = S_{AB} + S_A + S_B + S_e$$

source of variability	degrees of freedom	sum of squares	mean squares	value of the test statistic	p-value
row A	$df_A = a - 1$	S_A	$MS_A = \dfrac{S_A}{df_A}$	$F_A = \dfrac{MS_A}{MS_e}$	p_A
column B	$df_B = b - 1$	S_B	$MS_B = \dfrac{S_B}{df_B}$	$F_B = \dfrac{MS_B}{MS_e}$	p_B
interaction	$df_{AB} = (a-1)(b-1)$	S_{AB}	$MS_{AB} = \dfrac{S_{AB}}{df_{AB}}$	$F_{AB} = \dfrac{MS_{AB}}{MS_e}$	p_{AB}
residual	$df_e = n - ab$	S_e	$MS_e = \dfrac{S_e}{df_e}$		
total	$df_T = n - 1$	S_T			

When any null hypotheses is rejected, the multiple comparison usually follows.

1. Test the significance of the interactions using F_{AB}:
$$\text{if } F_{AB} \geq F_{1-\alpha}((a-1)(b-1), n-ab), \text{ reject } H_{0\,AB}.$$

2. Test the significance of the B (column) factor using F_B, at the same time, take into account effects of (row) factor A:
$$\text{if } F_B \geq F_{1-\alpha}(b-1, n-ab), \text{ reject } H_{0\,B}.$$

3. Test the significance of the A (row) factor using F_A, do not take into account effects of (row) factor B:
$$\text{if } F_A \geq F_{1-\alpha}(a-1, n-ab), \text{ reject } H_{0\,A}.$$

Or use the corresponding p-values from ANOVA table, **from the bottom up**.

Two-way ANOVA without interactions

The interactions can be omitted. (Advantage: fewer parameters.)

Then, $S_{AB} = 0$, $df_e = n - a - b + 1$, the testing procedure starts with $H_{0\,B}$.

Different parametrization of the one-way ANOVA is used by most of the statistical software (including *R*).

Definition

Random variables Y_{ijk} follow model

$$Y_{ijk} = \mu^* + \alpha_i^* + \alpha_j^* + \lambda_{ij}^* + \varepsilon_{ij},$$

for $i = 1, \ldots, a$, $i = j, \ldots, b$, $k = 1, \ldots, n_{ij}$,
where ε_{ijk} are *i.i.d.* random variables with $N(0, \sigma^2)$ probability distribution,

▶ $\mu^* = \mu_{11}$ = mean of the *top left* category $(A = 1, B = 1)$,

▶ $\alpha_1^* = 0$, $\beta_1^* = 0$, $\lambda_{ij}^* = 0$ are fixed,

▶ other effects $(i \geq 2$ or $j \geq 2)$ and interactions express the deviations in mean from the category $(A = 1, B = 1)$,

▶ $\mu_i = \mu^* + \alpha_i^* + \beta_j^* + \lambda_{ij}^*$ = mean of Y by the category $(A = i, B = j)$.

```
'data.frame': 24 obs. of  4 variables:
 $ Soil     : Factor w/ 2 levels "Acidic","Neutral": 2 2 2 2 2 2 2 2 2 2 ...
 $ Fertilizer: Factor w/ 3 levels "calcium","dung",..: 3 3 3 3 2 2 2 2 1 1 ...
 $ Yield    : num  2.8 3.2 3 3 3.7 3.6 3.9 3.6 3.4 3.8 ...

 aov.model <- aov(Yield~Soil+Fertilizer, data=dt)

             Df Sum Sq Mean Sq F value   Pr(>F)
 Soil         1  0.002  0.0017   0.027    0.871
 Fertilizer   2  3.182  1.5912  25.752 2.93e-06 ***
 Residuals   20  1.236  0.0618

    (Intercept)    SoilNeutral Fertilizerdung Fertilizernone
     3.84583333    -0.01666667    -0.28750000    -0.87500000
```

```
                                              Tables of means
         Tables of effects                 Grand mean
 Soil                                       3.45
Soil                                         Soil
   Acidic   Neutral                        Soil
 0.008333 -0.008333                         Acidic Neutral
 Fertilizer                                  3.458   3.442
Fertilizer                                  Fertilizer
calcium   dung    none                      Fertilizer
 0.3875  0.1000 -0.4875                     calcium   dung    none
                                             3.838   3.550   2.963
```

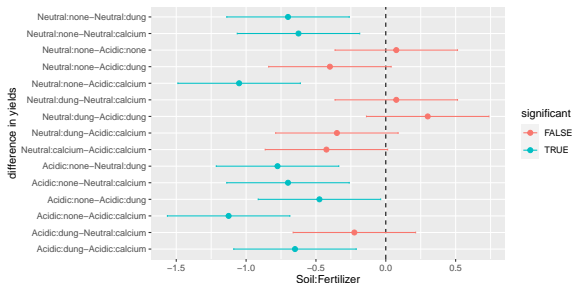Tukey's method:
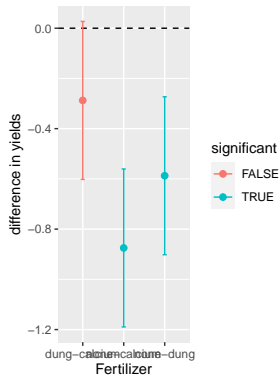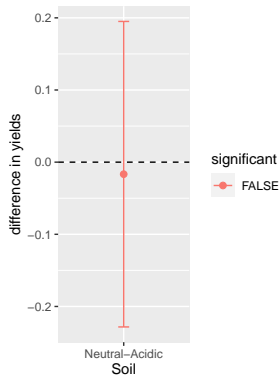
Scheffé's method:

```
              Yield groups
Acidic  3.458333        a
Neutral 3.441667        a

               Yield groups
calcium 3.8375         a
dung    3.5500         a
none    2.9625         b

                      Yield groups
Acidic:calcium  4.050        a
Neutral:dung    3.700        ab
Neutral:calcium 3.625        abc
Acidic:dung     3.400        bcd
Neutral:none    3.000        cd
Acidic:none     2.925        d
```

```
aov.model <- aov(Yield~Soil+Fertilizer+Soil:Fertilizer, data=dt) # or simply
aov.model <- aov(Yield~Soil*Fertilizer, data=dt)

                Df Sum Sq Mean Sq F value   Pr(>F)
Soil             1  0.002  0.0017   0.044  0.83658
Fertilizer       2  3.182  1.5912  41.814 1.72e-07 ***
Soil:Fertilizer  2  0.551  0.2754   7.237  0.00494 **
Residuals       18  0.685  0.0381

     (Intercept)                    SoilNeutral                  Fertilizerdung
           4.050                         -0.425                          -0.650
  Fertilizernone    SoilNeutral:Fertilizerdung  SoilNeutral:Fertilizernone
          -1.125                          0.725                           0.500
```

```
                                          Tables of means
Tables of effects                         Grand mean
 Soil                                     3.45
Soil                                       Soil
   Acidic    Neutral                      Soil
 0.008333 -0.008333                        Acidic Neutral
 Fertilizer                                 3.458   3.442
Fertilizer                                 Fertilizer
calcium    dung     none                  Fertilizer
 0.3875   0.1000  -0.4875                  calcium    dung    none
 Soil:Fertilizer                           3.838    3.550   2.963
        Fertilizer                          Soil:Fertilizer
Soil     calcium  dung     none                      Fertilizer
  Acidic   0.20417 -0.15833 -0.04583       Soil     calcium dung  none
  Neutral -0.20417  0.15833  0.04583        Acidic   4.050   3.400 2.925
                                            Neutral  3.625   3.700 3.000
```

Tukey's method:
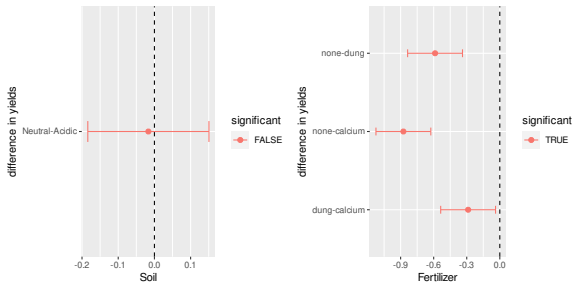
Scheffé's method:

```
            Yield groups
Acidic   3.458333       a
Neutral  3.441667       a

            Yield groups
calcium  3.8375         a
dung     3.5500         b
none     2.9625         c

                  Yield groups
Acidic:calcium    4.050        a
Neutral:dung      3.700        ab
Neutral:calcium   3.625        ab
Acidic:dung       3.400        bc
Neutral:none      3.000        c
Acidic:none       2.925        c
```

▶ Independence of particular random samples: very important assumption. Violation of the independence leads to change in the probability distribution of the test statistic and p-values.

▶ Normality of the data: verification by normal QQ-plot and statistical tests (Shapiro-Wilk, Lillieforse). ANOVA is not very sensitive to violation of the normality. Does not matter when when each group has at least 20 observations and the distributions are not very skewed. When strongly violated, use Kruskal-Wallis test.

▶ Homogeneity of the variances: verification by Levene and/or Bartlett test. Does not matter when slightly violated and all groups have similar number of observations. When strongly violated, use Kruskal-Wallis test.

- ▶ ANOVA – basic idea, typical examples
- ▶ assumptions – data types, normality and homogeneity of variances
- ▶ definition of ANOVA, model equation, parameters and their interpretation
- ▶ hypothesis and equivalent formulations
- ▶ test statistic in ANOVA, sums of squares
- ▶ ANOVA table – intepretation
- ▶ effects, group means, overall mean – calculation and interpretation
- ▶ methods of multiple comparison
- ▶ using ANOVA to compare nested linear regression models

Statistics II │ 3

Rank-based methods and tests

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

26 September 2022

- ▶ Most of the statistical test use a parametric family, e.g., the normal distribution $N(\mu, \sigma^2)$, for modeling random data.
- ▶ Based on the observed data $(X_1, \ldots, X_n)'$, we are able to calculate parameter estimates of the **parameters** (point estimates), e.g. $\widehat{\mu}, \widehat{\sigma}^2$, confidence intervals (interval estimates), and perform statistical tests on the parameters.
- ▶ All calculations are based on the assumption that the observed data comes from the specified parametric family.
- ▶ Typical assumptions of most parametric methods: interval or ratio type of data; normality of the sample; homogeneity of variances of random samples.
- ▶ A **probability model** describes ideas about the possible outcomes of a random event and the corresponding probabilities.
- ▶ **The model is essential** for determining the statistical uncertainty of an estimate (point as well as interval) or for deriving the critical region (and calculation of the p-value) of a test.

If these assumptions are not met, we use **nonparametric** methods.
Nonparametric statistics is about methods that do not make parametric
assumptions about the data generating process.

► Nonparametric regression models,
► nonparametric (distribution-free) tests,
► nonparametric density estimation (e.g., kernel estimates),
► …

Rank tests are nonparametric tests based on **ranks** of random variables in
random sample.

Definition (Random sample)

Vector of random variables $X = (X_1, \ldots, X_n)'$ is a random sample *(náhodný výběr)* of sample size *(rozsah)* n, if the random variables are i.i.d. = independent identically distributed, i.e., have the same probability distribution and are mutually independent.

Definition (Ordered random sample)

Ordered random sample *(uspořádaný náhodný výběr)* is random vector

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \qquad \text{where } X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

$X_{(i)}$ is the i-th order statistic *(pořádková statistika)*.

Definition (Rank)

Rank / rank statistic *(pořadí)* R_i of the random variable X_i is the order of X_i in the ordered random sample $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$.

If there are no ties in the sample,

$$R_i = \big|\{k : X_k \leq X_i\}\big|.$$

Otherwise, e.g., average ranks, are used,

$$R_i = \big|\{k : X_k < X_i\}\big| + 1 + \tfrac{1}{2}\big|\{k \neq i : X_k = X_i\}\big|.$$

Example

Consider random sample $X = (2.0, 1.8, 2.1, 2.4, 1.9, 2.1, 2.0, 1.8, 2.3, 2.1)'$.

i	1	2	3	4	5	6	7	8	9	10
X_i	2.0	1.8	2.1	2.4	1.9	2.1	2.0	1.8	2.3	2.1
$X_{(i)}$	1.8	1.8	1.9	2.0	2.0	2.1	2.1	2.1	2.3	2.4
R_i	4	1	6	10	3	7	5	2	9	8
average R_i	4.5	1.5	7	10	3	7	4.5	1.5	9	7

R

▶ Permutation: `order(X)`

▶ Ordered sample: `sort(X)`, `X[order(X)]`

▶ Ranks: `rank(x)`, `rank(x, ties.method = "average")`

Let $(X_1, \ldots, X_n)$ be random sample from a continouos probability distribution with median $\widetilde{x}$, i.e.,

$$P(X_i < \widetilde{x}) = P(X_i > \widetilde{x}) = \tfrac{1}{2}, \qquad i = 1, \ldots, n.$$

Is the median equal to chosen number $x_0 \in \mathbb{R}$?

$$H_0 : \widetilde{x} = x_0 \qquad H_1 : \widetilde{x} \neq x_0.$$

Calculate diferences $X_i - x_0$ of the observations from the tested value, and denote T the number of positive differences, $T^+ = \big|\{i : X_i > x_0\}\big|$.

Let us define indicator random variables *(indikátorové náhodné veličiny)*,

$$Z_i = \begin{cases} 1, & X_i > x_0, \\ 0, & X_i \leq x_0. \end{cases}$$

▶ Verify, that $T^+ = Z_1 + \cdots + Z_n$.
▶ Specify the probability distribution of T^+ under H_0.
▶ Calculate $\mathsf{E}(Z_i)$, $\mathsf{Var}(Z_i)$, $\mathsf{E}(T^+)$ and $\mathsf{Var}(T^+)$ under H_0.

Theorem (Sign test – for small sample size)

H_0 is rejected at the level of significance $\leq \alpha$, if
$$T^+ \leq k_\alpha \qquad \text{or} \qquad T^+ \geq n - k_\alpha.$$
Number k_α is the largest number from $\{0, \ldots, n\}$, for which
$$P(T^+ \leq k_\alpha) = \frac{1}{2^n} \sum_{i=0}^{k_\alpha} \binom{n}{i} \leq \frac{\alpha}{2} \text{ and } P(T^+ \geq n - k_\alpha) = \frac{1}{2^n} \sum_{i=n-k_\alpha}^{n} \binom{n}{i} \leq \frac{\alpha}{2}.$$

Moivre-Laplace theorem implies that $U = \dfrac{T^+ - \mathsf{E}(T^+)}{\sqrt{\mathsf{Var}(T^+)}} \overset{as.}{\sim} \mathsf{N}(0;1)$ as $n \to \infty$.

Theorem (Sign test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{2T^+ - n}{\sqrt{n}}$ has asymptotically standard normal distribution $\mathsf{N}(0;1)$ as $n \to \infty$.

H_0 is rejected at the asymptoticlevel of significance α, if $|U| \geq u_{1-\alpha/2}$.

Ten research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal.

Observations in seconds: $X = (53, 48, 45, 55, 63, 51, 66, 56, 50, 58)'$.

Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated.
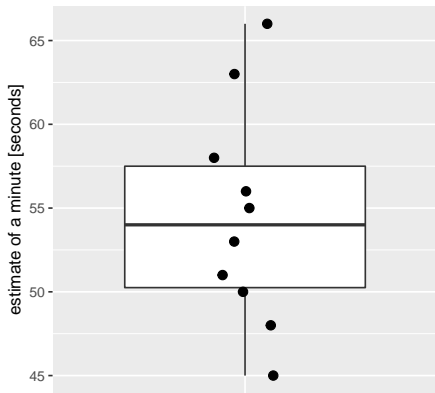
$$H_0 : \widetilde{x} = 60, \ H_1 : \widetilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$(X_i - 60)$	-7	-12	-15	-5	3	-9	6	-4	-10	-2

$n = 10, T^+ = 2, U = \frac{4-10}{\sqrt{10}} = -1.897, k_{0.05} = 1, u_{0.975} = 1.96, H_0$ is not rejected
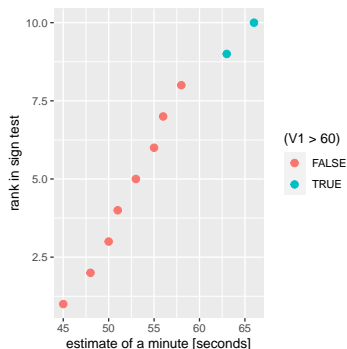

```
SIGN.test(dt$V1, md = 60)
  One-sample Sign-Test
data:  X
s = 2, p-value = 0.1094
alternative hypothesis: true median is not equal to
95 percent confidence interval:
 48.64889 61.37778
sample estimates:
median of x
        54
Achieved and Interpolated Confidence Intervals:
                 Conf.Level  L.E.pt  U.E.pt
Lower Achieved CI    0.8906  50.0000  58.0000
Interpolated CI      0.9500  48.6489  61.3778
Upper Achieved CI    0.9785  48.0000  63.0000
```

- ▶ Used especially in the case when the probability distribution of the observations X_i is significantly skewed, hence surely not gaussian. The t-test would be biased (incorrect p-values) in such a case.
- ▶ The test has low power. It is desirable to have a large sample.
- ▶ The asymptotical variant is sufficiently accurate when $n \geq 20$.
- ▶ Differences $X_i - x_0$ which are equal to zero are omitted, and the test is performed only for the remaining differences with reduced n.

Paired sign test *(párový znaménkový test)*

Let us assume **i.i.d.** pairs of (possibly dependent) observations $\big((Y_1, Z_1), \ldots, (Y_n, Z_n)\big)$ from a bivariate continuous probability distribution.

$$H_0 : \widetilde{z} - \widetilde{y} = x_0 \qquad H_1 : \widetilde{z} - \widetilde{y} \neq x_0.$$

Calculate differences $X_i = Z_i - Y_i$ and perform the sign test on the new sample $(X_1, \ldots, X_n)'$.

Let $(X_1, \ldots, X_n)$ be random sample from a continouos probability distribution with symmetrical probability density function $f(x)$,

$$P(X_i < \widetilde{x}) = \int_{-\infty}^{\widetilde{x}} f(x)\,dx = \int_{\widetilde{x}}^{\infty} f(x)\,dx = P(X_i > \widetilde{x}) = \tfrac{1}{2}, \qquad i = 1, \ldots, n.$$

Is the median equal to chosen number $x_0 \in \mathbb{R}$?

$$H_0 : \widetilde{x} = x_0, \qquad H_1 : \widetilde{x} \neq x_0.$$

1. Calculate diferences $Y_i = X_i - x_0$
2. and sort them in nondecreasing order according to their absolute value,

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \cdots \leq |Y|_{(n)}.$$

3. Denote R_i^+ the rank of $|Y_i|$ in this nondecreasing sequence.
4. Calculate the sum of R_i^+ ranks separately for positive and negative Y_i,

$$T^+ = \sum_{Y_i > 0} R_i^+, \qquad T^- = \sum_{Y_i < 0} R_i^+.$$

What is the sum $(T^+ + T^-)$ equal to?

Alternative calculation, signed ranks

$$T^+ = \frac{n(n+1)}{4} + \frac{T}{2}, \quad T^- = T^+ - T, \quad \text{where} \quad T = \sum_{i=1}^{n} R_i^+ \operatorname{sgn}(Y_i).$$

Theorem (Signed-rank Wilcoxon test)

H_0 is rejected at the level of significance α, if

$$\min\{T^+, T^-\} \leq w_\alpha(n),$$

where $w_\alpha(n)$ is *critical calue* of the Wilcoxon test.

Theorem (Signed-rank Wilcoxon test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{T^+ - \mathsf{E}(T^+)}{\sqrt{\mathsf{Var}(T^+}}$,

where $\mathsf{E}(T^+) = \frac{1}{4}n(n+1)$ and $\mathsf{Var}(T^+) = \frac{1}{24}n(n+1)(2n+1)$,

has asymptotically standard normal distribution $\mathsf{N}(0;1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

$H_0 : \widetilde{x} = 60,\ H_1 : \widetilde{x} \neq 60$										
i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$Y_i = (X_i - 60)$	-7	-12	-15	-5	3	-9	6	-4	-10	-2
R_i^+	6	9	10	4	2	7	5	3	8	1
$\mathrm{sgn} Y_i$	-1	-1	-1	-1	1	-1	1	-1	-1	-1

$T = -41,\ T^+ = 7,\ T^- = 48,\ n = 10,\ w_{0.05}(10) = 8,$
$\mathsf{E}(T^+) = 27.5,\ \mathsf{Var}(T^+) = 96.25,\ U = -2.09,\ u_{0.975} = 1.96,\ H_0$ is rejected
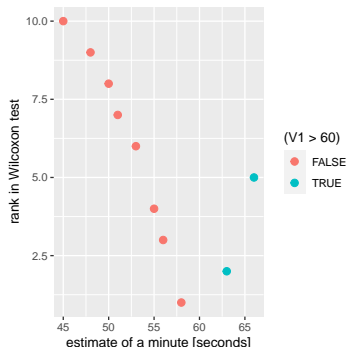

```
wilcox.test(dt$V1, mu = 60)
    Wilcoxon signed rank exact test
data:  X
V = 7, p-value = 0.03711
alternative hypothesis: true location is not equal
```

- ▶ Used especially to test whether data comes from a symmetric population with a specified median.
- ▶ T-test is parametric analogy of Wilcoxon signed-rank test for case of testing the mean of a sample from gaussian probability distribution.
- ▶ The Wilcoxon test assumes symmetry of the probability density of the observed variable around the median. In case of asymmetry of the probability density of the data, H_0 can be rejected even if $\widetilde{x} = x_0$ holds. In the case of asymmetry of the probability density, e.g., the sign test is used.
- ▶ Differences $X_i - x_0$ which are equal to zero are omitted, and the test is performed only for the remaining differences with reduced n.
- ▶ The asymptotical variant is sufficiently accurate when $n \geq 30$.

Paired Wilcoxon test *(párový Wilcoxonův test)*

Let us assume **i.i.d.** pairs of (possibly dependent) observations $\big((Y_1, Z_1), \ldots, (Y_n, Z_n)\big)$ from a bivariate continuous probability distribution.

$$H_0 : \widetilde{z} - \widetilde{y} = x_0 \qquad H_1 : \widetilde{z} - \widetilde{y} \neq x_0.$$

Calculate differences $X_i = Z_i - Y_i$ and perform the Wilcoxon signed-rank test on the new sample $(X_1, \ldots, X_n)'$.

Example

Two methods of fertilization were tested on a total of 13 experimental fields of the same soil quality: method A on 8 fields, method B on 5 fields.

fertilization	wheat yields in tons per hectare
A	5.7, 5.5, 4.3, 5.9, 5.2, 5.6, 5.8, 5.1
B	5.0, 4.5, 4.2, 5.4, 4.4

Does the fertilization method have an effect on wheat yields?

Comparison of two independent random samples:

- $(X_1, \ldots, X_m)$ coming from cumulative distribution function (c.d.f.) $F_X(x)$,
- $(Y_1, \ldots, Y_n)$ coming from c.d.f. $F_Y(y)$.

Test of the equality of c.d.f.s against an alternative of a **location shift**,

$$H_0 : F_X(x) = F_Y(x), \qquad H_1 : F_X(x) = F_Y(x - \Delta) \text{ for } \Delta > 0$$

1. Join both samples,

$$(Z_1, \ldots, Z_{m+n}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n),$$

2. and sort the combined sample in nondecreasing order,

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(m+n)}.$$

3. Denote $(R_1, \ldots, R_m)$ the ranks of $(X_1, \ldots, X_m)$ and $(S_1, \ldots, S_n)$ the ranks of $(Y_1, \ldots, Y_n)$ in the combined ordered sample.

4. Calculate the sums of ranks of X- and Y-sample,

$$T_1 = \sum_{i=1}^{m} R_i, \qquad T_2 = \sum_{j=1}^{n} S_j,$$

5. and corresponding Mann-Whitney's statistics

$$U_1 = T_1 - \frac{m(m+1)}{2}, \qquad U_2 = T_2 - \frac{n(n+1)}{2}.$$

6. The Mann-Whitney's test statistic is $U_{\mathrm{MW}} = \min\{U_1, U_2\}$.

▶ $U_1 =$ number of cases $X_i > Y_j$ out of all pairwise comparisons.
▶ $U_1 + U_2 = m\, n$.

Theorem (Mann-Whitney-Wilcoxon test)

H_0 is rejected at the level of significance α, if

$$U_{\mathrm{MW}} \leq w_\alpha(m, n),$$

where $w_\alpha(m, n)$ is *critical calue* of the Mann-Whitney-Wilcoxon test.

Theorem (Mann-Whitney-Wilcoxon test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{U_{\mathrm{MW}} - \mathsf{E}(U_{\mathrm{MW}})}{\sqrt{\mathsf{Var}(U_{\mathrm{MW}})}}$,

where $\mathsf{E}(U_{\mathrm{MW}}) = \frac{1}{2} m\, n$ and $\mathsf{Var}(U_{\mathrm{MW}}) = \frac{1}{12} m\, n(m + n + 1)$,

has asymptotically standard normal distribution $\mathsf{N}(0; 1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

Effect size $= \dfrac{U_1}{m\, n}$.
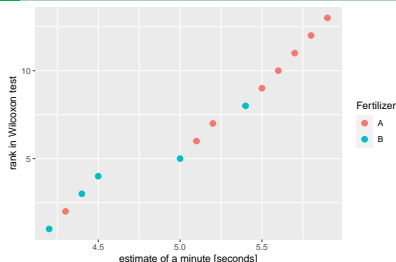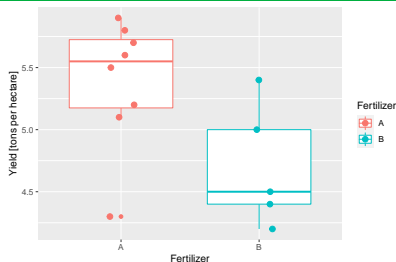
$$H_0 : F_A(x) = F_B(x), \qquad H_1 : F_A(x) = F_B(x - \Delta)$$

Z_k	4.2	4.3	4.4	4.5	5.0	5.1	5.2	5.4	5.5	5.6	5.7	5.8	5.9	rank sum
A: R_i		2				6	7		9	10	11	12	13	$T_1 = 70$
B: S_j	1		3	4	5			8						$T_2 = 21$

$$m = 8, n = 5, T_1 = 70, U_1 = 34, T_2 = 21, U_2 = 6, U_{\mathrm{MW}} = 6,$$

$$w_{0.05}(8, 5) = 6, \ U = \frac{6 - 20}{\sqrt{140/3}} = -2.049, \ u_{0.975} = 1.96, \ H_0 \text{ is rejected}$$

```
wilcox.test (dt|>filter(Fertilizer=="A")$Yield, dt|>filter(Fertilizer=="B")$Yield)
  Wilcoxon rank sum exact test
data:  pull(filter(dt, Fertilizer == "A"), Yield) and pull(filter(dt, Fertilizer ==
W = 34, p-value = 0.04507
alternative hypothesis: true location shift is not equal to 0
```

- Mann-Whitney U test = Mann–Whitney–Wilcoxon test = Wilcoxon rank-sum test = two-sample Wilcoxon test.
- The test assumes that the random samples are stochastically independent, and the data have of at least ordinal type.
- More general alternative is $H_1 : F_X(x) \neq F_Y(x)$.
- The more strict alternative $H_1 : F_X(x) = F_Y(x - \Delta)$ requires that the data comes from continuous probability distributions and is restricted to a shift in location. Rejection of H_0 then leads to a difference in medians.
- Mann–Whitney U test is preferable to the t-test when the data are ordinal but not of interval type.
- Mann–Whitney U test is more robust with respect to the presence of outliers.
- Mann-Whitney U test may have worse Type I error control when data are both heteroscedastic and non gaussian.
- The asymptotical variant is sufficiently accurate when $m, n > 10$.
- See also two-sample Kolmogorov-Smirnov test.

