Maria Králová
Michal Paleček

Multivariate Statistical Analysis

Brno 2015

Inovace studia ekonomických disciplín v souladu s požadavky znalostní ekonomiky
CZ.1.07/2.2.00/28.0227

Contents

Chapter 1

Problems of ordinal and discrete quantitative data

There are a lot of dealing with ordinal data in social sciences e.g. taking a questionnaire with a discrete ordered scale. For example, we can match a five-value scale with values of "absolute disagreement - rather disagreement - neutral approach - rather agreement - absolute agreement" with *scores* (i.d. numerical representaion) of 1 - 2 - 3 - 4 - 5. Whether the distance among the scores represents reality, the scores can be e.g. 1 - 10 - 20 - 50 - 200. We have to look out if the further mentioned statistical methods result in the same or not if the scores has been 1 - 2 - 3 - 4 - 5 and 1 - 10 - 20 - 50 - 20 respectively. From the mathematical point of view, the smoother the scale (a lot of values), the better the results, since it is possible to use more statistical methods. But from the practice; it is quite hard to assure a sufficient number of respondents willing to respond within a smooth scale.

We take discrete quantitative data also in discretization of the original continuous data. E.g. values of "salary" can be categorized with intervals with interval means.

original values	up to 10 000	10001-17500	17501-25000	25001-35000	35001-50000	more than 50000
new values	5000	13 750	21 250	30 000	42 500	75 000

We can then decide if the interval means would be considered as ordinal value scores or whether we retain the quantitative character regarding the problem being researched.

1.1 Descriptive statistics

If there is a need to explore the data of **one** ordinal or a discrete quantitative value X at a glance, we can use:

• **METHODS FOR NOMINAL VALUES**

Most of the time we are talking about frequency tables and relative frequency tables. They are not so useful if X represents "a lot of" values.

• **QUANTILES**

Quantiles are useful even for skewed data. The indicator of "position" is median. Thanks to

lower quartile ($Q1$) and upper quartile ($Q3$) we can demonstrate inter quartile range (IQR) =$Q3 - Q1$ that is useful for the indication of variability.

• **MEAN AND STANDARD VARIATION**
Only under obezretnos can be used for ordinal and quantitative discrete data, if there is a reasonable interpretation. It is quite problematic if the data stems from skewed distribution.

•**BOX-PLOTS**
It is suitable to set up the box plots for ordinal data as:
- median as box plot center
- upper and quartile as box plot endings
- minimum and maximum as whiskers

We can further categorize the ordinal or quantitative discrete data box plot into number of groups (e.g. men and women etc.). Box plot is also useful when assessing distribution of two or more variables but they have to be **measured in the same scale**.

Example 1.1. In the *Film.sta* data, there are answers of 1322 respondents to the question: "How do you assess impact of current movies on the youth?" The answers were made up on 5 value scale of: The impact I see as
Very positive(1) - Positive(2) - Neutral(3) - Negative(4) - Very negative(5).
Characterize the data by suitable descriptive statistics.

Solution
To be figure out during seminar. □

Example 1.2. In the data *Household_Marriage.sta*, there are answers of 1346 respondents to the question:
Question no. 1: "How important is the youth establish their own home and not live alongside parents?"
Question no. 2: "How important is to get married?" There has been a 5 value scale possible: Very important(1) - Quite important(2) - To same extent important(3) - Not so important(4) - Absolutely not important(5).
Characterize the data by suitable descriptive statistics and compare the answers.

Solution
To be figure out during seminar. □

1.2 Two sample problems

We will be dealing with ordinal and quantitative discrete variable X in the two mutually independent groups and will be exploring whether these two groups are different regarding the X. So that, in the problem 1.1 1.1 we can research if the spectrum of opinions about the movie impact differs between men and women. The problem to be figured out is **srovnání two independent samples**. We can alternatively formulate the problem through detection of independence between ordinal or quantitative discrete variable X and dichotomic nominal variable Y. So that, in the problem 1.1 1.1 we are researching if the variable "opinion about movie impact" and "sex" variable are associated. We are going to analyze **dependence of two variables**.

• TWO SAMPLE t-TEST SUITABILITY

If we want to test hypothesis by the t-test:

H_0 : The groups do not differ regarding controlled variable X
H_1 : The groups differ regarding controlled variable X,

there is a need of assumption that bot samples stem from normal distribution. This assumption is not met for ordinal or discrete data **always** and theoretically the two-sample t-test cannot be applied. From the practical point of view, we can use it with the reference to the central limit theorem if the number of cases is "sufficiently" large. Two-sample t-test does not need the normality; sufficient is if the sample means are normal, and with increasing number of cases the distribution of both means is nearing the normal distribution. The more the histograms of both groups differ from normal Gauss curve, the more cases are needed. The means of symetric distributions are going to converge to

• SUITABILITY OF THE WILCOXON RANK SUM TEST USING (EQUAL WITH THE MANN-WHITNEY TEST)

If we want to decide about non-dependence by the the Wilcoxon rank sum test:

H_0 : The variable distribution X is the same in both groups.
H_1 : The variable distribution X is not the same in both groups,

there is a need of assumption that both samples stem from continuous distributions. This assumption is not met for ordinal or discrete data **always** and theoretically the Wilcoxon rank sum test should not be used. From the practical point of view, this assumption can be neglected if the scale is "sufficiently" large. (The number of values that the X variable operates under must not be less than 4.) "The trick" how to make the scale more smooth despite being clumsy for respondents is this: we ask by let's say 5 different ways in the questionnaire on the same question and all the 5 "different" questions take the 5 value scale form. Subsequently, we add up the respondent's answers from these 5 questions. Now, we have got the smoother 21 value scale with the minimum of 5 and maximum of 25. The Wilcoxon rank sum test is good at revealing the differences between distributions of the two groups, mainly if the distributions differ only in shifting. If the histograms of the X variable are of the same or similar shape, we can set the hypothesis as:

H_0 : Medians of the X variable are the same in the both groups
H_1 : Medians of the X variable are not the same in both groups.

• SUITABILITY OF χ^2 TEST OF INDEPENDENCE

Were the ranges of the sample to be "small" or the scale is made of "few" values, then the t-test or the Wilcoxon rank sum test is not suitable. In this case it is recommended to neglect the ordinality and the data take as nominal. Subsequently we will turn the test of the agreement in the groups into the test of the independence between original X and the dichotomic factor Y and undergo the χ^2 test of independence in the contingency table.

Example 1.3. For the problem 1.1 1.1 decide whether the opinion about the movie impact differs in terms of women and men. Compare and interpret p-values of the all tests. In terms of t-test notice the confidence interval for the population mean difference.

Solution

To be figure out during seminar. □

1.3 k-sample problems

This section makes the section 1.2 more general for the $k \geq 3$ groups. Analogically, instead of the two sample t-test here comes ANOVA, instead of the Wilcoxon rank sum test there will be Kruskal-Wallis test, and concerning χ^2 test of the independence the factor Y is going to take k values.

Example 1.4. For the figure 1.2 1.2 consider next question (variable X). Divide the respondents into groups by the nominal variable Y representing education that takes following values: less than high school - high school or junior college - bachelor - graduate. The data are in *Household_Marriage.sta* file
Figure out whether the opinion spectrum about own domestic distribution differs in terms of particular groups with education. Compare and interpret p-values of all possible tests. As far as ANOVA is concerned, keep going with Tukey multi-comparing.

Solution
To be figure out during seminar. □

1.4 Paired data dependence

In this section, we will be dealing with two ordinal or quantitative discrete variables X a Y and ways to measuring their association. The sample $\binom{X}{Y}$ is two-dimensional, so that we are researching the values of X and Y in terms of all respondents and we are wondering about their association. So that, in the problem 1.2 1.2 we are researching whether (an how much) the opinion about own domestic importance and marriage importance. There lots of measures of association for ordinal data. The classic Pearson's correlation coefficient is absolutely unsuitable for ordinal and quantitative discrete data. This correlation coefficient is suitable only for linear association measurement and assumes the sample from two-sample normal distribution. The suitable measures of association between two ordinal or quantitative discrete variables are:

1. Spearm's ρ
2. Goodman a Kruskal's γ
3. Kendal's τ
4. Somers's d

All the upper-mentioned measures take the values between -1 and 1 with the interpretation similar to the Pearson's correlation coefficient. The nearer the 0, the association is weaker; and the nearer the 1 or -1, the association is stronger. The negative values represent "indirect" assotiation - with increasing values of scores of one variable the values of the second variable are decreasing. The positive values represent "direct" assotiation - with increasing values of scores of one variable the values of the second variable are also increasing. •**SPEARMAN'S** ρ
Details in the 11th lecture

- The figure r_S (r_S is sample estimation of ρ) is derived from the Pearson co-relation coefficient in the way the original values are replaced by their sequence.
- That's why the r_S is independent on absolute values of scores and the distance among them.
- Spearman's coefficient r_S is symmetric measure and do not differ the explained and explaining variable.

Thanks to r_S of the Spearman's ρ we can test hypotheses
$H_0: \ \rho = 0$ resp. $\rho \leq 0, \ \rho \geq 0$
$H_1: \ \rho \neq 0$ resp. $\rho > 0, \ \rho < 0$
Before we introduce next association measures, new terms *concordants* and *discordants* need to be launched. Imagine data of n respondents where every respondent answered two questions X and Y (we have scores for X and Y). Now we are going to pair that respondents. The number of pairs is $n(n-1)/2$.
Five situations are possible (indices i and j represent ith and jth respondent; $i \neq j$):

1. All the variables take the upper scores for one respondent than for the second respondent. These pairs we label as *concordant*. Obviously, when most of the pair are concordant, than X and Y associate "directly". The number of concordant pairs we tag by the C statistics.
2. In terms of variable X, there is a larger score value for the first respondent and score value of the variable Y is larger for the second respondent. These pair are called *discordant*. Obviously, when majority of pair are discordant, there is a negative association. We can sign the number of discordant pairs as D.
3. There are the same scores for both X and Y in terms of the both respondents. We can sign the number of such pairs as T_{XY}.
4. There are the same scores for X variable from the two respondents but scores for Y differ. We can sign number of such pairs as T_X.
5. There are the same scores for Y variable from the two respondents but scores for X differ. We can sign number of such pairs as T_Y.

By definition, the number of all pairs is $n(n-1)/2 = C + D + T_{XY} + T_X + T_Y$
•GOODMAN'S A KRUSKAL'S γ
The value of γ (we can sign the estimate of gamma as g) is derived from the probability difference of concordant and discordant pairs.

- The formula for g estimate, γ, is based on proportions of concordant and discordant pairs of respondents:
  $g = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}$.
- $g \in \langle -1, 1 \rangle$, and if all the pairs are concordant $g = 1$; and if all the pairs are discordant $g = -1$.
- The value of g does not depend on absolute values of scores, even on distance among them.
- g is a symetric measurement, it does not differ explaining and explained variable.

Since the Goodman's and Kruskal's γ are only based on concordant and discordant pairs, the association of X and Y variables is upper-estimated by the game. There are also asymptotic tests of independence for γ.

•SOMERS'S d
The construction is similar to γ, however, even equalities are concerned. d is featured in three types depending on which equal pairs is based on.

- Asymmetric coefficient $d(Y|X) = \frac{C-D}{C+D+T_Y}$ only includes pairs that are unequal in X. It is suitable when we are wondering if Y is dependent on X. (Y is explained variable and X is explaining one. When X changes, what makes Y?)

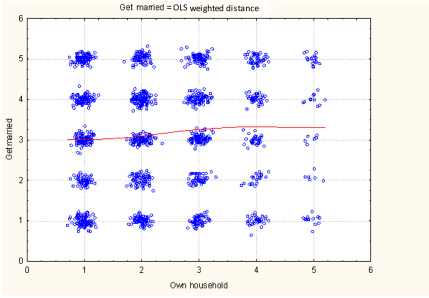
- Asymmetric coefficient $d(X|Y) = \frac{C-D}{C+D+T_X}$ includes only pairs without equality in Y. It is suitable when we are wondering about dependence X on Y.
- Symmetric coefficient $d = \frac{C-D}{C+D+(T_X+T_Y)/2}$
- $d \in \langle -1, 1 \rangle$; In terms of asymmetry, the coefficient can take the form of 1 or -1 only if the explained variable contains the same or more values than the explaining one.
- The value d is independent both on absolute values of scores and distances among.

•**KENDALL'S τ**
Kendall's τ is known in three versions:

- Coefficient $\tau_b = \frac{C-D}{\sqrt{(C+D+T_X)(C+D+T_Y)}}$
- τ_b is symmetric, it does not differ explaining and explained variable.
- $\tau_b \in \langle -1, 1 \rangle$; coefficient takes values of 1 or -1) only if not-zero frequencies on the major (or minor) diagonal of squared contingency table are the same. So that, this coefficient is suitable for squared tables if both variables X and Y are of the same nember of values.
- In the contingency table, there is a relation between Cramer's V and Kendall's τ_b: $|\tau_b| = V$
- The value of τ_b does not depend on absolute values of score or distances among.

Example 1.5. In problem 1.2 1.2 demonstrate all upper-mentioned associations between variables "opinion about own home importance" and "opinion about marriage importance". Test the independence by the Spearman's rank correlation coefficient. Regarding r_S value consider the significance of the test and its application. Compare the p-value of this test with the p-value of the χ^2 test.

Solution
To be figure out during seminar.



□

1.5 Equality of distributions within paired data

In this and past section 1.4 we are having paired data stemming from a two-dimensional sample $\binom{X}{Y}$ with n cases, where X and Y are ordinal or quantitative discrete variables. Unlike the past section, where we were researching whether the X and Y are associated, in this section we will be researching whether the distribution of these two variables differs. There is an assumption of the **same scale**. The usual applications are: Respondents filled out a form (X variable). Subsequently, some intervention came to fruition (respondents underwent training

or operation, time passes etc.) and after that the same respondents filled out a questionnaire of the same scale (Y variable). The question is apparent. Has the situation changed? From the statistical point of view the question is if the distributions of X and Y variables are different. We are measuring the same by two different methods and we wonder about if the both methods are equivalent. (For example, the same student is examined by two professors - do the professors examine alike? The patient's condition is assessed by two physicians - are the findings the same?) In the problem 1.2 1.2, we can ask if the opinions about own home living importance and about marriage importance differ - so if the distributions of the answer depend on the question. For better orientation withing the data, we can use descriptive statistics for both variables (attention for mean; rather median and spread between quartiles), histograms and box plots. These descriptive statistics are useful but they fail to deal with "the pairs". The important piece of information is hidden in the differences in the respondents' answers. By dealing with differences in the respondents' answers only we "erase" the impact of a particular respondent - his personal "features" that we do not care about. We care about "intervences". So that, we set up a sample $Z_i = X_i - Y_i,\ i = 1, \ldots, n$. If there is no difference in X and Y variable distribution, the majority of Z scores should be of zero value or nearing zero

• **SUITABILITY OF t-TEST**

In case we want to decide about a hypothesis using a pair t-test:

H_0 : The distribution of X and Y variable does not differ
H_1 : the distribution of X and Y variable does differ,

there is a need of an assumption that the two-dimensional sample $\binom{X}{Y}$ stems from two-dimensional normal distribution. So that, $Z = X - Y$ stems from one-dimensional normal distribution and we can rewrite the hypothesis as:
H_0 : the expected value of Z variable is zero
H_1 : the expected value of Z variable is not zero.

However, the assumption of the two-dimensional normality is not met for ordinal or discrete data always and theoretically the paired t-test should not be used. From the practical point of view, we can use it with the reference to the central limit theorem if the number of cases is "sufficiently" large, that means the more the histogram of Z variable differs from the normal Gauss curve, the more cases are needed.

• **SUITABILITY OF THE WILCOXON SIGNED RANK TEST FOR PAIRED DIFFERENCE**

Should we want to test a hypothesis that distributions of X and Y do not differ with a Wilxcox pair test properly, the hypothesis must take this form:

H_0 : The distribution of $Z = X - Y$ variable is symmetric around zero
H_1 : The distribution of $Z = X - Y$ variable is not symmetric around zero.

In order to use this test properly, two assumptions should be met:

1. The sample of Z variable should stem from continuous distribution.
2. This distribution should be symmetric around a particular value.

ad 1. As in the section 1.2 1.2, the assumption of continuous distribution is not met for ordinal or discrete data always and theoretically the Wilcoxon signed rank pair test should not be used. From the practical point of view, we can that ignore if the scale is "sufficiently" large.

ad 2. Whether this assumption is not met, the null hypothesis might be declined not for different distributions of X and Y variables but thanks to skewed distribution of the Z value. If the distribution of the Z is symmetric around a value, this value must be the median in case H_0 and the hypothesis can be formulated as:
H_0 : median of Z variable is zero
H_1 : median of Z variables is not zero.

