Statistics II | seminar 1

Work with data frames, descriptive statistics, linear regression model, t-test

Markéta Zoubková, Ondřej Pokora

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

9. 9. 2022

Example (1)

1. Load data from file `people.csv` into a table. Notice column separator and encoding (UTF-8). Work with this table for the rest of the seminar.
2. Explore data types of variables in the table.
3. Check the data types and change them if needed. Data type `factor` is used for categorical variables and `numeric` or `integer` for numerical variables. Correct data types or/and values of variables *Weight*, *Sex* and *BMI*.
4. Recall exploratory data analysis (EDA). Numerical variables: compute sample means, standard deviations, variances and medians, draw boxplots and histograms. Categorical variables: compute frequency tables, draw frequency bar charts.
5. Draw scatter plots of Height (x), Weight (y) and Height (x), BMI (y) with coloring based on the variable Sex.
6. Variable *BMI* shows some missing (`NA`) or mistaken observations. Find and return these rows of the table. Using the BMI formula and variables *Weight* and *Height* add new variable *BMI.new* with correctly computed values.

Example (2)

7. Use linear regression model with regression line to analyze the dependence of *Weight* on *Height* and *Sex*. Which regression coefficients are statistically significant? (Use significance level $\alpha = 0{,}05$.) Interpret the meaning of each coefficient. Data and model represent graphically, including 95% prediction intervals.
8. Recall (two-sample) t-test. Test a hypothesis about differences in mean BMI of men and women. Formulate null and alternative hypothesis and decide the result of the test.
9. Test previous hypothesis for another binary variables. E.g., test a hypothesis about differences in mean BMI when grouped by *iPhoneOwner* variable and explain the result.

Statistics II | seminar 2

One-way analysis of variance (ANOVA)

Markéta Zoubková, Ondřej Pokora

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

19. 9. 2022

Example (1)

Total weight of potatoes grown from one plant was measured for four potato varieties (denoted 1–4).

Variety	Weight (in 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		

Data file `potatoes.csv`: At a significance level of 5 %, test that the expected value of weight is not distinct across the varieties. Count effects and arithmetic means of the potato varieties. Recall ANOVA table and the meaning of the variables in it. Determine which pairs of varieties differ in case of rejecting the null hypothesis. Use both Tukey and Scheffe test.

Overall: Remember to use descriptive statistics and visualisation to get to know your data. Make sure that all of the conditions related to the chosen testing method are satisfied.

Example (2)

Data file salesman.csv contains monthly sales (in thousands Kč) for three salesmen in half-year period.

At a significance level of 5 %, test null hypothesis that the expected values of the monthly sales for all salesmen are equal. Determine pairs of salesmen exhibiting different sales in case of rejecting the null hypothesis.

Example (3)

Data file coagulation.csv: Clinical study observed relationship between a diet and time needed for blood coagulation. The research was performed on 24 animals fed with four different diets (A–D).

At a significance level of 0.05, test null hypothesis that the expected values of the coagulation time for all diets are equal. In case of rejecting the null hypothesis, determine pairs of diets that have significantly different coagulation time.

Example (4, manual calculation, R only as calculator and quantile table)

It is given $a = 5$ independent random samples of sizes $n_i = 5; 7; 6; 8; 5$, where i-th sample is of a distribution $N(\mu_i, \sigma^2)$, $i = 1, \ldots, 5$.
The total sum of squares $S_T = 15$ and the sum of squares of the residual error $S_e = 3$ are known.

Compute the treatment sum of squares S_A and at a significance level of 5 %, test hypothesis that expected values are equal in all five considered samples.

Example (5)

Three species of mouse were tested on aggression based on their behavior in a maze. At the start each mouse was placed to the center of the square maze with sides equal to 1 m and divided to 49 identical squares. While the mouse was trying to escape the maze for 5 minutes, the number of crossed squares was observed.
Data file `mice.csv` contains the number of crossed squares for each species of mouse noted in separated columns. There are some missing values (`NA`) for the third species.
At a significance level of 0,05, test hypothesis that the expected values of the number of crossed squares for all three species of mouse are equal. Determine which pairs of species significantly differ in case of rejecting the null hypothesis.

Example (6)

Consider linear regression models

$$\begin{aligned} M_1: &\quad y = \beta_0 + \beta_1 x, \\ M_2: &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2, \\ M_3: &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \end{aligned}$$

for values from file `data01.csv`. Using the analysis of variance compare these models and select the most suitable one for the data.

Example (7)

Annual depth data (ft) for lake Huron between the years 1875 and 1972 are stored in variable `LakeHuron` in *R* environment. Using the linear regression model fit the polynomial function of degree 8 to the data. Then using the analysis of variance look into the possibility of reducing the degree of the polynomial.

Example (8)

Sugar beet yield (in hundredweight per hectare) was monitored for 126 factories in Czechia in relation to the used amount of fertilizer K_2O (in kilograms per hectare).

Interval data are in file `sugarbeet.csv` in 4 columns:

1. lower limit of an interval K_2O (kg/ha),
2. upper limit of an interval K_2O (kg/ha),
3. number of factories that use K_2O in given interval,
4. average yield for sugar beet (q/ha).

Consider linear regression models

$$\begin{aligned} M_1: &\quad y = \beta_0 + \beta_1 x, \\ M_2: &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2. \end{aligned}$$

Choose the midpoint of the interval for the values of x.

Using the analysis of variance select the most suitable model for the data.

Results

1. ANOVA rejects H_0. Distinct pairs: Tukey: 2-1, 3-1. Scheffe: 3-1.
2. ANOVA rejects H_0. Distinct pairs: 3-1, 3-2.
3. ANOVA rejects H_0. Distinct pairs: A-B, A-C, B-D, C-D.
4. ANOVA rejects H_0. $S_A = 12$, $F_A = 26$.
5. ANOVA rejects H_0. Distinct pairs: all.
6. M_2
7. Polynomial function of degree 7.
8. M_2

Statistics II | seminar 3

Two-way analysis of variance (ANOVA)

Markéta Zoubková, Ondřej Pokora

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

26 September 2022

Example (1)

Data file hay.csv:

Analyze the hay yields (in tons per hectare) in relation to the soil type (neutral / acidic) and fertilizer (none / dung / calcium fertilizer). Each combination was realized four times and independently of each other. Perform two-way analysis of variance with and without interactions and one-way analysis of variance. Test null hypothesis at a significance level of 0.05:

- soil type and fertilizer are independent, i.e. they do not interact.
- soil type has no effect on yields;
- fertilizer has no effect on yields;

Determine distinct groups in case of rejecting the null hypothesis. Which factor causes the difference in expected value of the yields?

Example (2)

Work with the tabel in variable ToothGrowth (in *R* environment) that contains the results of the study on the effect of the vitamin C on guinea-pig's teeth growth:
len = length of odontoblasts (cells of the outer surface of the tooth);
supp = administration of vitamin C (OJ = in orange juice, VC = in ascorbic acid);
dose = dose of vitamin C (3 groups: 0.5; 1.0 a 2.0 mg/day).

Perform two-way analysis of variance with interactions in relation to the administration and dose of vitamin C. Test corresponding null hypotheses at a significance level of 0.05. Determine distinct groups in case of rejecting the null hypothesis that the expected values of odontoblasts length are equal. Which factors and their interaction cause the difference in expected value of odontoblasts length?

Example (3)

Data file goods.csv:

Certain goods sales were observed over the same period of time in relation to two factors. The goods were sold packed in a bag or box and with the support of advertising campaign only in the press or in the press and also on television or without any advertising at all. File contains profit data in millions Kč from the goods sales under the stated conditions. Perform two-way analysis of variance with and without the interactions. Test null hypotheses at a significance level of 0.05:

- there is no interaction between the type of the packaging and the type of the advertisement;
- expected values of the profit do not depend on the type of the packaging;
- expected values of the profit do not depend on the type of the advertisement.

Determine distinct groups in case of rejecting the null hypothesis that the expected values are equal. Which factor causes the difference in expected value of the profit?

Example (4)

Data file `corn.csv`:
When determining the yield of corn, measurements were made on three different types of seeds and five different methods of fertilization. Two measurments were made for each combination. Perform two-way analysis of variance with and without interactions.. Test null hypotheses at a significance level of 0.05:

- there is no interaction between the type of seed and fertilization method;
- expected values of the yields are equal for all types of seeds;
- expected values of the yields are equal for all fertilization methods.

Determine distinct groups in case of rejecting the null hypothesis that the expected values are equal. Which factor causes the difference in expected value of the yields?