The test statistic

$$T = \frac{m}{2} + \frac{1}{2} \sum_{i=1}^{m} \mathrm{sgn}\left(R_i - \frac{m+n+1}{2}\right)$$

is equal to the number of X_i observations that are greater the the median of the combined sample. If the total sample size $(m+1)$ is odd, $\frac{1}{2}$ is added.

Theorem (Median test)

Under H_0, test statistic $U = \dfrac{T - \mathsf{E}(T)}{\sqrt{\mathsf{Var}(T)}}$,

where $\mathsf{E}(T) = \dfrac{m}{2}$ and $\mathsf{Var}(T) = \begin{cases} \frac{mn}{4(m+n-1)}, & \text{for } m+n \text{ odd,} \\ \frac{mn}{4(m+n)}, & \text{for } m+n \text{ even,} \end{cases}$ has

asymptotically standard normal distribution $\mathsf{N}(0;1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

The test is particularly suitable in the case of so-called censored observations, when for some extreme values we only know that that they are smaller or larger than some limit, but we do not know their exact values.

Kruskal-Wallis test = nonparametric analogy of one-way ANOVA and generalization of Mann-Whitney-Wilcoxon test.

Assumptions
- ▶ The factor A has $a \geq 3$ levels.
- ▶ The ith level has n_i observations $(Y_{i1}, \ldots, Y_{in_i})$, which form a random sample of at least ordinal type, coming from cumulative distribution function $F_i(x)$.
- ▶ Y_{ij}: first index – group by the level of the factor, second index – order in the group.
- ▶ The particular random samples are stochastically independent.

Hypothesis

A hypothesis that the factor A has no influence on the probability distribution of the observed variable Y, i.e., so-called *non-dominance* of cumulative distribution functions:

$H_0 : F_1(x) = F_2(x) = \cdots = F_a(x),$
$H_1 : \exists i \neq j : F_i(x) > F_j(x), \quad \text{or} \quad F_i(x) < F_j(x).$

1. Join all observations, $(Y_{11}, \ldots, Y_{a\,n_a})$,

2. and sort the combined sample in nondecreasing order,

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}, \qquad n = \sum_{i=1}^{a} n_i.$$

3. Denote R_{ij} the (average) rank of Y_{ij} in the combined sample.

4. Calculate the sums of ranks in each category,

$$T_i = R_{i.} = \sum_{j=1}^{n_i} R_{ij}, \qquad i = 1, \ldots, a.$$

A	observations	ranks	size	sum of ranks
1	$(Y_{11}, \ldots, Y_{1\,n_1})$	$(R_{11}, \ldots, R_{1\,n_1})$	n_1	T_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$(Y_{i1}, \ldots, Y_{i\,n_i})$	$(R_{i1}, \ldots, R_{i\,n_i})$	n_i	T_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$(Y_{a1}, \ldots, Y_{a\,n_a})$	$(R_{a1}, \ldots, R_{a\,n_a})$	n_a	T_a
total			n	$\frac{n(n+1)}{2}$

Theorem (Kruskal-Wallis test)

Denote $\quad Q = \dfrac{12}{n(n+1)} \sum_{i=1}^{a} \dfrac{T_i^2}{n_i} - 3(n+1).$

H_0 is rejected at the level of significance α, if $Q \geq h_\alpha(a-1)$, where $h_\alpha(a-1)$ is *critical value* of the test.

Under H_0, test statistic Q has asymptotically chi-squared probability distribution $\chi^2(a-1)$ with $(a-1)$ degrees of freedom, $E(Q) = a-1$, and $h_\alpha(a-1) \approx \chi^2_{1-\alpha}(a-1)$.

H_0 is rejected at the asymptotic level of significance α, if $Q \geq \chi^2_{1-\alpha}(a-1)$.

For more than ca. 25 % of ties in data, following correction is used, $Q_K = \dfrac{Q}{K}$, where $K = 1 - \dfrac{\sum_k m_k(m_k^2 - 1)}{n(n^2 - 1)}$, and m_k denotes the number of ties.

Analysis of 4 varieties of potatoes based on the weights of the clusters of potato tubers.
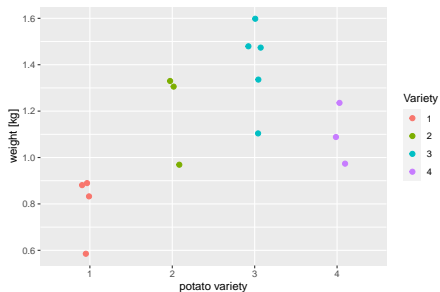
A	weight Y_{ij}	rank R_{ij}	n_i	T_i
1	0.9, 0.8, 0.6, 0.9	3.5, 2.0, 1.0, 3.5	4	10
2	1.3, 1.0, 1.3	11, 5.5, 11	3	27,5
3	1.3, 1.5, 1.6, 1.1, 1.5	11, 13.5, 15, 7.5, 13.5	5	60,5
4	1.1, 1.2, 1.0	7.5, 9.0, 5.5	3	22
total			15	120

$Q = 10.523,\ K = 1 - \frac{48}{3360},\ Q_K = 10.676 > \chi^2_{0,95}(3) = 7.815,\ H_0$ is rejected

Kruskal-Wallis test

```
KWtest <- with (dat, kruskal (Weight, Variety))
KWtest
```

```
$statistics
     Chisq    p.chisq
  10.67585 0.01361427
$parameters
  Df ntr  t.value
   3   4 2.200985
$rankMeans
  Variety    Weight r
1       1  2.500000 4
2       2  9.166667 3
3       3 12.100000 5
4       4  7.333333 3
$groups
  trt      means M
1   3 12.100000  a
2   2  9.166667 ab
3   4  7.333333  b
4   1  2.500000  c
```
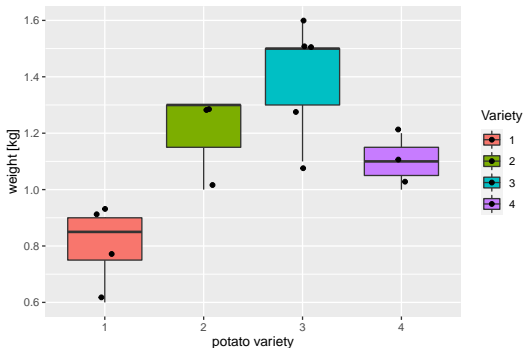
Median test uses the test statistic Denote A_i denotes the number of observations $(Y_{i1}, \ldots, Y_{in_i})$ from the i-th category that are greater than the median $\tilde{Y}$ of the joined sample,

$$A_i = \big|\{j : Y_{ij} > \tilde{Y}\}\big|, \qquad i = 1, \ldots, a.$$

If the total sample size n is odd, the A_i for which the median $\tilde{Y}$ of the joined sample belongs to the corresponding category i, is increased by $\frac{1}{2}$.

Theorem (Median test)

Under H_0, test statistic

$$Q_{\mathrm{M}} = 4 \sum_{i=1}^{a} \frac{A_i^2}{n_i} - n$$

has asymptotically ($\min\{n_1, \ldots, n_a\} \to \infty$) chi-squared probability distribution $\chi^2(a-1)$ with $(a-1)$ degrees of freedom.

H_0 is rejected at the asymptotic level of significance α, if $Q_{\mathrm{M}} \geq \chi_{1-\alpha}^2(a-1)$.

Median test

```
Mtest <- with (dat, Median.test (Weight, Variety))
Mtest
```

```
$statistics
     Chisq    p.chisq Median
  6.428571 0.09252244    1.1
$parameters
  Df ntr
   3   4
$Medians
  trt Median grather lessEqual
1   1   0.85       0         4
2   2   1.30       2         1
3   3   1.50       4         1
4   4   1.10       1         2
$comparison
          Median      Chisq      pval
1 and 2     0.90 7.0000000 0.0081509
1 and 3     1.10 5.7600000 0.0163950
1 and 4     0.90 7.0000000 0.0081509
2 and 3     1.30 2.8800000 0.0896860
2 and 4     1.15 0.6666667 0.414216178
3 and 4     1.25 4.8000000 0.028459737  *
```
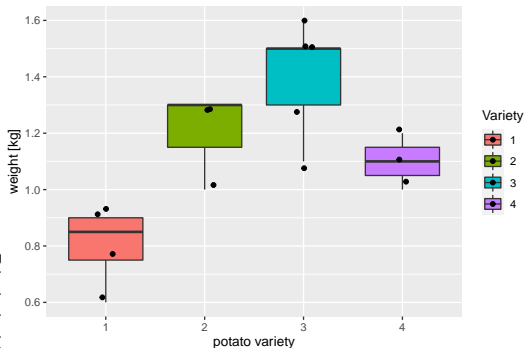
▶ Block design of the data,

▶ a levels of factor A,

▶ b blocks for each $i = 1, \ldots, a$,

▶ Y_{ij} stands for observation of j-th block at the i-th level of factor A,

▶ Y_{ij} comes from a distribution with cumulative distribution function $F_{ij}(x)$.

▶ Friedman test is nonparametric analogy of block-design two-way ANOVA with one observation Y_{ij} in each group.

Hypothesis

$H_0 : F_{1j}(x) = \cdots = F_{aj}(x)$, i.e., the c.d.f. is identical in each block, but does not need to be identical across the factor levels,

$H_1 :$ c.d.f.s differ also across factor levels.

1. Calculate ranks R_{ij} of Y_{ij} **separately in each block**,

2. sum the ranks for each level of factor A, $R_{i\cdot} = \sum_{j=1}^{b} R_{ij}$, $i = 1, \ldots, a$.

3. Test statistic is $Q_F = \dfrac{12\,b}{a(a+1)} \sum_{i=1}^{a} \left(\dfrac{R_{i\cdot}}{b} - \dfrac{a+1}{2} \right)^2$.

Theorem (Friedman test)

Under H_0, test statistic Q_F has asymptotically $\chi^2(a-1)$ probability distribution .

H_0 is rejected at the asymptotic level of significance α, if $Q_F \geq \chi^2_{1-\alpha}(a-1)$.

The $\chi^2(a-1)$ approximation is accurate enough when $n > 15$ or $k > 4$.

Numbers of sold pieces of products: $a = 5$ products, $b = 7$ blocks.

B	1	2	3	4	5	6	7
$A = 1$	5	1	16	5	10	19	10
$A = 2$	4	3	12	14	9	18	7
$A = 3$	7	1	22	3	7	28	6
$A = 4$	10	6	22	5	13	37	8
$A = 5$	12	2	35	4	10	58	7
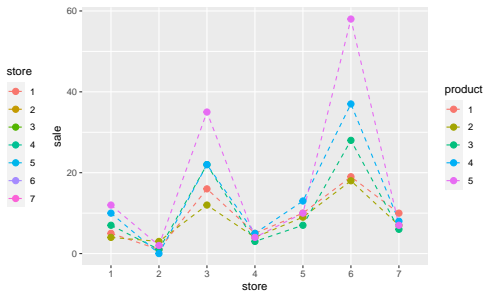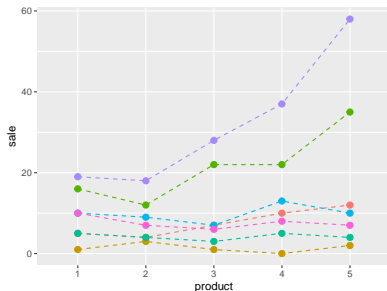

```
friedman.test(X$sale, X$product, X$store)
  Friedman rank sum test
data:  X$Y, X$A and X$B
Friedman chi-squared = 8.3284, df = 4, p-value = 0.08026
```

Van der Waerden test test the same hypothesis as in the Kruskal-Wallis test but converts the ranks R_{ij} to quantiles of the standard normal distribution $N(0;1)$.

1. $A_{ij} = \Phi^{-1}\left(\dfrac{R_{ij}}{n+1}\right),$

2. $A_{i\cdot} = \sum\limits_{j=1}^{n_i} A_{ij}, \qquad i = 1, \dots, a,$

3. Test statistic is $\qquad Q_{\mathrm{W}} = \dfrac{\sum\limits_{i=1}^{a} \dfrac{A_{i\cdot}^2}{n_i}}{\dfrac{1}{n-1}\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n_i} A_{ij}^2}.$

Theorem (Van der Waerden test)

Under H_0, test statistic Q_{W} has asymptotically $\chi^2(a-1)$ probability distribution.

H_0 is rejected at the asymptotic level of significance α, if $Q_{\mathrm{W}} \geq \chi^2_{1-\alpha}(a-1)$.

The test is robust as Kruskal-Wallis test, but efficient also for normal data.

When the null hypothesis H_0 is rejected, multiple comparison usually follows. Factor levels $A = k$ and $A = l$ are significantly different in their probability distributions (particularly in the location shift), if

$$|T_k - T_l| > \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) h_\alpha(a-1)}.$$

In the case of balanced design, i.e. when $n_i = b$ for all $i = 1, \ldots, a$, so called Neményi test is preferred. It is based on the Tukeys's idea from ANOVA. Factor levels $A = k$ and $A = l$ are significantly different in their probability distributions (particularly in the location shift), if

$$\sqrt{2\,b}\, \left| \overline{Z}_{k\cdot} - \overline{Z}_{l\cdot} \right| > q_\alpha,$$

where $\quad Z_{ij} = \begin{cases} 1, & Y_{ij} > \tilde{Y}, \\ 0, & Y_{ij} \leq \tilde{Y}, \end{cases}$ with group means $\overline{Z}_{i\cdot} = \dfrac{1}{m} \sum\limits_{j=1}^{m} Z_{ij}.$

- ▶ Sign test
  SIGN.test(X, md = x0)   library("BSDA")
- ▶ Wilcoxon signed-rank test
  wilcox.test(X, mu = x0)
- ▶ Wilcoxon rank-sum test
  wilcox.test(X, Y)
- ▶ Paired Wilcoxon test
  wilcox.test(X, Y, paired = TRUE)
- ▶ Kruskal-Wallis test
  kruskal(Y, group)   library("agricolae")
- ▶ Median test
  Median.test(Y, group)   library("agricolae")
- ▶ Van der Waerden test
  waerden.test(Y, group)   library("agricolae")
- ▶ Friedman test
  friedman.test(Y, group, block)

- parametric and nonparamwtric methods, assumptions, comparison
- ordered random sample, order statistic, rank
- rank tests (sign test, Wilcoxon's tests): hypotheses, principles of tests, calculation of ranks and test statistics
- Kruskall-Wallis test (median test, Van der Waerden test, Friedman test): model, hypothesis, test statistic, multiple comparison, comprarison with standard ANOVA

Statistics II │ 4

Goodness-of-fit test, testing probability distribution

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

3 October 2022

Often, we need to test whether the random sample $(X_1, \ldots, X_n)$ comes from a specific probability distribution of a family of probability distributions:

▶ with given parameters, e.g., $N(10, 4)$, $Ex(3.5)$, $Po(2)$, $Bi(10, 0.6)$;
▶ with unkown parameters, e.g., gaussian, exponential, Poisson, binomial;
▶ with given probability mass function $p(x)$ or probability density function $f(x)$.

Some examples:

▶ Parametric tests require a specific probability distribution of the random sample, e.g., t-test requires normality of the data.
▶ ANOVA requires normality of the data.
▶ Quality control and compliance with the prescribed probability distribution.
▶ Random number generators.
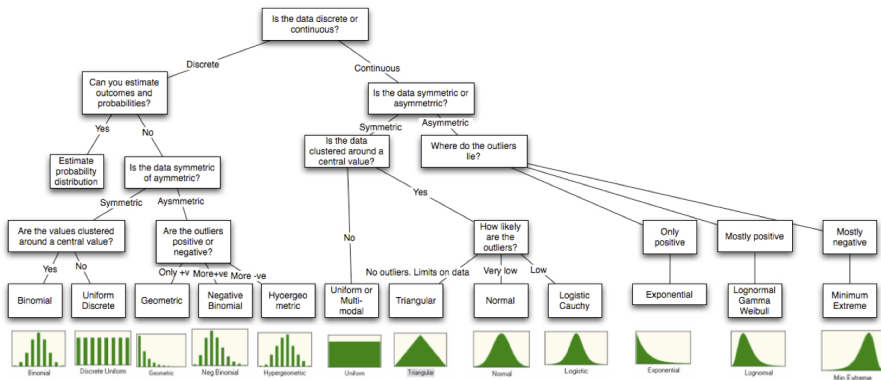
(Aswath Damodaran: *Probabilistic approaches to risk*)

▶ frequency barplot compared with the probability mass function
▶ histogram compared with the probability density function
▶ quantile-quantile (QQ) plot for given probability distribution
▶ empirical cumulative distribution function compared with the theoretical cumulative distribution function

Discrete:

▶ Bernoulli (alternative)
▶ binomial
▶ hypergeometric
▶ Poisson
▶ geometric
▶ negative binomial
▶ uniform discrete

Aboslutely continuous:

▶ gaussian (normal)
▶ chi-squared
▶ Student
▶ Fisher-Snedecor
▶ lognormal
▶ exponential
▶ gamma
▶ Weibull
▶ logistic
▶ beta
▶ uniform continuous

--- **Example (1)** ---

84 families was chosen randomly from a set of families of 5 children and the number of boys was detected for each family.

number of boys	0	1	2	3	4	5
number of familes	3	10	22	31	14	4

At a significance level of 0.05, test hypothesis that number of boys in families of 5 children has binomial distribution Bi $(5, 0.5)$.

--- **Example (2)** ---

Waiting time (in minutes) was observed for 70 clients of a certain company that they spent waiting for service (from the moment of taking their ticket).

waiting time	(0, 3]	(3, 6]	(6, 9]	(9, 12]	(12, 15]	(15, 18]	(18, 21]	(21, 24]
# of clients	14	16	10	9	8	5	3	5

At a significance level of 0.05, test hypothesis that waiting time has exponential distribution.

Pearson's chi-squared test

Assume that n objects $(X_1, \ldots, X_n)$ are distributed into k disjoint categories, $A_1, \ldots, A_k$, while each subject corresponds to exactly one category.

Theoretical probability distribution P assigns the probability p_j that randomly chosen object X is a member of the category A_j, $p_j = P(X \in A_j)$.

Definition

▶ Empirical / observed frequencies *(empirické četnosti)* are the numbers $N_1, \ldots, N_k$ of objects $(X_1, \ldots, X_n)$ in individual categories.

▶ Theoretical / expected frequencies *(teoretické četnosti)* are the expected numbers $n_1, \ldots, n_k$ of objects in individual categories,

$$n_j = n\, p_j.$$

---

category	A_1	A_2	$\cdots$	A_k	sum
empirical freqs.	N_1	N_2	$\cdots$	N_k	n
probabilities	p_1	p_2	$\cdots$	p_k	1
theoretical freqs.	n_1	n_2	$\cdots$	n_k	n

Definition

The joint probability distribution of empirical frequencies $(N_1, \ldots, N_k)$ is

$$P(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!}\, p_1^{n_1} \cdots p_k^{n_k}$$

for $n_j = 0, 1, \ldots, n$ and $n_1 + \cdots + n_k = n$.

This probability distribution is called k-variate multinomial *(multinomické)*,

$$(N_1, \ldots, N_k) \sim \mathsf{M}_k(n; p_1, \ldots, p_k).$$

Theorem

For $(N_1, \ldots, N_k) \sim \mathsf{M}_k(n; p_1, \ldots, p_k)$, it holds:

$$N_j \sim \mathsf{Bi}(n, p_j), \quad \mathsf{E}(X_j) = n\,p_j, \quad \mathsf{Var}(X_j) = n\,p_j(1 - p_j), \qquad j = 1, \ldots, k.$$

Moivre-Laplace theorem: For large n, large k and none *large* category,

$$\frac{N_j - n\,p_j}{\sqrt{n\,p_j(1 - p_j)}} \approx \frac{N_j - n\,p_j}{\sqrt{n\,p_j}} = \frac{N_j - n_j}{\sqrt{n_j}} \overset{as.}{\sim} \mathsf{N}(0, 1)$$

H_0: empirical distribution = theoretical distribution, i.e., all $N_j = n_j$,
H_1: empirické and theoretical distribution differ

Idea behind the test statistic K:

1. $N_1 - n_1, \ldots, N_k - n_k \longrightarrow 0$, better $\sum_{j=1}^{k} (N_j - n_j) \longrightarrow 0$,

2. $\sum_{j=1}^{k} |N_j - n_j| \longrightarrow 0$, better $\sum_{j=1}^{k} (N_j - n_j)^2 \longrightarrow 0$,

3. $\sum_{j=1}^{k} \frac{(N_j - n_j)^2}{p_j}$, better $\sum_{j=1}^{k} \frac{(N_j - n_j)^2}{n_j} \longrightarrow 0$.

$$K = \sum_{j=1}^{k} \frac{(N_j - n_j)^2}{n_j} = \sum_{j=1}^{k} \frac{N_j^2}{n_j} - 2 \underbrace{\sum_{j=1}^{k} \frac{N_j n_j}{n_j}}_{=n} + \underbrace{\sum_{j=1}^{k} \frac{n_j^2}{n_j}}_{=n} = \sum_{j=1}^{k} \frac{N_j^2}{n_j} - n = \frac{1}{n} \sum_{j=1}^{k} \frac{N_j^2}{p_j} - n$$

The test statistic K is the sum of squares of k independent random variables with standard normal distribution with one binding condition $\sum_{j=1}^{k} N_j = n$. What is the probability distribution of K and its expectation $\mathsf{E}(K)$?

Theorem (Pearson's chi-squared test *(Pearsonův test dobré shody)*)

Under H_0, the test statistic K has asymptotically chi-squared distribution with $(k-1)$ degrees of freedom,

$$K = \sum_{j=1}^{k} \frac{(N_j - n\,p_j)^2}{n\,p_j} = \frac{1}{n} \sum_{j=1}^{k} \frac{N_j^2}{p_j} - n \stackrel{as.}{\sim} \chi^2(k-1).$$

H_0 is rejected at the level of significance α, if $K \geq \chi^2_{1-\alpha}(k-1)$.

Assumptions *(Podmínka dobré aproximace)*

▶ $n_j \geq 5$, $j = 1, \ldots, k$, **or**

▶ $n_j \geq 5\,q$, where $q = \frac{1}{k} \cdot \left|\{j : n_j < 5\}\right|$ (so called Yarnold's criterion).

When these conditions are violated, it is necessary to appropriately merge some adjacent categories.

▶ $K = \sum\limits_{j=1}^{k} \dfrac{(O_j - E_j)^2}{E_j},$

where $O_j = N_j$ = **O**bserved frequencies, $E_j = n_j$ = **E**xpected frequencies.

▶ Deviation statistic / likelihood-ratio test *(deviační statistika, test poměrem věrohodnostní)*:

$$G = 2 \sum_{j=1}^{k} N_j \ln \frac{N_j}{n\, p_j} = 2 \sum_{j=1}^{k} O_j \ln \frac{O_j}{E_j}$$

has, under H_0, asymptotically chi-squared distribution with $(k-1)$ degrees of freedom, $G \sim \chi^2(k-1)$.

H_0 is rejected at the level of significance α, if $G \geq \chi^2_{1-\alpha}(k-1)$.

When testing the conformity of the probability distribution of random sample $(X_1, \ldots, X_n)$ with a theoretical probability distribution P with **unknown parameter(s)**, so-called modified minimum χ^2 method is used.

Modified minimum χ^2 method

Let us denote the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. The system of following equations is solved,

$$\sum_{j=1}^{k} \frac{N_j}{p_j(\boldsymbol{\theta})} \frac{\partial p_j(\boldsymbol{\theta})}{\partial \theta_i} = 0, \qquad i = 1, \ldots, m.$$

Theorem (Pearson's chi-squared test with unknown parameters)

When m parameters $\boldsymbol{\theta}$ estimated by the modified minimum χ^2 method are substitued into p_j and theoretical frequencies n_j, the degrees of freedom of the test statistic K is equal to $(k - 1 - m)$,

$$K \overset{as.}{\sim} \chi^2(k - 1 - m).$$

Degrees of freedom are reduced by the number of estimated parameters.

1. The categories $A_1, \ldots, A_k$ must cover all possible outcomes $t_1, \ldots, t_k$ of the considered discrete probability distribution.

2. Calculate empirical frequencies,
$$N_j = \left|\{X_i = t_j\}\right|,$$

3. and theoretical frequencies using the probability function of the theoretical distribution,
$$n_j = n\, p_j = n\, P(X = t_j);$$
the theoretical cumulative distribution function $F(x)$ can also be used,
$$n_j = n\, p_j = n\bigl[F(t_j) - \lim_{t \to t_j^-} F(t)\bigr].$$

4. Verify the assumptions (e.g., Yarnold's criterion) and modify (merge) the categories unless the conditions are met.

5. Calculate the value of the test statistic K and decide to reject or not to reject H_0.

1. The categories $A_1, \ldots, A_k$ are defined as intervals

$$A_j = (t_{j-1}, \, t_j], \qquad j = 1, \ldots, k,$$

covering the entire range of possible outcomes of the considered absolutely continuouso probability distribution. Sufficient number of values of the random sample $(X_1, \ldots, X_n)$ has to be in each interval. Recommended number of categories, i.e., intervals (as in the histogram construction): $k \approx \sqrt{n}$ for small n; sometimes $k \approx 1 + \log_2 n$; $k \approx 15 \left(\frac{n}{100}\right)^{2/5}$ for large n.