• **SUITABILITY OF THE TEST FOR NOMINAL VARIABLES**

If the two-dimensional sample has "few" cases or the scale contains "few" values, the pair t-test or the Wilcoxon signed rank pair test is not suitable. In this case, it is recommended to ignore the ordinality and "descend" with the data as being nominal.

Chapter 2

Multivariate analysis introduction

We are talking about multivariate methods when all the n objects are measured by $p \geq 2$ variables. These data take usually matrix form $n \times p$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix},$$

where lines are particular cases and columns are particular variables. Multivariate methods are usually exploratory and their mission is generating new hypotheses. Mainly, they help us to comprehend associations among variables and among individuals.

2.1 Main objectives of multivariate methods

• REDUCTION OF MULTI-DIMENSIONS
When two or more variables are correlated, we can replace them by a common variable (factor). That leads to a reduction in dimension of the original p-dimensions and thus better interpretation of the data. When we can reduce the number of variables even to the dimension of 2 or 3, we can visualize the data in 2D or 3D, which is better for orientation in the data. There are number of possibilities of the multidimensional data visualization that we will be talking about further.

• DATA STRUCTURE REVELATION
Are there groups of similar objects in the data? On the basis of which variables they are possible to differentiate? Is is possible to take an object into a group?

• DEVIATE OBJECTS DETECTION
Should we are having 2-3 variables, we can use our intuition for the deviations detection. When $p > 3$ that is tricky. Multivariate statistical methods help us to find out deviations and to ascertain which variables causes the deviations.

2.2 Data visualization

It is better for us to "scan" pictures than tables or numbers because of evolution reasons. By "graphics" we can better comprehend even complex associations or problems concerned. So that, it is suitable to visualize the data not only for findings enrichment but also as an essential for analysis. If we are having two or three variables, we can depict the data geometrically. If there are more variables, there are number of possibilities:

• **DEPICTION OF VARIABLES IN PAIRS**
We usually use:
- *matrix of plots*,
  that is a table of the plots between variables.
- *correlation matrix* of variables,
  that is a table of correlation coefficients between variables. Since the correlation coefficient is dimensionless, the table is not dependent on a scale. It is very useful in the beginning of some explorations of some multivariate methods.
- *matrix of covariances* of variables,
  that is a table of covariances between variables. Since the covariance is dependent on quantity, the values in this tables are dependent on the scale and the linear dependence is not apparent at the first sight. Also, it is suitable in the beginnings of some multivariate methods' explorations.

All the upper-mentioned matrices demonstrates only a piece of information from the the multidimensional data since they are always dealing with associations between two variables and not with the complexity.

• **MULTIDIMENSIONAL VISUALIZATION**
E.g. Chernoff faces, starred graphs, case profiles etc..

Example 2.1. In the *Criminality.sta* data, there are values of 7 variables for particular states in the USA: ratio of violence over 100 000 inhabitants, ratio of murders,..., percentage of families under the poverty line, percentage of incomplete families. Demonstrate the data using graphs.

Solution
To be figured out during seminar. □

2.3 multidimensional normality

Some multivariate methods assume multidimensional normality. Unlike the one-dimensional normality case, where there are a lot of tests, there is no perfect test for testing multidimensional normality. It is not sufficient when all the parts of the multidimensional normal distribution is normal but also all the linear combinations of them must be normal. Since we are using multivariete statitical methods for explorative tasks, we can take this as met and the findings with "reserve".

Chapter 3

Principal Component Analysis

Very often we are dealing with situation in which number of p variables is too high and too confusing for data processing and model construction. The main aim the Principal Component Analysis is to find a new system of k uncorrelated variables (even substitutes), that are possible to replace the original variables and demonstrate associations between original variables. These new variables are called components and sometimes they have also an interpretation. By new variables (components) it is more simple to identify deviance in multi-dimension. The method can also precede other methods for analysis needs (it can reveal deviant observations for ANOVA or regression, in terms of regression it can also reduce multi-collinearity) From the mathematical point of view it is a space transformation of the original variables to new variables with added requirements: 1) new system's axes are orthogonal; 2) axes are put into the direction of the "maximum possible variability". (Axes in directions of negligibly low variability can be deleted, so that the dimension can be reduced.) See figure 3.1

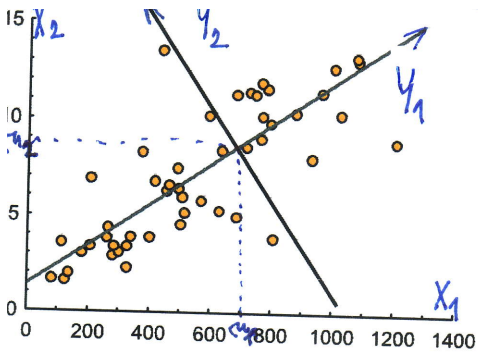


Figure 3.1: *The observed points can be demonstrated by the original coordinates X_1, X_2 or by new coordinates Y_1, Y_2. The highest variability is in the Y_1 direction. The orthogonal axis Y_2 in the direction of the highest variability.*

3.1 Principal Components

Let assume a random vector of original variables $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ with the vector of estimated values $E(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$ and variability matrix $var(\mathbf{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & & & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}$ with a rank of $r(\boldsymbol{\Sigma}) = p$.
The eigenvalues are signed as $\boldsymbol{\Sigma}$ sign $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ (and are ordered from highest to lowest and different from each other). The correspondent normalized eigenvectors are signed as $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_p}$ and are mutually uncorrelated. ("Normalized" means of length of 1, "uncorrelated" means that the vectors are orthogonal.

Eigenvectors p-dimensional, so that $\mathbf{v_r} = \begin{bmatrix} v_{r1} \\ v_{r2} \\ \vdots \\ v_{rp} \end{bmatrix}$ for $r = 1, \ldots, p$.

The variability matrix trace $tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \ldots + \sigma_p^2 = D(X_1) + \ldots + D(X_p)$ is called *total variability* of $\mathbf{X}$.

New variables $Y_1, Y_2, \ldots, Y_p$ are called *principal components*:

∘ **The first principal component** $Y_1 = \mathbf{v_1}' \mathbf{X} = v_{11}X_1 + v_{12}X_2 + \ldots + v_{1p}X_p$
- The new random variable Y_1 has arised by a linear combination of the all original variables; coefficients of the linear combination are the coefficients of the first eigenvector.
- The variability of the first component is equal to the first (the highest) eigenvalue; $\boxed{D(Y_1) = \lambda_1}$
- From the geometric point of view Y_1 is the vector of direction demonstrated by the original system of coordinates $X_1, \ldots, X_p$; it is the direction of the "highest possible" variability $\mathbf{X}$
- The vector length $\boldsymbol{v_1}$ is equal to 1.

∘ **the second principal component** $Y_2 = \mathbf{v_2}' \mathbf{X} = v_{21}X_1 + v_{22}X_2 + \ldots + v_{2p}X_p$
- The new random variable Y_2 arised from a linear combination of the all original variables; the coefficients of the linear combination are the parts of the second eigenvector.
- The variance of the second component is equal to the second eigenvalue. So that $\boxed{D(Y_2) = \lambda_2}$
- Y_1 a Y_2 are uncorrelated since the eigenvectors were uncorrelated.
- Geometrically: Y_1 and Y_2 are mutually orthogonal,Y_2 is the direction vector into the direction of the highest possible remaining variability $\mathbf{X}$ (part of the variability $\mathbf{X}$ has been depleted by the first component).
- The vector length $\boldsymbol{v_2}$ is one.

$\vdots$

∘ p**-th principal component** $Y_p = \mathbf{v_p}' \mathbf{X} = v_{p1}X_1 + v_{p2}X_2 + \ldots + v_{pp}X_p$

- analogically

□

Have a look on the total variability. The total variability of the original variables $X_1, \ldots, X_p$ is the same as the total variability of the new variables $Y_1, \ldots, Y_p$.
$(D(X_1) + \ldots + D(X_p) = Tr(\mathbf{\Sigma}) = \lambda_1 + \ldots + \lambda_p = D(Y_1) + \ldots + D(Y_p))$
Furthermore, in case of new variables $D(Y_1) > \ldots > D(Y_p)$.
All the p principal components explains total variability of the original variables without information loss. However, from the geometric point of view, we only rotate the original p-dimensional coordinate system to the new p-dimensional coordinate system. (The beginning of the new coordinate system is shifted to $[\mu_1, \mu_2, \ldots, \mu_p]$.) If we want to reduce the original dimension, we take only first k components. If we label all the components Y_j according to its significance as $\frac{\lambda_j}{tr(\mathbf{\Sigma})},\ j = 1, 2, \ldots, p$, the first k components explain $\sum_{j=1}^{k} \lambda_j / tr(\mathbf{\Sigma}) \cdot 100$ per cent of the original variablity. As far as visualization is concerned, it is reccomended to take first two or three components, which can be graphically demonstrated. We can also be interested in the ratio of the one variable X_j that is explained by the first $k \leq p$ components on the original variability. We call this ratio *j-th communality*. From the mathematical point of view it is $\sum_{r=1}^{k} [R(X_j, Y_r)]^2$.

3.2 Component Score

Whether we want to use the method for data set analysis where the vector $X_1, \ldots, X_p$ is observed on n objects, we need to convert every object onto values in the new coordinate system of the $k \leq p$ components.
In terms of i-th object we have observed the values

$$\mathbf{x}_i = x_{i1}, x_{i2}, \ldots x_{ip};\ i = 1, 2, \ldots, n$$

j-th coordinate in the new coordinate system for the i-th object is:

$$y_{ji} = \mathbf{v_j}' \mathbf{x}_i = v_{j1}x_{i1} + v_{j2}x_{i2} + \ldots + v_{jp}x_{ip}\ ,\ j = 1, \ldots, k; k \leq p$$

These values of particular objects in the new coordinate system are called *component score*.

3.3 Standardized Principal Components

Variability of particular components differs. It can be problem for an interpretation of the components since they are not "co-measurable". Since that we usually standardize the components in a way that all have the variability of the size one. We tag the r-th standardized component as Y_{rS}. Then for $r = 1, 2, \ldots, p$:

$$Y_{rS} = \frac{1}{\sqrt{\lambda_r}} \mathbf{v_r}' \mathbf{X} = \frac{1}{\sqrt{\lambda_r}} (v_{r1}X_1 + v_{r2}X_2 + \ldots + v_{rp}X_p) = \frac{1}{\sqrt{\lambda_r}} Y_r$$

3.4 Correlation between components and variables

The principal components "represents" mainly the variables that are correlated with the components mostly. That is why we are interested in the correlations. For the correlation of the

original X_j variable with the r-th component Y_r that goes:

$$R(X_j, Y_r) = \frac{\sqrt{\lambda_r} v_{rj}}{\sigma_j} = R(X_j, Y_{rS})$$

3.5 "Unmeasurable" original variables

If the original variables $X_1, X_2, \ldots, X_p$ are measured in different units or the are of different variability, then there is no good of measuring the principal components from the variability matrix $var(\mathbf{X}) = \mathbf{\Sigma}$. In that case we will be applying complete procedure on the standardized variable $Z_1, Z_2, \ldots, Z_p$, where $Z_i = \frac{X_i - \mu_i}{\sigma_i}$, $i = 1, 2, \ldots, p$. The variability matrix of the standardized variables is the same as the correlation matrix of the original variables, so that $var(\mathbf{Z}) = cor(\mathbf{X})$.

3.6 Application of principal components

We have been dealing with random variables $X_1, X_2, \ldots, X_p$ with known $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ so far. Were we are finding principal components and facing only random sample where we are discovering $X_1, X_2, \ldots, X_p$ values on n objects, then the unknown vector $\boldsymbol{\mu}$ is supposed to be replaced by the vector of sample means and the unknown matrix $\mathbf{\Sigma}$ is supposed to be replaced by the sample variance matrix. The eigenvalues and the correspondent eigenvectors are the consistent estimates of eigenvalues $\lambda_1, \ldots, \lambda_p$. PCA is useful only when original variables are correlated. That is why the sample covariance (correlation) matrix should not have zeros beyond main diagonal.

3.7 Number of principal components

We are facing a problem of optimization number of components. These are common methods solving:

- *scree plot*
  Graphical method, personal assessment of the scree plot appearance. The plot depicts the ranks of the descending eigenvalues of the sample covariance matrix. We tag k the order number of the last "acceptable" eigenvalue. The question to answer is "When do the stones stop rolling?"
- *stated required variability*
  We state the border of an acceptable variability in advance. We usually state approx. We usually require approx. 70-80%.

- *eigenvalues* > 1
  This can be used only during analysis of standardized data when the principal components are derived from the correlation matrix. k is then number of eigenvalues higher than 1.
- *Sufficient reproduction of the sample variance (correlation) matrix*
  It will be explained during the session about factor analysis.

Example 3.1. In the *Countries.sta* data set, there are data concerning employment percentage in particular sectors from 1979. Analyze the associations between employment in particular sectors using PCA. Assess the particular country differences in the economy.

Solution
To be figured out during seminar. □

Chapter 4

Factor analysis

Factor Analysis (FA) is another multivariate method aiming at reducing the original p variables. Unlike PCA that is doing its best at interpretation of the variance of the original variables, FA focuses on correlations. The idea behind FA is that dependence among original variables come from interactions of hidden factors that are considered as the ones that are considered as causes of mutually correlated original variables. Finding of these latent factors is the main aim of the factor analysis. However, the FA has a problem - it does not have a unique solution. The system of explaining factors can be of infinite number of suggestions and there is no algorithm for finding the good merge of the original variables and is very subjective. We define each original variable as a linear combination of the common factors plus one factor specific for the particular variable. The first problem is that we do not know how many the latent factors we are dealing with. So that, we state this number subjectively. Then we are going to find a matrix with the linear combinations. However, there are an infinity number of these matrices. The choice of the one is dependent on experience and subjective values. There is not exact algorithm.

4.1 Orthogonal model of the factor analysis

It is recommended to proceed the complete analysis on standardized data $z_1, \ldots, z_p$ since not very often the random variables $x_1, \ldots, x_p$ are in the same measurements and due to interpretation suitability. (From the last section we know that $var(\mathbf{z}) = cor(\mathbf{x})$ and we will be sign this matrix as $\mathbf{\Sigma}$ furthermore.) "Ortogonal" means that we are going to find system of factors that are mutually uncorrelated.

• **ORTHOGONAL MODEL SPECIFICATION:**

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & & & \\ a_{p1} & a_{p2} & \ldots & a_{pk} \end{bmatrix} \cdot \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

or

$$\underset{p \times 1}{\mathbf{z}} = \underset{p \times k}{\mathbf{A}} \cdot \underset{k \times 1}{\mathbf{f}} + \underset{p \times 1}{\boldsymbol{\varepsilon}}$$

$\mathbf{z}$ is a vector of the original standardized variables

$\mathbf{f}$ is a vector of new variables that we will call *common factors*
$\boldsymbol{\varepsilon}$ is a vector of the *specific factors* unique for particular variables
$\mathbf{A}$ is called *factor matrix*

• **MODEL ASSUMPTIONS:**

∗ Common factors $f_r,\ r = 1, 2, \ldots, k$ are mutually independent in the orthogonal model, each of them has the null expected value and unit variance. So that:
$E(f_r) = 0$
$D(f_r) = 1; C(f_r, f_s) = 0,\ r \neq s$
$var(\mathbf{f}) = \mathbf{I} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

∗ Specific factors $\varepsilon_j,\ j = 1, 2, \ldots, p$ are mutually independent, each of them has null expected value and a variance u_j^2 that is called *unicity*. So that:
$E(\varepsilon_j) = 0$
$D(\varepsilon_j) = u_j^2; C(\varepsilon_j, \varepsilon_i) = 0,\ j \neq i$
$var(\boldsymbol{\varepsilon}) = \begin{bmatrix} u_1^2 & & 0 \\ & \ddots & \\ 0 & & u_p^2 \end{bmatrix}$

∗ Specific and common factors are mutually independent:
$C(f_r, \varepsilon_j) = 0,\ r = 1, 2, \ldots, k;\ j = 1, 2, \ldots, p$

• **CONSEQUENCES AND COMMENTARY**

(i) Variance matrix of the original variables vector $z_1, \ldots, z_p$ can be written as:
$\boldsymbol{\Sigma} = var(\mathbf{z}) = AA' + var(\boldsymbol{\varepsilon})$
AA' matrix is called *reduced correlation matrix* and is different from the matrix $\boldsymbol{\Sigma}$ in the diagonal.

(ii) The unit variance of the each original variable z_j stems from two "phenomenons": the part of the variability can be explained by the common factors and is called j-th *communality* and we sign it h_j^2. The remaining variability stems from unicity, the variability of the specific facotr ε_j.

Thus $D(z_j) = 1 = h_j^2 + u_j^2$ and $h_j^2 = \sum_{r=1}^{k} a_{jr}^2$.

a_{jr}^2 demonstrates how the r-th factor contributes to the variability of the original variable z_j.