Results

1. Two-way ANOVA does not reject H_0 for Soil, but it does for Fertilizer, therefore Fertilizer has effect on Yield. Distinct pairs: none-calcium, none-dung.
   Two-way ANOVA with interactions moreover rejects H_0 for interactions. Distinct pairs for Fertilizer: all. Distinct pairs for interactions: 8 (Tukey), 7 (Sheffe).
2. Two-way ANOVA rejects H_0 for both factors. Distinct pairs: all.
   Two-way ANOVA with interactions moreover rejects H_0 for interactions. Distinct pairs for interactions: 11.
3. Two-way ANOVA rejects H_0 for both factors. Distinct pairs for Advert: PressTV-None, PressTV-Press.
   Two-way ANOVA with interactions cannot be performed due to the low number of observations.
4. Two-way ANOVA rejects H_0 for both factors. Distinct pairs for Variety: C-A, C-B. Distinct pairs for Fertilizer: 2-1, 4-2, 5-2.
   Two-way ANOVA with interactions does not reject H_0 for interactions, therefore two-way ANOVA without interactions is sufficient.

Statistics II | seminar 4

Rank-based tests and methods

Markéta Zoubková, Ondřej Pokora

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

3 October 2022

Example (1)

Data file `minute.csv`. Ten research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal. Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated. Check functions `sort` and `rank`. Perform the sign test and one-sample Wilcoxon signed-rank test.

Example (2)

Data file `field.csv`. Two fertilization methods were tested on 13 fields of the same soil quality. The first method A was tested on 8 fields and the second method B was tested on the remaining fields. Data file contains wheat yields (in tons per hectare) of tested fields. Determine if the fertilization method has effect on the wheat yields. Perform two-sample Wilcoxon rank-sum test (Mann-Whitney test).

Example (3)

Use data file `potatoes.csv` from the previous seminar. Test hypothesis using the Kruskal-Wallis and the median test that the expected value of weight of the potatoes grown from one plant does not depend on the variety. Furthermore, determine significantly different pairs of varieties.

Example (4)

Data file octane.csv. Octane rating was observed for 10 samples of gasoline. Test $H_0 : \widetilde{x} = 98$ using the sign test and the Wilcoxon signed-rank test.

Example (5)

Data file bloodpressure.csv. Systolic blood pressure was measured for 8 randomly chosen people before and after a medical procedure. Test hypothesis that the medians of the systolic blood pressure before and after the medical procedure are equal.

Example (6)

Data file machines.csv. There are 6 machines from 3 different manufacturers in the factory. Perform the Kruskal-Wallis test for the hypothesis that the expected value of machine efficiency does not depend on the manufacturer.

Example (7)

Data file activesubstance.csv. The amount of active substance was observed for products of two suppliers. Test hypothesis that the medians of the amount of active substance are equal for both suppliers.

Example (8)

Data file minute2.csv. Thirty research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal. Data file contains measured times in seconds. Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated. Use the sign test and one-sample Wilcoxon signed-rank test with one-sided and two-sided H_1.

Example (9)

Data file nickel.csv. Four laborants measured the containment of nickel in steel. Perform non-parametric one-way analysis using the Kruskal-Wallis test and the median test. Determine the pairs of laborants whose results significantly differ.

Example (10)

Data file potatoes2.csv. Potato yields of 4 potato varieties were observed. The yields in tons per hectare are recorded in the file. Perform the Kruskal-Wallis test amd the median test. Determine the pairs of varieties whose expected yields significantly differ.

Example (11)

Data file paper.csv. Four labs measured the smoothness of the paper. Perform non-parametric one-way analysis using the Kruskal-Wallis test and the median test. Determine the pairs of labs whose results significantly differ.

Example (12)

Data file IQvitaminB.csv. The effect of vitamin B on IQ was observed in 24 pairs of children. One child out of a pair was administered the vitamin B and the other one was administered the placebo. Compute the difference between the IQ in pairs. Test hypothesis that the medians of the differences in IQ between both groups of children are equal. (Try also one-sided alternatives.)

Example (13)

Data file rats.csv. One of the methods of wound sealing was chosen after surgery for twenty rats – the stitches or a compression bandage. Data on tension in the wound edges were collected. Test hypothesis that the medians of tension for both methods of wound sealing are equal.

Example (14)

Data file sales.csv: numbers of sold pieces of 5 different products in 7 different stores. Use Friedman test to test the hypothesis that the product type has no effect on the sales in particular stores (blocks).

Results

1. The sign test does not reject H_0 and the Wilcoxon signed-rank test does.
2. Wilcoxon rank-sum test rejects H_0.
3. Kruskal-Wallis test rejects H_0 and median test does not. Distinct pairs of varieties: 3-4, 3-1, 2-1, 4-1.
4. Both test do not reject H_0.
5. Paired Wilcoxon signed-rank test does not reject H_0.
6. H_0 that medians are equal is not rejected.
7. Wilcoxon rank-sum test does not reject H_0.
8. Both tests reject H_0. After repeating the test for one-sided H_1 we show that median is < 60.
9. Both methods reject H_0. Distinct pairs: D-A, D-C, B-C, (B-A).
10. Both methods reject H_0. Distinct pairs: B-A, B-D, C-D, A-D, (B-C).
11. Both methods reject H_0. Distinct pairs: all pairs except B-C, (A-B).
12. Paired Wilcoxon signed-rank test does not reject H_0.
13. Wilcoxon test does not reject H_0.
14. Friedman test does not reject H_0.

Statistics II | seminar 5

Goodness-of-fit tests

Markéta Zoubková, Ondřej Pokora

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

10 October 2022

Example (1)

Data file families.csv. It was randomly selected 84 families from a set of families of 5 children and the number of boys was detected for each family. 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)

Data file line.csv. 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). At a significance level of 0.05, test hypothesis that waiting time has exponential distribution using Pearson's chi-squared test and test for exponential distribution.

Example (3)

Data file trains.csv. The number of arriving trains during 1 hour was observed at the station for 5 days (i.e. 360 hours). At a significance level of 0.05, test hypothesis that the number of arriving trains during 1 hour follows Poisson distribution using Pearson's chi-squared test and test for Poisson distribution.

Example (4)

Data file Brno.csv contains a number of residents of Brno (data from the year 2001) sorted by birth month. At a significance level of 0.05, test hypothesis that the number of births in each month corresponds to the number of days in the month (consider a non-leap year).

Example (5)

Data file dice.csv. The absolute frequencies of the numbers thrown for 60 dice rolls were observed. At a significance level of 0.05, test hypothesis that the dice is homogeneous (fair).

Example (6)

Data file emergency.csv contains a number of patients in the emergency room during 8 hour shift in a total of 75 days. Test hypothesis that the number of patients has Poisson distribution using Pearson's chi-squared test and test for Poisson distribution.

Example (7)

Data file `dice2.csv`. The number of sixes getting in 12 dice was observed for 4096 rolls with 12 dice. At a significance level of 0.05, test hypothesis that the number of sixes in one roll has binomial distribution $\mathrm{Bi}(12;\ \frac{1}{6})$.

Example (8)

Data file `pond.csv`. Five traps was set into a pond glowing with white, yellow, blue, green and red light. The number of fish caught in each trap is contained in the data file. At a significance level of 0.05, test hypothesis that the color of light has no effect on the number of fish caught.

Example (9)

Data file `supermarkets.csv`. 300 customers of a supermarket chain were questioned which day of the week they shop the most. Data file contains these numbers. At a significance level of 0.05, test hypothesis that the customers shop evenly every day of the week.

Example (10)

Data file `football.csv`. Total number of goals scored in a total match time was observed for 84 football matches. Data file contains these numbers. At a significance level of 0.05, test hypothesis that the number of goals scored has Poisson distribution using Pearson's chi-squared test.

Example (11)

Variables `JANT` and `JULT` in data file `pollution.csv` contain January and July temperatures in °F. Convert them to °C and use the Lilliefors test (also the Kolmogorov-Smirnov test) and Pearson's chi-squared test at a significance level of 0.05 to test that these variables have normal distribution. Plot an empirical and a teoretical distribution functions.

Example (12)

Variable `Rainfall` in data file `Hawaii.csv` contains amount of precipitation in Hawaii. Use the Kolmogorov-Smirnov test and Pearson's chi-squared test at a significance level of 0.05 to test that the amount of precipitation has chi-squared distribution with 18 degrees of freedom. Plot an empirical and a teoretical distribution functions.