2. Calculate empirical frequencies,

$$N_j = \left| \{ t_{j-1} < X_i \leq t_j \} \right|$$

3. and theoretical frequencies using the theoretical cumulative distribution function $F(x)$,

$$n_j = n \, p_j = n \, P(t_{j-1} < X \leq t_j) = n \left[F(t_j) - F(t_{j-1}) \right].$$

4. Verify the assumptions (e.g., Yarnold's criterion) and modify (merge) the categories unless the conditions are met.

5. Calculate the value of the test statistic K and decide to reject or not to reject H_0.

Example 1

14/31

```
   t.j N.j       p.j      n.j
1    0   3 0.03125   2.625
2    1  10 0.15625  13.125
3    2  22 0.31250  26.250
4    3  31 0.31250  26.250
5    4  14 0.15625  13.125
6    5   4 0.03125   2.625

q <- sum (n * p.j < 5) / k
[1] 0.3333333
n * p.j >= 5 * q
[1] TRUE TRUE TRUE TRUE TRUE TRUE

K <- sum (N.j^2 / (n * p.j)) - n
[1] 3.12381

qchisq (0.95, df = k - 1)
[1] 11.0705

K >= qchisq (0.95, df = k - 1)
[1] FALSE

1 - pchisq (K, df = k - 1)
[1] 0.6809048

chisq.test (N.j, p = p.j)
   Chi-squared test for given probabilities
data:  N.j
X-squared = 3.1238, df = 5, p-value = 0.68
```

Example 1   15/31
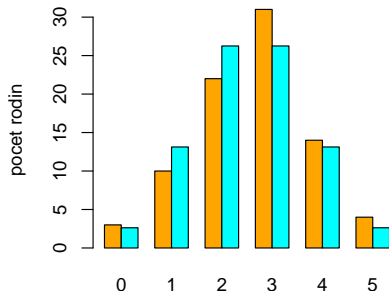
```
n * p.j >= 5
[1] FALSE  TRUE  TRUE  TRUE  TRUE FALSE

    t.j2 N.j2   p.j2  n.j2
1  0-1   13 0.1875 15.75
2    2   22 0.3125 26.25
3    3   31 0.3125 26.25
4  4-5   18 0.1875 15.75
n * p.j2 >= 5
[1] TRUE TRUE TRUE TRUE

K <- sum (N.j2^2 / (n * p.j2)) - n
[1] 2.349206

qchisq (0.95, df = k2 - 1)
[1] 7.814728

K >= qchisq (0.95, df = k2 - 1)
[1] FALSE
```

At asymptotic level of significance of
0.05, we do not reject the null
hypothesis, that the number of boys in
families with 5 children has binomial
distribution Bi$(5, 0.5)$.

Example 2

16/31
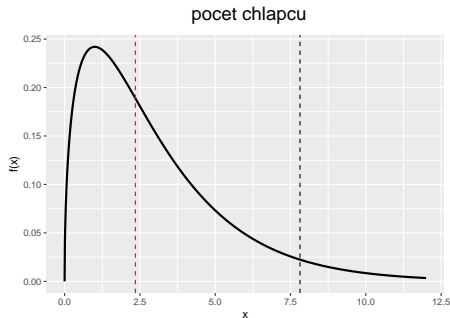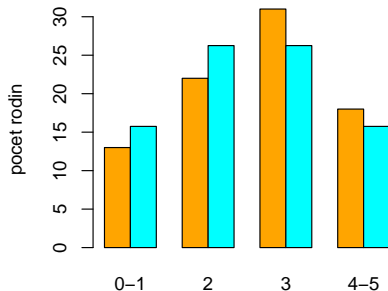
The *intensity* λ of the exponential distribution is estimated by the maximum likelihood method as $\widehat{\lambda} = \dfrac{1}{\overline{X}}$.

doba cekani

	A.j	N.j	p.j	n.j
1	(0; 3]	14	0.28576159	20.003311
2	(3; 6]	16	0.20410190	14.287133
3	(6; 9]	10	0.14577742	10.204419
4	(9; 12]	9	0.10411983	7.288388
5	(12; 15]	8	0.07436638	5.205647
6	(15; 18]	5	0.05311533	3.718073
7	(18; 21]	3	0.03793701	2.655591
8	(21; 24]	5	0.02709607	1.896725
9	> 24	0	0.06772447	4.740713
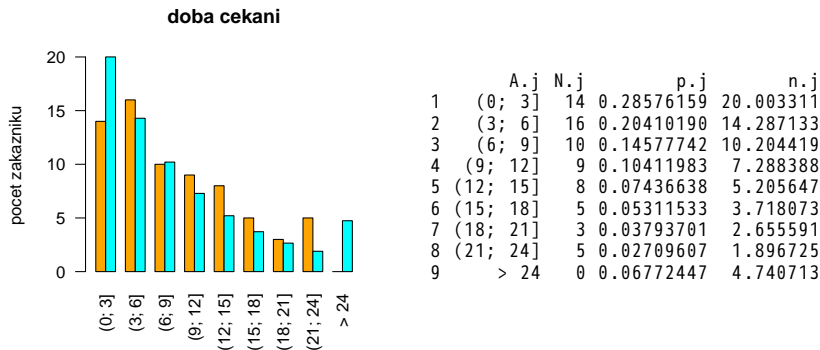
Note the last category, i.e., interval $(24, \infty]$, with no observation.

Example 2   17/31

assumptions, Yarnold's criterion

```
n * p.j >= 5
[1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE

q <- sum (n * p.j < 5) / k
[1] 0.4444444

n * p.j >= 5 * q
[1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

It is necessary to merge at least the last 2 categories. Then ...

doba cekani
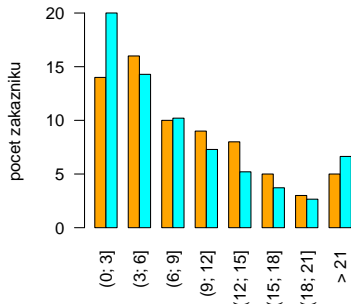
	A.j2	N.j2	p.j2	n.j2
1	(0; 3]	14	0.28576159	20.003311
2	(3; 6]	16	0.20410190	14.287133
3	(6; 9]	10	0.14577742	10.204419
4	(9; 12]	9	0.10411983	7.288388
5	(12; 15]	8	0.07436638	5.205647
6	(15; 18]	5	0.05311533	3.718073
7	(18; 21]	3	0.03793701	2.655591
8	> 21	5	0.09482054	6.637438

Example 2 18/31

Yarnold's criterion

```
q <- sum (n * p.j2 < 5) / k2
[1] 0.25

n * p.j2 >= 5 * q
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Pearson's chi-squared test, remeber $m = 1$ estimated parameter

```
K <- sum (N.j2^2 / (n * p.j2)) - n
[1] 4.803687

qchisq (0.95, df = k2 - 1 - 1)
[1] 12.59159

K >= qchisq (0.95, df = k2 - 1 - 1)
[1] FALSE
```

At the asymptotic level of significance of 0.05, we do not reject the null hypothesis, that the waiting times follow the exponential distribution.

Kolmogorov-Smirnov test and Lillieforse test

Definition (ECDF *(empirická distribuční funkce)*)

Let $(X_1, \ldots, X_n)$ be a random sample. Empirical cumulative distribution function / ECDF *(empirická distribuční funkce)* is defined

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{X_i \leq x\}, \quad \text{where } \mathbf{I}\{X_i \leq x\} = \begin{cases} 1, & X_i \leq x; \\ 0, & X_i > x. \end{cases}$$

ECDF is right-continuous step function. The steps correspond to the observed values $(X_1, \ldots, X_n)$.
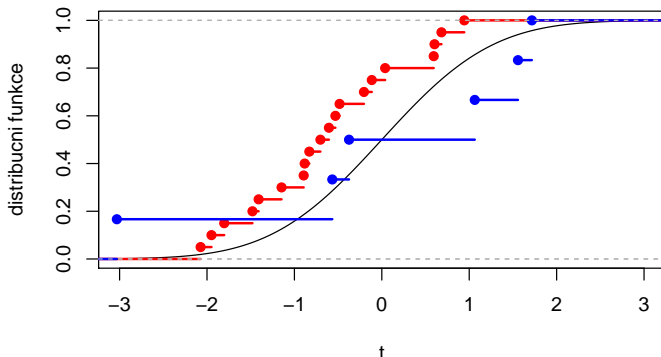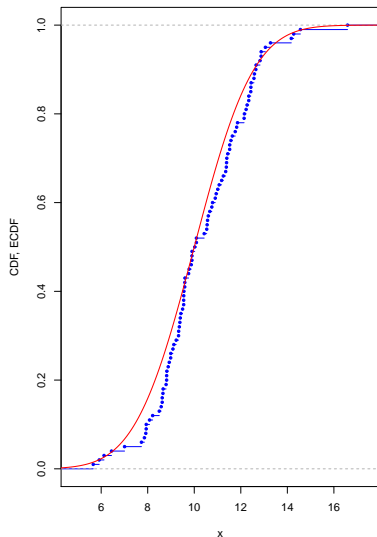
```
mu <- 10
sigma <- 2
X <- rnorm (100, mean=mu, sd=sigma)

F.emp <- ecdf (X)
plot (F.emp)

x <- seq (0, 20, by=0.1)
F <- pnorm (x, mean=mu, sd=sigma)
lines (t, F)
```

As the range n of the random sample increases, the empirical cumulative distribution function $\widehat{F}(x)$ approaches the true cumulative distribution function $F(x)$.

Let $(X_1, \ldots, X_n)$ be a random sample from an **absolutely continuous probability distribution** with cumulative distribution function $F(x)$. Let $F_0(x)$ be the cumulative distribution function beeing tested,

$$H_0 : F(x) = F_0(x) \qquad H_1 : F(x) \neq F_0(x).$$

The test statistic is

$$D = \sup \left\{ \left| \widehat{F}(x) - F_0(x) \right|; \quad -\infty < x < \infty \right\}.$$

Theorem (One-sample Kolmogorov-Smirnov test)

H_0 is rejected at the level of significance α, if $D \geq D_\alpha(n)$,
where $D_\alpha(n)$ is *critical value* of the one-sample Kolmogorov-Smirnov test.

For large n, ($n \geq 30$), we use $D_\alpha(n) \approx \sqrt{\dfrac{1}{2\,n} \ln \dfrac{2}{\alpha}}$.

What is the geometric meaning of the test statistic D?

Remark: The test statistic K has the same probability distribution as the supremum of *Brownian bridge*.

Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be two independent random samples from **absolutely continuous probability distributions** with cumulative distribution functions $F_X(x)$ and $F_Y(x)$. The equality of the cumulative distribution functions is tested,

$$H_0 : F_X(x) = F_Y(x) \qquad H_1 : F_X(x) \neq F_Y(x).$$

The test statistic is

$$D = \sup \left\{ \left| \widehat{F}_X(x) - \widehat{F}_Y(x) \right|; \quad -\infty < x < \infty \right\}.$$

Theorem (Two-sample Kolmogorov-Smirnov test)

H_0 is rejected at the level of significance α, if $D \geq D_\alpha(n.m)$,
where $D_\alpha(n.m)$ is *critical value* of the two-sample Kolmogorov-Smirnov test.

For large m, n, $(m + 1 \geq 35)$, we use $D_\alpha(n, m) \approx \sqrt{\dfrac{n + m}{2\, m\, n}\, \ln \dfrac{2}{\alpha}}$.

What is the geometric meaning of the test statistic D?

H_0: random sample $(X_1, \ldots, X_n)$ comes from a gaussian (normal) probability distribution $N(\mu, \sigma^2)$ with unknown parameters;

H_1: the random sample comes from non-gaussian (non-normal) probability distribution.

1. Parameters are estimated, $\widehat{\mu} = \overline{X}$, $\widehat{\sigma} = \sqrt{S^2}$,

2. and the K-S-like statistic is calculated,

$$L = \sup\left\{ \left| \widehat{F}(x) - \Phi\left(\frac{x - \widehat{\mu}}{\widehat{\sigma}}\right) \right|; \ -\infty < x < \infty \right\},$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal $N(0, 1)$ distribution.

Theorem (Lilliefors test)

H_0 is rejected at the level of significance α, if $L \geq L_\alpha(n)$, where $L_\alpha(n)$ is *critical value* of the Lilliefors test.

Some other specific tests

Let us remind: random variable X with Poisson distribution $X \sim \mathsf{Po}(\lambda)$ has equal expectation and variance,

$$\mathsf{E}(X) = \mathsf{Var}(X) = \lambda.$$

The test statistic uses this specific feature.

Theorem

Under the hypothesis of Poisson distribution of the observations,

$$\text{test statistic} \qquad Q = (n-1)\frac{S^2}{\overline{X}}$$

has asymptotically chi-squared distribution $\chi^2(n-1)$.

The hypothesis of the Poisson distribution of the sample is rejected at the asymptotic level of significance α, if

$$Q \leq \chi^2_{\alpha/2}(n-1) \quad \text{or} \quad Q \geq \chi^2_{1-\alpha/2}(n-1).$$

Let us remind: the variance of random variable X with exponential distribution $X \sim \mathsf{Ex}(\lambda)$ is equal to the square of its expectation,

$$\mathsf{Var}(X) = [\mathsf{E}(X)]^2 = \frac{1}{\lambda^2}.$$

The test statistic uses this specific feature.

Theorem

Under the hypothesis of exponential distribution of the observations,

$$\text{test statistic} \qquad Q = (n-1)\frac{S^2}{\overline{X}^2}$$

has asymptotically chi-squared distribution $\chi^2(n-1)$.

The hypothesis of the exponential distribution of the sample is rejected at the asymptotic level of significance α, if

$$Q \leq \chi^2_{\alpha/2}(n-1) \quad \text{or} \quad Q \geq \chi^2_{1-\alpha/2}(n-1).$$

```
X <- rep (prumer.j, N.j)
Q <- (n-1) * var (X) / (mean (X))^2
[1] 35.72647

q1 = qchisq (0.025, n-1)
q2 = qchisq (0.975, n-1)
[1] 47.92416 93.85647

Q <= q1 | Q >= q2
[1] TRUE
```

At the asymptotic level of significance of 0.05, we reject the null hypothesis, that the waiting times follow the exponential distribution.

$$H_0 : F(x) = F_0(x) \qquad H_1 : F(x) \neq F_0(x).$$

Theorem (Anderson-Darling test)

H_0 is rejected at the level of significance α, if

$$A = -n - \frac{1}{n} \sum_{i=1}^{n} (2\,i - 1) \Big[\ln F_0(X_{(i)}) + \ln \Big(1 - F_0(X_{(n-i+1)}) \Big) \Big] \geq A_\alpha(n),$$

where $A_\alpha(n)$ is *critical value* of the one-sample Anderson-Darling test.

Theorem (Cramér-von Mieses test)

H_0 is rejected at the level of significance α, if

$$T = \frac{1}{12\,n} + \sum_{i=1}^{n} \left[\frac{2\,i - 1}{n} - F_0(X_{(i)}) \right]^2 \geq T_\alpha(n),$$

where $T_\alpha(n)$ is *critical value* of the Cramér-von Mieses test.

Typically, both tests are used for **testing of normality** of random sample, i.e, F_0 is the distribution function of gaussian (normal) distribution.

Remark: Idea of test statistics is $\int_{-\infty}^{\infty} [\widehat{F}(x) - F_0(x)]^2 \, w(x) \, f_0(x) \, \mathrm{d}x$.

Shapiro-Wilk test

Shapiro-Wilk test is another very frequently used normality test, which uses order statistics $X_{(i)}$ and their specific properties in the normal probability distribution to calculate the test statistic.

Normality tests based on skewness and kurtosis

For random variable X, we define

▶ skewness *(šikmost)*: $\alpha_3 = \dfrac{\mathsf{E}(X^3)}{[\mathsf{Var}(X)]^{3/2}}$,

▶ kurtosis *(špičatost)*: $\alpha_4 = \dfrac{\mathsf{E}(X^4)}{\mathsf{Var}(X)}$.

Specifically for gaussian $X \sim \mathsf{N}(\mu,\sigma^2)$, it holds $\alpha_3 = 0$ and $\alpha_4 = 3$.

Tests of normality using these features of the nornmal distribution:

▶ **D'Agostino's** K-squared **test**,

▶ **Jarque-Bera test**.
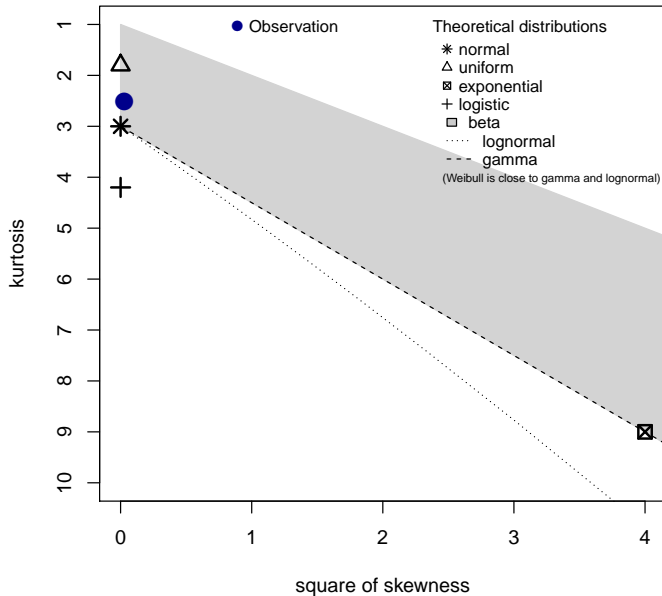
```
library("fitdistrplus")
descdist(X)
```

Pearson's chi-squared	`chisq.test(X, p=...)`	
Kolmogorov-Smirnov	`ks.test(X, "pnorm", mean=..., sd=...)`	
Lillieforse	`lillie.test(X)`	$\star$
Pearson	`pearson.test(X)`	$\star$
Anderson-Darling	`ad.test(X)`	$\star$
Cramér-von Mieses	`cvm.test(X)`	$\star$
Shapiro-Wilk	`shapirotest(X)`	
D'Agostino	`agostino.test(X)`	$*$
Jarque-Bera	`jarque.test(X)`	$*$
quantile-quantile plot	`qqnorm(X), qqline(X)`	

$*$ library("moments"), $\star$ library("nortest")

- ▶ empirical and theoretical frequencies, calculations
- ▶ Pearson's chi-squared test, calculation of test statistics, adjustment of the degrees of freedom
- ▶ algorithms for discrete and absolutely continuos random variables
- ▶ empirical distribution function, Kolmogorov-Smirnov test, geometric interpration, Lilliefors test
- ▶ specific tests for Poisson exponential and normal distribution

---
Think about …

1. Assume random sample $(X_1, \ldots, X_n)$ from an absolutely continuous probability distribution with cumulative distribution function $F_0(x)$.

2. Transform the observations, $Y_i = F_0(X_i)$.

3. What is the probability distribution of the transformed random sample $(Y_1, \ldots, Y_n)$?

Statistics II │ 5

Correlation coefficients,
multiple linear regression

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

**Pearson's correlation,
variance-covariance matrix,
correlation matrix**

Probability theory:

▶ random variable *(náhodná veličina)* X having some probability distribution *(rozdělení pravděpodobnosti)* with non-random parameters (numbers)

▶ cumulative distribution function (c.d.f.) *(distribuční funkce)* $F(x) = P(X \leq x), P(a < X \leq b) = F(b) - F(a)$

▶ discrete X: probability mass function (p.m.f.) *(pravděpodobnostní funkce)* $p(x) = P(X = x)$

▶ absolutely continuous X: probability density function (p.d.f.) *(hustota pravděpodobnosti)* $f(x), P(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$

Statistics:

▶ random sample *(náhodný výběr)* $X = (X_1, \ldots, X_n)$

▶ So-called sample statistics *(výběrové statistiky)* are functions of the random sample. They are random variables, i.e., have probability distributions.

- expected value / mean *(střední / očekávaná hodnota)*:
$$\mathsf{E}(X) = \sum_x x\, p(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$$

- sample mean *(výběrový průměr)*: $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

- variance *(rozptyl)*:
$$\mathsf{Var}(X) = \mathsf{E}[X - \mathsf{E}(X)]^2 = \sum_x [x - \mathsf{E}(X)]^2\, p(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} [x - \mathsf{E}(X)]^2\, f(x)\, \mathrm{d}x$$

- sample variance *(výběrový rozptyl)*: $S_X^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

- standard deviation *(směrodatná odchylka)*: $\sigma_X = \sqrt{\mathsf{Var}(X)}$

- sample standard deviation *(výběrová směrodatná odchylka)*: $S_X = \sqrt{S_X^2}$

▶ covariance *(kovariance)*: $\mathsf{C}(X, Y) = \mathsf{E}\left([X - \mathsf{E}(X)][Y - \mathsf{E}(Y)]\right)$

▶ sample covariance *(výběrová kovariance)*:
$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

▶ Pearson's correlation coefficient *(korelační koeficient)*:
$$\rho(X, Y) = \frac{\mathsf{C}(X, Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}} \qquad \in [-1; 1]$$

▶ sample Pearson's correlation *(výběrový korelační koeficient)*:
$$r_{XY} = r(X, Y) = \frac{S_{XY}}{S_X \cdot S_Y} \qquad \in [-1; 1]$$

Remember: Sample statistics $\overline{X}$, S_X^2, S_X, S_{XY}, r_{XY}, etc. are random variables. They are estimates of the corresponding theoretical parameters.

$H_0 : \rho_{XY} = 0$, i.e., random variables are uncorrelated *(nekorelované)*;
$H_1 : \rho_{XY} \neq 0$, i.e., random variables are correlated *(korelovanbé)*

Theorem

Under H_0, for $n \geq 3$, test statistic T has Student t$(n-2)$ distribution,

$$T = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \sim \text{t}(n-2).$$

H_0 is rejected at the level of significance α, if $|T| \geq t_{1-\alpha/2}(n-2)$.

▶ Pearson's correlation coefficient measures the strength of association between random variables X, Y and the direction of the relationship.

▶ **Stochastic independence of X, Y implies the uncorrelation.**

▶ H_0 not rejected, $r_{XY} \approx 0 \Rightarrow X$, Y are uncorrelated
But: **Uncorrelation does not imply stochastic independence of X, Y!**

▶ H_0 rejected, $r_{XY} \approx \ \ 1 \Rightarrow X$, Y correlated, positive relationship

▶ H_0 rejected, $r_{XY} \approx -1 \Rightarrow X$, Y correlated, negative relationship

$H_0 : \rho_{XY} = \rho_0$ for a given fixed value $\rho_0 \in [-1; 1]$,
$H_1 : \rho_{XY} \neq \rho_0$

Theorem (R. A. Fisher)

Under H_0, test statistic Z, so-called Z-transformation, has asymptotically normal distribution,

$$Z = \frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}} \overset{as.}{\sim} \mathsf{N}\left(\frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0}, \ \frac{1}{n - 3}\right).$$

H_0 is rejected at the level of significance α, if

$$\frac{|Z - \mathsf{E}(Z)|}{\sqrt{\mathsf{Var}(Z)}} = \sqrt{n - 3} \left| Z - \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0} \right| \geq u_{1 - \alpha/2}.$$

Example (1)

Expenses (E) of 7 households (thousands CZK per 3 months) for food and beverages were observed depending on the number of household members (M) and the net income (I) of the household (thousands CZK per 3 months).

E	40	30	40	10	60	40	50
M	4	2	4	1	5	3	4
I	100	80	120	30	150	120	130

Analyze and quantify the association (correlation) of the variables.

Example (2)

20 children of different ages underwent pedagogical-psychological research, during which, among other things, they answered test questions and were weighed. Surprising was the value 0.968 of Pearson's correlation coefficient between the children's weight and the number of points achieved in the test. Does this mean that obesity has a positive effect on learning ability?

Sample Pearson's correlation: $r_{ME} \doteq 0.942$

$$T = 0.942\sqrt{\frac{5}{1 - 0.942^2}} = 6.326 > t_{0.975}(5) = 2.571 \Rightarrow \rho_{ME} \text{ is significant}$$

$$Z: \sqrt{4}\left|0 - \frac{1}{2}\ln\frac{1 + 0.942}{1 - 0.942}\right| = 3.526 > u_{0.975} = 1.96 \Rightarrow \rho_{ME} \text{ is significant}$$

Expanses and number of members of household are correlated, with positive relationship.

```
cor(dt$expense, dt$members)
[1] 0.9762737

cor.test(dt$expense, dt$members)
  Pearson's product-moment correlation
data:  dt$expense and dt$members
t = 6.3263, df = 5, p-value = 0.001455
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544368 0.9917452
sample estimates:
      cor
0.9428374
```

Let us assume l random variables $X_1, \ldots, X_l$ are examined.
We have an l-dimensional random sample of size n,

$$M = \begin{pmatrix} X_{11}, & \cdots, & X_{1l} \\ \vdots & & \vdots \\ X_{n1}, & \cdots, & X_{nl} \end{pmatrix},$$

where X_{ij} denotes the i-th observation of X_j, $i = 1, \ldots, n$, $j = 1, \ldots, l$.

Definition

▶ Matrix S of sample covariances $S_{X_i X_j}$ of all pairs of random variables is called sample variance-covariance matrix *(výběrová kovarianční matice)*,

$$S = \left\{ S_{X_i X_j} \right\}_{i,j=1}^{l}$$

▶ Matrix R of sample correlation coefficients $r_{X_i X_j}$ of all pairs of random variables is called sample correlation matrix *(výběrová korelační matice)*,

$$R = \left\{ r_{X_i X_j} \right\}_{i,j=1}^{l}$$

▶ Sample variance-covariance matrix S is squred $l \times l$, symmetrical positive definite matrix.

▶ The diagonal of S consists of sample variances $S_{X_1}^2, \ldots, S_{X_l}^2$.

▶ Sample correlation matrix R is squred $l \times l$, symmetrical matrix.

▶ The diagonal of R consists of l ones.

Graphical tools:

▶ Scatterplot is a matrix of two-dimensional point plots of the relationships of variable X_j on variable X_i, for each i, j.