$h_j^2 = \sum_{r=1}^{k} a_{jr}^2$ demonstrates how the part of the variable z_j variability is explained by the k factors.

(iii) If we write down the original standardized variables by the linear combination of the factors, we get:
$C(z_i, z_j) = \sum_{r=1}^{k} a_{ir} a_{jr} = R(x_i, x_j)$

(iv) The last task is to derive how much the original variables correlates with the factors:
$C(z_j, f_r) = a_{jr} = R(x_j, f_r)$
So that, we can interpret the a_{jr} as a correlation between the j-th original variability and the r-th factor. The higher the a_{jr} value, the better is the explanation of the j-th original variable by the r-th factor.

Now we have the factor analysis model specified. The matrix $\mathbf{\Sigma}$ is known. We have to find the model parameters (parameter estimates) - all the cells of the A matrix and $var(\boldsymbol{\varepsilon})$ matrix. Now we will be dealing with the "good" choice of the factor system:

* methods of model parameters finding (finding of A and $var(\boldsymbol{\varepsilon})$ matrices' cells)
* choosing k
* factor rotation
* factor score set

4.2 Methods of model parameters finding

These methods are also called as *extraction factors methods* and their aim is to estimate the factor loadings (weights) $a_{ij}, i = 1, \dots p,\ j = 1, \dots k$ and the specific factors $\varepsilon_i, i = 1, \dots p$ on a basis of a random sample of the n objects. Matrix $\mathbf{Z}$ shaped $n \times p$ is going to have values measured on the n objects in p columns for $z_1, \dots, z_p$ variables. This matrix is observable. Then the $\mathbf{F}$ matrix is going to be of the $n \times k$ shape and in every row there are k unobservable values for the factors $f_1 \dots, f_k$. Matrix $\mathbf{E}$ shaped $n \times p$ is going to have p columns for the specific factors $\varepsilon_1, \dots \varepsilon_p$ and the unobservable values on the n objects.
we are finding $\mathbf{A}$ and $\mathbf{E}$ in order to suit the equation: $\mathbf{Z} = \mathbf{F} \cdot \mathbf{A}' + \mathbf{E}$.

4.3 Common factors number choice

Rarely can we know the number of factors from the problem nature. But more often we need to decide about the number of the factors.

- It is good to begin the factor analysis with the PCA and use the criteria from the PCA chapter (scree plot, eigenvalues, sample correlation matrices $\mathbf{R}$ that are higher than 1, factors that explains at least 70-80 per cent total variability cumulative).
- We can acquire the residual matrix when we subtract the estimate of the reduced matrix that has been calculated for the k factors from the sample correlation matrix $\mathbf{R}$. If there are *more* beyond-diagonal values in the residual matrix high (higher than 0.2), then it is good to increase the number of k. If the values are low, the common factors explains the associations in the $\mathbf{R}$ matrix very well and we can consider decresing the number of k.
- The number of factors should be $< p/2$ and $\geq$ number of eigenvalues higher than 1.

4.4 Factors rotation

We are able to rotate the geometric system in an infinity number of ways. Geometrically: we are rotating the system of k factors (coordinates) in the p-dimensional space and we remain the orthogonality intact. The goal is to rotate the system in the way that every rotated factor is now correlated with a small group of the original variables only. So that, some factor loadings are meximalized and the others are minimalized by the rotation. These rotated factors is now possible to better interpret. See picture 4.1. Most of the rotate algorithms are aiming to get the most loadings into +/-1 or 0. We usually use the *normalized varimax* rotate method.

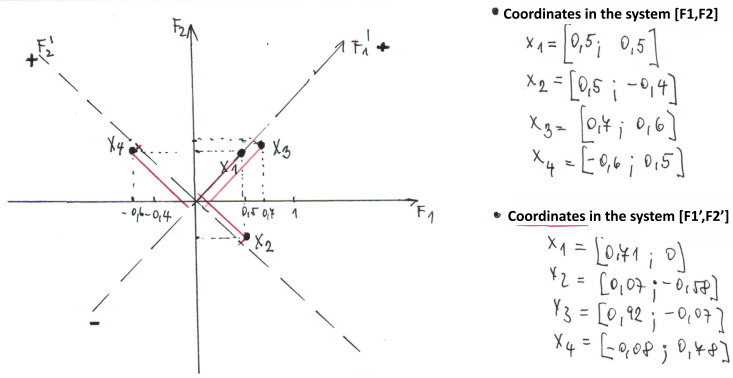


Figure 4.1: *The original 4-dimensional vector of the original variables $x_1, \ldots, x_4$ is now depicted in the 2-dimensional space and is represented by either f_1, f_2, or the pair of f_1', f_2'. f_1' "stands in for" mainly variables x_1, x_3 and f_2' stands in for variables x_2, x_4.*

4.5 Factor score determination

The main objective of the FA is to determine new variables - factors which interpret relations among original variables. Should we want to use the findings of the FA as a beginning of the further analysis, we need to estimate the values of the n objects in the new variables - common factors. These estimates are called a *factor score*. This is tricky since we are estimating values of the *unobservable* variables. What is more, there are more unobservable variables than the observable. Mostly, we use a weighted method of least squares and a regression method for solving.

Chapter 5

Canonical Correlation Analysis

Example 5.1. Imagine we are trying to find out if the happiness in the personal life and in work are associated through 3 questions regarding happiness in work (X_1: Are you satisfied with your boss? X_2: Are you satisfied with your colleagues?) and 7 questions regarding happiness in a personal life. The data are in the *Spokojenost.sta* file. □

So we are having two sets of variables and try to figure out whether they are associated. We are trying to find out something like a correlation coefficient but not between two variables, but between two sets. Whether we calculate only classical correlation coefficients between each pairs $R(X_i, Y_j)$, we would be given values concerning the pairs only and not between the sets. So that, we will be searching for a new pair of U and V that are going to "suitably" represent the two sets and then we find out the correlation between the U and V. But the question is what is the "appropriate" representation? The first thing you maybe come up with is that in each set, there can be a simple sum of X and Y variables. This would lead to an information loss and misrepresentation. Respondent that is having a good relationship with his wife but is not happy with money would not be different from the one who is not happy with his wife but is happy with money. So that, it is suitable to give some variables different weights. These weights will be found out in a way that the items from the both groups are correlated as much as possible.

5.1 Canonical Correlation Model

Let sign $X_1, X_2, \ldots, X_p$ the left side of the variables and $Y_1, Y_2, \ldots, Y_q$ the right side of the variables. Let assume that $p < q$.

• **MODEL SPECIFICATION:**

$$
\begin{array}{lll}
I. & a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p = U_1 & V_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q \\
II. & a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p = U_2 & V_2 = b_{21}Y_1 + b_{22}Y_2 + \ldots + b_{2q}Y_q \\
\vdots & \vdots & \vdots \\
p. & a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p = U_p & V_p = b_{p1}Y_1 + b_{p2}Y_2 + \ldots + b_{pq}Y_q
\end{array}
$$

terminology:

U_r	$\star$ r-th *Canonical variable* for the left side, $r = 1, \ldots, p$ is a linear combination of the original variables from the left set.
V_r	analogically for the right side, $r = 1, \ldots, p$
$a_{r1}, a_{r2}, \ldots, a_{rp}$	$\star$ *Canonical weights* (r-th canonical variable for the left side), $r = 1, \ldots, p$ Canonical weights are easy to interpret if the variables $X_1, \ldots, X_p$ are standardized; then the weights are contributions of the original variables from the left set to the rth canonical variable.
$b_{r1}, b_{r2}, \ldots, b_{rq}$	analogically for the right side
$\boxed{R_{(C)r} = R(U_r, V_r)}$	$\star$ r-tý *canonical correlation coefficient*, $r = 1, \ldots, p$ Demonstrate the rate of correlation between the rth pair of the canonical variables.
$R^2_{(C)r}$	demonstrates the co-variability.

• **MODEL ASSUMPTIONS AND COMMENTS:**

1. **Requirements for the first pair of the canonical variables**
   We are trying to find the $a_{11}, a_{12}, \ldots, a_{1p}$ and $b_{11}, b_{12}, \ldots, b_{1q}$ in a way that the correlation of U_1, V_1 variables is *maximal*.

This requirement help us to find weights only for the first pair of the canonical variables. Let have a look at the problem 5.1 again. We have both a question dealing with finance and relationship in both sets. Let assume that the "salary hapiness" has the biggest weight in the left set and "financial situation hapiness" in the right set (and data are standardized). It is apparent that the first pair of the canonical variables has been "very" influenced mainly by finance and the correlation coeficient does not represent the association within the sets. The thing is that the correlation coeficient has not captured all from the total common variability. So that, we are going to process next pair of the canonical variables whose correlation might capture even this. We can go further until we capture total variability of the smaller set. So that whether $p < q$, the number of the canonical variable pairs is p. In our case, that is 3 but there is a question if we need all 3 □

There are 3 questions for the model evaluation:

1. What is the value of the canonical correlation coefficient (correlation coefficients)? Are they statistically significant?
2. To what extent do the new pairs reproduce the variability of the original sets?
3. How much of the variability of one set can we explain by the variability of the second set (so-called redundancy)?

• **CONSEQUENCES, TERMINOLOGY AND COMMENTS:**

$R(X_i, U_r)$ — $*$ *structural correlation coefficient* of the i-th variable with the r-th canonical variable. (Analogically for the right side)
Canonical variable represents mainly the original variables that are correlated with it.

$R^2(X_i, U_r)$ — the quadratic form of the structural correlation coefficient; shows the part of the X_i variable's variability that is explained by the r-th canonical variable U_r.
(Analogically for the right side.)

$\sum_{i=1}^{p} R^2(X_i, U_r)$ — the part of the variability of $X_1, \ldots, X_p$ that is explained by the canonical variable U_r. (Analogically for the right side.)
Let remind that the left set "X" is of a smaller scale than the right "Y" set, so that $p < q$. Thus $\sum_{r=1}^{p} \sum_{i=1}^{p} R^2(X_i, U_r) = 100\%$, but $\sum_{r=1}^{p} \sum_{i=1}^{q} R^2(Y_i, V_r) < 100\%$.

$\sum_{i=1}^{p} \frac{R^2(X_i, U_r)}{p}$ — If the sets are standardized, we are talking about the ratio of variability of $X_1, \ldots, X_p$ variables that is explained by the canonical variable U_r. (Analogically for right set, there is q in the denominator.) p (or q) in the denominator represents the total variability of the standardized vector of the left (or right) set.
These are means of the structural correlation coefficients squares.

$\sum_{i=1}^{p} \frac{R^2(X_i, U_r)}{p} \cdot R^2_{(C)r}$ — $*$ *redundancy* - the ratio of the variability of variables $X_1, \ldots, X_p$ that is explained by the canonical variable V_r.
So that we are explaining the variability in the left set by the canonical variable of the right set.

$\sum_{r=1}^{p} \sum_{i=1}^{p} \frac{R^2(X_i, U_r)}{p} \cdot R^2_{(C)r}$ — $*$ *total redundancy*
That is the ratio of the variability of the left side that is explained by the variability of the right side - by its part that we succeeded in "hiding" in to the canonical variables $V_1, \ldots, V_p$. (Analogically for the right set)

By the canonical weights we can also determine a *canonical score* that are values of particular observations by the new variables. That can be useful in data visualization.

5.2 significance test of the canonical correlation coefficient

We have found out p canonical correlation coefficients for the $p < q$s but not all must be statistical significant. We will be testing
$H_0 : R_{(C)1} = 0$; $H_0 : R_{(C)2} = 0$; ... ; $H_0 : R_{(C)p} = 0$.
Algorithm of the *Bartlett's* χ^2 test of significance of the canonical correlation coefficients is, however, different. Let's remind that $R_{(C)1} > R_{(C)2} > \ldots > R_{(C)p}$.

Firstly, we are going to test a hypothesis that vector of all p canonical correlation coefficients is a zero vector. If we do not turn down the hypothesis, we can claim that the sets are associated. If we turn it down, we come to conclusion that at least the first coefficient $R_{(C)1}, \ldots, R_{(C)p}$ is not zero (so that significant) and we are going to research more how

about the other $p-1$ coefficients.

Now the zero hypothesis is formulated in a way that the vector $R_{(C)2}, \ldots, R_{(C)p}$ is a zero vector. If we do not turn it down, the conclusion is that only the first coefficient $R_{(C)1}$ is significant and the others are insignificant. If we do turn it down, we know that both coeficients $R_{(C)1}, R_{(C)2}$ are significant and we keep going.

The sense of the testing is to ascertain from which point we can consider the canonical correlations as zero.

The test assumes that the random sample $X_1, \ldots, X_p, Y_1, \ldots, Y_q$ comes from a $p+q$-dimensional normal distribution.

5.3 algorithm for the canonical weights search

The principle is shown on sample characteristics.

$var(\mathbf{x})$	is estimated by the	sample variance matrix $S_{\mathbf{x}}$ of the $\mathbf{x} = (x_1, \ldots, x_p)'$ vecor type $p \times p$
$var(\mathbf{y})$	is estimated by the	sample variance matrix $S_{\mathbf{y}}$ of the $\mathbf{y} = (y_1, \ldots, y_q)'$ vector type $q \times q$
$cov(\mathbf{x}, \mathbf{y})$	is estimated by the	sample covariance matrix $S_{\mathbf{xy}}$ of the $\mathbf{x} = (x_1, \ldots, x_p)'$, $\mathbf{y} = (y_1, \ldots, y_q)$

We assume rank of matrix $cov(\mathbf{x}, \mathbf{y}) = p$ and $p < q$. we are searching for matrices A and B in a way that the model assumptions from 5.1 are met, and

$$A = \begin{bmatrix} a_{11} & \ldots & a_{1p} \\ \vdots & & \\ a_{p1} & \ldots & a_{pp} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & \ldots & b_{1q} \\ \vdots & & \\ b_{p1} & \ldots & b_{pq} \end{bmatrix}.$$

The maximal values from the 5.1 assumptions we are getting through Lagrange multipliers (subjecting to $D(U),\ D(V)$ are unit), that leads to a system of homogenous equations. Their roots are the eigenvalues and eigenvectors of matrices:

1. matrix: $S_{\mathbf{x}}^{-1} \cdot S_{\mathbf{xy}} \cdot S_{\mathbf{y}}^{-1} \cdot S_{\mathbf{yx}}$
2. matrix: $S_{\mathbf{y}}^{-1} \cdot S_{\mathbf{yx}} \cdot S_{\mathbf{x}}^{-1} \cdot S_{\mathbf{xy}}$

Both matrices have the same eigenvalues. If we order them in a decreasing way, they corresponds with the quadratic forms of the canonical correlation coefficients: $\lambda_r = R^2_{(C)r},\ r = 1, \ldots, p$. The eigenvectors (are mutually orthogonal) in the first matrix corresponds with the searched raws of the matrix A and correspondent eigenvectors corresponds with searched raws of the matrix B.

5.4 final remarks

Remark. CCA is for quantitative data analysis and has the biggest sense when all the pairs of the original variables are in a linear relation. If the vector of the all $p+q$ variables is from the $p+q$-dimensional normal distribution, then the relation between all the pairs is linear. (We can assess that from the matrix of the plots)

Remark. Let summarize the reasons for using CCA: 1) for testing independence or dependence between two groups of variables 2) searching for groups correlating mutually at most 3) Generalize a regression analysis in terms of following: If there are more than one variable being correlated, regression separate functions would not keep this complexity. Thus we can

see the groups of the "Y-ových" variables as a group of variables dependent on the group of the "X-ových" varaiables. In fact, the canonical correlation coeficients are symmetric which we need to look for when writing findings. 4) Ascertain whether measurements by two methods of different groups lead to the same resul.

Remark. Concerning the problem: $*$ Deviant values misrepresent correlation coeficients and needs repairing before analysis. $*$There can be some variables useless in the model and these can be detected by the CCA. $*$It is important to assure suficient number of observations since when a lot of variables, there can be some incompletenesses in the dataset. This cases need leaving.

Remark. What are the CCA's findings? 1) values of the canonical coeficients whose significance can be tested. Number of significant canonical coeficients represent number of canonical variables' pairs that sufficiently represent dependency between groups. These variables are sensible to analyze 2) A and B matrices of the canonical weights that represent how particular variables of a particular group contribute to particular canonical variable. 3) values of correlations (canonical weights) $R(x_i, U_r)$ between the particular original variable and the canonical variable. 4) total redundancy value (in both directions) which is a ratio of variables' of one set variability that is explained by the canonical variables from the second set. 5) values of new canonical variables for particular observations.

Chapter 6

Cluster Analysis

The aim of the cluster analysis is to classify observed objects into some groups (clusters) according to their similar appearance. The clustering classification can be successful only if the objects tend to associate in clusters. If we are doing in the 2D, we can the structure see from the data and visualize. For higher dimensions we need to use an appropriate algorithm. But there are lots of possible methods and it is hard to choose a good one, which depends on a problem concerned.