Results

1. $K = 3.1$, $\chi^2_{0.95}(5) = 11.1$. Assumptions $\Rightarrow$ merge outer pairs of categories $\Rightarrow K = 2.4$, $\chi^2_{0.95}(3) = 7.8$. We do not reject the null hypothesis.
2. $K = 4.8$, $\chi^2_{0.95}(6) = 12.6 \Rightarrow$ Pearson's chi-squared test does not reject the null hypothesis.
   $Q = 35.7 \rightarrow$ Simple test rejects the null hypothesis.
3. $K = 9.7$, $\chi^2_{0.95}(6) = 12.6$; $Q = 331.1$. We do not reject the null hypothesis.
4. $K = 1506.2$, $\chi^2_{0.95}(11) = 19.7$. We reject the null hypothesis, month of birth distribution is not uniform distribution.
5. $K = 2.8$, $\chi^2_{0.95}(5) = 11.1$. We do not reject the null hypothesis.
6. Assumptions $\Rightarrow$ merge the last 2 categories $\Rightarrow K = 8.7$, $\chi^2_{0.95}(8) = 15.5$; $Q = 1158.6 \rightarrow$ Both tests do not reject the null hypothesis.
7. $K = 5.5$, $\chi^2_{0.95}(7) = 14.1$. We do not reject the null hypothesis.
8. $K = 14.1$, $\chi^2_{0.95}(4) = 9.5$. We reject the null hypothesis, color has effect on the number of fish caught.
9. $K = 78$, $\chi^2_{0.95}(6) = 12.6$. We reject the null hypothesis, preferences of the shopping days are not uniform.
10. $K = 2.1$, $\chi^2_{0.95}(3) = 7.8$. We do not reject the null hypothesis.

Statistics II | seminar 6

Correlation analysis

Markéta Zoubková, Ondřej Pokora

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

17. 10. 2022

Perform correlation analysis using the Pearson correlation coefficients, partial correlation coefficients, their significance tests and visualizations (scatterplots and correlograms). Interpret the results. Calculate also Spearman's and Kendall rank correlation coefficients.

Example (1)

Data file `households.csv`. The following variables were observed in 7 households: number of members, income and food and drink expenses (in thousands Kč per 3 months). Calculate sample multiple correlation between `expense` and remaining variables.

Example (2)

20 children in different ages underwent pedagogical-psychological research within which among others they wrote a dictation. Data file `children.csv` contains children's weight, age and points earned in dictation. Calculate sample multiple correlation between `Points` and remaining variables.

Example (3)

Data file `children2.csv`. Memory capacity, IQ and speed reading ability were observed for 20 children of different ages. Calculate sample multiple correlation between `IQ` and remaining variables.

Example (4)

Data file enrollment.csv. The number of filed university applications (ROLL) was observed in one of the states of USA for 29 years (YEAR) in relation to the average unemployment rate (UNEM, in %), the average wage (INC, in USD) and the number of students (HGRAD) who successfullly graduated from high school in a given year. Focus on the dependence of the number of university applications (the coefficient of multiple correlation) on the rest of the variables. Calculate sample multiple correlation between HRAD and remaining variables.

Example (5)

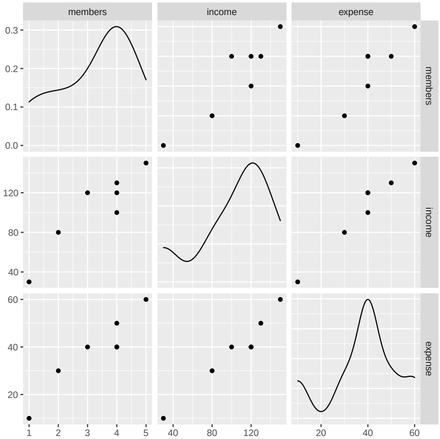
Data file heptathlon.csv contains heptathletes' results from a competition. Perform correlation analysis of given variables of results of 7 disciplines and column score. Examine their associations and interpret the results. Calculate sample multiple correlation between score and disciplines' results.

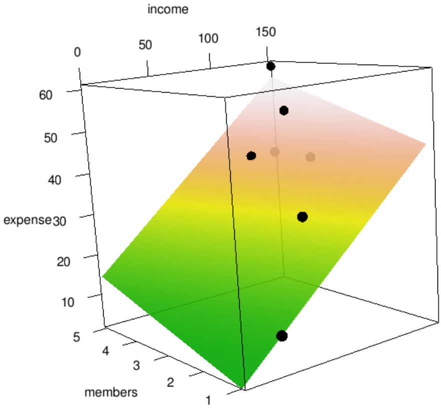
Example (6)

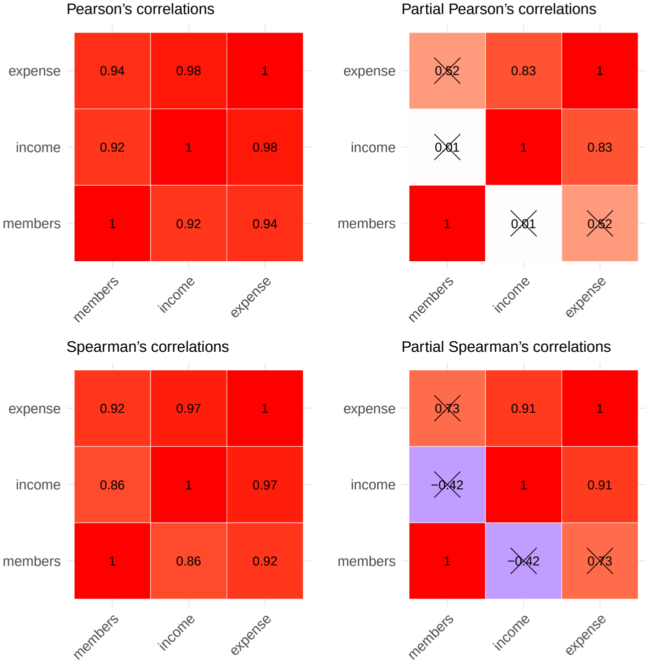
Data file decathlon.csv contains decathlets' results from competitions. Perform correlation analysis of given variables of results of 10 disciplines and column Points. Examine their associations and interpret the results. Calculate sample multiple correlation between Points and disciplines' results.

Results

Sample multiple correlations:
1. 0.983; **2.** 0.990; **3.** 0.731; **4.** 0.959; **5.** 0.997; **6.** 1.



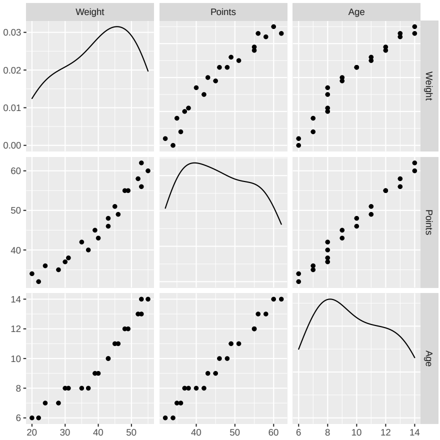


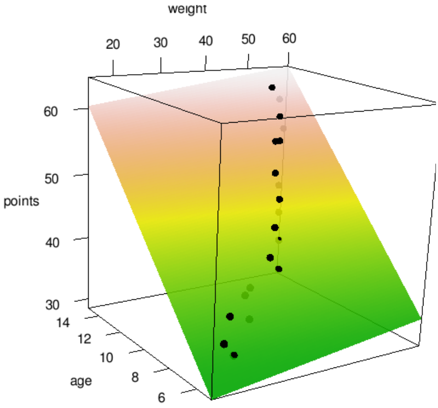

Pearson’s correlations

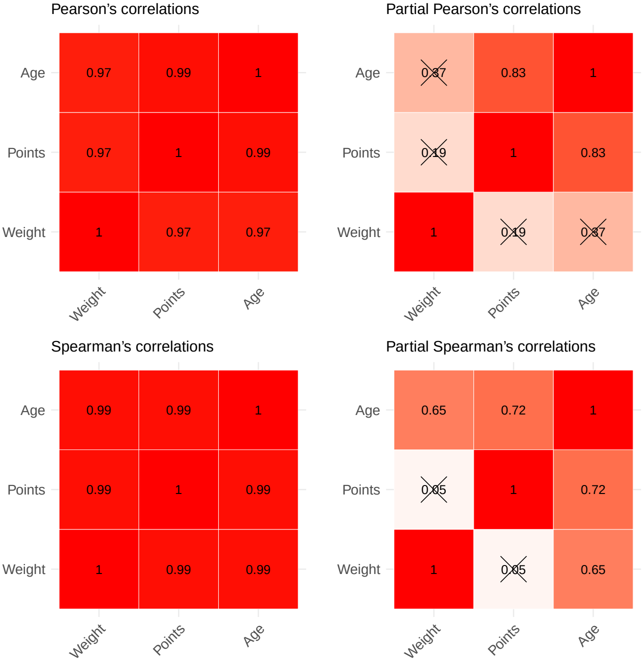
Partial Pearson’s correlations

Spearman’s correlations

Partial Spearman’s correlations


Weight
Points
Age
0.03
0.02
0.01
0.00
60
50
40
14
12
10
8
6
20
30
40
50
40
50
60
6
8
10
12
14
Weight
Points
Age