▶ Correlogram is a visual representation of the sample correlation matrix R. The color or size of a symbol indicates the value of the corresponding sample correlation coefficient. Correlogram can also display information about significance of the correlation coefficients.

```
cov(M) # Variance-covariance matrix          cor(M) # Correlation matrix
          members    income    expense                members    income    expense
members   1.904762   50.2381   20.47619       members 1.0000000 0.9210550 0.9428374
income    50.238095  1561.9048 607.14286      income  0.9210550 1.0000000 0.9762737
expense   20.476190  607.1429  247.61905      expense 0.9428374 0.9762737 1.0000000
```
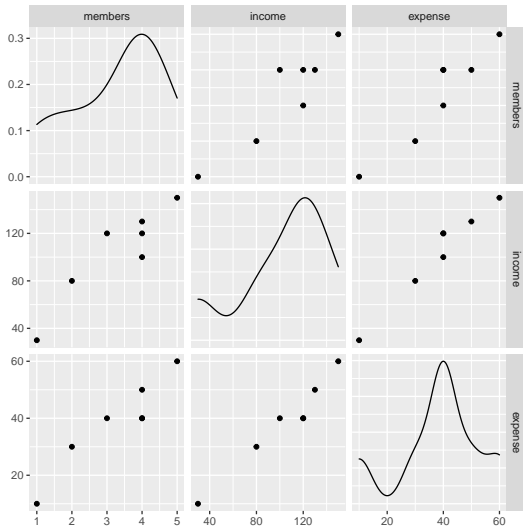
```
# Correlation matrix                        # p-values of tests of correlation
           members     income    expense                members        income        expense
members  1.0000000  0.9210550  0.9428374    members  0.000000000  0.0032221604  0.0014548012
income   0.9210550  1.0000000  0.9762737    income   0.003222160  0.0000000000  0.0001644319
expense  0.9428374  0.9762737  1.0000000    expense  0.001454801  0.0001644319  0.0000000000
```


Pearson's correlations
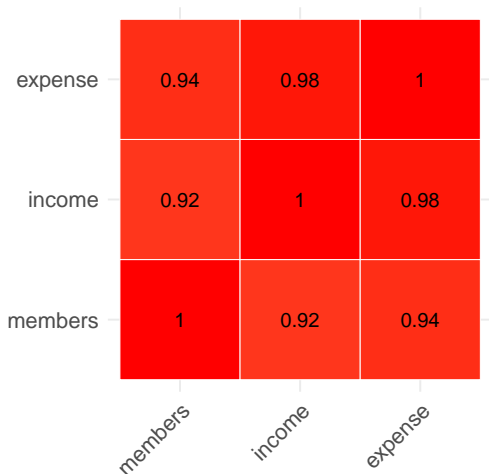
sample mean	$\overline{X}$	`mean(X)`
sample variance	S_X^2	`var(X)`
sample standard deviation	S_X	`sd(X)`
sample covariance	S_{XY}	`cov(X, Y)`
sample correlation	r_{XY}	`cor(X, Y), cor.test(X, Y)`
sample variance-covariance matrix	S	`cov(M)`
sample correlation matrix	R	`cor(M), Hmisc::rcorr(M)`
scatterplot		`GGally::ggpairs`
correlogram		`ggcorrplot::ggcorrplot`

Rank-based correlation coefficients

Definition (Spearman's correlation)

Assume a pair of random samples, $X = (X_1, \ldots, X_n)$, $X = (Y_1, \ldots, Y_n)$.
Denote $R = (R_1, \ldots, R_n)$, $S = (S_1, \ldots, S_n)$ the ranks of particular samples.
Sample Spearman's rank correlation coefficient *(Spearmanův výběrový pořadový korelační koeficient)* of random variables X and Y is defined as Pearson's correlation of vectors of their ranks,

$$r_S(X, Y) = r(R, S).$$

If the ranks are not averaged, then $r_S = 1 - 6 \dfrac{\sum_{i=1}^n (R_i - S_i)^2}{n(n^2 - 1)}$.

▶ Spearman's correlation $r_S \in [-1; 1]$ quantifies rank correlation between random variables X and Y;

▶ is nonparametric analogy of Pearson's correlation;

▶ typically used for ordinal or non-gaussian data.

Test of rank correlation *(Test pořadové korelovanosti)*

$H_0 : r_S = 0$,

$H_1 : r_S \neq 0$, i.e., random variables X and Y are rank-correlated *(pořadově korelované)*

Theorem

▶ Under H_0, test statistic $T_S = r_s \sqrt{\dfrac{n-2}{1-r_S^2}} \sim \mathsf{t}(n-2)$.

H_0 is rejected at the level of significance α, if $|T_S| \geq t_{1-\alpha/2}(n-2)$.

▶ Under H_0, test statistic $Z_S = \sqrt{\dfrac{n-3}{1.06}} \cdot \dfrac{1}{2} \ln \dfrac{1+r_S}{1-r_S} \overset{as.}{\sim} \mathsf{N}(0;1)$.

H_0 is rejected at the asymptotic level of significance α, if $|Z_S| \geq u_{1-\alpha/2}$.

`cor(..., method="spearman")`, `rcorr(..., type="spearman")`, `cor.test(..., method="spearman")`

E_i	40	30	40	10	60	40	50
M_i	4	2	4	1	5	3	4
I_i	100	80	120	30	150	120	130
R_i = rank of V_i	4	2	4	1	7	4	6
S_i = rank of C_i	5	2	5	1	7	3	5
T_i = rank of P_i	3	2	4.5	1	7	4.5	6

$$r_S(M, E) = r(\boldsymbol{R}, \boldsymbol{S}) = \frac{\sum_{i=1}^n (R_i S_i) - n \overline{R}\,\overline{S}}{\sqrt{\sum_{i=1}^n R_i^2 - n\overline{R}^2}\sqrt{\sum_{j=1}^n S_j^2 - n\overline{S}^2}} \doteq 0.923$$

```
[1] 0.5174525
# Spearman's correlation by definition
cor(rank(dt$expense), rank(dt$members))
[1] 0.9230769
cor(dt$expense, dt$members, method = "spearman")
[1] 0.9230769
cor.test(dt$expense, dt$members, method = "spearman") # Test of rank order
  Spearman's rank correlation rho
data: dt$expense and dt$members
S = 4.3077, p-value = 0.003023
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
```
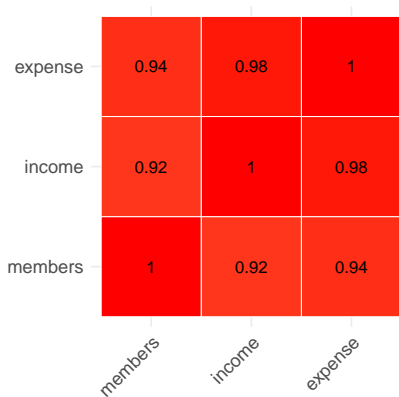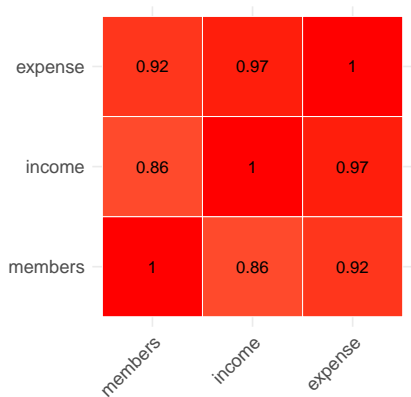
Definition (Kendall's tau)

Sample Kendall's rank correlation coefficient / Kendall's tau *(Kendallův výběrový korelační koeficient)* of random variables X and Y is

$$\tau(X, Y) = \frac{n_+ - n_-}{\sqrt{n_0 - n_X}\sqrt{n_0 - n_Y}}, \qquad \text{where}$$

▶ $n_0 = \frac{1}{2}n(n-1) = \binom{n}{2} =$ number of all pairs,

▶ $n_+ =$ number of concordant pairs,

▶ $n_- =$ number of discordant pairs,

▶ $n_X = \frac{1}{2}\sum_i u_i(u_i - 1)$, $n_Y = \frac{1}{2}\sum_j v_j(v_j - 1)$,
and u_i, v_j are numbers of particular ties in X and Y.

Pairs (X_i, Y_i) and (X_j, Y_j) are called:

▶ concordant *(konkordantní)*, if $\quad X_i < X_j \ \& \ Y_i < Y_j \quad$ or $\quad X_i > X_j \ \& \ Y_i > Y_j$;

▶ discordant *(diskordantní)*, if $\quad X_i < X_j \ \& \ Y_i > Y_j \quad$ or $\quad X_i > X_j \ \& \ Y_i < Y_j$.

- ▶ Kendall's tau $\tau \in [-1; 1]$ quantifies ordinal association between random variables X and Y;
- ▶ in contrast to Spearman's correlation, it does not consider the distance between ranks;
- ▶ typically used for ordinal data, e.g., for two types of ranking.

Test of ordinal association *(Test ordinální asociace)*

$H_0 : \tau = 0$,
$H_1 : \tau \neq 0$, i.e., random variables X and Y are ordinally associated

---
Theorem

In the case of no ties in the data, H_0 is rejected at the asymptotic level of significance α, if $\sqrt{\dfrac{9n(n-1)}{2(2n+5)}} \, |\tau| \geq u_{1-\alpha/2}$.

For small n or in the presence of ties, corrected test statistics can be used.

```
cor(..., method="kendall"), rcorr(..., type="kendall"), cor.test(..., method="kendall")
```

E_i	40	30	40	10	60	40	50
M_i	4	2	4	1	5	3	4
I_i	100	80	120	30	150	120	130

- $n_0 = \frac{1}{2} 7 \cdot 6 = \binom{7}{2} = 21$ pairs in total
- $n_+ = 16$ concordant pairs (all except 1-3, 1-6, 1-7, 3-6, 3-7)
- $n_- = 0$ discordant pairs
- 3×4 in $M \Rightarrow n_M = \frac{1}{2} 3 \cdot 2 = 3$; 3×40 in $E \Rightarrow n_E = \frac{1}{2} 3 \cdot 2 = 3$
- $\tau(M, E) = \dfrac{n_+ - n_-}{\sqrt{n_0 - n_M}\sqrt{n_0 - n_E}} = \dfrac{16 - 0}{\sqrt{21 - 3}\sqrt{21 - 3}} \doteq 0.889$

```
cor(dt$expense, dt$members, method = "kendall")
[1] 0.8888889
cor.test(dt$expense, dt$members, method = "kendall") # Test of ordinal association
  Kendall's rank correlation tau
data:  dt$expense and dt$members
z = 2.6146, p-value = 0.008933
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8888889
```

Multiple linear regression model

Assume random variable Y dependens on l random variables $X_1, \ldots, X_l$,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_l X_l + \varepsilon = \beta_0 + \sum_{j=1}^{l} \beta_j X_j + \varepsilon$$

with random error ε.

n observations $(X_{i1}, \ldots, X_{il}, Y_i)$, $i = 1, \ldots, n$, are collected, where X_{ij} denotes the i-th observation of random variable X_j.

Multiple linear regression model:

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \cdots + \beta_l X_{1l} + \varepsilon_1, \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \cdots + \beta_l X_{nl} + \varepsilon_n,
\end{aligned}
$$

written in matrix form as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & X_{11} & \cdots & X_{1l} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nl} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_l \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}, \quad \text{i. e., } Y = X\beta + \varepsilon.$$

$$Y = X\beta + \varepsilon$$

- $\beta = (\beta_0, \beta_1, \ldots, \beta_l)'$ = vector of $k = l + 1$ regression coefficients *(regresní koeficienty)*,
- X = regression / design matrix *(matice plánu)* of type $(n \times k)$ consists of a column of ones and of l columns of regressors *(regresory)*, $(X_{11}, \ldots, X_{n1})', \ldots, (X_{1l}, \ldots, X_{nl})'$,
- $n > k$,
- $r(X) = k = l + 1$, i. e., the design matrix has full rank *(plná hodnost)*, its columns are linearly independent.

Random errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$:

- are nonsystematic: $E(\varepsilon_i) = 0$, i.e., $E(\varepsilon) = \mathbf{0}$ and $E(Y) = X\beta$,
- have homogeneous variance: $\text{Var}(\varepsilon_i) = \sigma^2 > 0$,
- are mutually uncorrelated: $C(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$;
- variance-covariace matrix *(kovarianční matice)* of the vector of observations is $\text{Var}(Y) = \text{Var}(\varepsilon) = \sigma^2 I_n$.
- Hence, observations are uncorrelated and have homogeneous variance.

Optimization: find such $\boldsymbol{\beta}$ which minimizes the sum of quadratic deviations,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Big[Y_i - \beta_0 - \sum_{j=1}^{l}\beta_j x_{ij}\Big]^2 = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \longrightarrow \min.$$

▶ Ordinary Least Squares (OLS) estimate *(odhad metodou nejmenších čtverců)*
$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_l) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y},$$

▶ predicted / fitted values: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, i.e., $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{l}\widehat{\beta}_j x_{ij}$,

▶ residuals *(rezidua)* $r_i = Y_i - \widehat{Y}_i$,

▶ residual sum of squares *(reziduální součet čtverců)*
$$S_e = S(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sum_{i=1}^{n}\Big[Y_i - \widehat{\beta}_0 - \sum_{j=1}^{l}\widehat{\beta}_j X_{ij}\Big]^2 = \sum_{i=1}^{n}r_i^2,$$

▶ coefficient of determination *(index determinace)* R squared:
$$R^2 = \frac{S_{\widehat{Y}}}{S_T} = 1 - \frac{S_e}{S_T}, \text{ where } \quad S_{\widehat{Y}} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2, \ S_T = \sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

▶ adjusted - R bar squared: $\overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2).$

Theorem (Gauss-Markov)

OLS estimate $\widehat{\beta}_{OLS}$ is BLUE = Best Linear Unbiased Estimate *(nejlepší nestranný lineární odhad)* of vector β and its variance-covariance matrix *(kovarianční matice)* is $\mathsf{Var}(\widehat{\beta}_{OLS}) = \sigma^2\,(X'X)^{-1}$.

Theorem

$\widehat{\sigma^2}_{OLS} = \dfrac{S_e}{n - (l+1)} = \dfrac{S_e}{n-k}$ is an unbiased estimate of the variance σ^2 of random errors.

Theorem

Additionally, let us assume that the observations have n-dimensional gaussian (normal) disribution $Y \sim \mathsf{N}_n\Big(X\beta,\, \sigma^2 I_n\Big)$. Then:

▶ OLS estimate has gaussian distribution, $\widehat{\beta}_{OLS} \sim \mathsf{N}_k\Big(\beta,\, \sigma^2(X'X)^{-1}\Big)$,

▶ statistic $K = (n-k)\dfrac{\widehat{\sigma^2}_{OLS}}{\sigma^2} \sim \chi^2(n-k)$ has chi-square distribution,

▶ OLS estimate $\widehat{\beta}_{OLS}$ and statistic K are independent.

> **Definition**
>
> **Random variable**
>
> $$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_l,$$
>
> where $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_l}$ are the OLS-estimates, is called best linear approximation *(nejlepší lineární aproximace)* of random variable Y using random variables $(X_1, \ldots, X_l)$.
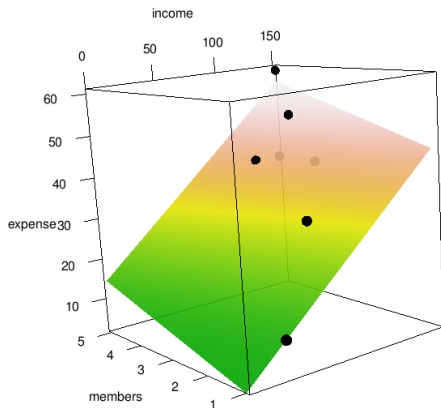
The graph of the best linear approximation $\widehat{Y}$ in dependency on variables $(X_1, \ldots, X_l)$ is **l-dimensional hyperplane in a k-dimensional vector space**, $k = l + 1$.

The household expenses are modeled in dependency on income and the number of household members.

Multiple linear regression model: $E = \beta_0 + \beta_1 \cdot I + \beta_2 \cdot M + \varepsilon$

```
lm(formula = expense ~ income + members, data = dt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.74142    4.08119  -0.427   0.6916
income       0.28320    0.09473   2.990   0.0404 *
members      3.28056    2.71254   1.209   0.2931
---
Residual standard error: 3.571 on 4 degrees of freedom
Multiple R-squared:  0.9657,	Adjusted R-squared:  0.9485
F-statistic: 56.25 on 2 and 4 DF,  p-value: 0.001179
```

OLS-estimates: $\widehat{\beta_0} = -1.741$, $\widehat{\beta_1} = 0.283 > 0$, $\widehat{\beta_2} = 3.281$; $R^2 = 0.966$
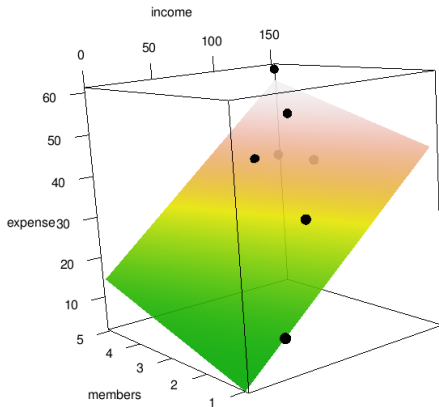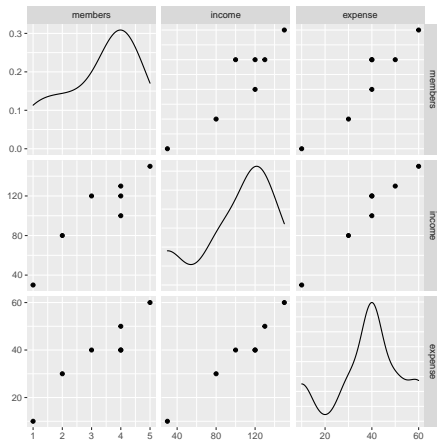
Best linear approximation $\widehat{E}$ of E using I and M is **regression plane** given by
$$\widehat{E} = -1.741 + 0.283 \cdot I + 3.281 \cdot M$$

Multiple and partial correlation

Multiple correlation coefficient *(koeficient mnohonásobné korelace)* $\rho_{Y \cdot X_1 \ldots X_l}$ is Pearson's correlation between variable Y and its best linear approximation $\widehat{Y}$ using variables $X_1, \ldots, X_l$,

$$\rho_{Y \cdot X_1 \ldots X_l} = \rho(Y, \widehat{Y}) \in [0; 1].$$

It quantifies the association between Y and the vector $(X_1, \ldots, X_l)$. It is the **largest correlation** between Y and any linear combination of variables $X_1, \ldots, X_l$.

Sample multiple correlation coefficient $\quad r_{Y \cdot X_1 \ldots X_l} = r(Y, \widehat{Y}) \in [0; 1]$ is sample Pearson's correlation between the observation vector Y and **fitted values** $\widehat{Y}$ using predictors $X_1, \ldots, X_l$ in multiple linear regression model

$$M : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Theorem (Test of significance of multiple correlation)

$H_0 : \rho_{Y \cdot X_1 \ldots X_l} = 0$ against one-sided alternative is rejected at the level of significance α, if $\quad F = \dfrac{n - l - 1}{l} \cdot \dfrac{r_{Y \cdot X_1 \ldots X_l}^2}{1 - r_{Y \cdot X_1 \ldots X_l}^2} \geq F_{1-\alpha}(l, \ n - l - 1).$

Partial correlation coefficient *(koeficient parciální korelace)* $\rho_{XY \cdot Z_1 \ldots Z_m}$ is Pearson's correlation between variables $(X - \widehat{X})$ and $(Y - \widehat{Y})$, where $\widehat{X}, \widehat{Y}$ are the best linear approximations of X, Y using variables $Z_1, \ldots, Z_m$,

$$\rho_{XY \cdot Z_1 \ldots Z_m} = \rho(X - \widehat{X}, \ Y - \widehat{Y}) \in [-1; 1].$$

It quantifies the association between variables X and Y **excluding the effect of variables** $Z_1, \ldots, Z_m$.

Sample partial correlation coefficient
$$r_{XY \cdot Z_1 \ldots Z_m} = r(\ \underbrace{X - \widehat{X}}_{\text{residuals in } M_X}, \ \underbrace{Y - \widehat{Y}}_{\text{residuals in } M_Y}\ ) \in [-1; 1],$$

is sample Pearson's correlation between the vectors of **residuals** $(X - \widehat{X})$ and $(Y - \widehat{Y})$ in multiple linear regression models

$$M_X : X = \alpha_0 + \alpha_1 Z_1 + \cdots + \alpha_m Z_m; \qquad M_Y : Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_m Z_m.$$

Theorem (Test of significance of partial correlation)

$H_0 : \rho_{XY \cdot Z_1 \ldots Z_m} = 0$ against two-sided alternative is rejected at the level of significance α, if $\quad T = r_{XY \cdot Z_1 \ldots Z_m} \sqrt{\dfrac{n - m - 2}{1 - r_{XY \cdot Z_1 \ldots Z_m}^2}} \geq t_{1-\alpha/2}(n - m - 2).$
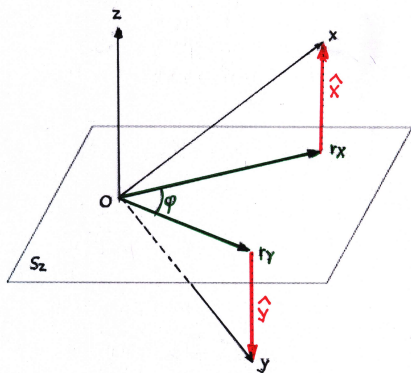
- x, y, z = three observations in 3D
- S_z = hyperplane perpendicular to z
- $\widehat{x}$ = best linear approximation of x using z
- $\widehat{y}$ = best linear approximation of y using z
- $r_x = x - \widehat{x}$ = residuals of x = projection of x into S_z
- $r_y = y - \widehat{y}$ = residuals of y = projection of y into S_z
- partial correlation = cosine of the angle of residuals in S_z,

$$r_{XY.Z} = \cos \varphi = \cos |\sphericalangle\, r_X,\, r_Y|$$

mutiple correlation $r_{Y \cdot X_1 \dots X_l}$

```
model.M <- lm(Y ~ X1 + ... + Xl) # model M
# multiple correlation between Y and X1, ..., Xl
# as correlation between Y and best linear approximation of Y using X1, ..., Xl in m
cor(Y, fitted.values(model.M))
```

partial correlation $r_{X Y \cdot Z_1 \dots Z_m}$

```
model.MX <- lm(X ~ Z1 + ... + Zm) # model MX
model.MY <- lm(Y ~ Z1 + ... + Zm) # model MY
# partial correlation between X and Y excluding the effect of Z1, ..., Zm
cor(residuals(model.MX), residuals(model.MY))

library("ppcor")  # or using functions from "ppcor" library
# partial correlations of each pair of variables exluding all other variables
pcor(M)
# test of significance of partial correlation between X and Y excluding Z
pcor.test(X, Y, Z)
```
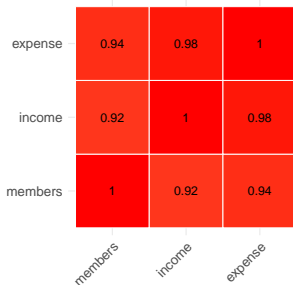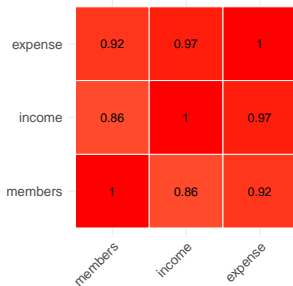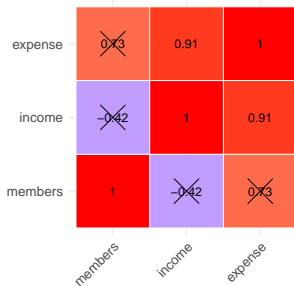
Number of point achieved in the test are modeled in dependency on weight and ages of the children.

Multiple linear regression model: $Points = \beta_0 + \beta_1\,Weight + \beta_2\,Age + \varepsilon$

```
lm(formula = Points ~ Weight + Age, data = dt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.06490    1.23693   8.945 7.72e-08 ***
Weight       0.09466    0.12090   0.783    0.444
Age          3.19203    0.51058   6.252 8.77e-06 ***
---
Residual standard error: 1.377 on 17 degrees of freedom
Multiple R-squared:  0.9806,  Adjusted R-squared:  0.9784
F-statistic: 430.4 on 2 and 17 DF,  p-value: 2.753e-15
```

OLS-estimates: $\widehat{\beta_0} = 11.065 > 0$, $\widehat{\beta_1} = 0.095$, $\widehat{\beta_2} = 3.192 > 0$; $R^2 = 0.981$
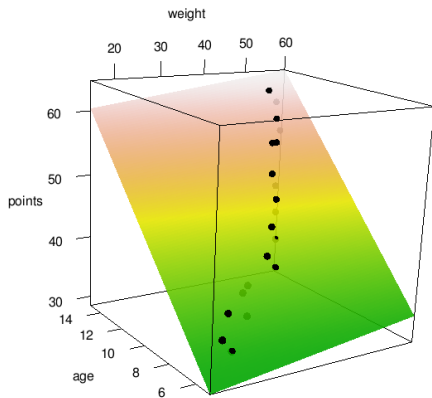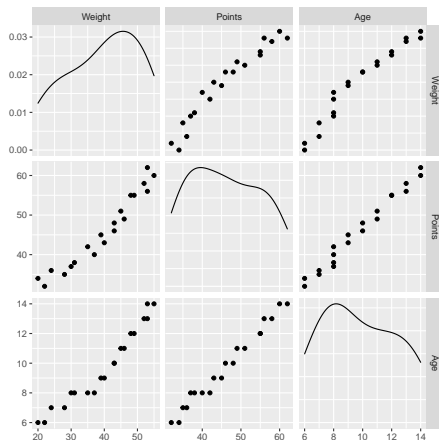
Best linear approximation $\widehat{Points}$ of *Points* using *Weight* and *Age* is **regression plane** given by

$$\widehat{Points} = 11.065 + 0.095\,Weight + 3.192\,Age$$
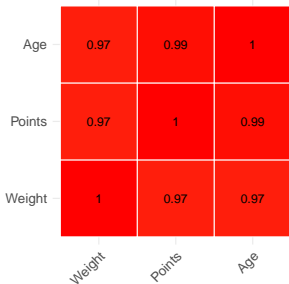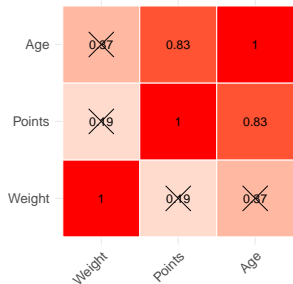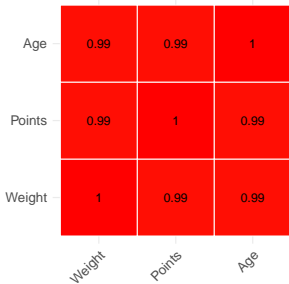
Eample 2: Correlograms
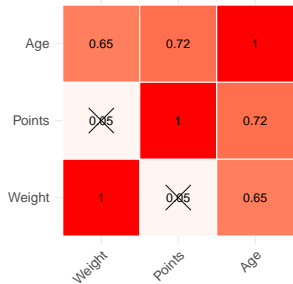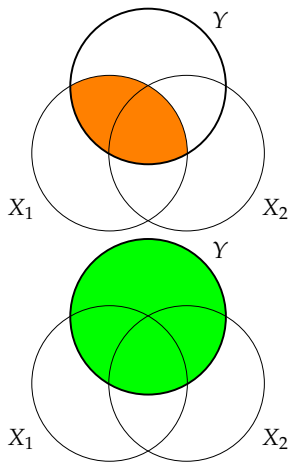
- variable Y is modeled by variable X_1
- Pearson's correlation r_{YX_1} quantifies the association between Y and X_1
- coefficient of determination $R^2_{Y \cdot X_1}$ in model
  `Y ~ X1`
  quantifies the ratio of variability of Y which is explained by X_1
- the square $r^2_{YX_1}$ of the Pearson's correlation is equal to the coefficient of determination,

$$r^2_{YX_1} = \frac{R^2_{YX_1}}{1} = R^2_{YX_1}$$
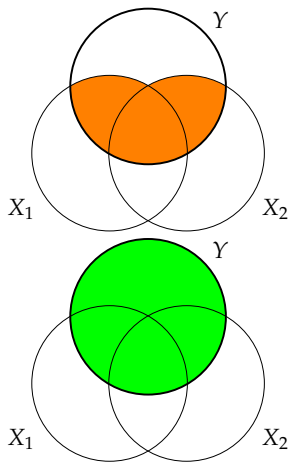
- ▶ variable Y is modeled by variables X_1, X_2
- ▶ multiple correlation $r_{Y \cdot X_1 X_2}$ quantifies the association between Y and its best linear approximation $\widehat{Y}$ using X_1, X_2
- ▶ coefficient of determination $R^2_{Y \cdot X_1 X_2}$ in model
  Y ~ X1 + X2
  quantifies the ratio of variability of Y which is explained by X_1 and X_2
- ▶ the square $r^2_{Y \cdot X_1 X_2}$ of the multiple correlation is equal to the coefficient of determination,

$$r^2_{Y \cdot X_1 X_2} = \frac{R^2_{Y \cdot X_1 X_2}}{1} = R^2_{Y \cdot X_1 X_2}$$

Y ~ X1 + X2
$R^2_{Y \cdot X_1 X_2}$

Y ~ X2
$R^2_{Y \cdot X_2}$

$R^2_{Y \cdot X_1 X_2} - R^2_{Y \cdot X_2}$
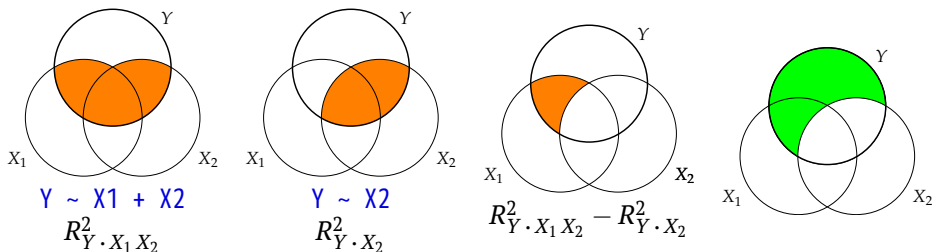
- ▶ variable Y is modeled by variable X_1 and excluding variable X_2
- ▶ partial correlation $r_{Y X_1 \cdot X_2}$ quantifies the association between Y and X_1 excluding the effect of X_2
- ▶ coefficient of determination $R^2_{Y \cdot X_2}$ in model Y ~ X2 quantifies the ratio of variability of Y which is explained by X_2
- ▶ the square $r^2_{Y X_1 \cdot X_2}$ of the partial correlation is equal to the ratio ov variability of Y which is explained by X_1 independently on X_2,

$$r^2_{Y X_1 \cdot X_2} = \frac{R^2_{Y \cdot X_1 X_2} - R^2_{Y \cdot X_2}}{1 - R^2_{Y \cdot X_2}}$$

- ▶ Note that Y ~ X2 is a submodel of Y ~ X1 + X2.