As in the preceding chapters we are beginning with a data matrix where we are measuring values of p variables $X_1, \ldots, X_p$ on n objects, so $\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$.

• **METHODS OF THE CLUSTERS FINDING:**

Hierarchy methods: The hierarchy algorithm is either *agglomerative* or *divisive*. In the first case, all the objects are considered as no cluster in the begging. Then we join the two clusters that are nearest. The process ends when all the objects are in one cluster.
In the divisive case, vice versa.

The graphical result of the hierarchy methods is a *dendrogram*. It is a binary tree depicted horizontally or vertically. In a dendrogram, each bundle represents a cluster. For example, in a horizontal dendrogram, the horizontal direction represents a *connection level* that is a distance between clusters in time when they have been connected into one cluster. Vertical cuts in a dendrogram represent classification ob objects into clusters. See picture 6.1

In the hierarchy methods, the number of clusters does not need to be determined in advance. (Algorithm of the clustering can be "stopped" when the levels of connection are sufficiently "big".)

Unhierarchic methods: There is a need to determine the number of clusters in the beginning.
Then the objects are classified into disjunctive groups "optimally" according a criterion. In the most cases, we are using the method of K means.

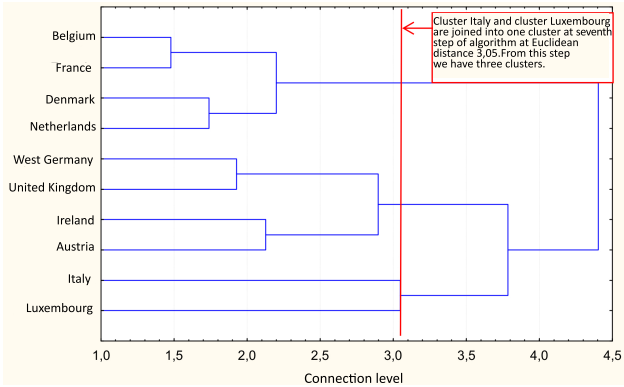


Figure 6.1

6.1 Objects distance

Since we are trying to determine the similar objects, we need to define "similarity". We do this by the distance among the objects - the nearest the distance, the more similar objects. We sign the values of the p variables measured on the ith and the jth object as $(x_{i1}, \ldots, x_{ip})$ and $(x_{j1}, \ldots, x_{jp})$. The distance of the ith and the jth object will be signed as d_{ij}. Finally, the distances of the each pair will be written in the *distance matrix* $D = \begin{bmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \vdots & & & \\ d_{n1} & d_{n1} & \ldots & 0 \end{bmatrix}$.

Below-mentioned measures of objects distances can be used only for *quantitative* variables[1].

• **EUCLIDEAN METRIC:**
- "common-known" distance of two points in a space.
- $d_{ij} = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2}$
- Euclidean metrics is hardly influenced by the variables of big values. Then the clustering is made mainly with regard to these variables
- So that, it is important to transform the data by dividing e.g. by the standard deviation
- Euclidean distance is also not suitable when the variables are correlated. Then these variables are having bigger weight than they should have.[2]

• **MANHATTAN DISTANCE (CITY-BLOCK):**
- The street on the Manhattan are orthogonal. If I want to go from a point i to a point j (see picture 6.2), I have to come round a corner house.
- $d_{ij} = \sum_{s=1}^{p} |x_{is} - x_{js}|$

[1] Measures for qualitative variable also exist.

[2] This problem is coped with the*Mahalanobis distance* that we are not dealing here

• **CHEBYCHEV DISTANCE:**
- We are trying to find the variable where the difference between the i-th and the j-th object is biggest. This difference is called the Chebychev distance of the ith and the jth object, so that:
- $d_{ij} = \max_s |x_{is} - x_{js}|$ through all $s = 1, ..., p$
- It strengthen the variables with big differences between objects.

See picture 6.2.

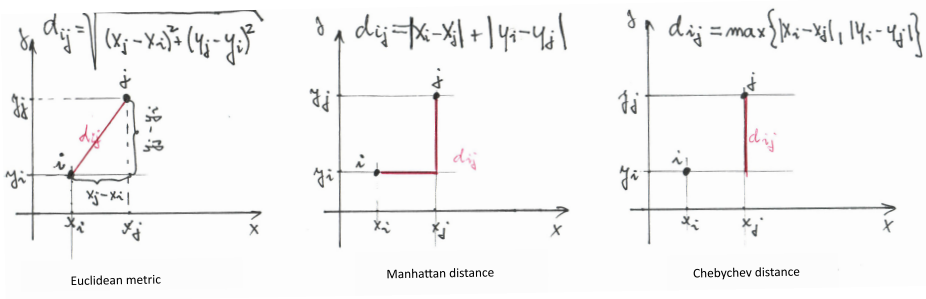


Figure 6.2: *We have only two variables X and Y ($p = 2$) and their values measured on the ith and the jth object. The distance d_{ij} of these objects is demonstrated by a length of the red line.*

6.2 cluster distances

If we have objects in clusters, we need to define the cluster distance. For the next methods, it is possible to use an idea that we assign each cluster a "representative" and then will be measure the distance between the representatives (according to selected method). After that, we will be merging the clusters that are nearest.

• **SINGLE METHOD (METHOD OF A NEAREST NEIGHBOR):**
- The distance between two clusters is the minimum of the all distances between objects.
- With clustering by this method, there is a tendency to create new objects as a snowball upon an existing cluster.

• **COMPLETE METHOD (METHOD OF THE FURTHEST NEIGHBOR):**
- The maximum of all distances between the clusters' objects is the distance between two clusters.
- With clustering by this method, there is a tendency to create clusters with similar number of objects.

• **AVERAGE METHOD:**
- The mean of the distances between all object pairs from the first and the second cluster is the distance.
- This method has also its weighted variant.
- The results are similar to results obtained from the method of the furthest neighbor.

• **CENTROID METHOD:**

- This method is suitable only if Euclidean metrics.
- The distance of the centroids is the distance between two clusters.
- The centroid is a fictive object of the cluster whose coordinates are means for particular variables that are computed for all cluster's objects.
- This method has also its weighted variant.

• **WARD METHOD:**
- Very effective, tending to create small clusters.
- Suitable only if Euclidean metrics.
- Principle based on ANOVA[3]

Schematic depiction of distances can be seen on the picture 6.3.

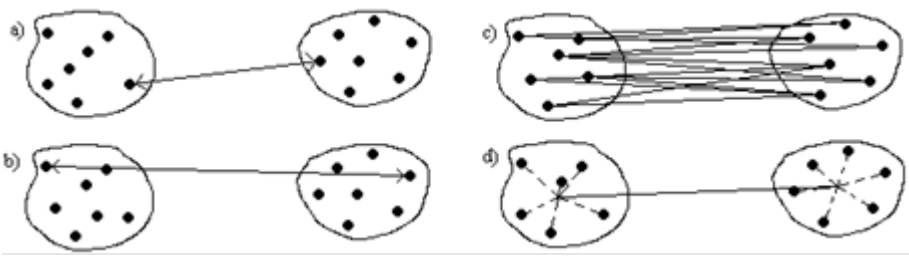


Figure 6.3: *a) method of the nearest neighbor, b) method of the furthest neighbor, c) method of a mean linkage, d) centroid method. The distance in the Ward method can be introduced as a centroid method case multiplied by a coefficient depending on a cluster sizes.*

6.3 Cofenetic coefficient of correlation

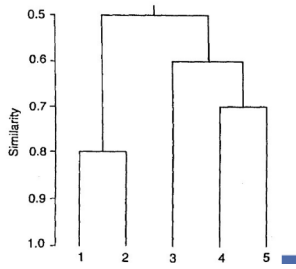
Different clustering procedures can be giving different findings. What is the right one? We can use an "empiric visual method", so we create dendrograms using some clustering procedures and if the findings are similar, the clustering should be considered as good.

More objective is the assessment by the "cofenetic coefficient correlation". It evaluates the rate of equivalency of the matrix of object distances D and the *cofenetic matrix* that is a result of a particular clustering method. (i, j)th cell of this matrix is defined as a cluster distance where the cluster containing ith object is joined with the cluster containing jth object. The more similar the cofenetic matrix to the original matrix of object distances, the better job has the method done in retaining the object distances. (We can see an example of a cofenetic matrix on the picture 6.4.)

Cofenetic coefficient of correlation is a "common" correlation coefficient computed from the "twodimensional data set" where one variable is represented by cells upper the diagonal in the distance matrix D and the second variable is represented by cells upper diagonal in

[3] Ward method is going to connect the ith cluster and the jth cluster into a new cluster h in a way that the sum of the squares in the new cluster is lower by the sum of squares in the fading clusters and that number minimize. Let sign SS_i (resp. SS_j, resp. SS_h) the sum of the all objects' deviations from the particular centroids squares in the ith (jth, hth) cluster. Then we are going to minimize $SS_h - SS_i - SS_j$.

the cofenetic matrix (see picture 6.5.). We choose the method with the highest cofenetic correlation coefficient.



	1	*2*	*3*	*4*	*5*
1	1.0	0.8	0.5	0.5	0.5
2	0.8	1.0	0.5	0.5	0.5
3	0.5	0.5	1.0	0.6	0.6
4	0.5	0.5	0.6	1.0	0.7
5	0.5	0.5	0.6	0.7	1.0

Figure 6.4: *There is a cofenetic matrix particular to a determined clustering on the picture. The values on the y axis of the dendrogram are misleading. The bigger the similarity, the lower the distance. 1 points to a perfect accordance - zero distance.*

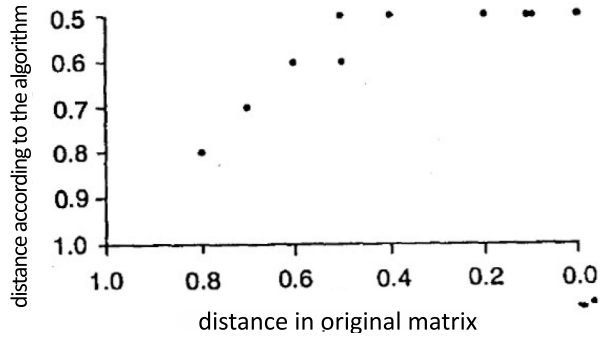


Figure 6.5: *If there is a perfect accordance of a matrix of distances with a cofenetic matrix, the points would be in a line.*

6.4 Agglomerative hierarchic clustering algorithm

In the subsection 6.2 we have introduced methods of clusters' distances measuring that are correct to use during the hierarchic clustering. Now we are going to introduce a cyclic algorithm of the clustering:

1. We calculate the distance matrix D according to the chosen "distance" from the 6.1
2. We consider each object as a cluster.
3. We search for the two nearest clusters, the ith and jth (according to chosen method from the 6.2) in the matrix D.
4. We connect the ith and the jth cluster into a new the hth cluster. We delete the ith and jth raw and column from the matrix D and replace them by a raw and column for the new hth cluster.
5. We make a note in which cycle the connection of the i-th and j-th cluster has become and level of their connection. (This is depicted by the dendrogram.)

6. If all the objects are not joined into one cluster, we go back to step no. 3.

The result of the clustering depicted by a dendrogram allows us to evaluate in which cycle the optimal distribution of objects happened.

6.5 unhierarchic clustering methods

We introduce only a principle of the *method of the k means*. Firstly, this method distribute the cases into the k clusters randomly. Then it is replacing objects among clusters in a way that the variability within the groups is minimized and the variability among groups is maximized. In other words, it classifies objects into groups in a way that we get the highest possible significance ANOVA test. Con of this method is that it highly influenced by the initial choice of the clusters. The initial choice of the clusters is determined regarding some criteria or it is determined on the problem basis. We can also take the clusters as a result of a having been proceeded clustering that we want to verify or enhance.

The algorithm:

1. We determine the initial analysis of the set of n objects into the k clusters.
2. We determine the centroids of the clusters in particular.
3. We calculate the distances from all sample centroids for all objects and the object classify in the cluster to which the distance to its sample centroid is the nearest. If we keep doing but there are no changes, we consider the result as definite; otherwise we go back to the step no. 2.

6.6 Final remarks

Remark. All the upper-mentioned methods are suitable only for quantitative variables.

Remark. The distance depends on a variable scaling. If the scaling is not the same, standardization is recommended.

Remark. If there is not a clear structure in the data, the different methods lead to different findings. On the other hand, when the different approaches are giving the same results, we can consider the structure in the data as “clarified”.

Remark. Findings of the clustering can be highly influenced by outlying observations.

Chapter 7

Discriminant Analysis

Discriminant analysis is a method for differentiating objects and classifying them into groups according to p observed variables $X_1, \ldots, X_p$. Imagine there is a bank that aims to group clients according to their salary, age etc. for the risk of providing credit. The groups are to be: safe clients, clients with acceptable risk, unacceptable clients. Based on observations of the vector $x_1, \ldots, x_p$ the bank aiming to classify the clients (objects) into groups I, II, III. Discriminant analysis is searching for a rule that is able to run this classification facing a problem that an object may be classified into more groups (with regard to $x_1, \ldots, x_p$ values). The rule must be designed in such a way that minimalize probability of mistaken classification. This rule is being specified onto the *training* data which have known to which groups the are belong to. On a basis of these data we can observe which values of $x_1, \ldots, x_p$ are typical for particular groups. After tuning the rule upon the training data, we will use them for the objects which have unknown their belong to a particular group and classify them in accordance with the having designed rule. So that, DA is being realized in two steps.

The main objectvie of the DA is to classify the objects into groups. Another goal is to ascertain which variables from the $X_1, \ldots, X_p$ vector are having information useful for discrimination. For example, for a physician it is useful to know which variables describing state of a pacient are important for determination if to initiate a special care or not.

• **SIGNING AND COMMENTS**:

p	number of variables in the vector $X_1, \ldots, X_p$
$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$	particular observations of vector $X_1, \ldots, X_p$
n	number of objects of the training set
k	number of groups In each group, the objects stem from the same $p$dimensional distribution with a density of $f_j(\mathbf{x})$, $j = 1, \ldots, k$. Densities are different among groups (otherwise the classification would be useless).
π_j	$*$ *apriori* citizenship into the jth group. We are talking about known probability that a random object stems from the jth goups in advance, $j = 1, \ldots, k$. This probability reflects not-the-same-frequent representation of the particular groups in the sample. (In the group of credit clients, there probably should be more wealthy people than the poors.)
$\frac{f_j(\mathbf{x}) \cdot \pi_j}{\sum_{j=1}^{k} f_j(\mathbf{x}).\pi_j}$	$*$*aposteri probablity* of the citizenship into the jth group We are talking about a bayesian probability of a citizenship into the j group after a random vector $\mathbf{X}$ realized by a vector of values $\mathbf{x}$, thus $P(\text{object belongs to the } j\text{th group} \mid \mathbf{X} = \mathbf{x})$
$f_j(\mathbf{x}) \cdot \pi_j$	$*$*jth diskriminant score* It is a numerator in the aposteri probability. Since aposteri probabilities are having the same denominator for all the groups, the numerator is the only thing sufficient to discriminant.
$f(\mathbf{x}) = \sum_{j=1}^{k} f_j(\mathbf{x}).\pi_j$	$*$ *The mix density* where we do not distinguish from which group an observation is stemming from. We will measure values of $\mathbf{x}$ with the "probability" that is defined by a density $f(\mathbf{x})$ on a chosen object.
n_j	number of objects in the jth group $j = 1, \ldots, k$
$\boldsymbol{\mu}_j$	$p$dimensional column vector of expected values in the jth group $j = 1, \ldots, k$
Σ_j	$p \times p$ variance matrix of the $\mathbf{X}$ in the jth group $j = 1, \ldots, k$

7.1 Summary of criteria for the rules

The rule classifies an object with observation $\mathbf{x}$ into group j where

1. **CRITERIUM OF THE MINIMAL MAHALANOBISAL DISTANCE:**

   • $j = \underset{i=1,\ldots,k}{\arg\min} (\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$

•Thus j is the group whose centroid is the nearest to the observed object.

2. **CRITERIUM OF THE MAXIMUM LIKEHOOD**

•$j = \underset{i=1,\ldots,k}{\arg\max}\, f_i(\mathbf{x})$

•We have an observation $\mathbf{X} = \mathbf{x}$. For what density is this observation most probable? - It is the "highest" density which we can gain this observation for. Insex of this density id the index of the group to which we classify the object with the observation $\mathbf{x}$. An illustration of the principle can be seen in the picture 7.1.