weight
20
30
40
50
60
points
60
50
40
30
age
14
12
10
8
6

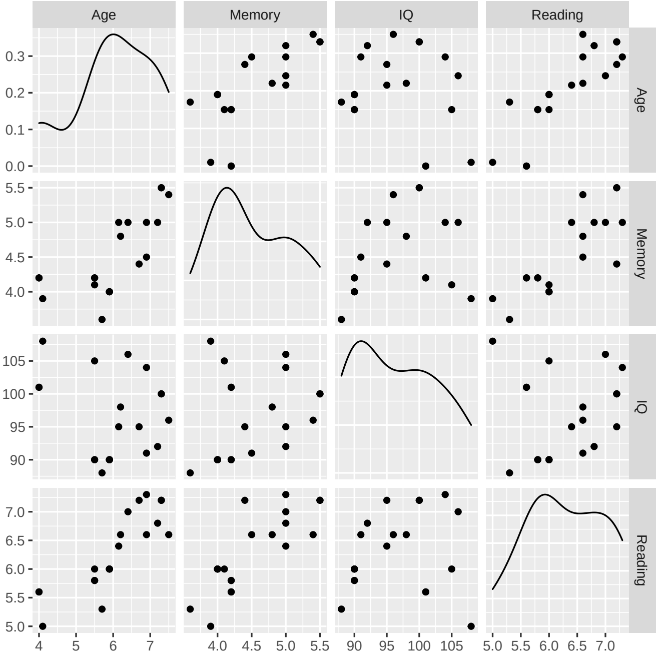


Pearson’s correlations

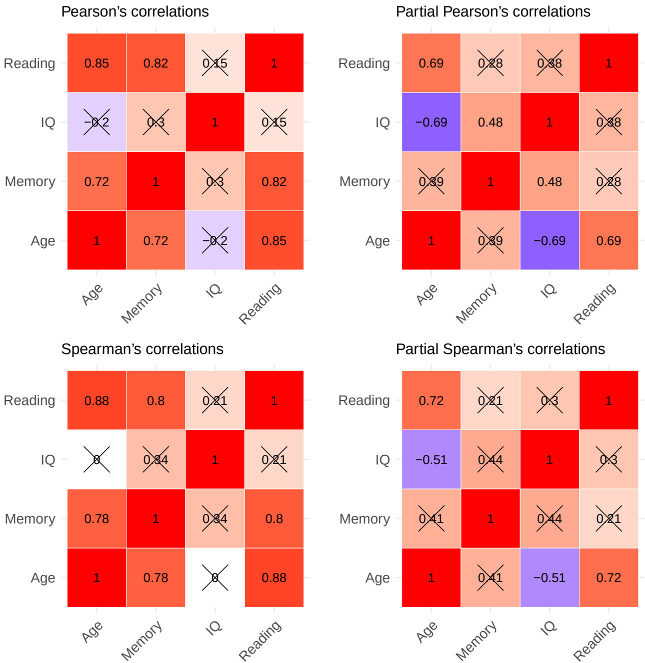
Partial Pearson’s correlations

Spearman’s correlations

Partial Spearman’s correlations



Age
Memory
IQ
Reading
Age
Memory
IQ
Reading
0.3
0.2
0.1
0.0
5.5
5.0
4.5
4.0
105
100
95
90
7.0
6.5
6.0
5.5
5.0
4
5
6
7
4.0
4.5
5.0
5.5
90
95
100
105
5.0
5.5
6.0
6.5
7.0

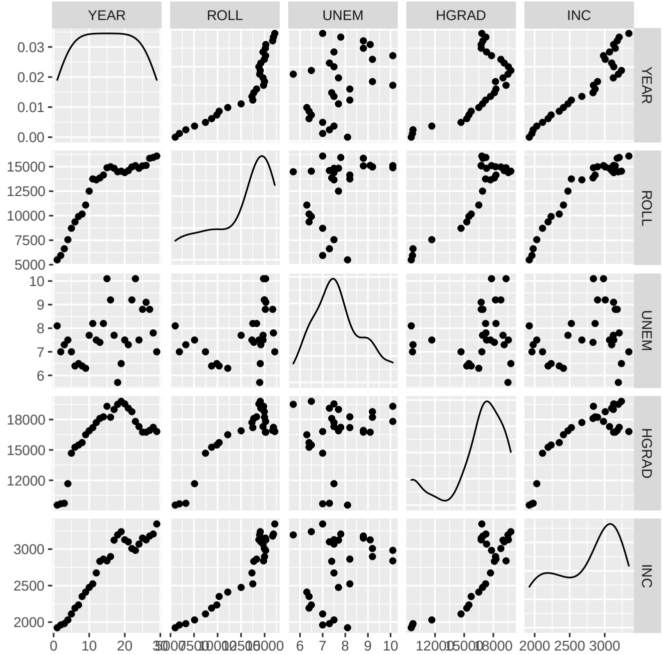


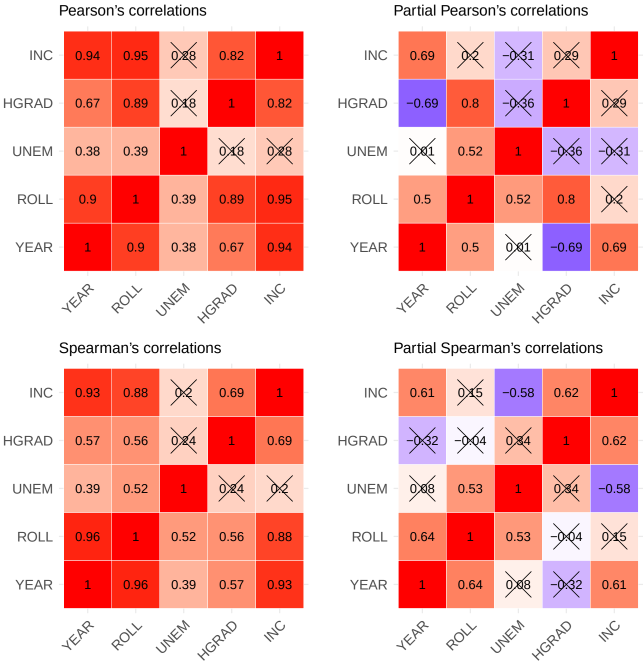
Pearson’s correlations

Partial Pearson’s correlations

Spearman’s correlations

Partial Spearman’s correlations


YEAR
ROLL
UNEM
HGRAD
INC
0.03
0.02
0.01
0.00
15000
12500
10000
7500
5000
10
9
8
7
6
18000
15000
12000
3000
2500
2000
0
10
20
30
5000
7500
10000
12500
15000
6
7
8
9
10
12000
15000
18000
2000
2500
3000