Mutiple correlation $\rho_{E \cdot MI}$

```
model.M <- lm(expense ~ income + members, data = dt)
cor(dt$expense, fitted.values(model.M)) # by definition, as correlation between var
sqrt(summary(model.M)$r.squared)        # or, as square root of R squared
[1] 0.9826827
```

$$r_{E \cdot IM} = 0.983, \ R^2_{E \cdot IM} = r^2_{V \cdot CP} = 0.966$$

Partial correlation $\rho_{EM \cdot I}$

```
R.partial$estimate["expense", "members"] # from the partial-correlation matrix
model.M1 <- lm(expense ~ income, data = dt)
model.M2 <- lm(members ~ income, data = dt)
cor(residuals(model.M1), residuals(model.M2)) # by definition, as correlation betwe
sqrt(( summary(model.M)$r.squared - summary(model.M1)$r.squared ) /
  ( 1 - summary(model.M1)$r.squared )) # or, using relationship with R squared
[1] 0.5174525
```

$$\rho_{EM \cdot I} = 0.517$$

- ▶ Pearson's correlation: definition, calculation, test of significance, interpretation
- ▶ Spearman's and Kendall's rank correlation: definition, calculation, test, interpretation, concordant and discordant pairs
- ▶ Correlation matrix, correlogram, scatterplot
- ▶ Multiple linear regression: model, best linear approximation, geometric interpretation
- ▶ Coefficients of multiple and partial correlation: definition, calculation, interpretation, relation to R-squared

Statistics II │ 6

Autocorrelation and multicollinearity
in linear regression model

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

17 October 2022

Autocorrelation

$$Y = X\beta + \varepsilon, \qquad \text{i.e.,} \quad Y_i = \beta_0 + \sum_{j=1}^{l} \beta_l x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

Assumptions on random errors ε_i:

▶ nonsystematic, $\quad \mathsf{E}(\varepsilon_1) = \cdots = \mathsf{Var}(\varepsilon_n) = 0$,

▶ homogeneous variance, $\quad \mathsf{Var}(\varepsilon_1) = \cdots = \mathsf{Var}(\varepsilon_n) = \sigma^2$,

▶ uncorrelated, $\quad \rho(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$

▶ i.e., the variance-covariance matrix is
$$V = \mathsf{Var}(\varepsilon) = \mathsf{Var}(Y) = \sigma^2 \, I_n = \mathrm{diag}\left(\sigma^2, \ldots, \sigma^2\right),$$

▶ normal (gaussian) probability distribution, $\quad \varepsilon \sim \mathsf{N}_n\left(\mathbf{0}, \sigma^2 \, I_n\right)$

If the variance-covariance matrix V does not have the required form? …

If random errors are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, but they **do not have homogeneous variance**, their variance-covariance matrix V is diagonal,

$$V = \text{Var}(\varepsilon) = \text{Var}(Y) = \text{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right).$$

Then, the OLS-estimate $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$ is not correct.

Inverse to the variance-covariance matrix is diagonal, too,

$$V^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right),$$

and correct estimate of the vector of regresion coefficients is

Definition (weighted least squares (WLS) estimate)

$$\widehat{\beta}_{\text{WLS}} = (X'V^{-1}X)^{-1}X'V^{-1}Y.$$

it is called estimation by the weighted least squares (WLS) method *(vážená metoda nejmenších čtverců)*.

The WLS-method actually assigns weight $\frac{1}{\sigma_i^2}$ to the i-th observation.

Autocorrelation *(autokorelace)* of random errors means that the random errors are correlated, $\rho(\varepsilon_i, \varepsilon_j) \neq 0$.

Consequences of autocorrelation:

▶ Variance-covariance matrix of random errors is not of diagonal form with homogeneous variance, $\text{Var}(\varepsilon) \neq \sigma^2 I_n$.

▶ Using the ordinary least squares (OLS) method to estimate the vector of regression coefficients $\widehat{\beta}_{\text{OLS}}$ is incorrect.

▶ Variances of the OLS-estimates are biased, underestimated. This can induce a **false impression of the significance of the regression coefficients**.

▶ With a more complex structure of the covariance matrix, it is necessary to consider another, so-called extended linear model, in which the variance-covariance matrix can generally be **any symmetric and positive definite matrix**, $\text{Var}(\varepsilon) = \sigma^2 V$.

Definition (Extended linear regression model)

Extended linear regression model is
$$Y = X\beta + \varepsilon, \qquad \mathsf{E}(\varepsilon) = 0, \qquad \mathsf{Var}(\varepsilon) = \sigma^2\, V,$$
where variance-covariance matrix $\mathsf{Var}(\varepsilon) = \sigma^2\, V$ of random errors is symmetric and positive definite matrix.

Theorem (Generalized least squares method)

The estimation of the vector of regression coefficients β by generalized least squares method (GLS) *(zobecněná metoda nejmenších čtverců)* in the extended linear regression model $Y = X\beta + \varepsilon$ with variance-covariance matrix $\mathsf{Var}(\varepsilon) = \sigma^2\, V$ of random errors is
$$\widehat{\beta}_{\mathsf{GLS}} = (X'V^{-1}X)^{-1}X'V^{-1}Y.$$
It is also called Aitken's estimate. Its variance-covariance matrix is
$$\mathsf{Var}(\widehat{\beta}_{\mathsf{GLS}}) = \sigma^2(X'V^{-1}X)^{-1}.$$

The least squares estimate $\widehat{\beta}_{\mathsf{OLS}}$ corresponds to the special choice $V = I_n$.

Autoregression *(autoregrese)* **AR(p)** of random errors is a special form of relationship between consecutive errors,

$$\varepsilon_i = \theta_1 \, \varepsilon_{i-1} + \theta_2 \, \varepsilon_{i-2} + \cdots + \theta_p \, \varepsilon_{i-p} + w_i,$$

where p is the order of autoregression *(řád)*, $\theta_1, \ldots, \theta_p$ are autoregressive parameters, and w_i is white noise (WN) *(bílý šum)*,

$$\mathsf{E}(w_i) = 0, \qquad \mathsf{Var}(w_i) = \sigma_w^2, \qquad \rho(w_i, w_j) = 0 \text{ for } i \neq j.$$

Different variance-covariance matrices correspond to particular **AR(p)** autoregressions.

Examples:
AR(1) = ARMA(1, 0): $\quad \varepsilon_i = \theta \, \varepsilon_{i-1} + w_i, \qquad |\theta| < 1.$
AR(2) = ARMA(2, 0): $\quad \varepsilon_i = \theta_1 \, \varepsilon_{i-1} + \theta_2 \, \varepsilon_{i-2} + w_i.$

- $\varepsilon_i = \theta\varepsilon_{i-1} + w_i = \theta(\theta\varepsilon_{i-2} + w_{i-1}) + w_i = \theta^2\varepsilon_{i-2} + \theta w_{i-1} + w_i =$
  $= \theta^2(\theta\varepsilon_{i-3} + w_{i-1}) + \theta w_{i-1} + w_i = \cdots = \sum_{l=0}^{\infty} \theta^l w_{i-l}$

- $\mathsf{E}(\varepsilon_i) = \mathsf{E}\left(\sum_{l=0}^{\infty} \theta^l w_{i-l}\right) = \sum_{l=0}^{\infty} \theta^l \underbrace{\mathsf{E}(w_{i-l})}_{0} = 0$

- $\sigma^2 = \mathsf{Var}(\varepsilon_i) = \mathsf{Var}\left(\sum_{l=0}^{\infty} \theta^l w_{i-l}\right) = \sum_{l=0}^{\infty} \theta^{2l} \underbrace{\mathsf{Var}(w_{i-l})}_{\sigma_w^2} = \dfrac{\sigma_w^2}{1-\theta^2}$

- $\mathsf{C}(\varepsilon_i, \varepsilon_{i-r}) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \theta^k\theta^l \mathsf{C}(w_{i-k}, w_{i-r-l}) = \theta^r\sigma_w^2 \sum_{l=0}^{\infty} \theta^{2l} = \dfrac{\theta^r\sigma_w^2}{1-\theta^2} = \theta^r\,\sigma^2$

- $\mathsf{Var}(\varepsilon) = \underbrace{\dfrac{\sigma_w^2}{1-\theta^2}}_{\sigma^2} \underbrace{\begin{pmatrix} 1 & \theta & \theta^2 & \ldots & \theta^{n-1} \\ \theta & 1 & \theta & \ldots & \theta^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \theta \\ \theta^{n-1} & \ldots & \theta^2 & \theta & 1 \end{pmatrix}}_{V} = \sigma^2\,V$

Durbin-Watson test is used to detect the autoregression. It uses so called Durbin-Watson test statistic $D \in [0, 4]$.

---

in AR(1)

$H_0 : \theta = 0$, i.e., random errors are uncorrelated,
$H_1 : \theta \neq 0$, i.e., random errors are correlated,

Durbin-Watson test in AR(1)

Durbin-Watson statistic is $\quad D = \dfrac{\sum\limits_{i=2}^{n} (r_i - r_{i-1})^2}{\sum\limits_{i=1}^{n} r_i^2} \quad \in [0, 4]$,

where r_i is the ith residual (i.e., the realization of the ith random error ε_i).
The decision according is by critical values of D-W test, or by the p-value.

- $D \approx 2 \Rightarrow$ uncorrelated errors,
- $D \to 0 \Rightarrow$ positively correlated errors,
- $D \to 4 \Rightarrow$ negatively correlated errors.

Consider so-called weakly stationary *(slabě stacionární)* random series *(časová řada)* of random errors $\varepsilon_1, \ldots, \varepsilon_n$, where Pearson's correlations $\rho(\varepsilon_i, \varepsilon_j)$ between random errors depends on the difference $l = j - i$.

Autocorrelation function (ACF) *(autokorelační funkce)* is
$$\rho(l) = \rho(\varepsilon_i, \varepsilon_{i-l}), \quad l = 0, 1, 2, \ldots$$
ACF is estimated by sample Pearson's correlations $r(r_i, r_{i-l})$ of the residuals $r_1, \ldots r_n$. The variable l is called lag.

- $\rho(0) = 1$
- For uncorrelated errors, $\rho(l) = 1$ for $l = 1, 2, \ldots$
- Asymptotic confidence bounds when errors are uncorrelated:
$$|\rho(l)| < \frac{u_{1-\alpha/2}}{\sqrt{n}}.$$
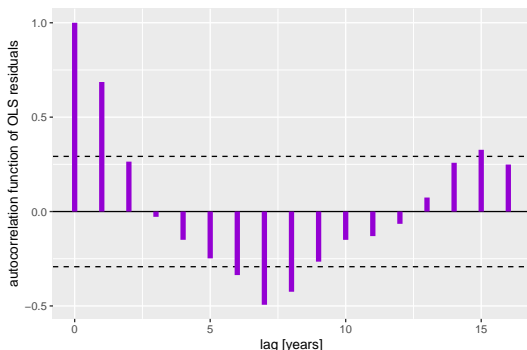
1. Consider linear regresion model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

and calculate residuals r_i using OLS method.

2. Perform Durbin-Watson test on the residuals r_i and calculate the estimation $\widehat{\theta}$ of the autoregressive parameter.

3. Introduce new variables by shifting,

$$Z_i = Y_i - \widehat{\theta} Y_{i-1}, \qquad v_i = x_i - \widehat{\theta} x_{i-1}, \qquad i = 2, 3, \ldots, n.$$

4. Find OLS-estimations $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ in linear regression model

$$Z_i = \alpha_0 + \alpha_1 v_i + w_i, \qquad i = 2, 3, \ldots, n.$$

5. Adjust the estimations,

$$\widehat{\beta}_0 = \frac{\widehat{\alpha}_0}{1 - \widehat{\theta}}, \qquad \widehat{\beta}_1 = \widehat{\alpha}_1.$$

6. Calculate residuals r_i in the model from **step 1** using adjusted coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

7. Repeat steps 2–7 until reaching convergence.

Population of certain bird's species in Hawaii during 1956–2003. Build a linear regression model for square root of the population in dependency on time.
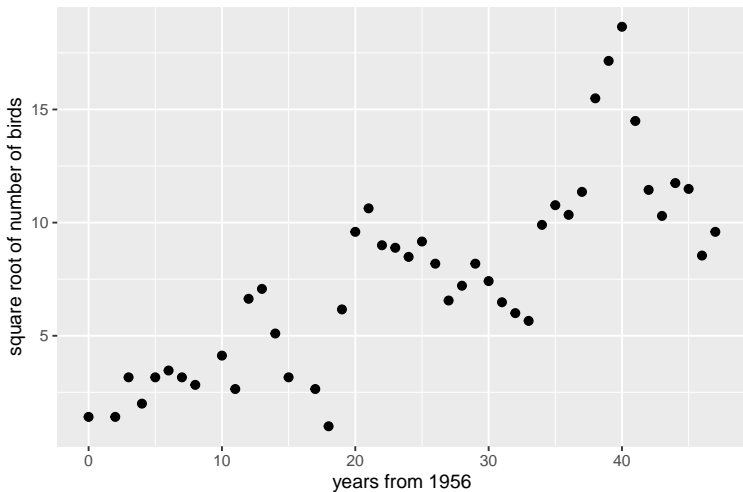
```
lm(formula = Y ~ t, data = dt)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.59739    0.78892   2.025   0.0491 *
t            0.24500    0.02814   8.708 4.83e-11 ***
---
Residual standard error: 2.578 on 43 degrees of freedom
Multiple R-squared:  0.6381,	Adjusted R-squared:  0.6297
F-statistic: 75.83 on 1 and 43 DF,  p-value: 4.825e-11

durbinWatsonTest(model.OLS, max.lag = 5)
 lag Autocorrelation D-W Statistic p-value
   1      0.68612915      0.5842435   0.000
   2      0.26434545      1.3608104   0.030
   3     -0.02797201      1.9385399   0.950
   4     -0.14967170      2.1793829   0.358
   5     -0.24803468      2.3638970   0.064
 Alternative hypothesis: rho[lag] != 0
# or
lmtest::dwtest(model.OLS)
  Durbin-Watson test
data:  model.OLS
DW = 0.58424, p-value = 2.932e-09
alternative hypothesis: true autocorrelation is greater than 0
# estimates of the autoregression coefficient
1 - unname(lmtest::dwtest(model.OLS)$statistic) / 2
0.7078783
cor(dt$r.OLS[-n], dt$r.OLS[-1])
0.7018654
as.numeric( (t(dt$r.OLS[-n]) %*% dt$r.OLS[-1]) / (t(dt$r.OLS[-n]) %*% dt$r.OLS[-n]))
0.7172438
```
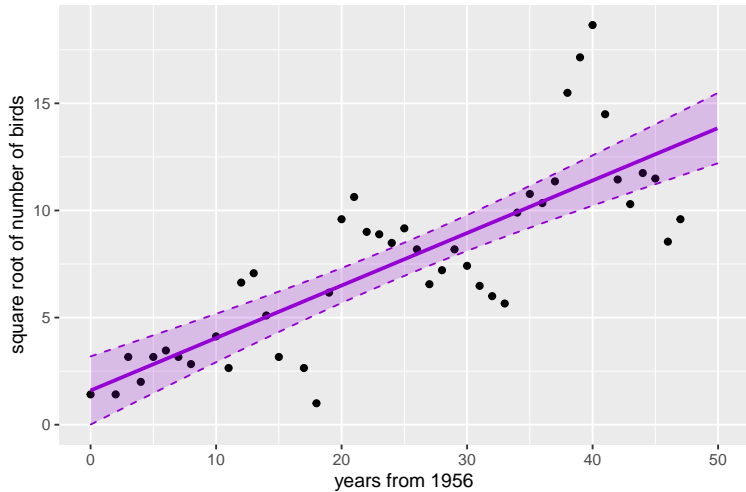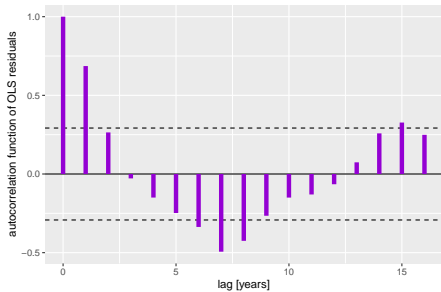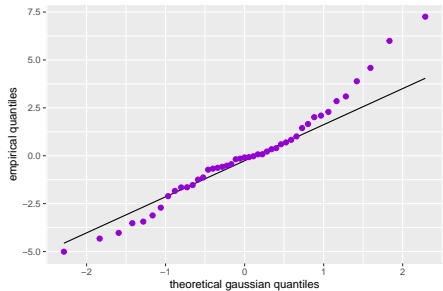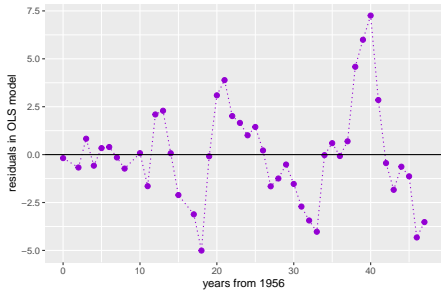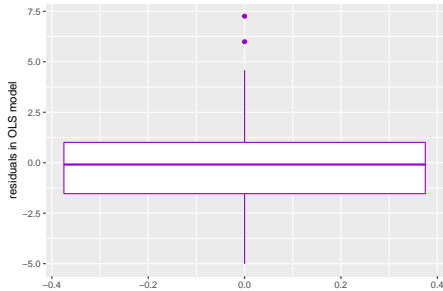
```
gls(Y ~ t, data = dt, correlation = corAR1()
Generalized least squares fit by REML
  Model: Y ~ t
  Data: dt
       AIC      BIC    logLik
  190.7945 197.8393 -91.39723
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.7898044
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 1.9725737 2.1993607 0.896885  0.3748
t           0.2187078 0.0754442 2.898934  0.0059
Residual standard error: 3.003914
Degrees of freedom: 45 total; 43 residual
```

```
gls(Y ~ t, data = dt, correlation = corARMA(p = 2))
Generalized least squares fit by REML
  Model: Y ~ t
  Data: dt
       AIC      BIC    logLik
  187.5679 196.3739 -88.78393
Correlation Structure: ARMA(2,0)
 Formula: ~1
 Parameter estimate(s):
      Phi1       Phi2
 0.9922084 -0.3519464
Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 1.6765439 1.3760182 1.218402  0.2297
t           0.2402667 0.0486775 4.935888  0.0000
Residual standard error: 2.672751
Degrees of freedom: 45 total; 43 residual
```

```
orcutt::cochrane.orcutt(model.OLS)
Cochrane-orcutt estimation for first order autocorrelation
 number of interaction: 6
 rho 0.720767
Durbin-Watson statistic
(original):     0.58424 , p-value: 2.932e-09
(transformed): 1.47511 , p-value: 2.567e-02
 coefficients:
(Intercept)           t
   2.251169    0.214595
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.251169   2.347424   0.959 0.343049
t           0.214595   0.076367   2.810 0.007491 **
---
Residual standard error: 1.8544 on 42 degrees of freedom
Multiple R-squared:  0.1583 , Adjusted R-squared:  0.1382
F-statistic: 7.9 on 1 and 42 DF,  p-value: < 7.491e-03
```
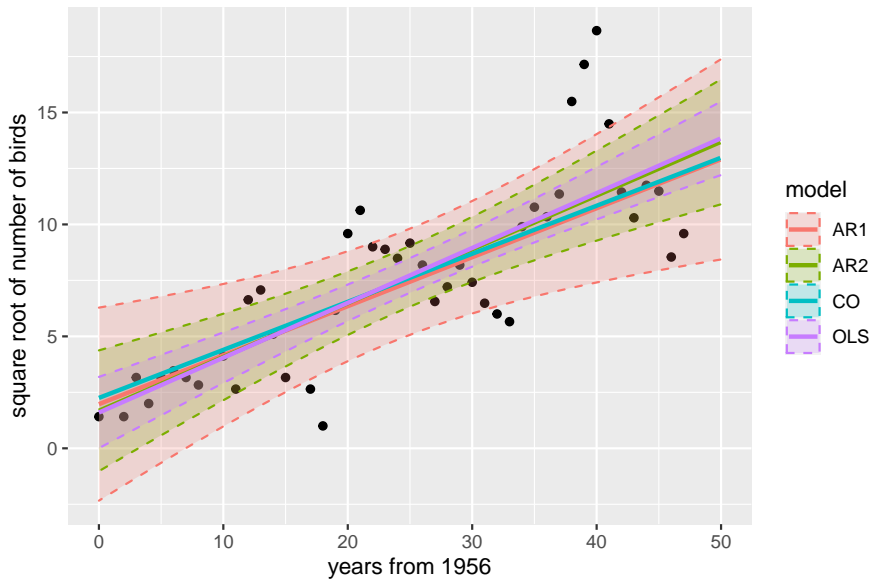
Compare the 95% confidence intervals for regression coefficients β_0 and β_1.

```
confint(model.OLS)
                  2.5 %      97.5 %
(Intercept) 0.006377231 3.1883945
t           0.188262984 0.3017425

confint(model.AR1)
                 2.5 %     97.5 %
(Intercept) -2.33809407 6.2832414
t            0.07083986 0.3665758

confint(model.AR2)
                2.5 %    97.5 %
(Intercept) -1.0204022 4.3734901
t            0.1448605 0.3356728

coefficients(model.CO) + t(c(-1, 1) %*% t(model.CO$std.error * qnorm(1 - alpha/2)))
                 [,1]       [,2]
(Intercept) -2.34969652 6.8520350
t            0.06491762 0.3642719
```

Notice the short confidence intervals for incorrectly used OLS estimates, inducing a **false impression** of their accuracy and significance of β_0.

Multicollinearity

Definition

Multicollinearity *(multikolinearita)* is a mutual linear dependence of predictors (explanatory variables) in a multiple regression model, i.e. when columns X_j, $j = 1, \ldots, l$, of the design matrix X without the first column of ones are linearly dependent, i.e. $c_1 X_1 + \cdots + c_l X_l = \mathbf{0}$ for at least one nonzero c_j.

In practice, multicollinearity denotes a case when the determinant of matrix $X'X$ is close to zero, the smallest eigenvalue is close to zero, and the matrix $X'X$ is almost singular *(skorosingulární)*.

▶ **Redundant explanatory variables** or incorrect selection of them can lead to multicollinearity.

▶ It is difficult to avoid the multicollinearity when the predictors are *hiddenly* linearly dependent, caused by unconsidered quantities or by the design of statistical experiment. E.g. in time series, similar *dependence* of observed variables on time is a usual reason for the presence of multicollinearity.

▶ A serious reason for multicollinearity is the **true (linear) relationship** between some predictors.

In the case of exact multicollinearity, matrix $X'X$ is singular, inverse $(X'X)^{-1}$ does not exist and OLS-estimates $\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$ can not be computed.

In the case of approximate multicollinearity, the inverse matrix $(X'X)^{-1}$ has large eigenvalues, which causes:

▶ OLS-estimates $\widehat{\beta}_{OLS}$ have large variance, i.e., there are some large numbers in the variance-covariance matrix of OLS-estimates,

$$\mathsf{Var}(\widehat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1},$$

▶ numerical calculations could be inaccurate or numerically unstable.

Consequences:

▶ The interpretation of the influence of individual predictors can be difficult. In t-tests of significance, the particular regression coefficients tend to be not significant. Large variances of the OLS-estimates cause wide confidence intervals.

▶ The estimates are computationally unstable and unreliable. Even a small change in data leads to a relatively big change in results (estimates).

Multicollinearity is related to the **correlation of predictors** X_j in the design matrix X (except for the first column of ones).

Definition (Variance inflation factors)

Variance Inflation Factors (VIF) are diagonal elements of matrix $(X'X)^{-1}$,
$$(a_1, \ldots, a_l) = \mathrm{diag}\big((X'X)^{-1}\big).$$
VIF is equal to $\quad a_j = \dfrac{1}{(1 - r_j^2)\, X_j' X_j},\quad$ where $\quad r_j = r_{X_j \cdot X_1 \ldots X_{j-1} X_{j+1} \ldots x_l}$
is coefficient of sample mutiple correlation between the jth predictor X_j and all other predictors $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_l$.