3. **BAYES CRITERIUM**

•$j = \underset{i=1,\ldots,k}{\arg\max}\, \pi_i \cdot f_i(\mathbf{x})$

•Objekt will be classified into the group for which the score $\pi_i \cdot f_i(\mathbf{x})$ is highest.
•It generalizes the maximum likehood criterium in the sense that regards to apriori probabilities of the object frequencies in the groups.
•Maximum likehood is a special case of the Bayes criterium if we plug $\pi_1 = ... = \pi_k = \frac{1}{k}$. in to the aposteri probabilities
•Bayes criterium minimalize total expected loss arised from misconduct classification.
•If the vector $X_1, \ldots, X_p$ is having the $p$dimensional normal distribution $N_p(\mu_j, \mathbf{\Sigma}_j)$ in each group $j = 1, \ldots, k$, then the classification of the objects with the Bayes criterium is called a quadratic discriminant analysis. Moreover, if the matrices $\mathbf{\Sigma}_j$ are the same in each group, we are talking about a linear discriminant analysis.

4. **FISHER CRITERIUM**

•$j = \underset{i=1,\ldots,k}{\arg\min}\, |\mathbf{v}'\mathbf{x} - \mathbf{v}'\boldsymbol{\mu}_i|$,

where for the matrix of the variablity between groups $\mathbf{B}$ and matrix of within group variability $\mathbf{W}$ is

the $p$dimensional vector $\mathbf{v} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\max}\, \frac{\mathbf{w}'\mathbf{B}\mathbf{w}}{\mathbf{w}'\mathbf{W}\mathbf{w}}$.

•Fisher criterium is searching for the transformation of the original vector of variables $X_1, \ldots, X_p$ onto the new vector of variables $Y_1, \ldots, Y_l$ when the differences between groups are highlighted) Klasifikace objektů probíhá tak, že se nejdříve přepočítají souřadnice jednotlivých objektů pro nové proměnné (ty již budou kolmé), přepočítají se i střední hodnoty jednotlivých proměnných a pozorování zařadíme do té skupiny v níž se přepočítaný vektor středních hodnot liší nejméně od přepočítaného vektoru pozorování.
•Toto kritérium při klasifikaci objektů do skupin předpokládá stejné aprioní pravděpodobnosti ve všech skupinách.

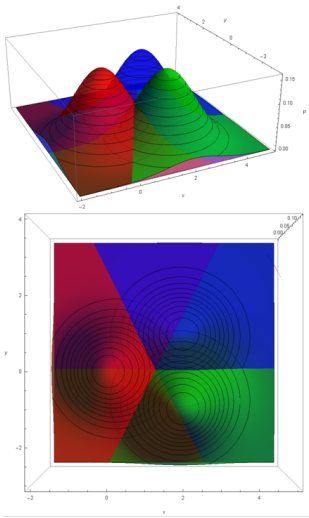


Figure 7.1: *There are densities of the twodimenional normal distributions $p = 2$ with the same unit variance matrix and with different vectors of the expected values. The red, green and the blue density f_1, f_2, f_3 correspond with three groups of objects. $f_1 \sim N_2(\binom{0}{0}, \binom{10}{01})$; $f_2 \sim N_2(\binom{2}{-1}, \binom{10}{01})$; $f_3 \sim N_2(\binom{2}{1}, \binom{10}{01})$ When having look from the upper side, we can see color-distinguished particular parts of the plane $\mathbb{R}^2$ which we have separated using maximum likehood criterium.*

7.2 Fisher criterium - canonical discriminant analysis

Fisher criterium has been "designed" that it is able to distinguish the groups better. The original $p$dimensional vector of $X_1, \ldots, X_p$ variables would be transformed onto new l variables $Y_1, \ldots, Y_l$ that the differences between groups would be maximized using this transformation (information about differences between groups can be found in the B "between" matrix) and the differences within the groups would be minimalized (information about differences within groups can be found in the W "within" matrix). We are having $l = \min\{(k-1), p\}$ new variables, and we call them *discriminants* (or canonical variables) and we define them as a linear combination of the original variables, so that
$Y_r = v_{1r}x_1 + v_{2r}x_2 + \ldots + v_{pr}x_p = \mathbf{v_r}' \cdot \mathbf{x}, \quad$ pro $r = 1, \ldots, l$. Coeficients of the linear combination creates the parts of $\mathbf{v_r}$ vector in the direction of the best differentiation of the transformed densities of the all k groups. Consequently, the observed objects are classified better. Fisher discriminant criterium then classifies an object with the values of $\mathbf{x}$ into the group in which the reflect of the expected value of the variables in the transformation $\mathbf{v_r}$ is nearest to the $\mathbf{x}$. We can see how one discriminant help us to distinguish objects from the two groups in the picture 7.2.

It is possible to prove that the vectors $\mathbf{v_1}, \ldots \mathbf{v_l}$ are eigenvectors of the $W^{-1} \cdot B$ matrix and correspond with the eigenvalues of this matrix $\lambda_1 > \lambda_2 > \ldots > \lambda_l$.

There is a question if we need all l new variables (discriminatns) for discrimination. For this the tests based on *Wilks* Λ are useful tool.
$\Lambda = \frac{\det \mathbf{W}}{\det \mathbf{T}} = \frac{\det \mathbf{W}}{\det (\mathbf{W}+\mathbf{B})} \in (0, 1)$ represent a ratio of the within-group variability and

the total variability. If the varibalility within the groups is little, a big deal from the toatl variability is concerned with the variability between the groups and that points to a good discriminant ability of the analysis. The value $\Lambda = 0$ reflects "perfect discrimination" of the objects and the value $\Lambda = 1$ tells us that the result of the DA is useless as the objects are not possible to be distinguished on a basis of the observed variables.

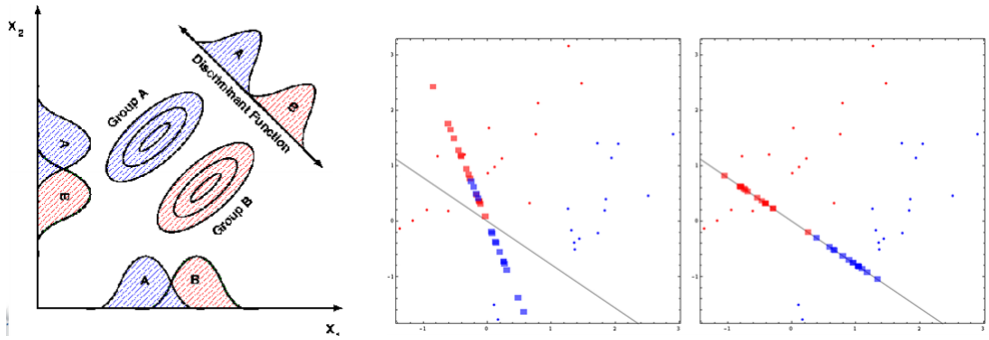


Figure 7.2: *We observe two variables X_1, X_2, ($p = 2$) upon objects stemming from two groups $k = 2$. Theretical picture in the left vlevo shows marginal densities that overlap each other both at the x_1 and x_2 axis concerning both groups. So that some realizations of (x_1, x_2) would be difficult to classify into the groups. In the direction vector of the new variable (discriminant) the densities of the both groups differ more between (not overlapping) and the variability for the particular groups is reduced. There are measured data in the pictures in the right. The middle-one picture shows a projection onto a line with a random chosen direction. Picture in the right shows projection onto a line with the direction corresponding with the vector $\mathbf{v_1}$. Projection in the right within all possible projections enable us distinguish the group origin of the measured object better.*

Canonical DA (DA using Fisher criterium) is suitable for evaluating which variable from $X_1, \ldots, X_p$ are needed for the discrimination and which are useless. For this reason, it is suitable to have a look at the coordinates of the eigenvectors and correaltion betwen the original variables and the discriminants. In order to interpret the coordiantes of the eigenvectors well, it is useful to standardize the eigenvectors $\mathbf{v}$ onto the $\mathbf{v}^*$. Then the value of the ith coordinate of the rth standardized eigenvector v^*_{ir} informs us about the rate that theX_i variable contributed to the rth discriminant. Similarly, the correaltion between the ith varible and the rth discriminant is also interesting. Then, we can ascertain what would be the change in the Wilk Λ whether we an original variable leave behind etc.

The classifiacation of the objects into the groups is another aim of the analysis. The assessment of the clasification success we will cover in the 7.5.

Example 7.1. In the dataset *dovolena.sta*, there are information about 50 families that can be considered a random sample from a population. The variable *ID* represent if a family traveled to some resort in the last 2 years (value 0 is an answer *no*, value 1 is answer *yes*), variable X_1 states an annual family income in thousands USD; variable X_2 states an attitude towards traveling (9value scale, 1 = absolutely rejecting, 9 = absolutely accepting); variable X_3 states the significance of the family trip (9value scale, 1 = lowest, 9 = highest); proměnná X_4 states number of family members; proměnná X_5 states an age of the agest member of the family and the V variable states whether a family want to spend a little (1), averaged (2), or a lot of

money (3) for the family trip.

a) The task aim is to ascertain which properties (variables $X_1, \ldots, X_5$) are typical for a certain population of families going tripping to a resort (variable *ID* classifies families into two groups). At the beginning ascertain how many discriminants are significant, then establish and interpret discriminant coeficients, diskriminant correlation with the original variables (Which variables could be left behind?), diskriminant score and classify observed families into groups. Ascertain how many families would have been identified right with regard to the group.

b) The same tasks as in a); the variable distinguishing families into groups will be *V*. Besides that, draw a chart of classified observations in the plane of the two discriminants.

Solution
To be figured out during seminar □

• MATHEMATICAL DESCRIPTION OF THE CDA PRINCIPLE:

We know (from the multivariate analysis) the total variability of the vector $\mathbf{X} =(X_1, \ldots, X_p)$', $var\mathbf{X} =: \mathbf{T}$ can be analyzed onto the sum of the matrix of the variability between groups $\mathbf{B}$ and the matrix of the variability within the groups $\mathbf{W}$. Tedy $\mathbf{T} = \mathbf{B} + \mathbf{W}$. So for the variability of the onedimensional random variable $\mathbf{Y} = \mathbf{v}^{'} \cdot \mathbf{X}$ is valid following:
$Q_T := \mathrm{var}\,(\mathbf{v}^{'}\mathbf{X}) = \mathbf{v}^{'}\mathbf{T}\,\mathbf{v} = \mathbf{v}^{'}\mathbf{W}\,\mathbf{v} + \mathbf{v}^{'}\mathbf{B}\,\mathbf{v} =: Q_W + Q_B$
Fisher linear diskriminant function is the function that maximilize the Fisher ratio. In other words it meets

$$\mathbf{v} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\max}\,\frac{Q_B}{Q_W} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\max}\,\frac{\mathbf{w}^{'}\mathbf{B}\,\mathbf{w}}{\mathbf{w}^{'}\mathbf{W}\,\mathbf{w}}.$$

Needed matrices are being estimated from the data matrix as follows:
$\mathbf{T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \mathbf{M})(\mathbf{X}_{ij} - \mathbf{M})^{'}$,
$\mathbf{B} = \sum_{i=1}^{k} n_i(\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^{'}$,
$\mathbf{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \mathbf{M}_i)(\mathbf{X}_{ij} - \mathbf{M}_i)^{'}$, kde
•$\mathbf{X}_{ij}$ is an object from the ith group $i = 1, ..., k$; we index the objects through $j = 1, ..., n_i$ within this group. (We measure values of the variables $X_1, \ldots, X_p$, thus $\mathbf{X}_{ij}$ is a $p$dimensional vector.)
•$\mathbf{M} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ is a $p$dimensional vector of the means of all observations
•$\mathbf{M}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{X}_{ij}$ is a $p$dimensional vector of means in the ith group, $i = 1, ..., k$
If there is $\mathbf{v}_1$ an eigenvector corresponding with the highest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$, then we can show that the $\mathbf{v}_1$ maximalize the Fisher ratio. If the rank of the matrix $rank(\mathbf{W}^{-1}\mathbf{B}) = l$ the number of inequal-to-zero eigenvalues (and corresponding eigenvectors) is l and we can order them in a descending way $\lambda_1 > \ldots > \lambda_l$. We can standardize the eigenvectors according the following: $\mathbf{v}_r^* = \frac{1}{\sqrt{n-k}}\,F\,\mathbf{v}_r$, where F is a diagonal matrix with roots of the diagonal objects of the matrix $\mathbf{W}$. (We standardize in a way that a variable $Y = \mathbf{v}_r^{*'} \cdot \mathbf{x}$ is having a unit variability.)

Fisher diskriminant criterium classifies objects with the value of $\mathbf{x}$ into the group whose reflect of the expected value of properties in the transformation $\mathbf{v}$ is nearest to the reflect of

$\mathbf{x}$ or into group

$$j = \arg\min_{i=1,\ldots,k} |\mathbf{v}'\mathbf{x} - \mathbf{v}'\mathbf{M}_i|.$$

7.3 "Economic" assessment of the rule

Imagine again a client trying to be given a credit, and a bank deciding whether classify them into the group of potential "solvents" or potential "insolvents". If the bank classified them right, there is no big deal of losing money. But if it is not right, there is a deal with losing money.

When we do not need to be interested in the losses (they are the same in all directions "směrech"), we do not prefer none of the criteria mentioned in 7.1 (with regard to possible loss). When we need to be interested in the different losses, then we can prove that the Bayes criterium minimalize the fault of the mistaken classification. This property of the Bayes criterium can be generalized onto more than two groups.

7.4 linear and quadratic DA (LDA, QDA)

Bayes criterium classifies the observed object into the group in which the value of the jth diskriminant score $S_j(\mathbf{x}) = f_j(\mathbf{x}) \cdot \pi_j$ is maximal, $j = 1, \ldots, k$. For practical reasons the knowledge of densities $f_j(\mathbf{x})$ is needed. (The densities are $p$dimensional, index j is related to the jth group.)

So that we are going to introduce a formula for the calculation of the jth score with an assumption that the vector $X_1, \ldots, X_p$ is having a $p$dimensional normal distribution in each group, $\mathbf{X} \sim f_j(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, \ldots, k$
1) for cases when the variance matrices $\boldsymbol{\Sigma}_j$ can be different for all groups $j = 1, \ldots, k$
2) for cases when the variance matrices $\boldsymbol{\Sigma}$ are the same concerning all the groups:

1. **quadratic DA**

   $$\boxed{S_j(\mathbf{x}) = \ln \pi_j - \tfrac{1}{2}\ln(\det \Sigma_j) - \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)}$$

   -$S_j(\mathbf{x})$ is called *quadratic discriminant score*.
   -By this criterium we separate the $p$dimensional space onto areas bordered with quadrics. (For $p = 1$ we are dealing with conic section)

2. **linear DA**

   $$\boxed{S_j(\mathbf{x}) = \ln \pi_j - \boldsymbol{\mu}_j' \Sigma^{-1} \mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_j' \Sigma^{-1} \boldsymbol{\mu}_j}$$

   -$S_j(\mathbf{x})$ are called *linear diskriminant scores*.
   -By this criterium we seperate the $p$dimensional space onto areas bordered with supplanes. (For $p = 1$ we are dealing with a point, for $p = 2$ we are dealing with a line, for $p = 3$ we are dealing with a plane.)

7.5 Probability estimates of the right classification

Two methods of evaluating the quality of the decision-making rule follows.

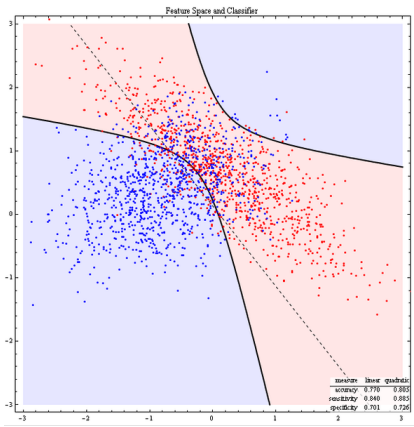


Figure 7.3: *There is a result of the QDA (a LDA) classification for the data stemming from the twodimensional distributions ($p = 2$) in two groups ($k = 2$). Values of the variables from the first group are being in red; from the second group in blue.*

.

- **RESUBSTITUTE METHOD**
  This method consists in using the designed decision-making rule onto the traning data. We are walking throuh all n objects of the training set and we save number of objects stemming from the ith group and classified into the jth group into the n_{ij} statistics. Thus we will get a matrix of relative frequencies

$$\frac{1}{n}\begin{bmatrix} n_{11} & \dots & n_{1k} \\ \vdots & & \\ n_{k1} & \dots & n_{kk} \end{bmatrix}$$

  Cells of this matrix are giving estimates of the particular probabilities of the right or bad classifications. There are ordered probability estimates of the right classification for teh particular groups at the main diagonal. So that, the matrix trace demonstrates the estimate of the total probability of the right classification of the random chosen object. This estimate is actually upper-estimated good as the dicision-making rule is being tested in the same data from which we have developed it. Since each cell of the training set contributes to the rule design, the chances of right classifications are being increasing. For $k = 2$ see table 7.1. If its cells are being devided by the scale of n training set, we would be given an upper-mentioned matrix.

		classification		
		group I	group II	
reality	group I	n_{11}	n_{12}	$n_{1.}$
	skupina II	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

Table 7.1: *table of right and bad classification frequencies*

- **CROSS VALIDATION**
  This method consists in random distribution of the training set onto two subsets. We apply the discriminant part of the DA by wchich we will be given a dicision-meking rule onto the first subset. In the second subset, we are reviewing the quality of the gained rule by the resubstitute method. This setup do not favour objects unlike the upper-mentioned procedure, but requires higher training set data for the similar value.