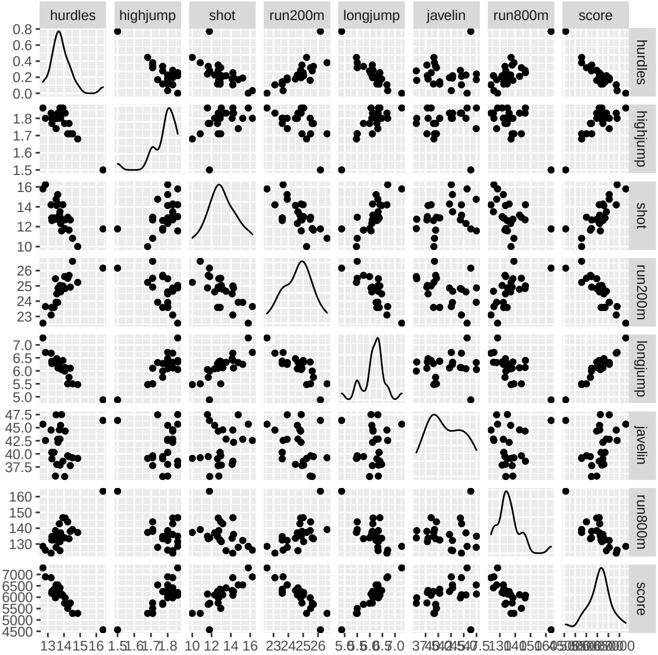


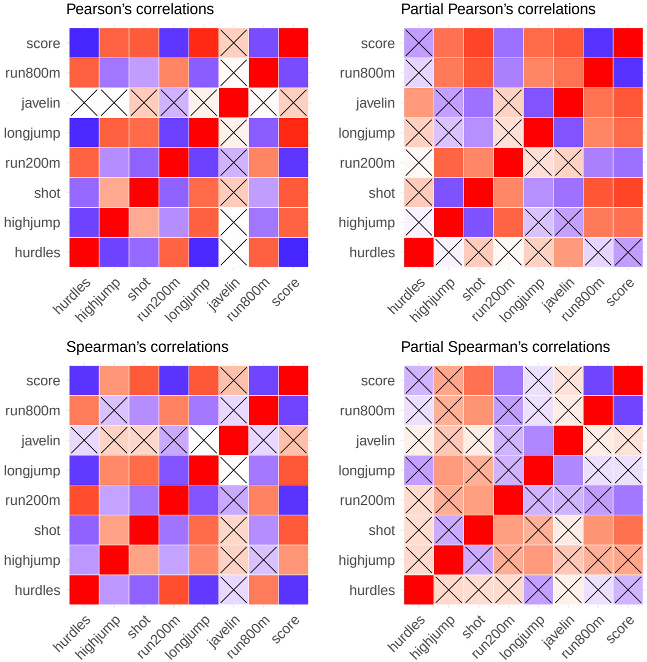
Pearson's correlations

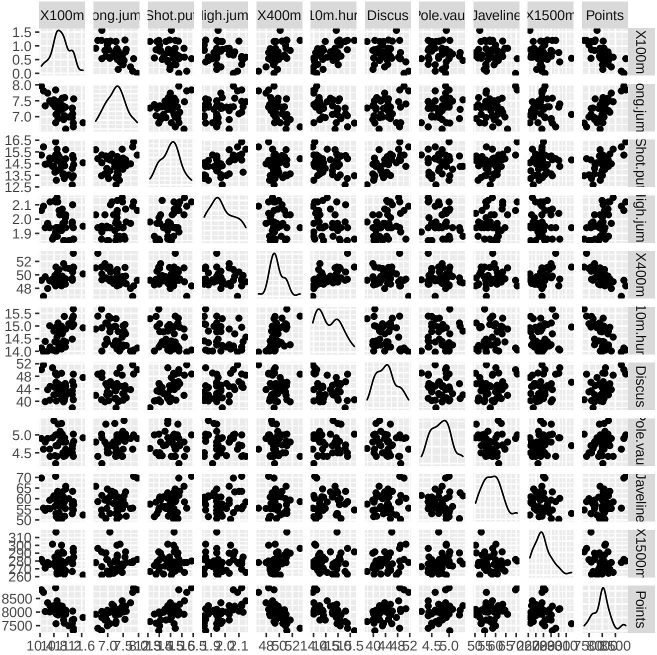
Partial Pearson's correlations

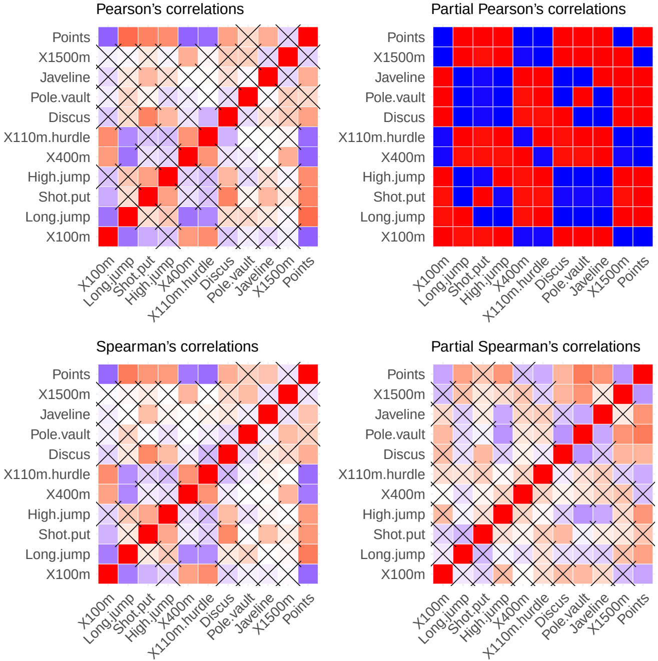
Spearman's correlations

Partial Spearman's correlations


hurdles
highjump
shot
run200m
longjump
javelin
run800m
score
0.8
0.6
0.4
0.2
0.0
1.8
1.7
1.6
1.5
16
14
12
10
26
25
24
23
7.0
6.5
6.0
5.5
5.0
47.5
45.0
42.5
40.0
37.5
160
150
140
130
7000
6500
6000
5500
5000
4500
13 14 15 16
1.5 1.6 1.7 1.8
10 12 14 16
23 24 25 26
5.5 5.5 6.6 6.7 7.0
37 40 42 45 47.5
130 140 150 160
4500 5000 5500 6000 6500 7000


Pearson's correlations
Partial Pearson's correlations
Spearman's correlations
Partial Spearman's correlations
score
run800m
javelin
longjump
run200m
shot
highjump
hurdles
hurdles
highjump
shot
run200m
longjump
javelin
run800m
score


X100m
ong.jum
Shot.pu
ligh.jum
X400m
10m.hur
Discus
Pole.vau
Javeline
X1500m
Points
1.5
1.0
0.5
0.0
8.0
7.5
7.0
16.5
15.5
14.5
13.5
12.5
2.1
2.0
1.9
52
50
48
15.5
15.0
14.5
14.0
52
48
44
40
5.0
4.5
70
65
60
55
50
310
300
290
280
270
260
8500
8000
7500
7.0
7.5
1.9
2.0
2.1
48
50
52
40
44
48
52
4.5
5.0
7500
8000
8500


Pearson’s correlations
Partial Pearson’s correlations
Spearman’s correlations
Partial Spearman’s correlations
Points
X1500m
Javeline
Pole.vault
Discus
X110m.hurdle
X400m
High.jump
Shot.put
Long.jump
X100m
X100m
Long.jump
Shot.put
High.jump
X400m
X110m.hurdle
Discus
Pole.vault
Javeline
X1500m
Points

Statistics II | seminar 7

Autocorrelation, multicollinearity, stepwise regression

Markéta Zoubková, Ondřej Pokora

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

24 October 2022

Example (1)

Data file Hawaii.csv contains a certain bird's species numbers on Hawaii between years 1956 and 2003. Using the linear regression model with AR(1), model the dependence of the square root of the number of birds on time (in years). Check for autoregression of order 1 in the data and in case of finding it, remove it.

Example (2)

Data file expenses.csv: data collected from 20 randomly selected households. Columns in the table contains the following variables: food and drink expenses (Y), number of members (X_1), number of children (X_2), average age of workers (X_3) a household income (X_4). Select model with the most well-conditioned regressors using stepwise regression.

Example (3)

Data file water.csv: The water losses in distribution networks were determined in the years 1953–1983. Model dependece of the water loss (Y) on the produced amount of water (x). Check for autoregression of order 1 in the data and in case of finding it, remove it.

Example (4)

Data stored in the variable mtcars in *R* are for modeling the dependence of passenger car fuel consumption (mpg, number of miles per gallons) on the engine characteristics:

cyl	number of cylinders
disp	cylinders' volume (cubic inches)
hp	engine power (horsepower)
drat	differential ratio
wt	vehicle weight (in hundreds lbs)
qsec	acceleration (seconds from 0 to 1/4 mile)
vs	cylinder arrangement (1 = 'V'-type, 0 = inline)
am	gearbox (0 = automaict, 1 = manual)
gear	number of gears
carb	number of carburetors

Test for multicollinearity in the model. Select a suitable model using stepwise regression. Check for normality of residuals.