The multicollinearity is showed by high values of coefficients of mutiple correlations r_j and thus by high values of VIFs a_j.

Denote R_X the sample **correlation matrix of predictors**, and
$d_1, \ldots, d_l$ the diagonal elements in inverse matrix, $(d_1, \ldots, d_l) = \operatorname{diag}(R_X^{-1})$.

$H_0 : R_X = I_l$, i.e., the correlation matrix of predictors is identity matrix,
$H_1 : R_X \neq I_l$

Theorem (Test of multicollinearity)

Under H_0, test statistic
$$K = -\left[n - 1 - \frac{1}{6}(2\,l + 7)\right] \ln \det R_X \ \sim \ \chi^2\left(\frac{l(l-1)}{2}\right).$$
H_0 is rejected at the level of significance α, if $K > \chi^2_{1-\alpha}\left(\frac{p(p-1)}{2}\right)$.

Theorem (Identification of suspected predictors)

If the predictor X_j does not cause multicollinearity, then
$$F_j = \frac{n - l}{l - 1}\,(d_j - 1) \ \sim \ \mathsf{F}(l - 1, \ n - l).$$
If $F_j > F_{1-\alpha}(l - 1, \ n - l)$, predictor X_j contributes to multicollinearity.

- ▶ It can be difficult to find a suitable model when having many predictors, due to the correlations and relationships between them.
- ▶ There is no algorithm that finds the best model in general.
- ▶ **We need:**
    1. criterion for comparing a pair of models,
    2. strategy – algorithm for construction and traversal of models.
- ▶ There are many different criteria, some sensitive to a specific data type. Some criteria are relative, they do not directly quantify the strength (quality) of the prediction.
- ▶ A frequently used algorithm is stepwise regression method *(metoda postupné regrese)*.
- ▶ Model selection in case of very large number of predcitors is mommon task of *data mining* and *machine learning*, too.

- residual sum of squares, $S_e \to \min$, **inappropriate**
- coefficient of determination, $R^2 \to \max$, **inappropriate**
- adjusted coefficient of determination, $\bar{R}^2 = 1 - \dfrac{n-1}{n-l}(1 - R^2) \to \max$
- sample coefficients of partial correlations (selection of a predictor) and F_j statistics (test for stopping the algorithm)
- Akaike information criterion (AIC), $AIC \to \min$, **popular**,
$$AIC = -2\,\ell(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}, \widehat{\sigma}^2_{\mathsf{ML}}) + 2\,l$$
- Schwarz bayesian information criterion (BIC), $BIC \to \min$,
$$BIC = -2\,\ell(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}, \widehat{\sigma}^2_{\mathsf{ML}}) + l\ln n$$

AIC and *BIC* are relative measures of the model quality based on penalized log-likelihood. For the same data, they allow to compare also models that are not submodels.

▶ backward elimination:
  1. Start: full (maximal) model, all predictors included.
  2. Iteration: remove one non-significant (least significant) predictor; try all such submodels and select the one with the *best* criterion.
  3. End: all remaining predictors are significant, no submodel with *better* criterion can be build.

▶ forward selection:
  1. Start: null (minimal) model, no predictor included, only intercept.
  2. Iteration: add one significant predictor; try all such models and select the one with the *best* criterion.
  3. End: all included predictors are significant, no model with *better* criterion can be build.

▶ bidirectional stepwise:
  1. Start: full or null model.
  2. Iteration: remove one non-significant (least significant) or add one significant predictor; try all such (sub)models and select the one with the *best* criterion.
  3. End: all included predictors are significant; no model with *better* criterion can be build by removing or adding one predictor.

Consider **observed random sample** $(y_1, \ldots y_n)$ from probability distribution with density $f(y_i; \boldsymbol{\theta})$ depending on a vector of **parameters** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

▶ Likelihood (function) *(věrohodnostní funkce)* $L(\boldsymbol{\theta})$ is product of marginal probability densities $f(y_i; \boldsymbol{\theta})$,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}).$$

▶ Log-likelihood *(logaritmická věrohodnostní funkce)* $\ell(\boldsymbol{\theta})$ is natural logarithm of the likelihood,

$$\ell(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}).$$

▶ Maximum likelihood (ML) method *(metoda maximální věrohodnosti)*: maximize (log-)likelihood with respect to the vector of parameter $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) \longrightarrow \max, \quad \text{or} \quad \ell(\boldsymbol{\theta}) \longrightarrow \max.$$

▶ Maximum likelihood estimate (MLE) *(maximálně věrohodný odhad)* of $\boldsymbol{\theta}$ is a vector for which (log-)likelihood attains its maximum,

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \operatorname{argmax} L(\boldsymbol{\theta}) = \operatorname{argmax} \ell(\boldsymbol{\theta}).$$

- Assume $\quad Y_i \sim \mathsf{N}\Big(\beta_0 + \sum_{j=1}^{l} \beta_j x_{ij}, \ \sigma^2\Big)$

- $f(y_i; \, x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\dfrac{1}{2\sigma^2}\Big[y_i - \beta_0 - \sum_{j=1}^{l} \beta_j x_{ij}\Big]^2\Big)$

- $L(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \prod_{i=1}^{n} f(y_i; \, x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2)$

- $\ell(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \sum_{i=1}^{n} \ln f(y_i; \, x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2) =$

  $= -\dfrac{n}{2}\big[\ln(2\pi) + \ln\sigma^2\big] - \dfrac{1}{2\sigma^2} \sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{l} \beta_j x_{ij}\Big)^2$

- $\ell(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) \longrightarrow \max, \qquad \widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \operatorname{argmax} \ell(\boldsymbol{\theta})$

ML-estimates:

- $\widehat{\boldsymbol{\beta}}_{\mathsf{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}$

- $\widehat{\sigma^2}_{\mathsf{ML}} = \dfrac{S(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}})}{n} \ \neq \ \dfrac{S(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}})}{n - (l+1)} = \widehat{\sigma^2}_{\mathsf{OLS}}$

Statistics II │ 7

Principal component analysis (PCA)

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

24 October 2022

Principal component analysis (PCA) *(analýza hlavních komponent)* is a statistical method for reducing the dimension of mutivariate random sample. Recall, e.g., the multicollinearity issue in multiple linear regression model.

▶ PCA uses rotation of orthonormal basis of the vector space of the original random variables.

▶ The new basis random variables, called principal components (PC) *(hlavní komponenty)*, are uncorrelated.

▶ The choice of a new basis is optimal. The principal components are successively build in such a way to explain as much of the total variance of the data as possible using the lowest possible number of PCs. As a result, even the first few PCs explain most of the total variance of data and justify the reduction of the dimension.

▶ A typical disadvantage of PCA is the impossibility of interpreting the PCs (e.g., due to the incompatibility of the physical units used).

Consider n observations of l centered random variables, arranged in matrix

$$X = \begin{pmatrix} X_{11}, & \cdots, & X_{1l} \\ \vdots & & \vdots \\ X_{n1}, & \cdots, & X_{nl} \end{pmatrix}.$$

of size $n \times l$, $n > l$. The centering can be done by subtracting the column sample mean from all column values, $X_{ij} - \overline{X_j}$.

The ith observation $(X_{i1}, \ldots, X_{il})$ is geometrically represented by point x_i in k-dimensional vector space $\mathbb{R}^l$,

$$x_i = X_{i1}\, e_1 + \cdots + X_{il}\, e_l,$$

where $e_1, \ldots, e_l$ are unit vectors forming the standard orthonormal basis *(ortonormální báze)* of vector space $\mathbb{R}^l$, and the observations $(X_{i1}, \ldots, X_{il})$ are coordinates *(souřadnice)* of this point. Directions of the axes correspond to the random variables in columns of X.

Now, consider a general **orthonomal basis** $u_1, \ldots, u_l \in \mathbb{R}^l$ of vector space $\mathbb{R}^l$, i.e., with **scalar product** *(skalární součin)*

$$u_i' u_j = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

A matrix U is created by joining the basis vectors as columns,

$$U = (u_1, u_2, \ldots, u_l).$$

Then, matrix U is orthogonal *(ortogonální matice)*, i.e.,

$$U'U = I_k \quad \text{and} \quad U^{-1} = U'.$$

Every point x in vector space $\mathbb{R}^k$ is expressed as

$$x = a_1 u_1 + \cdots + a_l u_l = Ua$$

in terms of its coordinates $a = (a_1, \ldots, a_l)'$ in the basis $u_1, \ldots, u_l$ of columns of U. Vector of coordinates a of the point x is calculated as

$$a = U'x.$$
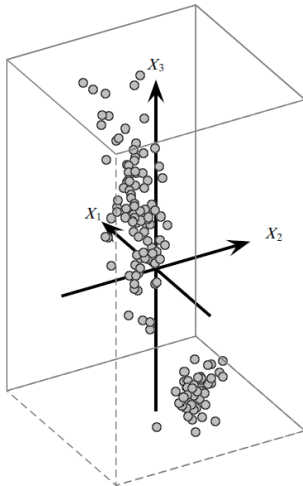
(a) Original Basis

(b) Optimal Basis

Zaki M. J., Meira W.: *Data Mining and Analysis: Fundamental Concepts and Algorithms*, 2014.
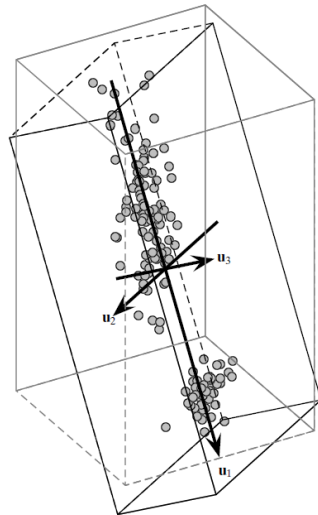
The orthonormal basis $u_1, \ldots, u_k$ should be chosen in order to have
$$x \approx x^* \qquad \text{for each point } x \in \mathbb{R}^k,$$
$$x = a_1 u_1 + \cdots + a_l u_l = Ua, \qquad x^* = a_1 u_1 + \cdots + a_r u_r = U^* a^*,$$
where x^* is an **approximation** formed by a linear combination of only the first r basis vectors, $r < l$.

The corresponding reduced *(redukovaná)* matrix U^* is formed by the first r columns of U, $U^* = (u_1, u_2, \ldots, u_r)$. The reduced vector of coordinates is
$$a^* = (a_1, \ldots, a_r)' = (U^*)' x.$$

Hence, we have
$$x^* = U^* a^* = U^* (U^*)' x,$$
that means the approximation x^* is a projection *(projekce)* of point x into the subspace generated *(podprostor generovaný)* by columns of U^*. Matrix $U^* (U^*)'$ is called projection matrix *(projekční matice)*.

Optimality of the basis

The basis vectors $u_1, \ldots, u_l$ of the optimal orthonormal basis are found in order to approximate x using x^* by **explaining as much variance of the data as possible**.

The first basis vector u_1 of the optimal orthonormal basis is found in order to approximate x using x^* by **explaining as much variance of the data as possible**.

Projection of the ith observation x_i into the direction of the first basis vector u_1 is $x_i^* = a_{i\,1}\,u_1$, and $a_{i\,1} = u_1' x_i^* = u_1' x_i$.

Population variance of the projected observation $x_1^*, \ldots, x_n^*$ is equal to

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^n a_{i\,1}^2 = \frac{1}{n}\sum_{i=1}^n (u_1' x_i)^2 = \frac{1}{n}\sum_{i=1}^n u_1' x_i x_i' u_1 = u_1' \underbrace{\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)}_{S} u_1 = u_1' S\, u_1,$$

where $S = \dfrac{1}{n}\sum_{i=1}^n x_i x_i'$ is **sample variance-covariance matrix** ($l \times l$) of centered random variables $X_1, \ldots, X_j$, i.e., of columns of matrix X.
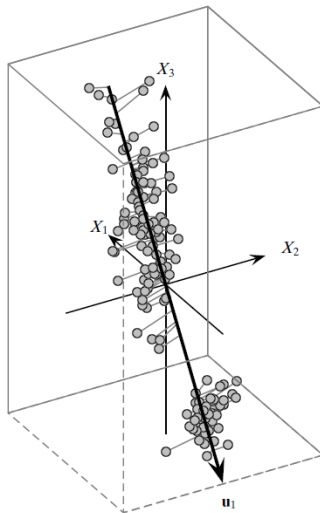
1. Find vector u_1 of unit length, in order to **maximize the variance** σ_1^2 of observations projected into the direction of u_1,

$$\sigma_1^2 = u_1' \, S \, u_1 \to \max, \quad \text{subject to a constraint} \quad u_1' u_1 = 1.$$

2. Find vector u_2 of unit length, in order to **maximize the variance** σ_2^2 of observations projected into the direction of u_2, perpendicular to u_1,

$$\sigma_2^2 = u_2' \, S \, u_2 \to \max, \quad \text{subject to constraints} \quad u_2' u_2 = 1, \quad u_2' u_1 = 0.$$

3. Find vector u_3 of unit length, in order to **maximize the variance** σ_3^2 of observations projected into the direction of u_3, perpendicular to u_1, u_2,

$$\sigma_3^2 = u_3' \, S \, u_3 \to \max, \quad \text{subject to} \quad u_3' u_3 = 1, \quad u_3' u_1 = 0, \; u_3' u_2 = 0.$$

...
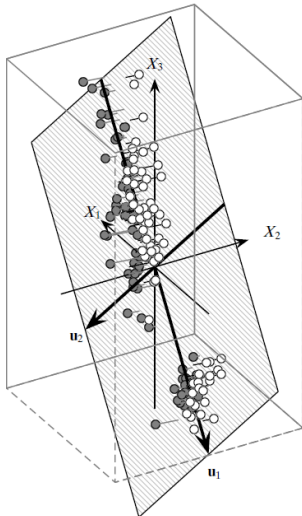
(a) Optimal basis
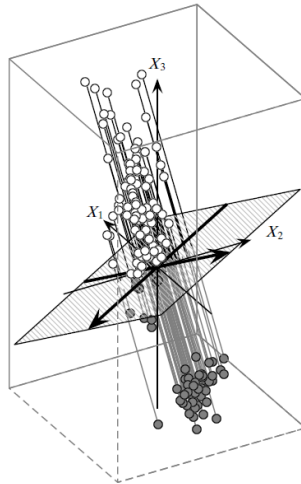
(b) Nonoptimal basis

Constrained extrema task *(vázaný extrém)*:
$$\sigma_1^2 = u_1' \, S \, u_1 \to \max, \quad \text{subject to a constraint} \quad u_1' u_1 = 1$$

▶ Build Lagrange function $L(u_1)$ with Lagrange multiplier λ_1,
$$L(u_1) = u_1' \, S \, u_1 - \lambda_1 (u_1' u_1 - 1),$$

▶ differentiate with respect to u_1 and set the derivative equal to zero,
$$\mathbf{0} = \frac{\partial L(u_1)}{\partial u_1} = 2 \, S \, u_1 - 2\lambda_1 u_1,$$

▶ solution: $\quad S \, u_1 = \lambda_1 u_1$.

▶ The variance explained by the first PC: $\sigma_1^2 = u_1' \, S \, u_1 = u_1' \lambda_1 u_1 = \lambda_1$.

▶ λ_1 is the **largest** eigenvalue *(vlastní číslo)* of sample variance-covariance matrix S and the first principal component (PC) *(hlavní komponenta)* u_1 is corresponding eigenvector *(vlastní vektor)*.

$$\sigma_j^2 = u_j' S u_j \to \max, \text{ subject to } u_j' u_j = 1, \ u_j' u_1 = \cdots = u_j' u_{i-1} = 0$$

▶ Lagrange function $L(u_j)$ with Lagrange multipliers $\lambda_j, \phi_1, \ldots, \phi_{i-1}$,

$$L(u_1) = u_j' S u_j - \lambda_j(u_j' u_1 - 1) - \sum_{i=1}^{j-1} \phi_i(u_j' u_i - 0),$$

▶ differentiate with respect to u_j and set the derivative equal to zero,

$$\mathbf{0} = \frac{\partial L(u_j)}{\partial u_j} = 2 S u_j - 2\lambda_j u_j - \sum_{i=1}^{j-1} \phi_i u_i,$$

▶ left multiply by eigenvector u_k, $\quad \mathbf{0} = 2u_k' S u_j - 2\lambda_j u_k' u_j - \sum_{i=1}^{j-1} \phi_i u_k' u_i,$

▶ solution: $\quad S u_j = \lambda_j u_j.$

▶ The variance explained by this PC: $\quad \sigma_j^2 = u_j' S u_j = u_j' \lambda_j u_j = \lambda_j.$

▶ λ_j is the *j*th **largest eigenvalue** of S and the *j*th PC u_j is corresponding **eigenvector**.

1. Center the particular columns of data matrix X of size $n \times l$, $n > l$, i.e., subtract the column sample mean from all column values, $X_{ij} - \overline{X_j}$.

$$X = \begin{pmatrix} X_{11}, & \cdots, & X_{1l} \\ \vdots & & \vdots \\ X_{n1}, & \cdots, & X_{nl} \end{pmatrix}.$$

2. Denote S the sample variance-covariance matrix of centered random variables in matrix X,

$$S = \left\{ S_{X_i X_j} \right\}_{i,j=1}^{l} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_{i1} \\ \vdots \\ X_{il} \end{pmatrix} \left(X_{i1}, \ldots, X_{il} \right).$$

3. Find eigenvalaues λ_j of the variance-covariance matrix S and order them in nondecreasing sequence. In the same order, place the corresponding eigenvectors u_j into matrix U of the basis of principal components,

$$\lambda_1 \geq \cdots \geq \lambda_l, \qquad U = \left(u_1 \mid \cdots \mid u_l \right).$$

4. Choose suitable r, $r < l$, and build the matrix U^* of reduced basis of principal components,

$$U^* = (u_1 \mid \cdots \mid u_r).$$

5. Coordinates of n observations in the reduced basis of principal components are rows of matrix

$$A^* = \begin{pmatrix} a_{11}, & \cdots, & a_{1r} \\ \vdots & & \vdots \\ a_{n1}, & \cdots, & a_{nr} \end{pmatrix} = X\,U^*.$$

Theorem

Let S be the variance-covariance matrix of l centered random variables, and let $u_1, \ldots, u_l$ be the principal components (PCs). Then,

- PCs are linear combinations of original random variables,
- PCs are eigenvectors of S,
- PCs have zero means,
- PCs are mutually uncorrelated and form an orthonormal basis of l-dimensional vector space,
- variance of data in the direction of jth PC u_j is equal to the corresponding eigenvalue λ_j of S, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l \geq 0$,
- variance of data is equal to $\displaystyle \sum_{j=1}^{l} \lambda_j = \operatorname{Tr} S$,
- variance explained by first r PCs is $\displaystyle \sum_{j=1}^{r} \lambda_j = \operatorname{Tr} S$,
- $\prod_{j=1}^{l} \lambda_j = \det S$.

▶ Smallest proportion $p \in (0; 1)$ of explained variance,
$$\frac{\sum_{i=1}^{r} \lambda_j}{\sum_{i=1}^{l} \lambda_j} = \frac{\sum_{i=1}^{r} \lambda_j}{\text{Tr } S} \geq p;$$
often, e.g., $p = 0.8$.

▶ Kaiser's rule,
$$\lambda_1 \geq \cdots \geq \lambda_r \geq \frac{1}{l} \text{Tr } S,$$
when the principal components eigenvalues greater than or equal to the averaged total variability of data are included.

Variance-covariance matrix S is symmetric and positive definite of size $l \times l$, hence its eigenvalues are nonnegative, $\lambda_j \geq 0$, and the number of positive eigenvalues is equal to the matrix rank $r(X)$.

Eigendecomposition of variance-covariance matrix

$$S = U \Lambda U', \qquad \text{where} \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_l)$$

Singular value decomposition (SVD) *(singulární rozklad)*

$$X = VDW',$$

where columns of V ($n \times n$) are left-singular vectors of X,
columns of W ($l \times l$) are right-singular vectors of X,
and $D = \text{diag}(d_1, \ldots, d_k)$ is rectangular ($n \times l$) diagonal matrix of singular values of matrix X.

$$\text{Then,} \qquad S = WD'DW'$$

▶ Data are often not only centered, but also scaled, i.e., standardized, to have zero mean and unit variance. Then, the sample variance-covariance matrix S is equal to the sample correlation matrix R.

▶ Loadings are coefficients of PCs in the basis of original variables.

▶ Scores are coordinates of observations in basis of (unscaled) PCs.

▶ Biplot is a plot of observations and original variables in the vector space of two selected PCs.

▶ Scree plot is a plot of the variance explained by particular PCs, often supplemented by the plot of cumulative explained variance.

PCA
```
pca <- prcomp(M, center = TRUE, scale. = TRUE)
```

Principal components = coefficients of rotation = loadings matrix
```
pca$rotation
                      PC1          PC2           PC3           PC4          PC5
GDP          -0.33763464  -0.18638818  -0.0846640804  -0.064437498  -0.20949401
GDPgrowth     0.29167110  -0.24879686   0.0057896059  -0.171364553  -0.21081885
Agriculture   0.29444454   0.04364901  -0.2047992461   0.039198396   0.21506457
```

Coordinates of observations in basis of PCs = scores
```
pca$x
                   PC1          PC2          PC3          PC4          PC5          PC6
Belgium    -3.00823329   1.079310521   0.94632553   0.28962579  -0.7123690  -0.20400896
Bulgaria    4.42477516   0.695687129  -1.70449250   1.80497077   0.7762390   0.66273074
Czechia     1.82168349  -1.896305208   1.48643413   0.53314200   0.4528168   0.13707352
```

Standard deviations in directions of PCs
```
pca$sdev
 2.630563815 1.604543390 1.231959330 1.159124780 1.045207719 0.858628136 0.746403144
sum(pca$sdev^2)
 16
```

Biplot
```
ggbiplot::ggbiplot(pca, choices = c(1, 2), labels = rownames(M))
```

Variances
```
dt.var <- data.frame(r = seq_along(pca$sdev),
  var = (pca$sdev**2) / sum(pca$sdev**2),
  cumvar = cumsum(pca$sdev**2) / sum(pca$sdev**2))
```

```
dt.var |> filter(cumvar >= 0.8)              dt.var |> filter(var >= mean(var))
   r          var    cumvar                      r        var    cumvar
1  5 6.827870e-02 0.8405112                   1  1 0.43249162 0.4324916
2  6 4.607764e-02 0.8865888                   2  2 0.16090997 0.5934016
3  7 3.481985e-02 0.9214087                   3  3 0.09485774 0.6882593
4  8 2.764757e-02 0.9490562                   4  4 0.08397314 0.7722325
5  9 1.438905e-02 0.9634453                   5  5 0.06827870 0.8405112
6  10 1.282605e-02 0.9762713
7  11 1.140343e-02 0.9876748
8  12 5.994460e-03 0.9936692
9  13 3.032757e-03 0.9967020
10 14 2.208960e-03 0.9989109
11 15 1.088419e-03 0.9999994
12 16 6.382461e-07 1.0000000
```
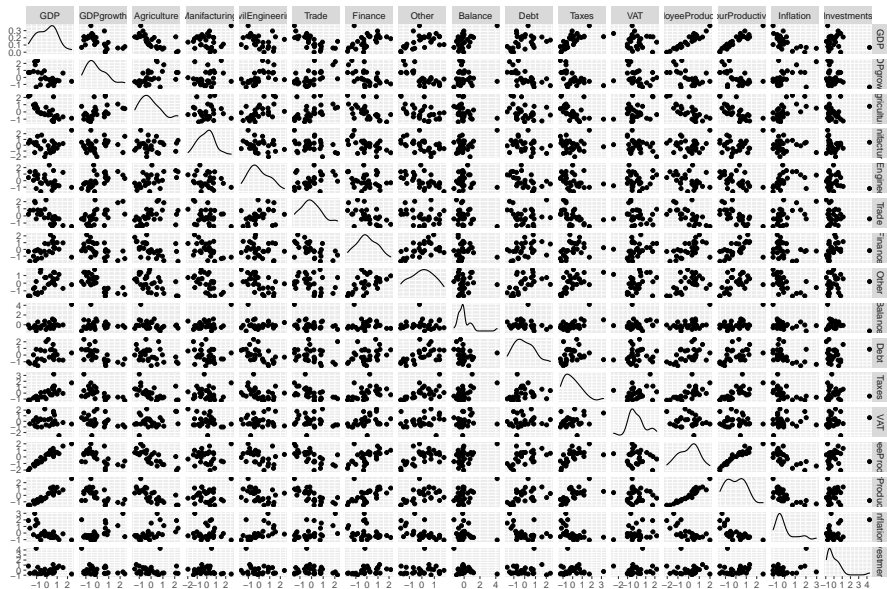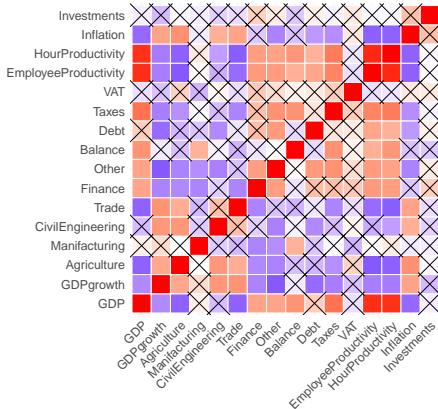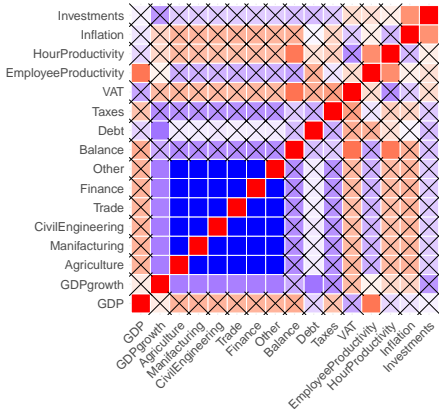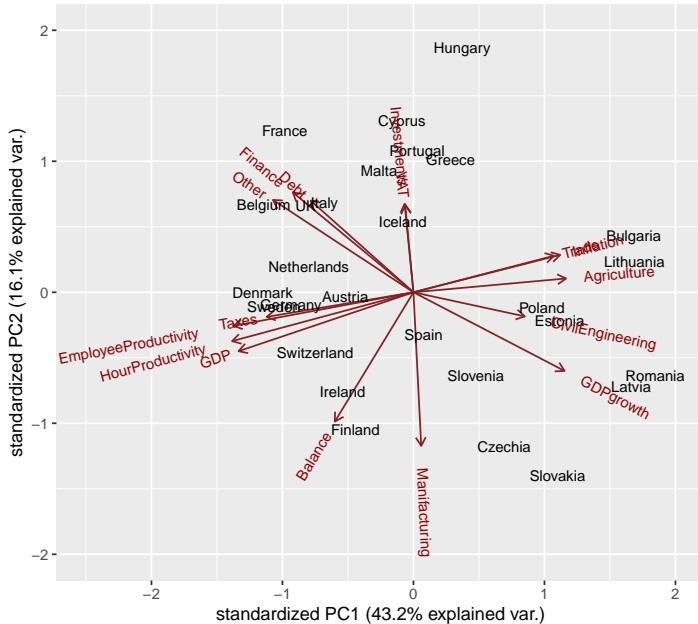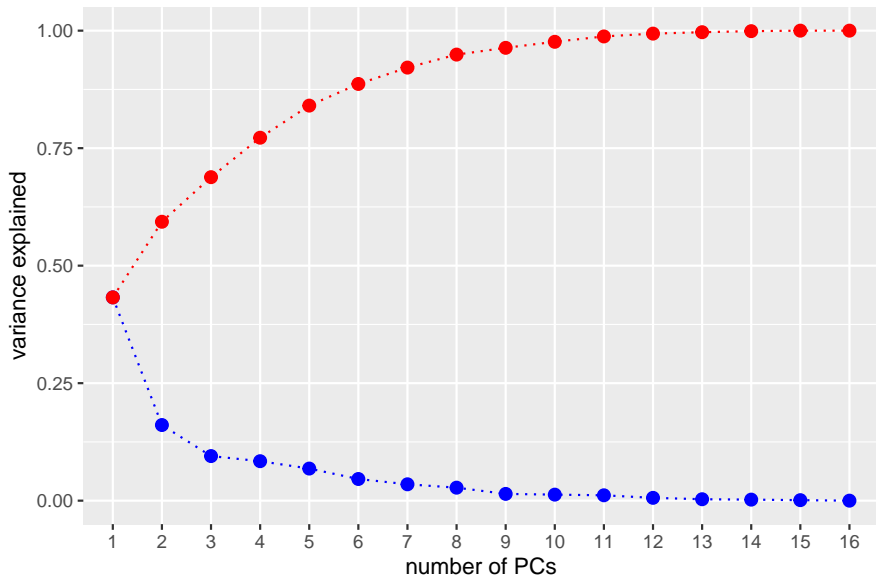
- Geometric principal of PCA, algorithm for finding the principal components
- Reasons for PCA in mathematical statistics, choice of the nummber of PCs, biplot and scree plot
- Properties of PCs and relationship to original random variables
- PCs and explained variance, eigenvaues and eigenvectors of varice-covariance matrix