  This procedure is widely used for probability estimates of the right classification only.

7.6 sample characteristics and assumption verification

As in the former chapters, there are usually unknown parameters of the distribution in the DA, thus are replaced with sample characteristics. For the $X_1, \ldots, X_p$, we estimate $p$dimensional vector of expected values $\boldsymbol{\mu}$ with vector of means, variance matrix $\boldsymbol{\Sigma}$ with sample variance matrix.
Apriori probabilities of the groups $\pi_1, ..., \pi_k$ are usually also unknown so they are replaced with the frequencies $\hat{\pi}_i = \frac{n_i}{n}$ for $i = 1, ..., k$.

Before introducing DA assumptions, let consider that this is an explorative technique so slight assumption break is not a problem. (e.g. When using CDA we need normality only for using χ^2 tests, but for the classification the normality is not required.)

• MULTIDIMENSIONAL NORMALITY
Whatever testing in DA or discriminant scores calculations are, we need to verify the multidimensional normality assumption in each group. It is not practically possible to test multidmensional normality so we at least consider the onedimensional normalities of the all p variables separately in each k groups. (SW test, KS test, or visually N-P plot,...) If the onedimensional normalities would not be turned down, we would "believe" in the multidimensional normality. Otherwise, we can keep conducting our analysis but must take the findings with reserve.

• VARIABILITY HOMOGENITY
LDA assumes the variance matrices conformity in all k groups, so $H_0 : \boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_k$. This hypothesis can be tested by e.g. Box test of variance matrices conformity. Nonetheless, this test is very sensitive to even a slight normality break so its decline of the variance conformity we should take seriously at all.

• TEST OF VECTORS OF THE EXPECTED VALUES
We test a hypothesis $H_0 : \mu_1 = \ldots = \mu_k$. If we have a reason of assumption that the observed objects stem from k groups, we want to turn down this hypothesis. (The procedure assumes variance matrices conformity.)

• UNDESIRABLE LINEAR COMBINATIONS OF VARIABLES $X_1, \ldots, X_p$
We have met the term of redundance before. Simply, if two or more variables are asking the same (at least one of the variables is a linear combination of the others), the inverse matrix to the matrix Σ which is needed to discriminant scores calculations cannot be found. Tolerancy is between null and one; the nearer null, the more is the certain variable useless in the model. If the value falls under the threshold, it is required to leave the variable from the model.

• UNCORRELATION OF $\boldsymbol{\mu}$ **AND** $\boldsymbol{\Sigma}$

Chapter 8

Correspondence analysis

Correspondence analysis is a descriptional and explorational technique suitable for analysis of multidimensional categorical data (should we want to use it for analysis of quantitative variables, they can be categorized) and is similar to factor analysis in methodology. The main aim of this method is to make clear the structure of the hidden associations withing the contingency tables and show it graphically, in the best case scenario. The clear strucutre enables:

1. Detection of similarity of catogories for particular variables and thus enables their optimal clustering. It discovers associations of categories within particular variables.

2. Reveal multual similarity among categories of different variables, thus associations among variables.

The main principles are going to be shown for two categorical variables which are described by a contingency table. In that case we are talking about the *simple CA*. If we are analyzing associations among three categorical variables at least, we are talking about *multivariete CA*.

8.1 Elementary analysis of contingency tables and χ^2 test of independence

We introduce a table of observed frequency, relative observed frequency and table of estimated relative frequency if both variables were independent. We assign the simultaneous parts of the tables the matrix assignment.

	$y_{[k]}$	$y_{[1]}$	$\cdots$	$y_{[s]}$	$n_{j\cdot}$
$x_{[j]}$	n_{jk}				
$x_{[1]}$		n_{11}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{[r]}$		n_{r1}	$\cdots$	n_{rs}	$n_{r\cdot}$
$n_{\cdot k}$		$n_{\cdot 1}$	$\cdots$	$n_{\cdot s}$	n

observed frequencies table

	$y_{[k]}$	$y_{[1]}$	$\cdots$	$y_{[s]}$	$p_{j\cdot}$
$x_{[j]}$	p_{jk}				
$x_{[1]}$		p_{11}	$\cdots$	p_{1s}	$p_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{[r]}$		p_{r1}	$\cdots$	p_{rs}	$p_{r\cdot}$
$p_{\cdot k}$		$p_{\cdot 1}$	$\cdots$	$p_{\cdot s}$	1

observed frequencies table, where

$p_{ij} = \frac{n_{ij}}{n} \quad p_{i\cdot} = \frac{n_{i\cdot}}{n} \quad p_{\cdot j} = \frac{n_{\cdot j}}{n}$

$i = 1, \ldots, r, \quad j = 1, \ldots, s$

We put the simultaneous frequencies into a matrix P of type $r \times s$.

	$y_{[k]}$	$y_{[1]}$	$\cdots$	$y_{[s]}$	$p_{j\cdot}$
$x_{[j]}$	q_{jk}				
$x_{[1]}$		q_{11}	$\cdots$	q_{1s}	$p_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{[r]}$		q_{r1}	$\cdots$	q_{rs}	$p_{r\cdot}$
$p_{\cdot k}$		$p_{\cdot 1}$	$\cdots$	$p_{\cdot s}$	1

estimated relative frequencies table if X and Y were independent where $q_{ij} = p_{i\cdot} \cdot p_{\cdot j}$

$i = 1, \ldots, r, \quad j = 1, \ldots, s$

We put the simultaneous frequencies into the matrix Q of the typer $r \times s$.

In accordance with the CA terminology it is common to call the vector of marginal frequencies $(p_{1\cdot}, \ldots, p_{r\cdot})'$ vector of *row weights* and vector $(p_{\cdot 1}, \ldots, p_{\cdot s})'$ as vector of *column weights*. The X variable is called *row variable*, variable Y is called *column variable*. Matrix P is called *Correspondence matrix*.

χ^2 test of independce compares the P matrix of the relative frequencies table with the Q matrix of the estiamted frequencies table. If independent, the P and Q tables should be "similar". When these two matrices are "significantly" different, the test will be stating that X ans Y are dependent. However, the test cannot point on a independency structure (which categories of particular variables cause significant difference between P and Q. With that we will be dealing further.

• χ^2 TEST OF INDEPENDENCE:

Let's get back to the χ^2 test. The null hypothesis states that X and Y are independent. Hypothesis is turned down if the matrix of P and Q are significantly different.

Test statistics $V = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \approx \chi^2((r-1)(s-1))$.

The V statistics can be also written as $V = n \cdot \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(p_{ij} - q_{ij})^2}{q_{ij}}$ that depicts which "relative frequencies table cell" broke the independency.

Big differences between tables lead to big values of V so the critical region on the right is of: $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$.

• CONDITIONAL RELATIVE FREQUENCIES TABLES:

$\frac{n_{11}}{n_{\cdot 1}}$	$\frac{n_{12}}{n_{\cdot 2}}$	$\ldots$	$\frac{n_{1s}}{n_{\cdot s}}$
$\frac{n_{21}}{n_{\cdot 1}}$	$\frac{n_{22}}{n_{\cdot 2}}$	$\ldots$	$\frac{n_{2s}}{n_{\cdot s}}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$
$\frac{n_{r1}}{n_{\cdot 1}}$	$\frac{n_{r2}}{n_{\cdot 2}}$	$\ldots$	$\frac{n_{rs}}{n_{\cdot s}}$
$\mathbf{c}_1$	$\mathbf{c}_2$	$\ldots$	$\mathbf{c}_s$

$\frac{n_{11}}{n_{1\cdot}}$	$\frac{n_{12}}{n_{1\cdot}}$	$\ldots$	$\frac{n_{1s}}{n_{1\cdot}}$	$\mathbf{r}_1$
$\frac{n_{21}}{n_{2\cdot}}$	$\frac{n_{22}}{n_{2\cdot}}$	$\ldots$	$\frac{n_{2s}}{n_{2\cdot}}$	$\mathbf{r}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\frac{n_{r1}}{n_{r\cdot}}$	$\frac{n_{r2}}{n_{r\cdot}}$	$\ldots$	$\frac{n_{rs}}{n_{r\cdot}}$	$\mathbf{r}_r$

Vector of raw conditioned frequencies $\mathbf{r}_i = (\frac{n_{i1}}{n_{i\cdot}}, \frac{n_{i2}}{n_{i\cdot}}, \ldots, \frac{n_{is}}{n_{i\cdot}})$ is called *raw profile*, $i = 1, \ldots, s$.

Vector of column conditioned frequencies $\mathbf{c}_j = (\frac{n_{1j}}{n_{\cdot j}}, \frac{n_{2j}}{n_{\cdot j}}, \ldots, \frac{n_{rj}}{n_{\cdot j}})$ is called *column profile*, $j = 1, \ldots, r$. The relation between profiles and weights can be seen in *CA_1_priloha*.

We can further enrich the table analysis with 3D column graphs and conditioned pie charts or with residual analysis (see *CA_2_priloha*). □

8.2 Simple CA

Example 8.1. In a marketing research, we are exploring impact of some criteria during purchasing behavior regarding juice brands. We have conducted a sample of respondents which were questionned which juice they are purchasing most and which of the criteria affect their purchasing choice mostly. The results follows in table 8.1. Interpret associations between variables "juice brand" and "choice criterium".

Observed Table (Frequencies) (dzusy.sta)
Row variables: Brand(5)
Column variables: Criterion(5)

	Packaging	Price	Quality	Taste	Tradition	Total
Cappy	13	14	27	23	28	105
Hello	19	22	14	12	10	77
Rauch	6	17	24	15	34	96
Relax	17	29	24	31	35	136
Toma	10	22	38	15	51	136
Total	65	104	127	96	158	550

Percentages of Row Totals (dzusy.sta)
Row variables: Brand(5)
Column variables: Criterion(5)

	Packaging	Price	Quality	Taste	Tradition	Total
Cappy	12,38095	13,33333	25,71429	21,90476	26,66667	100,0000
Hello	24,67532	28,57143	18,18182	15,58442	12,98701	100,0000
Rauch	6,25000	17,70833	25,00000	15,62500	35,41667	100,0000
Relax	12,50000	21,32353	17,64706	22,79412	25,73529	100,0000
Toma	7,35294	16,17647	27,94118	11,02941	37,50000	100,0000

Table 8.1: *The table on the left is a table of absolute frequencies, table on the right is a table with raw profiles.* ***If we are facing perfect independency of both variables, all the raw profiles would be the same.***

We will show the CA principle at the raw profiles and analogically can be used for column profiles. Notice that particular raw profiles correspond with particular categories of the raw variable of X. One raw profile is a vector with s objects and can be regarded as a point in a s-dimensional space whose coordinates correspond with profile values. In figure 8.1 we have been given 5 raw points (there are 5 juice brands) and every point is of 5-dimension (5 choice criteria). In general, we would be having r points in a s-dimensional space. The aim of the CA is to reduce the s-dimensional space with good repoduction of "distances" among points in the original s-dimensional space.

• **INERTION AND DIMENSION REDUCTION**

The world is still stemming from the matrix of raw profiles. We make an effort to reduce the s dimension onto lower dimension with full retain of information about distances between raw points. Maximal dimension of the "new" space is $\min\{r-1, s-1\}$. In the juices example we got reduce onto the fourth dimension space without losing information. But in terms of the graphical interpretation of associations, there is no advancement. So there is a question if it is possible to erase some axis of the new coordination system and with what price. We introduce a term *inertion* I.

$$\textit{Total intertion: } I = \frac{V}{n} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(p_{ij}-q_{ij})^2}{q_{ij}}$$

demonstrates how the the matrix P differs from the matrix Q.

If we fix index i in the inertia sum, the insider sum demostrates how the i-th row contributed to the total inertia; and this value is called *total inertia of the i-th row*. For example,

the row with the highest value of the inner sum in the formula for I has destructed the independence at most. That is why corresponding variables of the X variable are in association with the Y variable at most. (there is an analogy for the variability - particular objects constribute to total variability) *Relative inertion* of the i-th row is the ratio of the i-th row's contribution to the total inertia.

In the preceding paragraph we were dealing with contribution to the total inertia of the particular raw profiles. Now we will see how particular variables (axes of the new system) contributes to the total inertia. Each axis of the new coordinate system explains part of the total intertia and the axes are ordered that the explined parts of the inertia are decresing. [1] Whther we part of the information about "resources of the dependence" obey and depict the raw points into the new system of coordinates only by using the firt two new coordinates (and depict graphically), we have to ascertain how big part of the information we are going to lose. In other words, we have a look how many per cent of the total inertia are explained by the first two coordinates of the new system. If it is more than 90%, the depiction in the 2D is reproducing the orignail distances of the raw points very well and the associations interpretation based on distances in the 2D graph is having excellent ability of demonstrating associations among categories of the X variable.

Eigenvalues and Inertia for all Dimensions (dzusy.sta)
Input Table (Rows x Columns): 5 x 5
Total Inertia=,08354 Chi2=45,948 df=16 p=,00010

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	*0,254004*	*0,064518*	*77,22819*	*77,2282*	*35,48493*
2	*0,115648*	*0,013374*	*16,00912*	*93,2373*	*7,35590*
3	0,074520	0,005553	6,64719	99,8845	3,05426
4	0,009823	0,000096	0,11549	100,0000	0,05306

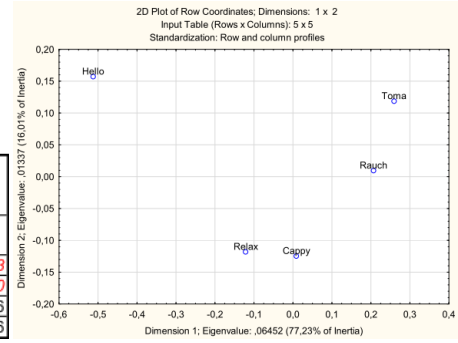


Figure 8.1: *In the table in the left we can see that the first axis of the new system of coordinates is explaining over 77% of the inertia. The second axis is explaining 16% of the inertia. So that, using the first two axes of the new system of coordinates the ratio of the explained inertia is over 93 %. Then even interpretation of the chart in the right is relevant. Should we are to merge the categories of the "brand" variable, the merge of the Relax and Cappy would not change the χ^2 value of the V statistics so much. These two brands seem not to be much different in the "choice criterium" variable association.*

Let sign k the maximal possible dimension in the new system and m the chosen dimension, $m \leq k$. *Depiction quality of the ith category* demosntrates how many per cent of the inertia of the ith row have remained explained by the depiction into m coordinated of the new system. See picture 8.2 for the next association.

Instead of the raw profiles we can stem from the columns profiles as well and analogically we will be gained the depiction of the points of the r-dimensional space in the spacce of the

[1] $I = \lambda_1 + \lambda_2 + \ldots + \lambda_k; \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ where k is number of variables (axes) and the lambas are eigenvalues of the ZZ' matrix. The explained intertion for particular axes in the new system is equal to the ordered eigenvalues.

Row Name	Row Coordinates and Contributions to Inertia (dzusy.sta) Input Table (Rows x Columns): 5 x 5 Standardization: Row and column profiles Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine2 Dim.1	Inertia Dim.2	Cosine2 Dim.2
Cappy	1	0,008629	0,124597	0,190909	0,479754	0,074301	0,000220	0,002290	0,221599	0,477465
Hello	2	0,512628	0,157271	0,140000	0,998574	0,482518	0,570232	0,912672	0,258912	0,085903
Rauch	3	0,206649	0,009405	0,174545	0,928852	0,096255	0,115530	0,926932	0,001154	0,001920
Relax	4	0,121478	0,118139	0,247273	0,802927	0,105848	0,056557	0,412648	0,258042	0,390279
Toma	5	0,259184	0,118653	0,247273	0,997610	0,241079	0,257460	0,824759	0,260293	0,172851

Figure 8.2: *Table is for $m = 2$, the row profiles (points in the 5dimensional space) are depicted into the new 2dimensinal space. Coordinates in the new system are in the first two columns of the table "Masa" demonstrates relative frequencies of the particular categories. In the "Cos2Dim1" column, there is a ratio of the inertia of the ith row that is explained by the first new axis; in the "Cos2Dim2" column, there is a ratio explained by the second new axis. These values can be interpreted as a correlation of the ith raw profile with the first (or second) new axis. The sum of these two values in the ith row can be found in the "Quality" column, which is a quality of depiction of the ith category when using two axes of the new system (analogy to the communalities in the FA). "Relative inertia" shows a relative contribution of the i-th row to the total inertia. "Inertia of the Dim.1" (or 2) shows the contribution of the particular category to the inertia of the first (or the second) new axis. We can interpret it as a relative rate of the impact of the category on the first (second) axis of the new system orientation.*

lower dimension.