Example (5)

Data file cement.csv: chemical composition of Portland cement:

y	heat of hydration in calories per 1 gram of cement
x1	*tricalcium aluminate* $3\,\mathrm{CaO.Al_2O_3}$ in %
x2	*tricalciam silicate* $3\,\mathrm{CaO.SiO_2}$ in %
x3	*tetracalcium alumino ferrite* $4\,\mathrm{CaO.Al_2O_3.Fe_2O_3}$ in %
x4	*dicalcium silicate* $2\,\mathrm{CaO.SiO_2}$ in %

Test for multicollinearity in a linear model on dependence of the heat of hydration on content of 4 given cement components. Select a suitable model for y using stepwise regression. After that, check for normality of residuals.

Example (6)

Macroeconomic data of the USA from 1947–1962 are stored in variable longley (built-in in v *R*). Using a linear regression model, examine the dependence of employment on GDP and the population older than 14 years: Employed ~ GNP + Population. Consider AR(1) in the data.

Example (Continuation of Examples 4–6 from Seminar 6)

Using stepwise procedure, build suitable linear models in Examples from **Seminar 6**. Model the following target variables using the other variables as predictors:

4. `ROLL`
5. `score`
6. `Points`

Results

2. Multicollinearity is not rejected, suitable model: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$, normality of residuals is not rejected.
4. Multicollinearity is not rejected. Suitable models are e.g. $\texttt{mpg} = \beta_0 + \beta_1\texttt{wt} + \beta_2\texttt{qsec} + \beta_3\texttt{am}$, $\texttt{mpg} = \beta_0 + \beta_1\texttt{wt} + \beta_2\texttt{cyl} + \beta_3\texttt{hp}$, normality of residuals is not rejected.
5. Multicollinearity is not rejected, suitable model: $\texttt{y} = \beta_0 + \beta_1\texttt{x1} + \beta_2\texttt{x2} + \beta_4\texttt{x4}$, normality of residuals is not rejected.

Statistics II | seminar 8

Principal component analysis

Markéta Zoubková, Ondřej Pokora

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

31. 10. 2022

Perform correlation analysis and principal component analysis in the following examples. Draw a correlogram, screeplot, proportion of variance explained and biplot. Identify principal components (first 2).

How many principal components are necessary to consider to explain at least 80 % of variance? What percentage of the data variance is explained by the first 2 principal components?

Example (1)

Data file `indicators.csv`: 16 economic and financial indicators for a set of 29 European countries from the year 2007 that was published by EUROSTAT in the year 2009.

Example (2)

Data file `pollution.csv`: USA climate and demographic data from the year 1960. Use the first 15 variables (do not consider variable `MORT`).

Example (3)

Data file `pain.csv`: 121 patients from the year 1991 complaining about a headache, responses are on a scale of 1 to 9. Do not consider variables `isub` and `dayslost`.

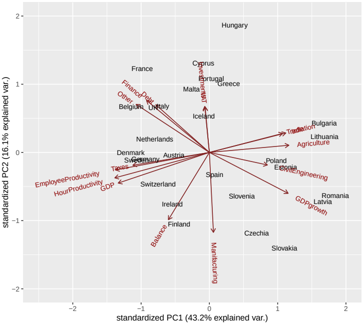
Example (4)

Data file heptathlon.csv: 25 heptathletes' results from 7 disciplines and total score in one athletic competition. Do not consider variable score. Find coefficients for the heptathlete Lajbnerová in the first two principal components.

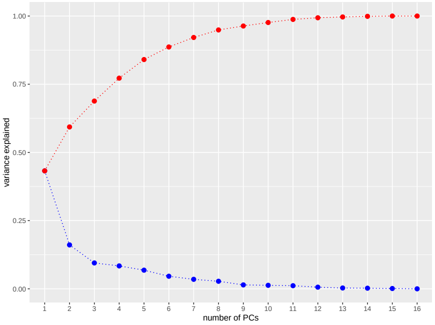
Example (5)

Data file decathlon.csv: results from individual disciplines, ranking and total score for 41 decathletes in one athletic competition. Use only ten variables of the results from individual disciplines. Find coefficients for the decathlete Šebrle in the first two principal components.

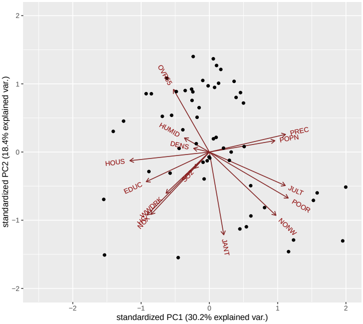
Results



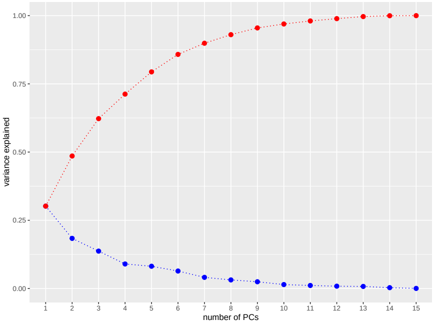
Hungary
France
Cyprus
Investment
Portugal
Greece
Malta
VAT
Finance
Debt
Other
Belgium
UK
Italy
Iceland
Bulgaria
Trade
Inflation
Lithuania
Agriculture
Netherlands
Denmark
Sweden
Germany
Austria
Poland
Estonia
CivilEngineering
Taxes
EmployeeProductivity
HourProductivity
GDP
Spain
Switzerland
Slovenia
Romania
Latvia
GDPgrowth
Ireland
Finland
Balance
Czechia
Slovakia
Manifacturing
2
1
0
−1
−2
−2
−1
0
1
2
standardized PC2 (16.1% explained var.)
standardized PC1 (43.2% explained var.)



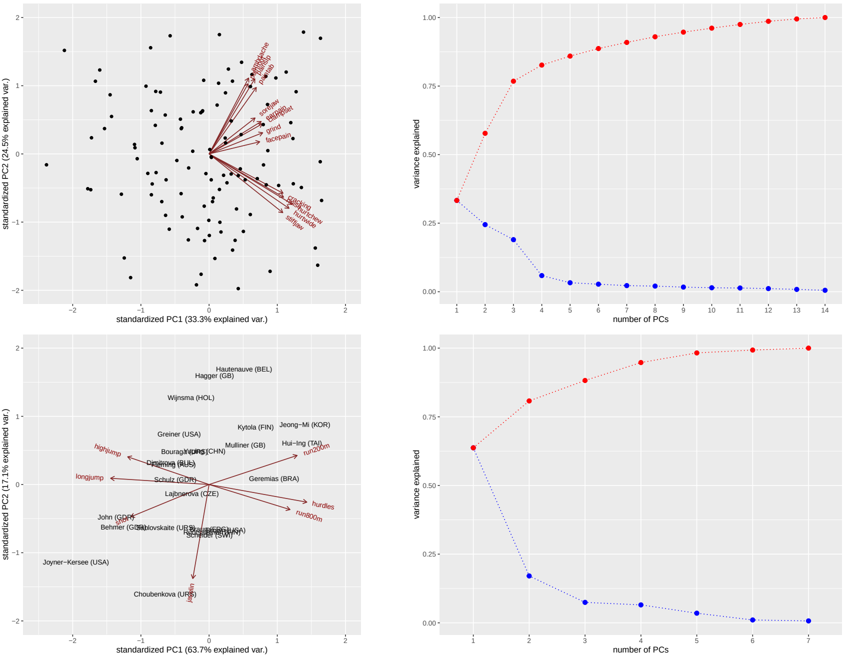
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
variance explained
number of PCs

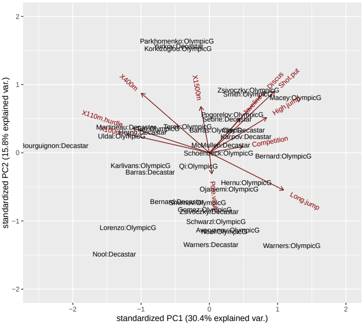


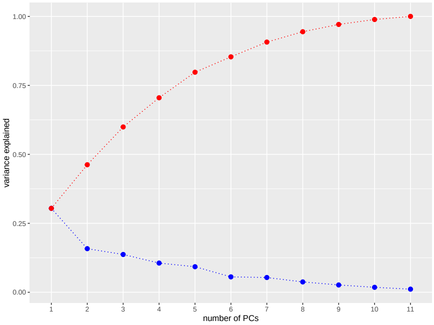
OVR65
HUMID
DENS
HOUS
EDUC
SO2
HC
WWDRK
NOX
PREC
POPN
JULT
POOR
NONW
JANT
2
1
0
−1
−2
−2
−1
0
1
2
standardized PC2 (18.4% explained var.)
standardized PC1 (30.2% explained var.)