Statistics II │ 8

Logistic regression
and other generalized linear models (GLM)

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

31 October 2022

Logistic regression

Consider random sample $Y_1, \ldots, Y_n$ of size n of dichotomous (binary) random variables $Y_i \in \{0, 1\}$ in dependency on predictors (numerical or dummy variables) $(X_1, \ldots, X_l)$,

observation i	predictors	observation
1	$(x_{1\,1}, \ldots, x_{1\,l})$	Y_1
2	$(x_{2\,1}, \ldots, x_{2\,l})$	Y_2
$\vdots$	$\vdots$	$\vdots$
n	$(x_{n\,1}, \ldots, x_{n\,l})$	Y_n

Random variables Y_i have alternative (Bernoulli) probability distribution with probability of success $p_i \in [0, 1]$, depends on the predictors $(x_{1\,1}, \ldots, x_{1\,p})$ and are mutually independent,

$$Y_i \sim \mathsf{A}(p_i), \quad i = 1, \ldots, n,$$

Probability mass functions *(pravděpodobnostní funkce)* are

$$f(y_i) = \mathsf{P}(Y_i = y_i) = \begin{cases} p_i, & y_i = 1, \\ 1 - p_i, & y_i = 0, \\ 0, & \text{otherwise,} \end{cases} = \begin{cases} p_i^{y_i}(1 - p_i)^{1-y_i}, & y_i = 0, 1, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, consider random sample $Y_1, \ldots, Y_n$ of number of successes $Y_i \in \{0, 1, \ldots, n_i\}$, or random sample $Z_1, \ldots, Z_n$ of relative number of successes, in N disjoint groups of sizes n_i,

group i	predictors	size	number of succeses	fraction of sucesses
1	$(x_{11}, \ldots, x_{1l})$	n_1	Y_1	$Z_1 = Y_1/n_1$
2	$(x_{21}, \ldots, x_{2l})$	n_2	Y_2	$Z_2 = Y_2/n_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$(x_{N1}, \ldots, x_{Nl})$	n_N	Y_N	$Z_N = Y_N/n_N$

Number of sucesses have independent binomial distributions,

$$Y_i \sim \mathsf{Bi}(n_i, p_i), \quad i = 1, \ldots, N,$$

Probability mass functions are

$$f(y_i) = \mathsf{P}(Y_i = y_i) = \mathsf{P}\left(Z_i = \frac{y_i}{n_i}\right) = \begin{cases} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i-y_i}, & y_i = 0, 1, 2, \ldots, n_i, \\ 0, & \text{otherwise.} \end{cases}$$

By substition, we have

$$y_i = z_i n_i \quad \text{for} \quad z_i = \frac{0}{n_i}, \frac{1}{n_i}, \frac{2}{n_i}, \ldots, \frac{n_i}{n_i}.$$

- ▶ **Task:** model the **probability of success** p_i in dependency on predictors $x_{i1}, \ldots, x_{il}, i = 1, \ldots, n$.
- ▶ Linear regression model $p_i = \beta_0 + x'_i \beta + \varepsilon_i$ **cannot be used**. (Why?)

Consider model

$$g(p_i) = \eta_i = \beta_0 + x'_i \beta = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij}, \qquad i = 1, \ldots, n,$$

- ▶ η_i is linear predictor *(lineární prediktor)*, the linear part of the model,
- ▶ $g(\cdot)$ is link function *(linkovací/spojovací funkce)*, suitable strictly monotonic (i.e., increasing or decreasing) and differentiable function, which models the non-linear relationship between the probability of success p_i and the linear predictor η_i.

Typical link functions for dichotomous or binomial data:

- ▶ logit,
- ▶ probit,
- ▶ CLogLog = complementary LogLog,
- ▶ LogLog.

Logit model *(logitový model)*, so-called logistic regression *(logistická regrese)* uses logit link function g,

---
Logit link function

$$g(p_i) = \ln \frac{p_i}{1 - p_i} = \eta_i = \beta_0 + x_i' \boldsymbol{\beta}$$

Then, probabilities of success and failure are

$$p_i = g^{-1}(\eta_i) = \frac{1}{1 + \exp(-\beta_0 - x_i' \boldsymbol{\beta})} = \frac{\exp(\beta_0 + x_i' \boldsymbol{\beta})}{1 + \exp(\beta_0 + x_i' \boldsymbol{\beta})},$$

$$1 - p_i = \frac{1}{1 + \exp(\beta_0 + x_i' \boldsymbol{\beta})}.$$

Probit model *(probitový model)* uses probit link function g, which is the quantile (inverse cumulative distribution) function $g = \Phi^{-1}$ of standard normal distribution N(0, 1),

---
Probit link function

$$g(p_i) = \Phi^{-1}(p_i) = \eta_i = \beta_0 + x_i' \boldsymbol{\beta}$$

Then, probability of success is

$$p_i = g^{-1}(\eta_i) = \Phi(\eta_i) = \Phi\left(\beta_0 + x_i' \boldsymbol{\beta}\right),$$

where Φ is cumulative distribution function of standard normal distribution N(0, 1).

In particular, for linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, i.e., one-predictor regression line, we have

$$p_i = \Phi\left(\beta_0 + \beta_1 x_i\right) = F(x_i),$$

where F is cumulative distribution function of $N\left(-\dfrac{\beta_0}{\beta_1}, \dfrac{1}{\beta_1^2}\right)$ distribution.

CLogLog model (complementary LogLog) uses link function

CLogLog link function

$$g(p_i) = \ln\left[-\ln(1 - p_i)\right] = \eta_i = \beta_0 + x_i'\,\boldsymbol{\beta}$$

Probability of success is

$$p_i = g^{-1}(\eta_i) = 1 - \exp\left[-\exp(\beta_0 + x_i'\,\boldsymbol{\beta})\right].$$

LogLog model uses link function

LogLog link function

$$g(p_i) = -\ln\left(-\ln p_i\right) = \eta_i = \beta_0 + x_i'\,\boldsymbol{\beta}$$

Probability of success is

$$p_i = g^{-1}(\eta_i) = \exp\left[-\exp(-\beta_0 - x_i'\,\boldsymbol{\beta})\right].$$
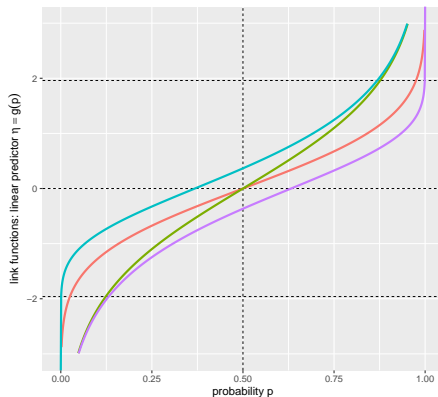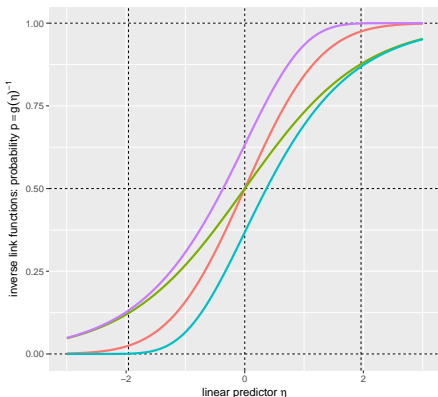
Notice the different shapes of the link functions.
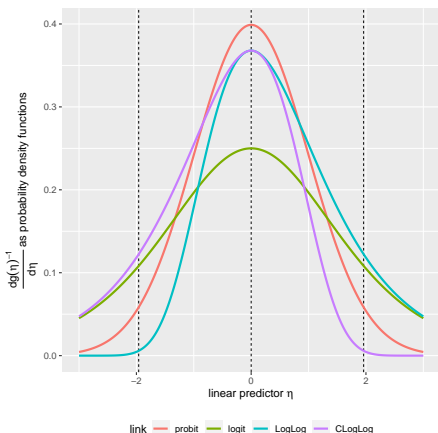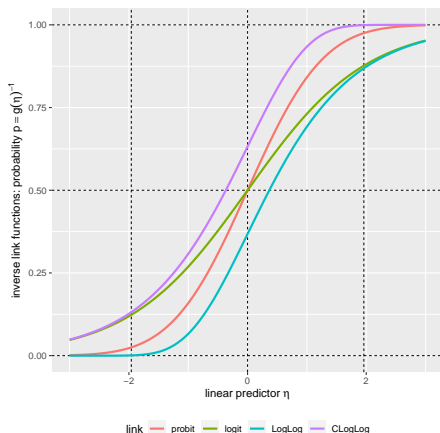
- ▶ logit: logistic probability distribution
- ▶ probit: standard normal distribution
- ▶ CLogLog: logarithmic Weibull = extreme-minimal-value distribution
- ▶ LogLog: Gumbel = extreme-maximal-value distribution

Consider **observed random sample** $(y_1, \ldots y_n)$ from probability distribution with density $f(y; \boldsymbol{\theta})$ depending on a vector of **parameters** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

▶ Likelihood (function) *(věrohodnostní funkce)* $L(\boldsymbol{\theta})$ is product of marginal probability densities $f(y_i; \boldsymbol{\theta})$,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}).$$

▶ Log-likelihood *(logaritmická věrohodnostní funkce)* $\ell(\boldsymbol{\theta})$ is natural logarithm of the likelihood,

$$\ell(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}).$$

▶ Maximum likelihood (ML) method *(metoda maximální věrohodnosti)*: maximize (log-)likelihood with respect to the vector of parameter $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) \longrightarrow \max, \quad \text{or} \quad \ell(\boldsymbol{\theta}) \longrightarrow \max.$$

▶ Maximum likelihood estimate (MLE) *(maximálně věrohodný odhad)* of $\boldsymbol{\theta}$ is a vector for which (log-)likelihood attains its maximum,

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \operatorname{argmax} L(\boldsymbol{\theta}) = \operatorname{argmax} \ell(\boldsymbol{\theta}).$$

The task

$$L(\boldsymbol{\theta}) \longrightarrow \max, \quad \text{or} \quad \ell(\boldsymbol{\theta}) \longrightarrow \max.$$

is typically expressed by a system of k nonlinear equations

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i; \boldsymbol{\theta})}{\partial \theta_j} = 0, \qquad j = 1, \ldots, k.$$

This system is solved numerically:

▶ by linearization using Taylor expansion and subsequently by Newton-Raphson method;

▶ by the so-called scoring method, where the matrix of second partial derivatives is approximated by the Fisher information matrix $J(\boldsymbol{\beta})$.

Theorem

Consider random sample $(Y_1, \ldots, Y_n)$ of random variable Y with probability mass function or probability density function $f(y; \boldsymbol{\theta})$ depending on a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and ML-estimate $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ of $\boldsymbol{\theta}$.

Under so-called regularity conditions *(podmínky regularity)*, we have:

▶ $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} \overset{as.}{\sim} \mathsf{N}_k \left(\boldsymbol{\theta}, \ \frac{1}{n} \, \boldsymbol{J}(\boldsymbol{\theta})^{-1} \right)$,

▶ $W = n(\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} - \boldsymbol{\theta})' \, \boldsymbol{J}(\boldsymbol{\theta}) \, (\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} - \boldsymbol{\theta}) \overset{as.}{\sim} \chi^2(k)$,

where $\boldsymbol{J}(\boldsymbol{\theta})$ is Fisher information matrix *(Fisherova informační matice)*

$$\boldsymbol{J}_{ab}(\boldsymbol{\theta}) = \int_{\mathbb{R}^n} \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_a} \cdot \frac{\partial \ln f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \theta_b} f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{y}, \qquad a, b = 1, \ldots, k.$$

▶ ML-estimates $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$ are asymptotically normal, asymptotically unbiased, and consistent.

▶ The variance-covariance matrix of ML-estimators is equal to the Fisher information matrix, $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}) = \boldsymbol{J}(\boldsymbol{\theta})$.

▶ ML-estimators may not be *optimal* for finite sample size.

▶ Assume $\quad Y_i \sim N\left(\beta_0 + \sum\limits_{j=1}^{l} \beta_j x_{ij}, \ \sigma^2\right)$

▶ $f(y_i; \ x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{1}{2\sigma^2}\left[y_i - \beta_0 - \sum\limits_{j=1}^{l} \beta_j x_{ij}\right]^2\right)$

▶ $L(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \prod\limits_{i=1}^{n} f(y_i; \ x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2)$

▶ $\ell(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) = \sum\limits_{i=1}^{n} \ln f(y_i; \ x_i, \beta_0, \beta_1, \ldots, \beta_l, \sigma^2) =$

$\quad = -\dfrac{n}{2}\left[\ln(2\pi) + \ln\sigma^2\right] - \dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n}\left(y_i - \beta_0 - \sum\limits_{j=1}^{l} \beta_j x_{ij}\right)^2$

▶ $\ell(\beta_0, \beta_1, \ldots, \beta_l, \sigma^2) \longrightarrow \max$

ML-estimates:

▶ $\widehat{\boldsymbol{\beta}}_{\mathsf{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}$

▶ $\widehat{\sigma^2}_{\mathsf{ML}} = \dfrac{S(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}})}{n} \ \neq \ \dfrac{S(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}})}{n - (l+1)} = \widehat{\sigma^2}_{\mathsf{OLS}}$

▶ Assume $\quad Y_i \sim \text{Bi}(n_i, p_i), \qquad p_i = \dfrac{1}{1 + \exp\left(-\beta_0 - \sum\limits_{j=1}^{l} \beta_j x_{ij}\right)}$

▶ $f(y_i; x_i, \beta_0, \beta_1, \ldots, \beta_l) = \dbinom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$

▶ $L(\beta_0, \beta_1, \ldots, \beta_l) = \prod\limits_{i=1}^{N} f(y_i; x_i, \beta_0, \beta_1, \ldots, \beta_l)$

▶ $\ell(\beta_0, \beta_1, \ldots, \beta_l) = \sum\limits_{i=1}^{N} \ln f(y_i; x_i, \beta_0, \beta_1, \ldots, \beta_l) =$

$= \sum\limits_{i=1}^{N} \left(\ln \dbinom{n_i}{y_i} + y_i \left(\beta_0 + \sum\limits_{j=1}^{l} \beta_j x_{ij}\right) - n_i \ln\left[1 + \exp\left(\beta_0 + \sum\limits_{j=1}^{l} \beta_j x_{ij}\right)\right] \right)$

▶ $\ell(\beta_0, \beta_1, \ldots, \beta_l) \longrightarrow \max$

ML-estimates $\widehat{\beta}_{\text{ML}}$, then estimates $\widehat{p}_i$ are calculated.

For individual β_j, $j = 0, 1, \ldots, l$:

$H_0 : \beta_j = 0$,
$H_1 : \beta_j \neq 0$

▶ Using the Wald statistic $W = \dfrac{\widehat{\beta_{\text{ML}j}}^2}{s_{jj}}$ having asymptotically $\chi^2(1)$ distribution under H_0.

H_0 is rejected at the level of significance α, if $W > \chi^2_{1-\alpha}(1)$.

▶ Using asymptotic normality of ML-estimate under H_0, $\widehat{\beta_{\text{ML}j}} \overset{as.}{\sim} \text{N}\left(\beta_j, s_{jj}\right)$.

H_0 is rejected at the level of significance α, if $\dfrac{\left|\widehat{\beta_{\text{ML}j}}\right|}{\sqrt{s_{jj}}} > u_{1-\frac{\alpha}{2}}$.

▶ The variance of the ML-estimate $\widehat{\beta_{\text{ML}j}}$ is approximated by value
$\text{Var}(\widehat{\beta_{\text{ML}j}}) \approx s_{jj} = \frac{1}{n}\left(J(\widehat{\boldsymbol{\beta}}_{\text{ML}})^{-1}\right)_{jj}$.

For testing the whole model or comparing with submodels, scale deviance or *AIC* criterion is used.

Definition (Scaled deviance)

Let random sample Y follows *maximal* model M^+. Consider its submodel M with k_1 parameters and further its submodel M^- with $k_2 < k_1$ parameters, with the same distribution type and same link function.

Scaled deviance *(škálová deviance)* of models M and M^-, respectively, is

$$D = \ln \left[\frac{L(\widehat{\beta^+_{\mathsf{ML}}}; Y)}{L(\widehat{\beta_{\mathsf{ML}}}; Y)} \right]^2, \qquad D^- = \ln \left[\frac{L(\widehat{\beta^+_{\mathsf{ML}}}; Y)}{L(\widehat{\beta^-_{\mathsf{ML}}}; Y)} \right]^2.$$

where $\widehat{\beta^+_{\mathsf{ML}}}, \widehat{\beta_{\mathsf{ML}}}, \widehat{\beta^-_{\mathsf{ML}}}$ are ML-estimated in corresponding models.

Hypothesis

H_0 : random sample Y follows model M^-.

Theorem

Let random sample Y follows model M. Under the regularity conditions and under H_0, the difference of scaled deviances $\Delta D = D^- - D$ has asymptotically $\chi^2(k_1 - k_2)$ distribution.

H_0 is rejected at the level of significance α, if $\Delta D = D^- - D > \chi^2_{1-\alpha}(p - q)$.

AIC criterion is used as relative measure of performance of model.

1. calculate estimate $\widehat{\eta}(x)$ of linear predictor as a function of predictor(s) x
2. calculate standard error $\text{SE}(\eta)(x)$, i.e., standard deviation of linear predictor, as a function of predictor(s) x
3. assume normal distribution of random errors, i.e., normal distribution of values of linear predictor,

$$\eta(x) \sim \text{N}\left(\widehat{\eta}(x),\ \text{SE}(\eta)^2(x)\right)$$

4. calculate lower and upper bound of the confidence intervals for linear predictor using quantiles of normal distribution,

$$\underbrace{\widehat{\eta}(x) - u_{1-\alpha/2}\,\text{SE}(\eta)(x)}_{\eta_{\text{L}}(x)} \quad \leq \quad \eta(x) \quad \leq \quad \underbrace{\widehat{\eta}(x) + u_{1-\alpha/2}\,\text{SE}(\eta)(x)}_{\eta_{\text{U}}(x)}$$

5. use the inverse link function, $p = g^{-1}(\eta)$, to calculate the estimate $\widehat{p}(x)$ and lower and upper bound for the probability of success,

$$\widehat{p}(x) = g^{-1}\left(\widehat{\eta}(x)\right),$$

$$g^{-1}\left(\eta_{\text{L}}(x)\right) \quad \leq \quad p(x) \quad \leq \quad g^{-1}\left(\eta_{\text{U}}(x)\right)$$
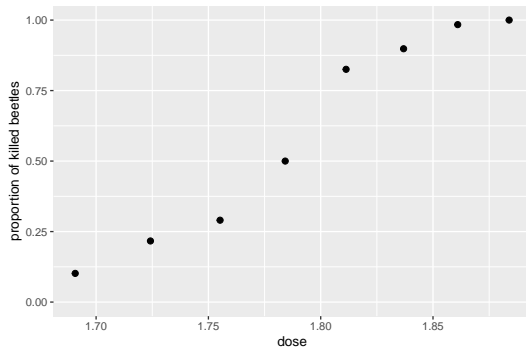
Mortality of the *confused flour beetle* (*Tribolium confusum*) *(potemník skladištní)*) due to exposure to gaseous carbon disulfide CS_2.
Model the dependence of mortality on the dose of CS_2.

logit model

```
m <- glm(cbind(killed,survived) ~ dose, data=dt, family=binomial(link="logit"))
summary(m)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.5941  -0.3944  0.8329  1.2592  1.5940
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43
Number of Fisher Scoring iterations: 4
```

probit model

```
m <- glm(cbind(killed,survived) ~ dose, data=dt, family=binomial(link="probit"))
```

CLogLog model

```
m <- glm(cbind(killed,survived) ~ dose, data=dt, family=binomial(link="cloglog"))
```

```
x.new <- data.frame(dose = seq(1.6, 2.0, by = 0.001))
q <- qnorm(1 - alpha / 2)
eta.logit <- predict(m, x.new, type = "link", se.fit = TRUE) |>
  as.data.frame() |>
  cbind(x.new) |>
  transmute(
    model = "logit",
    dose = dose,
    eta = fit,                          # linear predictor
    eta.lower = fit - q * se.fit,       # lower bound ( 2.5%) for lin. predictor
    eta.upper = fit + q * se.fit,       # upper bound (97.5%) for lin. predictor
    fit = 1 / (1 + exp(-eta)),          # fit
    fit.lower = 1 / (1 + exp(- eta.lower)),# lower bound ( 2.5%) for fit
    fit.upper = 1 / (1 + exp(- eta.upper)) # upper bound (97.5%) for fit
    )
```
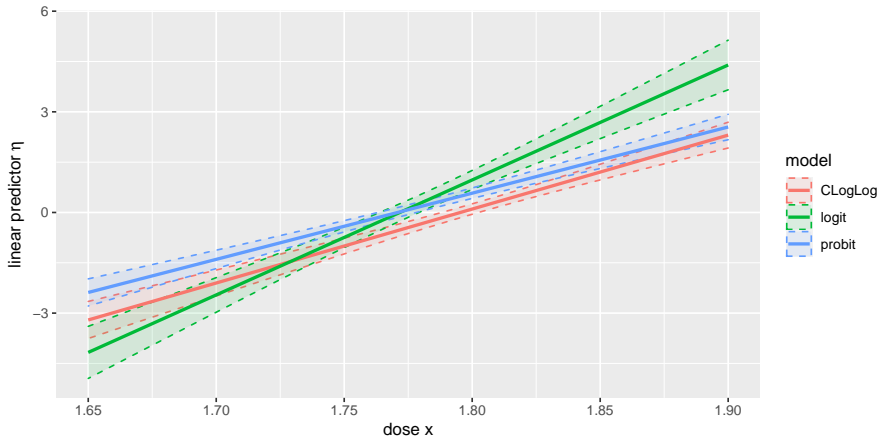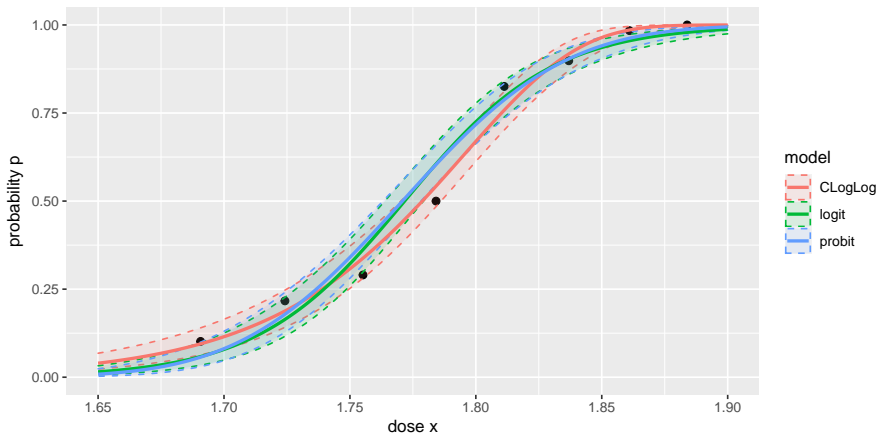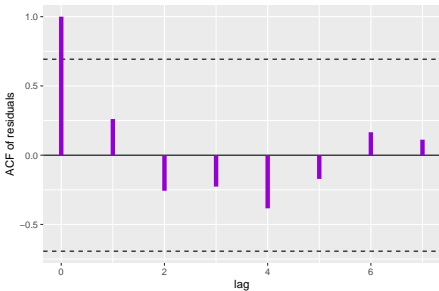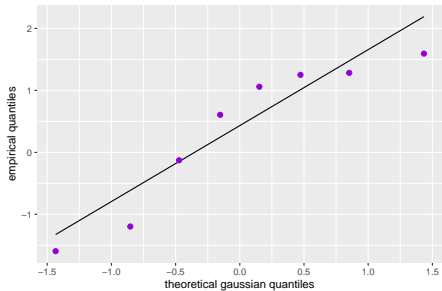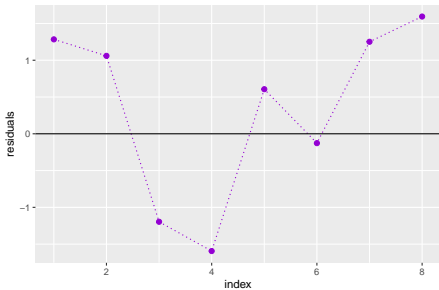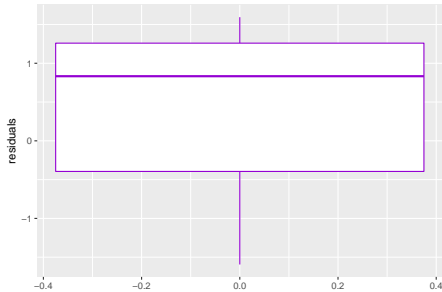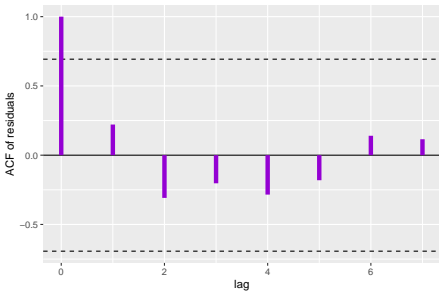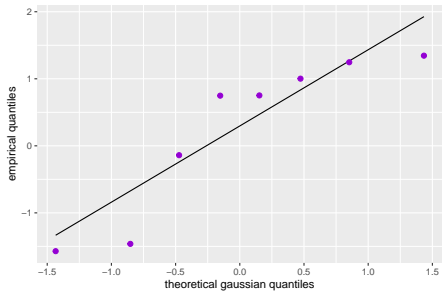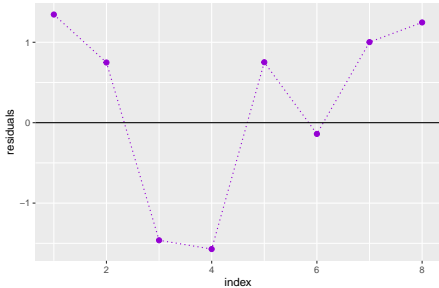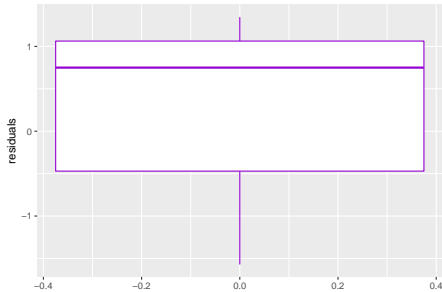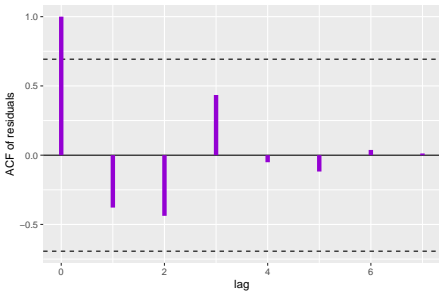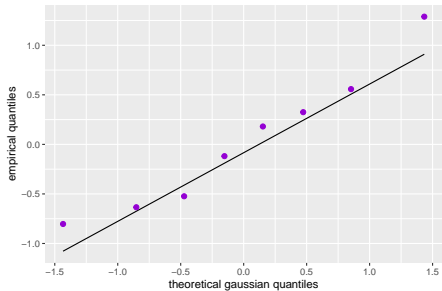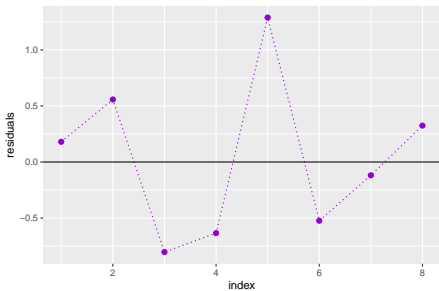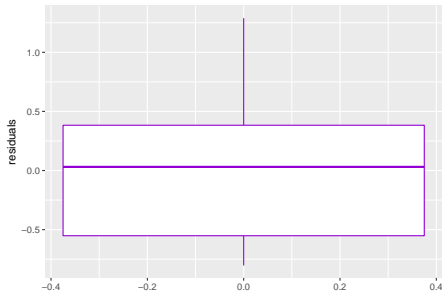
Definition (Odds)