• **CORRESPONDENCE MAP**

The main aim of the CA was a depiction of the raw or the column profiles into the lower dimension space and subsequent graphical depiction in 2D. If the coordinates are depicted in the graph using first two axes of the new system, we will get a *correspondence map*.

Now we have to emphasize that the coordiantes in the new system can be standardized differently. If we are interested in associations among categories of the raw variable X, we are choosing a method for *raw profiles analysis* for the calculation of the coordinates. [2]. If we are interested in the relations among categories of the column variable Y, we are choosing a method for *analysis of the column profiles* for the calculation of the coordinates.

Most often we want to evaluate the distance of categories through both variables. For this reason we are construing *symmetric correspondence maps*. Let introduce a method of *simultaneous analysis of the raw and column profiles*. In that case, there are categories of both variables in one graph, however, the coordinates for the categories of X are calculated by a method for raw profiles analysis and coordinates for the categories of Y are calculated by a method for analysis of the column profiles. Thus we are mapping two different spaces into the 2D. That is why the categories distances throughout variables**cannot** be interpreted as a rate of similarity among categories of the X and Y variables. The only sensible way is an orientation of points (both variable categories) regarding both axis of the new system of

[2]Coordinates of the column points in the new system are standardized that a sum of the squared distances from the centroid is equal to one.

coordinates. See picture 8.3.

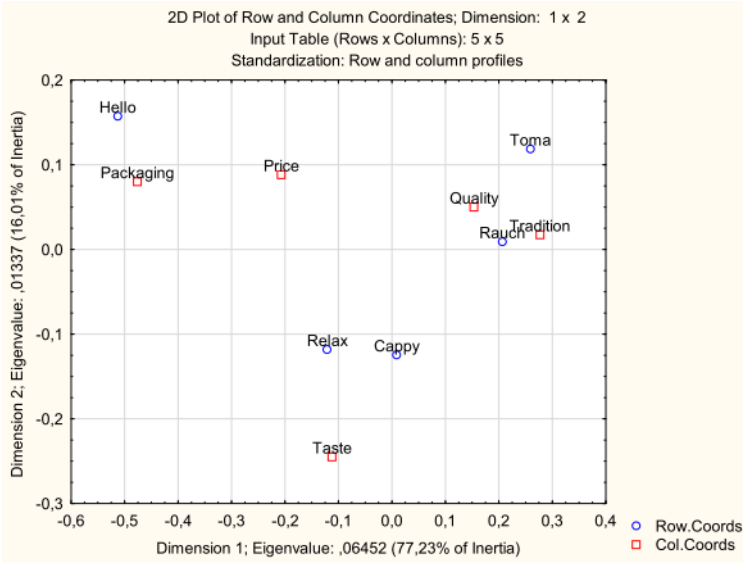


Figure 8.3: *The distance among Hello and Package or Price is not defined but, thanks to the fact that all these categories are depicted in the left domain of the horizontal axis, we can assume that majority of Hello juices is purchased due to package and price. Just in case, in the table of row profiles (tab. 8.1 in the right) there are Package and Price the highest values in the row of juice Hello. Similarly we can see that Rauch and Toma are purchased due to Trtadition and Quality at most. And the minor axis shows an association between Taste and Relax or Cappy.*

8.3 Multivariete CA

Multivariete CA generalizes the simple CA onto more variebles. Let's get back to problem 8.1 and add a new variable "gained education" to the "brand" and "choice criteria" variables of three categories: primary school, high school, college. The possibilities how analyze associations among multi-variables are following.

1. We conduct simple CA for pairs from the "primary" variable and other variables. In our case, we will analyze two tables: *brand - choice criterium* and *brand - gained education*.

2. We will conduct simple CA upon a merged table. The merged table arises that raw variable incorporates the categories of the "primary" variable and the column variable incorporates categories of all remaining variables. In our case, the raw variable incorporates the categories of the "brand" variable and the column variable incorporates categories of remaing variables "choice criterium" and "gained education".

3. We conduct simple CA upon *Burt* table. *Burt* table arises when raw and column variable incorporate categories of all variables. In our case, we will have a table of 5+5+3

raws and columns, see 8.4. Burt table is symmetric. There are marginal frequencies of original correspondence tables on the main diagonal. However, they increase the total inertia that make Burt tables less suitable ones.

	Observed Table (Frequencies) (dzusy.sta) Input Table (Rows x Columns): 13 x 13 (Burt Table)										
	Brand Cappy	Brand Hello	Brand Rauch	Brand Relax	Brand Toma	Criterion Packaging	Criterion Price	Criterion Quality	Criterion Taste	Criterion Tradition	Education Elementary
Brand:Cappy	105	0	0	0	0	13	14	27	23	28	36
Brand:Hello	0	77	0	0	0	19	22	14	12	10	30
Brand:Rauch	0	0	96	0	0	6	17	24	15	34	32
Brand:Relax	0	0	0	136	0	17	29	24	31	35	54
Brand:Toma	0	0	0	0	136	10	22	38	15	51	44
Criterion:Packag	13	19	6	17	10	65	0	0	0	0	29
Criterion:Price	14	22	17	29	22	0	104	0	0	0	51
Criterion:Quality	27	14	24	24	38	0	0	127	0	0	33
Criterion:Taste	23	12	15	31	15	0	0	0	96	0	29
Criterion:Traditio	28	10	34	35	51	0	0	0	0	158	54
Education:Eleme	36	30	32	54	44	29	51	33	29	54	196
Education:Secon	49	35	50	64	61	25	41	61	53	79	0
Education:Univer	20	12	14	18	31	11	12	33	14	25	0
Total	315	231	288	408	408	195	312	381	288	474	588

Figure 8.4: *Burt table for enriched figure 8.1*

Chapter 9

Higher order ANOVA

Higher order ANOVA is following the One-way ANOVA lecture where we were researching if one factor A influences a dependent quantitative variable Y. In other words if the values of Y are different for particular levels of A significantly. In this case we have one explaining variable (factor A) and one explained variable (Y, sometimes known as response variable), and particular factor levels of A are not random, i.e. fixed.

Nevertheless, there is a possibility that more variables influence the dependent variable, and, what is more, they can influence each other. In that case we are talking about the high order ANOVA (with or without interactions). The dependent variables can be more than one too. Whether we are researching impact of one or more factors on more variables simultaneuously, we are talking about MANOVA; Multidimensional Analysis of Variance. Furthermore, the effects of the factors can be fixed or random.

In this chapter we will be dealing with two-factor ANOVA with interactions. First of all, we will introduce a summary of mostly used design experiments.

Example 9.1. Imagine there are 18 similar squares for growing wheat. We are interested if the volume of fertilizer (factor A) have an impact on wheat crop (response Y). Factor A is of 3 levels: a lot of fertilizer, middle volume of fertilizer and little volume of fertilizer. Next we will assume weighted classification, i.e. for each level of factor will be randomly chosen 6 squares ($6 \times 3 = 18$). 18 squares are not the same, so the randomization help us to reduce influences of the unrevealed exogenous factors that can influence the crop too. □

Example 9.2. Let enrich the problem 9.1 with an assumption that the 18 squares are located on a hillside where the crop is increasing from the west to the east. But the experimenter is not having other squares disposable. While the levels of the fertilizer she can setup, the crop of the squares regarding the location she cannot setup. In this example the *randomized complete block design* is suitable where the 18 squares will be classified into e.g. six blocks regarding to the hill location and in each block of 3 very similar squares we will use a different amount of fertilizer factor.The choice of the square from the each threesome for levels of fertilizer is random again.

In this case the crop is dependent on two factors: volume of fertilizer and the block. In fact, we are interested only in factor of fertilizer and the block part is enforced in order to be part of the variability explained. □

The block classification is used when we expect an impact of variables whose values we cannot influence. Within one block the values of Y are being given in relatively homogenous conditions and the differences in values are mainly caused by different levels of the factor A. The best case scenario is when we can use all the levels of A within one block as in our problem 9.2.

Now we will be dealing with higher order design where we assume interactions among factors during influencing the Y variable.

Example 9.3. Let's get back to the 9.1. The dependent variable is still Y "wheat crop" but now we will assume even two factors influencing the crop. We will add a new factor "volume of irrigation" that can be of two levels: little water, a lot of water. Since the fertilizer effect can be influenced by the rate of irrigation, we have 6 groups totally in which we conduct a measurement of the variable "wheat crop ": group "a lot of fertilizer, little water", "middle volume of fertilizer, little water", ..., "little fertilizer, a lot of water". If there is the same number $n \geq 2$ of measuring Y variable, we are dealing with *factorial design* or *crossed design*. The number of response Y measurements total $N = 6 \cdot n$. □

If the respective groups are represented by normal random samples that are mutually independent, we can test hypothesis about significance of particular factors and interactions.

9.1 Factor design

We introduce a labeling for two factors:

•"raw" factor A contains R levels and symbol r signs particular levels of the A factor; $r = 1, \ldots, R$.
•"column" factor B contains S levels and symbol s signs particular levels of B variable; $s = 1, \ldots, S$.
•symbol n_{rs} signs number of observations of the Y variable for the rth level of A factor and the sth level of B factor. Since we require the weighted classification in the factor design, we sign $n = n_{rs}$ for all combinations of levels of both factors.
•symbol N signs number of observations of the Y variable; and regarding the weighted classification $N = RSn$.
•symbol y_{irs} signs value of the ith observation of the Y variable in the rsth cell; $i = 1, \ldots, n$ for the weighted classification (generally $i = 1, \ldots, n_{rs}$).

Example 9.4. We enrich the assignment of the problem 9.3 by stating $n = 10$. Factor A is the "irrigation volume" and this factor contains $R = 2$ levels. Factor B is "fertilizer volume" and this factor contains $S = 3$ levels. Totally we are having $R \cdot S = 6$ groups in which there are always 10 measurements of wheat crop. The groups are mutually different regarding the two factors' levels and thus their effects on Y. Total number of measurements is $N = 60$. The results can be seen in the table 9.1. □

Yield	little fertilizer	middle vol. fertilizer	a lot of fertilizer
little water	52	28	15
	48	35	14
	43	34	23
	50	32	21
	43	34	14
	44	27	20
	46	31	21
	46	27	16
	43	29	20
	49	25	14
a lot of water	38	43	23
	42	34	25
	42	33	18
	35	42	26
	33	41	18
	38	37	26
	39	37	20
	34	40	19
	33	36	22
	34	35	17

Table 9.1: *measured crop*

9.2 Sample effects in the factor design

First of all let us remind the concept of sample effects in ANOVA.

Let sign the mean of the measured values of Y for the rth level of A factor by a symbol $m_{r\cdot}$ and the total mean of the all measured values by a symbol $m_{\cdot\cdot}$. Then the difference $a_r = m_{r\cdot} - m_{\cdot\cdot}$ demonstrates an effect of the rth factor level of A on the values of Y variable. Furhermore, if we are considering a weighted classification (so that for n_r observations of Y variable in the rth group is valid $n_r = n$), and the condition $\sum_{r=1}^{R} n_r(m_{r\cdot} - m_{\cdot\cdot}) = 0$ can be rewritten as $\sum_{r=1}^{R} na_r = 0$, thus $\sum_{r=1}^{R} a_r = 0$.

Statistics S_A (sum of the squares between groups) characterizing the variability between groups can be rewritten as:
$S_A = \sum_{r=1}^{R} n \cdot a_r^2$
a_r is an unbiased estimate of the theoretical (population) effect α_r which describes the effect of the A factor in the rth sample (population).

Now let's get back to the factor design with the two or factors where we will distinguish more types of effects: *main effects, interaction effects* and *effects of the cells*. These effects are described by following means:

symbol $m_{..}$ signs mean of all N measured values of the Y variable
symbol $m_{r.}$ signs mean of all values of the Y variable measured during the rth level of A factor. (regardless a level of the B factor.)
symbol $m_{.s}$ sigs mean of the all measured values of the Y variable measured during the sth level of the B factor.
symbol m_{rs} signs mean of the all values of the Y variable measured during the rth level of the A factor and the sth level of the B factor.
For the example 9.4 all the mentioned means are in the table 9.2

m_{rs}	B_1	B_2	B_3	$m_{r.}$
A_1	46,4	30,2	17,8	31,4667
A_2	36,8	37,8	21,4	32
$m_{.s}$	41,6	34	19,6	31,7333

Table 9.2: *For the data from the table 9.1 we introduce means of the values of Y measured during all levels of the row factors, column factors and cells; total mean $m_{..} = 31,7333$. Symbol A_1 signs the first level of the factor A,...*

• CELLS EFFECTS:
The sense of the cell effect is clear when we realize that particular cells represent $R \cdot S$ independent groups of observation of the Y variable, where a different combination of the factor levels affects the Y. Thus the effect on particular observations of Y is determined by the citizenship towards particular group. The sample effect of the rsth cell is signed as $[ab]_{rs}$. Its value shows how the mean of the rs th cell m_{rs} is different from the total mean $m_{..}$ Thus $[ab]_{rs} = m_{rs} - m_{..}$
It is clear that the bigger the value of $[ab]_{rs}$ the more significant effect of the rsth combination of the factor levels on the Y values is. What is more, the effects also meet a condition[1]
$\sum_{r=1}^{R} \sum_{s=1}^{S} [ab]_{rs} = 0.$
The cells effects in the problem 9.4 are showed in the table 9.3.

The systematic part of the total variance of Y (the part that can be explained by the factors), is described by the cell effects. In other words, if the cell effects differ, the main row or column vector or the effects interaction or their combinations have an impact on the values

[1] this condition is valid also for unweighted classifications

$[ab]_{rs}$	B_1	B_2	B_3	a_r
A_1	14,6667	-1,5333	-13,9333	-0,2667
A_2	5,0667	6,0667	-10,3333	0,2667
b_s	9,8667	2,2667	-12,1333	0

Table 9.3: *We feature the cell effects and the main effects for the data from the table 9.1: a_r is the effect of the rth level of the A factor; b_s is the effect of the sth level of the factor B; $[ab]_{rs}$ is the effect of the rsth cell.*
E.g. $14,6667 = m_{11} - m_{..} = 46,4 - 31,7333$

of Y. Thus we introduce a statistics S_{AB} that characterize the variability between all groups.

$$S_{AB} = \sum_{r=1}^{R} \sum_{s=1}^{S} n \cdot [ab]_{rs}^2$$

• MAIN EFFECTS:
Let introduce an interpretation of effect of the row factor A. Analogically for the column factor B.
If we are interested in the question how particular levels of the A factor influence the values of the Y variable regardless the levels of B, we are talking about the *main effect* of the row factor A. This effect shows how the values of Y variable are deviated for the rth level of the A factor "deviated" from the common mean $m_{..}$ and can be described as $a_r = m_{r.} - m_{..}$
Again, there is a condition of $\sum_{r=1}^{R} a_r = 0$.

Since we do not care about the factor B, the total number of observations for the rth level of the A factor is Sn. Thus the statistics describing the part of the total variability of the Y variable is explained by an impact of different levels of the A factor regardless levels of B is

$$S_A = \sum_{r=1}^{R} Sn \cdot (a_r)^2$$

(Analogically for levels of the column factor)
• INTERACTION EFFECTS:
The particular values of both factors can influence each other. This effect is not incorporated either in the rth level of the row factor not in the sth level of the column factor. Thus we can describe the effect of the cell as a result of the main effects added to interaction effect.

$$\begin{aligned} (ab)_{rs} &= [ab]_{rs} - a_r - b_s = \\ &= (m_{rs} - m_{..}) - (m_{r.} - m_{..}) - (m_{.s} - m_{..}) = \\ &= m_{rs} - m_{r.} - m_{.s} + m_{..} \end{aligned}$$

Formally in this way:
Again, we have to meet conditions $\sum_{r=1}^{R} (ab)_{rs} = 0$; $\sum_{s=1}^{S} (ab)_{rs} = 0$; $\sum_{r=1}^{R} \sum_{s=1}^{S} (ab)_{rs} = 0$. The values of the sample effects of interactions for the problem 9.4 are in the table 9.4.

The explained part of the total variability of Y by the interactions of different factor levels

$(ab)_{rs}$	B_1	B_2	B_3	
A_1	5,0667	-3,5334	-1,53343	0
A_2	-5,0667	3,5334	1,53343	0
	0	0	0	0

Table 9.4: *We introduce interaction effects from the table 9.1 that are derived from the table 9.3.*

of A and B is signed

$$S_{A\times B} = \sum_{r=1}^{R}\sum_{s=1}^{S} n \cdot (ab)_{rs}^2$$

$$\boxed{y_{irs} = m_{..} + a_r + b_s + (ab)_{rs} + e_{irs}},$$

where e_{irs} is a realization of the random error (residuum) of the ith object in the rsth cell, $e_{irs} = y_{irs} - m_{rs}$. This model can be written as

$$y_{irs} = m_{..} + [ab]_{rs} + e_{irs},$$

and is equivalent with the model of the one way ANOVA for RS independent random samples.