1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
variance explained
number of PCs


standardized PC2 (24.5% explained var.)
standardized PC1 (33.3% explained var.)
2
1
0
−1
−2
−2
−1
0
1
2
ambdache
painup
painslp
paintab
sorejaw
earpain
clampset
grind
facepain
cracking
noise
hurtchew
hurtwide
stiffjaw
variance explained
number of PCs
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
standardized PC2 (17.1% explained var.)
standardized PC1 (63.7% explained var.)
Hautenauve (BEL)
Hagger (GB)
Wijnsma (HOL)
Kytola (FIN)
Jeong−Mi (KOR)
Greiner (USA)
Mulliner (GB)
Hui−Ing (TAI)
Bouraga (URS)
Yuping (CHN)
Dimitrova (BUL)
Fleming (AUS)
Schulz (GDR)
Geremias (BRA)
Lajbnerova (CZE)
John (GDR)
Behmer (GDR)
Sablovskaite (URS)
Braun (FRG)
Scheider (SWI)
Joyner−Kersee (USA)
Choubenkova (URS)
highjump
longjump
run200m
hurdles
run800m
shot
javelin
variance explained
number of PCs
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7





Coefficients in the first two principal components:

	PC1	PC2
Lajbnerova	-0.5302507	-0.1463219
Sebrle:Decastar	0.4795821	0.6534164
Sebrle:OlympicG	4.039407	1.333305

Statistics II | seminar 9

Logistic regression and other generalized linear models (GLM)

Markéta Zoubková, Ondřej Pokora

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

6. 11. 2022

Example (1)

Data file beetle.csv: mortality of the *confused flour beetle* (*Tribolium confusum*) due to exposure to gaseous carbon disulfide CS_2. Model the dependence of mortality on the amount of CS_2 using the (i) logistic regression, (ii) probit link function and (iii) CLogLog link function. Examine statistical significance of variables, plot final regression functions, compare models.

`dose`	the dose of CS_2 (in mg/l)
`population`	number of beetles tested
`killed`	number of beetles killed

Example (2)

Data file `car_income.csv`: information about a new car purchase during last 12 months in relation to the household income and the age of a previous car. (i) Plot the dependence of `purchase` on other variables and find a suitable logistic model. Is every variable statistically significant? (ii) Create a suitable factor from variable `age`, use it in the model and look at its statistical significance.

`purchase`	indicator of a new car purchase (1 = yes, 0 = no)
`income`	annual household income (in thousands dollars)
`age`	age of a previous car (years)

Example (3)

Data file `heart.csv`: presence of myocardial infarction in relation to the age of a patient. Model the dependence of `chd` on `age` using the (i) logistic regression, (ii) probit link function and (iii) CLogLog link function. A suitable model can be obtained by aggregating the data by (`age`). Examine statistical significance of variables and plot final regression function.

`age`	age of patient (years)
`chd`	indicator of myocardial infarction (1 = yes, 0 = no)

Example (4)

Data file hospital.csv: patient recovery in relation to the infection severity and the hospital. Find a suitable logistic model for variable Treatment_Outcome.

Infection_Severity	severity of infection
Treatment_Outcome	indicator of recovery (1 = recovered, 0 = death)
Hospital	code of the hospital (1, 2, 3)

Example (5)

Data file cancer.csv: the number of skin cancer cases in women in relation to the age and the city in the USA where the patients lived. Find a suitable logistic regression model for the variable Cases. Plot the results. Calculate probability of the disease occurence for 60 years old woman living in Minneapolis and the woman of the same age living in Dallas.

Cases	number of cases
Town	city (0 = Minneapolis, Minnesota, 1 = Dallas, Texas)
Age	age category
Population	total number of women of a given age category in the particular city

[Minneapolis: 0.00117, Dallas: 0.00276.]

Example (6)

Data file sharks.csv: the number of shark attacks in Florida between years 1946 and 1999. There are the following variables:

Year	year
Population	population size
Attacks	number of shark attacks
Fatalities	number of deaths by shark attack

Start with scatterplot of number of shark attacks per 1 million people depending on the time. Model number of shark attacks – use binomial and poissonian model with canonical link function. Consider a cubic polynomial for the design matrix for variable Year.

Draw predictions for both models with 95% confidence bands for regression function. Examine the problem with too high/low variance and redefine the model if necessary. Plot the results. Estimate the number of shark atacks (per 1 million people) in 2013 using the final model including 95% confidence interval.

Example (7)

Data file bees.csv. One of the important characteristics when investigating bee activity is the number of bees that leave the hive to work in the outside environment. The study dealt with the measurement of this variable during several sunny days depending on the time of day. Data set contains thee following variables:

`number`	number of bees that left the hive
`time`	time when the data was collected

Model the dependence of the number of bees that leave the hive on the time of day using the poissonian model with canonical link function. Consider a quadratic polynomial for the design matrix for variable `time`. Look for the *overdispersion* and alter the model if necessary.

Example (8)

Data file species.csv: 90 pastures of size 25×25 m were observed differing in biomass, soil pH and species richness (the number of plant species). It is known that species richness decreases with increasing biomass. The question remained whether the rate of decline is related to the soil pH. Therefore, each pasture was classified according to the pH value of soil to the three levels (low, mid and high) and 30 pastures for each level were selected for the experiment. Variable Biomass is the long-term average of June measurements of biomass values. Data set contains the following variables:

pH	pH value of soil (low, mid, high)
Biomass	amount of biomass
Species	number of plant species

Plot the dependence of variable species on other variables. Find suitable poissonian model with logarithmic, identity and square root link function. Which variables are statistically significant? Use non-significance to simplify the model. Plot final models. Estimate the number of plant species for Biomass$= 9$ and pH=mid using all three models.

Example (9)

Data file henharrier.csv: predator *hen harrier (Circus cyaneus)* hunting for a capercaillie depending on a capercaillie population. Let Y_i denote the proportion of capercaillie consumed and x_i the population of capercaillie in the given area. Theories dealing with the behavior of these predators suggest using this formula for modeling

$$E(Y_i) = \mu_i = \frac{\alpha x_i^3}{\delta + x_i^3},$$

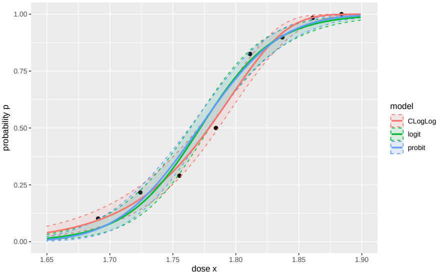
where Y_i is Gamma-distributed. It is necessary to estimate unknown parameters α and δ. Using the link function `inverse`, we get

$$\frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{\delta}{\alpha x_i^3}.$$

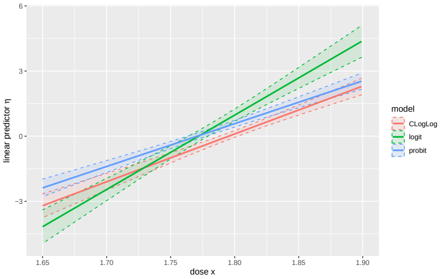
By defining the parameters $\beta_0 = 1/\alpha$ and $\beta_1 = \delta/\alpha$, we get a linear relationship

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 \frac{1}{x_i^3}.$$

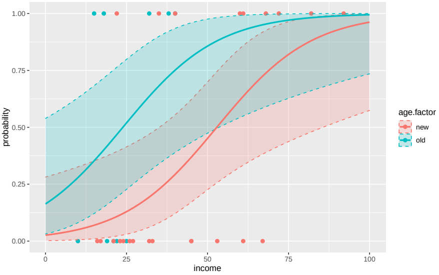
Results



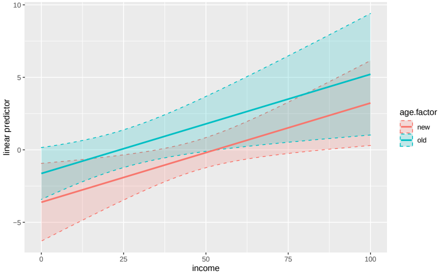
1.00
0.75
0.50
0.25
0.00
probability p
1.65
1.70
1.75
1.80
1.85
1.90
dose x
model
CLogLog
logit
probit