Odds *(šance)* is the proportion of the probability of success and the probability of failure in the binomial (or alternative) distribution,

$$odds(Y_i) = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{p_i}{1 - p_i}.$$

In logistic regression, the odds are equal to the exponetial of linear predictor, or, the logarithm of odds is equal to the linear predictor,

$$odds(Y_i) = \exp(\beta_0 + \boldsymbol{x_i}' \boldsymbol{\beta}) = e^{\eta_i}, \qquad \ln odds(Y_i) = \eta_i = \beta_0 + \boldsymbol{x_i}' \boldsymbol{\beta}.$$

Consider logistic regression of dichotomous $Y \in \{0, 1\}$ by dichotomous predictor $x \in \{0, 1\}$. The association of Y and x is measured by *OR* statistic.

Definition (Odds ratio)

Odds ratio *(podíl šancí, podíl rizik)* is $OR = \dfrac{odds(Y \mid x = 1)}{odds(Y \mid x = 0)} = \dfrac{\frac{P(Y=1 \mid x=1)}{P(Y=0 \mid x=1)}}{\frac{P(Y=1 \mid x=0)}{P(Y=0 \mid x=0)}}$.

▶ $OR \approx 1$: non-association
▶ $OR > 1$: positive association
▶ $OR < 1$: negative association

By substituting the odds, we obtain
$$OR = \frac{odds(Y \mid x = 1)}{odds(Y \mid x = 0)} = \frac{\exp(x'\beta)\mid_{x=1}}{\exp(x'\beta)\mid_{x=0}}.$$

Further, if the linear predictor has the form of a linear function of the dichotomous predictor x,

$$\eta = \beta_0 + \beta_1 x, \quad \text{i.e., where} \quad \ln \frac{p}{1-p} = \beta_0 + \beta_1 x,$$

we obtain

$$OR = \frac{\exp(\beta_0 + \beta_1 x)\,|_{x=1}}{\exp(\beta_0 + \beta_1 x)\,|_{x=0}} = e^{\beta_1}, \quad \text{i.e.,} \quad \beta_1 = \ln OR.$$

The logarithm of the odds ratio is equal to the linear coefficient β_1 of the dichotomous predictor x.

Generalized linear models (GLM)

▶ linear relationship of the response Y on the regression coefficients:
  - Many real events show a relationship different from linear. Reciprocal, power, and other non-linear relationships are used to explain processes in the natural sciences.
  - The probability of a person's survival in case of a certain disease and a certain method of treatment can, by definition, take on values only from $[0, 1]$. In economics, many relationships have a logarithmic dependence.

▶ normal distribution of the response Y:
  - Normality is characterized by independence of mean and variance. This is, in genetral, not true for other distributions.
  - Typically, for economic variables with an increasing mean value, the variance of the random variable also increases, and the probability distribution is asymmetric.

▶ In linear regression model:
  - the mean of the response is modeled, $\mathsf{E}(Y_i) = \mu_i$,
  - the response has normal distribution, $Y_i \sim \mathsf{N}$,
  - variance $\mathsf{Var}(Y_i) = \sigma^2$ does not depend on the mean $\mathsf{E}(Y_i) = \mu_i$.

1. generalization to other than normal probability distributions of the response Y_i, specifically to the class of distribution of exponential family,
2. generalization to nonlinear functions, which connect the unknown mean values μ_i of the chosen probability distribution with predictors $X_1, \ldots, X_l$ using regression coefficients $\beta_0, \beta_1, \ldots, \beta_l$.

These allow

▶ analyze observations from a non-gaussian probability distribution and with a limited range, e.g., $[0, 1]$, $\{0, 1\}$, $\{0, 1, \ldots, n\}$, $\mathbb{N}_0$,

▶ to model other parameters of the probability distribution than the mean μ_i,

▶ consider models where the variance $\text{Var}(Y_i)$ depends on the mean value $\text{E}(Y_i) = \mu_i$.

Definition (Exponential family)

Random variable Y_i has probability distribution of exponential family / exponential class *(exponenciálního typu)*, if its probability mass function or probability density function $f(y_i)$ can be written as

$$f(y_i; \theta_i) = \exp\Big[T(y_i)\,A(\theta_i) + B(\theta_i) + C(y_i)\Big],$$

i.e., $\quad \ln f(y_i; \theta_i) = T(y_i)\,A(\theta_i) + B(\theta_i) + C(y_i), \quad$ where

- θ_i is natural parameter *(přirozený parametr)*,
- $T(y_i)$, $A(\theta_i)$, $B(\theta_i)$, $C(y_i)$ are known functions.
- If A is identity function, i.e., $A(y_i) = y_i$, so-called canonical form *(kanonická forma)* is obtained.
- Other parameters except the natural θ_i are called nuisance parameters *(rušivé parametry)*.

Definition (Canonical scaled form with one nuisance parameter)

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{\theta_i y_i - b(\theta_i)}{\frac{\phi}{w_i}} + c(\phi, y_i)\right],$$

$$\text{i.e.,} \quad \ln f(y_i; \theta_i, \phi) = \frac{w_i}{\phi}\left[\theta_i y_i - b(\theta_i)\right] + c(\phi, y_i), \quad \text{where}$$

- θ_i is natural parameter *(přirozený parametr)*,
- the nuisance parameter ϕ is called dispersion parameter *(disperzní parametr)*,
- w_i is called prior weight *(apriorní váha)*, by default $w_i = 1$,
- $b(\theta_i)$, $c(\phi, y_i)$ are known functions.
- Function $V(\theta_i) = b''(\theta_i)$ is called variance function *(rozptylová funkce)*.

Theorem

Under the regularity conditions, for distributions of exponential family in canonical form $f(y_i; \theta_i, \phi)$ with one nuisance parameter, we have

$$\mathsf{E}(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}, \qquad \mathsf{Var}(Y_i) = \frac{\phi}{w_i}\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \frac{\phi}{w_i}V(\theta_i).$$

Random sample $(Y_1, \ldots, Y_n)$ follows generalized linear model (GLM) *(zobecněný lineární model)*, if

1. probability distribution of $(Y_1, \ldots, Y_n)$ is of exponential family with simultaneous probability mass function or probability density function

$$f(\boldsymbol{y}; \boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} \exp\left[\frac{\theta_i y_i - b(\theta_i)}{\phi/w_i} + c(\phi, y_i)\right],$$

2. paramater θ_i depends on predictors $\boldsymbol{x}_i$ and regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_l)$ through the linear predictor η_i,

$$\eta_i = \boldsymbol{x}_i' \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij},$$

3. strictly monotonic and differentiable function g, so-called link function, is given, which models the non-linear relationship between the linear predictpr η_i and the mean $\mu_i = \mathsf{E}(Y_i)$ of the response,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}, \qquad \mu_i = g^{-1}(\eta_i), \qquad i = 1, \ldots, n.$$

The link function g is called canonical, if the linear predictor η_i is the natural parameter θ_i, i.e., if

$$g(\mu_i) = \theta_i = \eta_i, \qquad i = 1, \ldots, n.$$

The linear predictor η_i includes information about predictors in GLM, it is a linear combination of unknown parameters β.

The link function g describes the relationship between the linear predictor η_i and the median value μ_i of the observed probability distribution. The link function can be any strictly monotonic and differentiable function; in practice, we try to consider such functions that have a domain equal to the set of multiple mean values. The canonical link function expresses the natural parameter using the mean value, $\theta_i = g(\mu_i)$.

For some probability distributions, the mean μ_i is directly a parameter of the given distribution. In such a case, the canonical link function is a function g that converts the probability mass function or probability density function into a canonical form, $\theta_i = g(\mu_i)$.

$$Y_i \sim \mathsf{N}(\mu_i,\, \sigma^2), \qquad \mu_i \in \mathbb{R},\ \sigma^2 > 0,\ y_i \in \mathbb{R}$$

$$\ln f(y_i; \mu_i, \sigma^2) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \mu_i)^2}{2\,\sigma^2} =$$

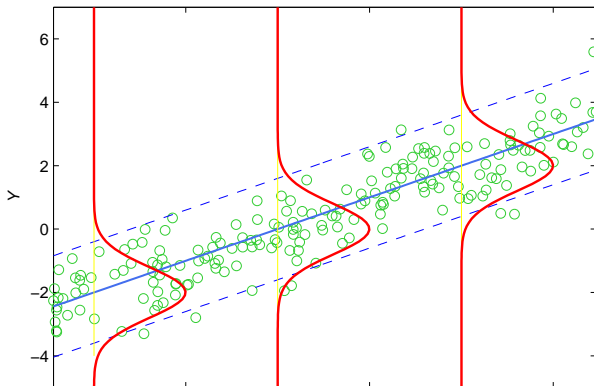$$= \frac{\mu_i\, y_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\,\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

▶ natural parameter is the mean: $\theta = \mu \in \mathbb{R}$

▶ dispersion parameter is the variance: $\phi = \sigma^2$, $w_i = 1$

▶ $b(\theta_i) = \frac{1}{2}\theta_i^2$

▶ variance function: $V(\theta_i) = b(\theta_i)'' = 1$

▶ $c(\phi, y_i) = -\frac{y_i^2}{2\,\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$

▶ $\mathsf{E}(Y_i) = b(\theta_i)' = \theta_i = \mu_i$

▶ $\mathsf{Var}(Y_i) = \frac{\phi}{w_i}\, V(\theta_i) = \sigma^2$

$$Y_i \sim \mathsf{N}(\mu_i, \sigma^2), \quad \mathsf{E}(Y_i) = \mu_i, \quad i = 1, \ldots, n.$$

Linka function in corresponding GLM is **identity**,

$$g(\mu_i) = \mu_i = \eta_i = \beta_0 + \beta_1 x_i,$$

regression coefficients β_0, β_1 and dispersion parameter σ^2 are unknown, x_i are known predictors.

$$Y_i \sim \mathsf{A}(p), \qquad p_i \in [0, 1], \; y_i \in \{0, 1\}$$

$$\ln f(y_i; p_i) = \ln p_i^{y_i}(1-p_i)^{1-y_i} = \ln \left(\frac{p_i}{1-p_i} \right)^{y_i} (1-p_i) =$$

$$= y_i \ln \frac{p_i}{1-p_i} + \ln(1-p_i) = y_i \theta_i - \ln(1 + \mathrm{e}^{\theta_i})$$

▶ natural parameter is logit transform of the probability of success:
$$\theta_i = \ln \frac{p_i}{1-p_i} \in \mathbb{R}, \qquad p_i = \frac{1}{1 + \mathrm{e}^{-\theta_i}}, \; 1-p_i = \frac{1}{1 + \mathrm{e}^{\theta_i}}$$

▶ $\phi = 1, w_i = 1$

▶ $b(\theta_i) = \ln(1 + \mathrm{e}^{\theta_i}) = -\ln(1-p_i)$

▶ variance function: $V(\theta_i) = b(\theta_i)'' = \frac{\mathrm{e}^{\theta_i}}{(1+\mathrm{e}^{\theta_i})^2} = p_i(1-p_i)$

▶ $c(\phi, y_i) = 0$

▶ $\mathsf{E}(Y_i) = b(\theta_i)' = \frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}} = p_i$

▶ $\mathsf{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = p_i(1-p_i)$

$$Y_i \sim \text{Bi}(n_i, p_i), \qquad n_i \in \mathbb{N}, \ p_i \in [0, 1], \ y_i \in \{0, 1, \ldots, n\}$$

$$\ln f(y_i; p_i) = \ln \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{1-y_i} = \ln \binom{n_i}{y_i} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) =$$

$$y_i \ln \frac{p_i}{1 - p_i} + n_i \ln(1 - p_i) + \ln \binom{n_i}{y_i} = y_i \theta_i - n_i \ln(1 + e^{\theta_i}) + \ln \binom{n_i}{y_i}$$

▶ natural parameter is logit transform of the probability of success:
$$\theta_i = \ln \frac{p_i}{1 - p_i} \in \mathbb{R}, \qquad p_i = \frac{1}{1 + e^{-\theta_i}}, \ 1 - p_i = \frac{1}{1 + e^{\theta_i}}$$

▶ $\phi = 1; w_i = 1$

▶ $b(\theta_i) = n_i \ln(1 + e^{\theta_i}) = -n_i \ln(1 - p_i)$

▶ variance function: $V(\theta_i) = b(\theta_i)'' = \frac{n_i e^{\theta_i}}{(1 + e^{\theta_i})^2} = n_i p_i (1 - p_i)$

▶ $c(\phi, y_i) = \ln \binom{n_i}{y_i}$

▶ $E(Y_i) = b(\theta_i)' = \frac{n_i e^{\theta}}{1 + e^{\theta}} = n_i p_i$

▶ $\text{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = n_i p_i (1 - p_i)$

$$\ln f(y_i; p_i, \phi) = \frac{y_i \ln \frac{p_i}{1-p_i} + n_i \ln(1-p_i)}{\phi} + c(\phi, y_i) =$$

$$= \frac{y_i \theta_i - n_i \ln(1 + e^{\theta_i})}{\phi} + c(\phi, y_i)$$

▶ natural parameter is logit transform of the probability of success:
$$\theta_i = \ln \frac{p_i}{1-p_i} \in \mathbb{R}, \qquad p_i = \frac{1}{1 + e^{-\theta_i}}, \ 1 - p_i = \frac{1}{1 + e^{\theta_i}}$$
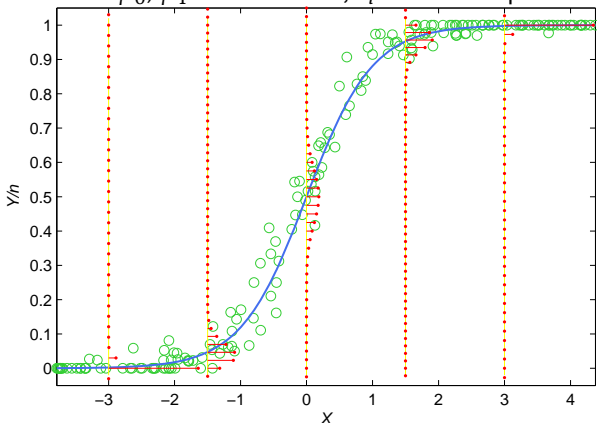
▶ dispersion paramater = ϕ; $w_i = 1$

▶ $b(\theta_i) = n_i \ln(1 + e^{\theta_i}) = -n_i \ln(1 - p_i)$

▶ variance function: $V(\theta_i) = b(\theta_i)'' = \frac{n_i e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i p_i (1 - p_i)$

▶ $c(\phi, y_i) = \ln \binom{n_i}{y_i}$

▶ $E(Y_i) = b(\theta_i)' = \frac{n_i e^{\theta}}{1+e^{\theta}} = n_i p_i$

▶ $\text{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = \phi \, n_i p_i (1 - p_i)$

▶ in R: `quasibinomial`

$$Y_i \sim \text{Bi}(n_i, p_i), \quad \text{E}\left(\frac{Y_i}{n_i}\right) = \mu_i = p_i, \quad i = 1, \ldots, n.$$

Link function in GLM is **logit function**,

$$g(p_i) = \ln \frac{p_i}{1 - p_i} = \eta_i = \beta_0 + \beta_1 x_i,$$

regression coefficients β_0, β_1 are unknown, x_i are known predictors.

$$Y_i \sim \text{Po}(\lambda_i), \qquad \lambda_i > 0, \ y_i \in \mathbb{N}_0$$

$$\ln f(y_i; \lambda_i) = \ln \frac{\lambda_i^{y_i}}{y_i!} \, e^{-\lambda_i} = y_i \ln \lambda_i - \lambda_i - \ln(y_i!) = \theta_i y_i - e^{\theta_i} - \ln(y_i!)$$
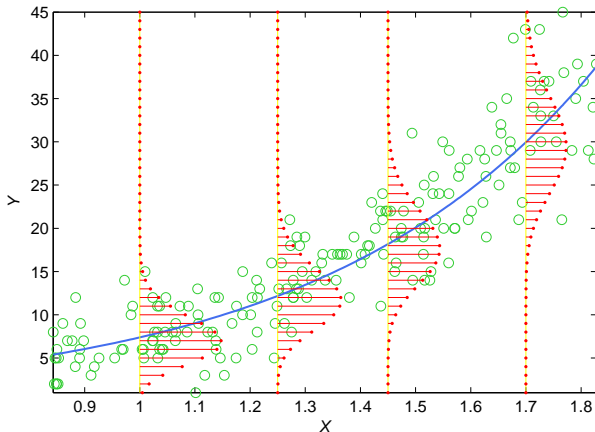
▶ natural parameter is logarithm of the intensity: $\theta_i = \ln \lambda_i \in \mathbb{R}$
▶ $\phi = 1; w = 1$
▶ $b(\theta_i) = e^{\theta_i} = \lambda_i$
▶ variance function: $V(\theta_i) = b(\theta_i)'' = e^{\theta_i} = \lambda_i$
▶ $c(\phi, y_i) = -\ln(y_i!)$
▶ $\text{E}(Y_i) = b(\theta_i)' = e^{\theta_i} = \lambda_i$
▶ $\text{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = \lambda_i$

$$Y_i \sim \mathsf{Po}(\lambda_i), \quad \mathsf{E}Y_i = \lambda_i, \quad i = 1, \ldots, n.$$

Link function in GLM is **logarithm**,

$$g(\mu_i) = \ln \mu_i = \ln \lambda_i = \eta_i = \beta_0 + \beta_1 x_i,$$

regression coefficients β_0, β_1 are unknown, x_i are known predictors.

$$\ln f(y_i; \lambda_i, \phi) = \frac{y_i \ln \lambda_i - \lambda_i}{\phi} + c(\phi, y_i) = \frac{\theta_i y_i - e^{\theta_i}}{\phi} + c(\phi, y_i)$$

- ▶ natural parameter is logarithm of the intensity: $\theta_i = \ln \lambda_i \in \mathbb{R}$
- ▶ dispersion paramater = ϕ; $w = 1$
- ▶ $b(\theta_i) = e^{\theta_i} = \lambda_i$
- ▶ variance function: $V(\theta_i) = b(\theta_i)'' = e^{\theta_i} = \lambda_i$
- ▶ $E(Y_i) = b(\theta_i)' = e^{\theta_i} = \lambda_i$
- ▶ $\text{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = \phi \lambda_i$
- ▶ in R: `quasipoisson`

$$Y_i \sim \mathsf{Ex}(\lambda_i), \qquad \lambda_i > 0, \ y_i \geq 0$$

$$\ln f(y_i; \lambda_i) = \ln \lambda_i \, e^{-\lambda_i y_i} = -\lambda_i y + \ln \lambda_i = \theta_i y_i + \ln(-\theta_i)$$

- ▶ natural parameter: $\theta_i = -\lambda_i < 0$
- ▶ $\phi = 1; w = 1$
- ▶ $b(\theta_i) = -\ln(-\theta_i) = -\ln \lambda_i$
- ▶ variance function: $V(\theta_i) = b(\theta_i)'' = \frac{1}{\theta_i^2} = \frac{1}{\lambda_i^2}$
- ▶ $c(\phi, y_i) = 0$
- ▶ $\mathsf{E}(Y_i) = b(\theta_i)' = -\frac{1}{\theta_i} = \frac{1}{\lambda_i}$
- ▶ $\mathsf{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = \frac{1}{\lambda_i^2}$

$$Y_i \sim \mathsf{Ex}(\mu_i), \qquad \mu_i > 0, \ y_i \geq 0$$

$$\ln f(y_i; \mu_i) = \ln \frac{1}{\mu_i} \, e^{-y_i/\mu_i} = -\frac{y_i}{\mu_i} - \ln \mu_i = \theta_i \, y_i + \ln(-\theta_i)$$
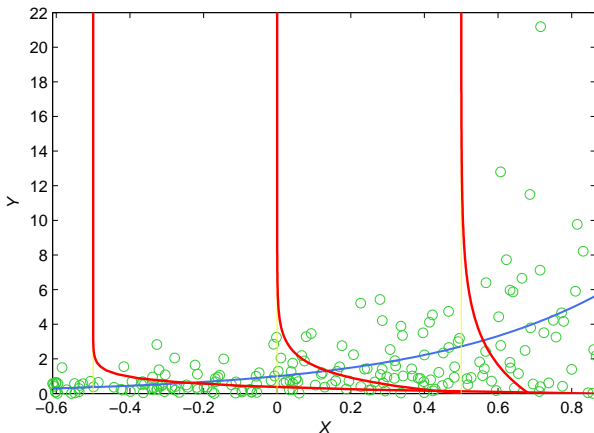
- natural parameter: $\theta_i = -\frac{1}{\mu_i} < 0$
- $\phi = 1; w = 1$
- $b(\theta_i) = -\ln(-\theta_i) = \ln \mu_i$
- variance function: $V(\theta_i) = b(\theta_i)'' = \frac{1}{\theta_i^2} = \mu_i^2$
- $c(\phi, y_i) = 0$
- $\mathsf{E}(Y_i) = b(\theta_i)' = -\frac{1}{\theta_i} = \mu_i$
- $\mathsf{Var}(Y_i) = \frac{\phi}{w_i} \, V(\theta_i) = \mu_i^2$

$$Y_i \sim \mathsf{Ex}(\mu_i) \equiv \mathsf{G}(1, \mu_i), \quad \mathsf{E}Y_i = \mu_i, \quad i = 1, \ldots, n.$$

Linka function in GLM is **logarithm**,

$$g(\mu_i) = \ln \mu_i = \eta_i = \beta_0 + \beta_1 x_i,$$

regression coefficients β_0, β_1 are unknown, x_i are known predictors.

$$Y \sim \mathsf{G}(k, \mu_i), \qquad k > 0, \ \mu_i > 0, \ y_i \geq 0$$

$$\ln f(y_i; \mu_i, k) = \ln \frac{1}{\Gamma(k)} \left(\frac{\mu_i}{k}\right)^{-k} y_i^{k-1} \exp\left[-\frac{k y_i}{\mu_i}\right] =$$

$$= \frac{-\frac{y_i}{\mu_i} - \ln \mu_i}{\frac{1}{k}} + k \ln k - \ln \Gamma(k) + (k-1) \ln y_i$$

- natural parameter: $\theta_i = -\frac{1}{\mu_i} < 0$
- $\phi = \frac{1}{k}; w = 1$
- $b(\theta_i) = -\ln(-\theta_i) = \ln \mu_i$
- variance function: $V(\theta_i) = b(\theta_i)'' = \frac{1}{\theta_i^2} = \mu_i^2$
- $c(\phi, y_i) = k \ln k - \ln \Gamma(k) + (k-1) \ln y_i$
- $\mathsf{E}(Y_i) = b(\theta_i)' = -\frac{1}{\theta_i} = \mu_i$
- $\mathsf{Var}(Y_i) = \frac{\phi}{w_i} V(\theta_i) = \frac{\mu_i^2}{k}$

$$Y_i \sim \mathsf{G}(k, \mu_i), \quad \mathsf{E}Y_i = \mu_i, \quad i = 1, \ldots, n.$$

Linka function in GLM is **logarithm**,

$$g(\mu_i) = \ln \mu_i = \eta_i = \beta_0 + \beta_1 x_i,$$

regression coefficients β_0, β_1 and dispersion parameter $\phi = 1/k$ are unknown, x_i are known predictors.