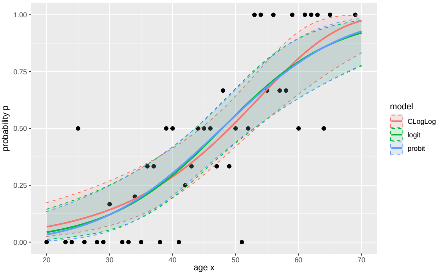
6
3
0
−3
linear predictor η
1.65
1.70
1.75
1.80
1.85
1.90
dose x
model
CLogLog
logit
probit



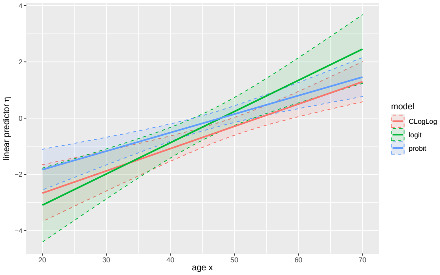
1.00
0.75
0.50
0.25
0.00
probability
0
25
50
75
100
income
age.factor
new
old



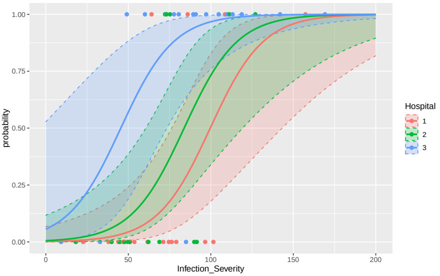
10
5
0
−5
linear predictor
0
25
50
75
100
income
age.factor
new
old



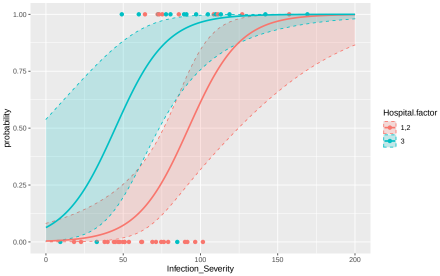
probability p
1.00
0.75
0.50
0.25
0.00
20
30
40
50
60
70
age x
model
CLogLog
logit
probit



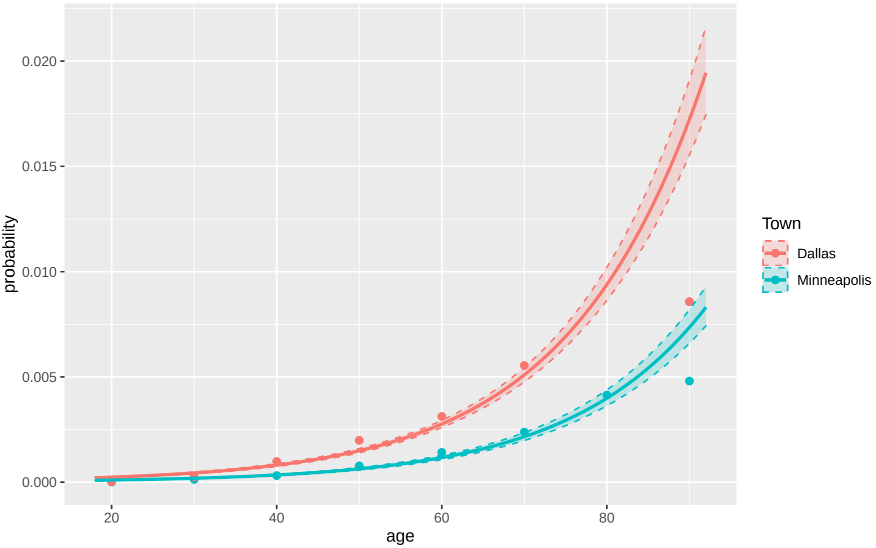
linear predictor η
4
2
0
−2
−4
20
30
40
50
60
70
age x
model
CLogLog
logit
probit

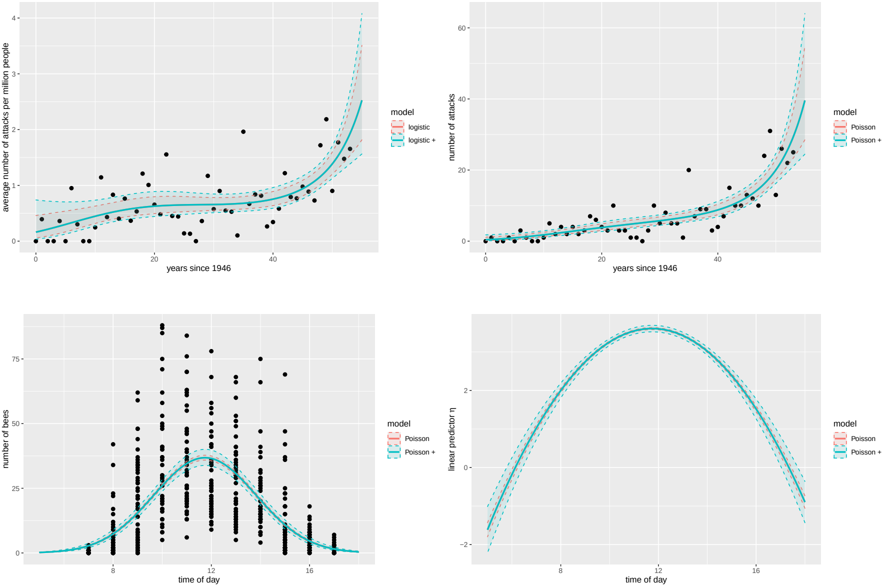


probability
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Infection_Severity
Hospital
1
2
3

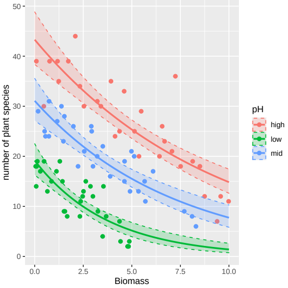


probability
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Infection_Severity
Hospital.factor
1,2
3

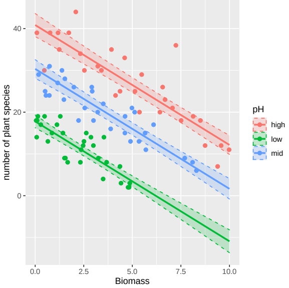

0.020
0.015
0.010
0.005
0.000
probability
20
40
60
80
age
Town
Dallas
Minneapolis



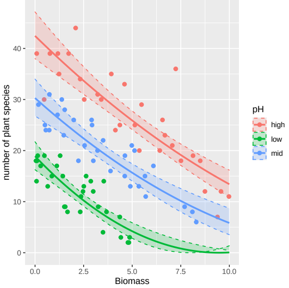
average number of attacks per million people
0
1
2
3
4
0
20
40
years since 1946
model
logistic
logistic +
number of attacks
0
20
40
60
0
20
40
years since 1946
model
Poisson
Poisson +
number of bees
0
25
50
75
8
12
16
time of day
model
Poisson
Poisson +
linear predictor η
−2
0
2
8
12
16
time of day
model
Poisson
Poisson +



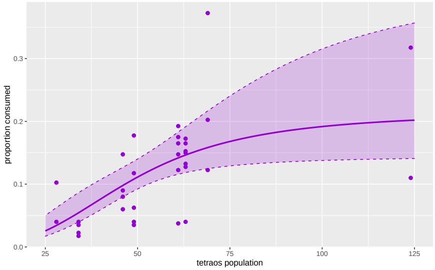
number of plant species
50
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
Biomass
pH
high
low
mid



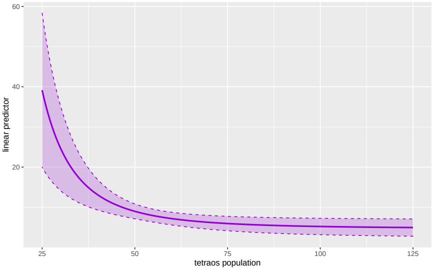
number of plant species
40
20
0
0.0
2.5
5.0
7.5
10.0
Biomass
pH
high
low
mid



number of plant species
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
Biomass
pH
high
low
mid



proportion consumed
0.3
0.2
0.1
0.0
25
50
75
100
125
tetraos population



linear predictor
60
40
20
25
50
75
100
125
tetraos population

2. *Hint*: age ≤ 3 and age > 3
6. *Overdispersion*. Estimation: 33.96 attacks per 1 million people, confidence interval: $[3.207; 359.55]$.
7. Residual deviance $4\,879.3$ is disproportionately higher than the number of degrees of freedom $501 \Rightarrow$ *overdispersion*.
8. The number of species estimates for logarithmic link: 8.895; identical link: 4.513; square root link: 7.414.