9.3 Variability analysis in the factor design

Besides least squares, there are also squares for the total variability and for the variability of particular groups

$S_T = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{n}(y_{irs} - m_{..})^2$ total sum of the squares

describing the total variability of the Y variability. (statistics S_T has $f_T = N - 1$ degrees of freedom)

$S_E = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{n} e_{irs}^2$ sum of the residual squares

describing the variability within particular groups (cells) (statistics S_E has $f_E = RS(n-1) = N - RS$ degrees of freedom)

The sums of the squares:

$S_{AB} = \sum_{r=1}^{R} \sum_{s=1}^{S} n \cdot [ab]_{rs}^2$ sum of the squares between groups
describing variability between all groups

$S_A = \sum_{r=1}^{R} Sn \cdot (a_r)^2$ sum of the squares describing the part of the total variability Y
that is caused by factor A impact only.
(statistics S_A has $f_A = R - 1$ degrees of freedom)

$S_B = \sum_{s=1}^{S} Rn \cdot (b_s)^2$ sum of the squares describing the part of the total variability Y,
that is caused by factor B impact only.
(statistics S_B has $f_B = S - 1$ degrees of freedom)

$S_{A\times B} = \sum_{r=1}^{R} \sum_{s=1}^{S} n \cdot (ab)_{rs}^2$ sum of squares describing the part of the total variability Y that is
caused by impact of intercations between levels of A and B.
(statistics $S_{A\times B}$ has $f_{A\times B} = (R-1)(S-1)$ degrees of freedom)

As in ANOVA the total variability of Y can be analyzed onto variability "within" particular groups and variability "between" groups, thus:

$$S_T = S_{AB} + S_E$$

Furthermore, the variability between groups S_{AB} can be further analyzed as:

$$S_{AB} = S_A + S_B + S_{A\times B}$$

Thus the variability between groups can be explained by the factors' A and B impacts and by the impact of the interactions between factors A and B. Analysis of the total variance of the Y variable can be described as:

$$\boxed{S_T = S_A + S_B + S_{A\times B} + S_E}$$

All is derived in the *Rozklad_Komentar.PDF* file.

9.4 Theoretical effects and hypothesis tests

For a random variable Y_{irs} we use a model:

$$\boxed{Y_{irs} = \mu + \alpha_r + \beta_s + (\alpha\beta)_{rs} + \varepsilon_{irs}}$$

Y_{irs} random variable stemming from the population corresponding with the rth level of the A factor and the sth level of $i = 1, \ldots, n$

μ common part of the expected value of the random variable Y.

α_r effect of the rth level of the A factor, $r = 1, \ldots, R$.

β_s effect of the sth level of the B factor, $s = 1, \ldots, S$.

$(\alpha\beta)_{rs}$ effect of the interaction of the rth level of the A factor and the sth level of the B factor, $r = 1, \ldots, R$, $s = 1, \ldots, S$

ε_{irs} stochastically independent (within the rsth population and between populations) random variables with distribution

Unbiased estimates of the parameters $\mu,\ \alpha_r,\ \beta_s,\ (\alpha\beta)_{rs}$ are (in sequence) $m..,\ a_r,\ b_s,\ (ab)_{rs}$.
The following equation is valid for a weighted classification:
$\sum_{r=1}^{R} \alpha_r = 0;\ \sum_{s=1}^{S} \beta_s = 0;\ \sum_{r=1}^{R} (\alpha\beta)_{rs} = 0;\ \sum_{s=1}^{S} (\alpha\beta)_{rs} = 0; \quad \sum_{r=1}^{R} \sum_{s=1}^{S} (\alpha\beta)_{rs} = 0;$
For sums of the squares is valid:

1. $E(\frac{S_E}{RS(n-1)}) = \sigma^2$
   Regardless significance of the main effects or interactions, the sum of squares within the group divided corresponding degree of freedom number
   $f_E = RS(n-1) = N - RS$ represents an unbiased estimate of the random error variability ε.

2. $E(\frac{S_A}{R-1}) \geq \sigma^2$
   We test a hypothesis that a row factor A does not have an impact on Y facing an alternative hypothesis that for at least one level of factor A the row main effect is significant.
   So that $H_0:\ \alpha_1 = \ldots = \alpha_R = 0$ against $H_1:\ \alpha_r \neq 0$ for at least one $r = 1, \ldots, R$.
   As a_r is an unbiased estimate of α_r, the S_A statistics is relevant for evaluating of the null hypothesis.
   If the null hypothesis is true, then $E(\frac{S_A}{R-1}) = \sigma^2$. Thus we are dealing with another unbiased estimate of the parameter σ^2 which is independent on a sum of squares between the groups estimate above that. We are using a statistics for testing
   $F_A = \frac{S_A/f_A}{S_E/f_E} \sim F((R-1),(N-RS))$. Big values of the F_A statistics are in favor of the alternative hypothesis.

3. $E(\frac{S_B}{S-1}) \geq \sigma^2$
   We are testing a hypothesis that a column factor B does not have an effect on Y against alternative that for at least one level of factor B the column factor is significant.
   So that $H_0:\ \beta_1 = \ldots = \beta_S = 0$ against $H_1:\ \beta_s \neq 0$ for at least one $s = 1, \ldots, S$.
   If the null hypothesis should be true, then $E(\frac{S_B}{S-1}) = \sigma^2$, thus we would be dealing with another unbiased estimate of parameter σ^2 which is independent on an estimate by the sum of squares withing the groups besides. We are using a statistics for testing
   $F_B = \frac{S_B/f_B}{S_E/f_E} \sim F((S-1),(N-RS))$. Big numbers of F_A statistics are against the null hypothesis.

4. $E(\frac{S_{A\times B}}{(R-1)(S-1)}) \geq \sigma^2$
   We are testing a hypothesis that interactions between factors A and B do not have an effect on Y variable against alternative that for at least one pair of levels r, s the interaction effect is significant.
   So that $H_0:\ (\alpha\beta)_{rs} = 0$ for all pairs of r, s against $H_1: (\alpha\beta)_{rs} \neq 0$ for at least one pair of r, s; $s = 1, \ldots, S$, $r = 1, \ldots, R$.
   If the null hypothesis were to be true, then $E(\frac{S_{A\times B}}{(R-1)(S-1)}) = \sigma^2$ so we would be dealing with another unbiased estimate of parameter σ^2 that is independent on an estimate by the sum of squares of the within groups furthermore. For testing we are using a statistics

$F_{A\times B} = \frac{S_{A\times B}/f_{A\times B}}{S_E/f_E} \sim F((R-1)(S-1),(N-RS))$. Big values of $F_{A\times B}$ statistics are against the null hypothesis. [2].

From the upper-mentioned can be seen that it is possible to conduct separated tests about insignificance of the row factor A, column factor B and insignificancy of interactions between A and B. The findings are usual to be written into a table:

Variability source	sum of squares	degrees of freedom	mean sum of squares	test statistics
factor A	S_A	$f_A = R-1$	S_A/f_A	$F_A = \frac{S_A/f_A}{S_E/f_E}$
factor B	S_B	$f_B = S-1$	S_B/f_B	$F_B = \frac{S_B/f_B}{S_E/f_E}$
factor of interactions	$S_{A\times B}$	$f_{A\times B} = (R-1)(S-1)$	$S_{A\times B}/f_{A\times B}$	$F_{A\times B} = \frac{S_{A\times B}/f_{A\times B}}{S_E/f_E}$
residuals	S_E	$f_E = N-RS$	S_E/f_E	
total	S_T	$f_T = N-1$		

9.5 Contrast and methods of high ordered comparing

Let sign $\mu_{r\cdot}$ expected value of the population, corresponding with the rth level of the A factor; $\mu_{\cdot s}$ expected value of the population corresponding with the sth level of the B factor and μ_{rs} as an expected value of population corresponding with the rth factor level of A and the sth level of B factor. If the hypothesis tests showed that the factor A is significant, there is a question which pairs of A levels are significantly different in terms of their impact on Y. In other words, we would be interested in results of separated tests of hypothesis
$H_0 : \mu_{r\cdot} = \mu_{l\cdot}$ for $r \neq l,\ r = 1,\dots R,\ l = 1,\dots R.$ If the factor A is significant, probably at least one pair of expected values $\mu_{r\cdot}, \mu_{l\cdot}$ should differ. However, separated tests do not have to show a significant difference related to all pairs. In this case the significance of the factor A has been caused by another linear combination of expected values which is called a *contrast* and signed ψ.

$$\psi = c_1\mu_{1\cdot} + \ldots + c_R\mu_{R\cdot}$$

and for coefficients c_r is valid $\sum_{r=1}^{R} c_r = 0$ and $\sum_{r=1}^{R} c_r^2 > 0$.
An unbiased estimate of the contrast is a linear combination $\hat{\psi}$ of sample means.

$$\hat{\psi} = c_1 M_{1\cdot} + \ldots + c_R M_{R\cdot}$$

If we are intered in the separated tests $H_0 : \mu_{r\cdot} = \mu_{l\cdot}$, it is sufficient to state $c_r = 1$, $c_l = -1$ and other constants c remain zero.

Whether we want to test more contrasts at the same time and remain simultaneous level of α, we are using *methods of high order comparing*. Tyhey are usually presented in a way that testing
$H_0 : \psi = 0$ is conducted by a confidence interval for a contrast ψ. If the null is outside the confidence interval, the hypothesis H_0 would be turned down and the contrast is significant.

[2] For quick assessment of interaction impacts, the chart of integrations is useful. If the "curves are crossing" or their slopes are clearly different, the interaction effect is clear to be significant.

In the statistical softwares, the most common are Bonferron method, Tukeyho method and Scheffé method. Details about Scheffé method are in the *Scheffe_Komentar.PDF* file. This method is most universal as it can be used even for unweighted classifications and is less sensitive to breaking normality and variability homogeneity.

9.6 Multifactor ANOVA

We briefly outline a factor design for three or more factors. Let assume three factors A, B, C where factor A has R levels, factor B has S levels, factor C has T levels and for each combination of r, s, t we have the same number of observations $n \geq 2$ thus total number of observations is $N = nRST$. We sign Y_{irst} the ith observation of the dependent variable Y in the rstth group. Then the model for Y looks like this:

$$Y_{irst} = \mu + \alpha_r + \beta_s + \gamma_t + (\alpha\beta)_{rs} + (\alpha\gamma)_{rt} + (\beta\gamma)_{st} + (\alpha\beta\gamma)_{rst} + \varepsilon_{irst}$$

where $(\alpha\beta), (\alpha\gamma), (\beta\gamma)$ represent interactions of the first order and the $(\alpha\beta\gamma)$ represents interactions of the second order.

The higher number of factors leads to harder requirements about number of observations. (For 10 factors with two levels there is a need of at least $n \cdot 2^{10}$ observations where$n \geq 2$.)

9.7 Generalized linear model and ANOVA

See *GLM_ANOVA.PDF*.

Chapter 10

Survival Analysis

Survival Analysis[1] is a set of statistical procedures that analyzes a behavior of a random variable *duration of survival*. This variable respresents time since the beginning of an observation of a object till an event. (Not be confused by terms like a "failure", "recidivism", "stop paying debt", "reaction to an offer" or "death" in the text meaning that event.) The beginning of an observation: a disease outbreak; begin of a medical treatment; marriage; release day of a prisoner; entrance of a patient into a study; begin of a tool functioning etc.

Why standard statistical methods are insufficient for behavior analysis of a "duration of a survival" variable? There is a problem with the "censored" data. Censoring is when a particular object does not hold full information about its survival duration.

For example, if an exploration of patients ended before a patient died, we do not exact time of the patient's survival. Moreover, she can stop attending the physician and we do not know how long does she live after her operation.

The reasons for censoring follow:

- We have not noticed the particular event.
- We have lost a connection with the individual.
- An individual has been removed from the exploration from some reason.

A graphical illustration of the upper-mentioned reasons can be seen in the picture 10.1. We assign two values to each object; the first is survival duration, the latter represent whether the first is censored or the real one. There is a graphical notation of a cross or a dot in the picture.

All censored durations of survival in the picture 10.1 are shorter than the real one in fact. This type of censoring is called the *censoring from the right*. Rarely do we deal with censoring from the left. Let have a group of drug addicts and we are measuring time to HIV outbreak. If the blood tests prove HIV, we do know that a patient is an HIV positive but do not know when the disease outbroke so the real time of "waiting for an outbreak" was shorter than the time to the positive blood test. In that case we are dealing with censoring from the left.

[1] The term "survival analysis" is widely used in biostatistics; in economics and sociology, there is often used the term "duration analysis"; and in engineering is often used term "reliability analysis"; all this is the same

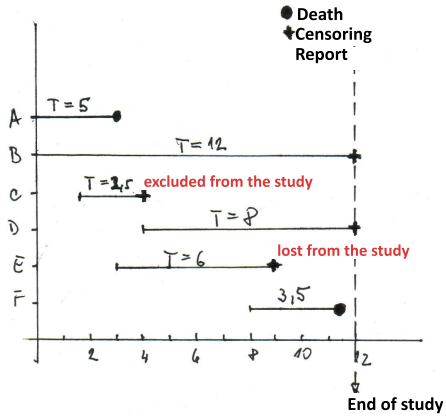


Figure 10.1: *Out of 6 observed persons, the persons A and F died, other values are censored.*

Following we will only assume censoring from the right; and all statistical methods assume so called *uninformative censoring* that means reasons for censoring do not have any association with the event we are researching.

10.1 Terminology and comments

T $*$ *survival duration*
A continuous random variable representing time since the beginning of the observation to default (or censored point in time).
t signs a realization of a T random variable.
holds non-negativity.
Probability distribution of this random variable is represented by a density $f(t)$ and distribution function $F(t)$.

δ $*$ *default indicator*
An alternative random variable that encodes "default" with value of 1 and "censoring" with value of 0.
$\delta = 1$ represent that there was a default in case of an individual;
$\delta = 0$ demostrates that the survival time is censored.

$S(t)$ $*$ *survival function*
$P(T > t)$; a probability that an individual survive the t point in time.
$S(t) = 1 - F(t)$
is cintinuous and not-increasing, and
$S(0) = 1$; in the beggining of the observation (study) in the $t = 0$ point in time all individuals are alive, so that a probability of a survival in the 0 point in time is equal to 1.
$\lim_{t\to\infty} S(t) = 0$; should the observation takes infinity time, all the individuals dead before the study termination.

$\hat{S}(t)$ $*$ *survival funtion estimate*
how to obtain will be explained in the section 10.4
Since a duration of the study is not infinite, all the individuals canot be dead when the study t_x ends so that $\hat{S}(t_x) \geq 0$. For illustration see picture 10.2.

$h(t)$ $*$*risk function*
$h(t) = \lim_{dt\to 0} \frac{P(t\leq T<t+dt \mid T\geq t)}{dt}$
Risk function is often interpreted as a "probability" of a default in the t point in time, which is wrong.
We'd rather talk about a "conditional risk of default rate in the t point in time, and a condition is that an individiual survived the t point in time.
Or simply: if an individual survived t, then $h(t)$ demonstrates a "tend to death/default right after t.
holds non-negativity and is not upper restricted.
Risk function values are dependent on time units. See problem 10.1.

Both functions describe a behavior of the same random variable T, which is survival duration. But their interpretations are different. Unlike survival function $S(t)$ deals with a probability of surviving t - that is probability of **not defaulting**, the risk function deals with **default**. Both functions are tied with following relations. That is why our use of one of the functions does not matter, both functions can be derived from the latter.

For the risk function: $h(t) = -\frac{1}{S(t)} \frac{\partial(S(t))}{\partial t}$. For survival function: $S(t) = \exp\{-\int_0^t h(u)du\}$

Both formulas are derived in *S(t)_h(t).PDF*.

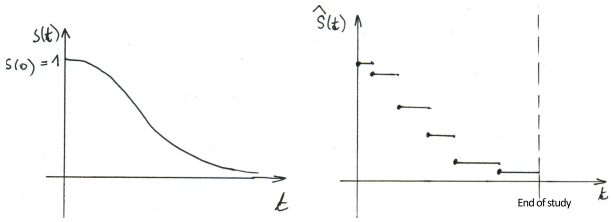
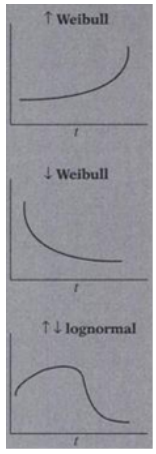


Figure 10.2: *Survival function $S(t)$ is not increasing, continuous, for $t = 0$ valids $S(t) = 1$ and its value limits zero with increasing t. Survival function estimate $\hat{S}(t)$ is not-increasing stairway function, $\hat{S}(0)$=1, but in the end of the study the function value can be not-zero).*

Example 10.1. Let $P(t \leq T < t + dt \mid T \geq t) = 1/3$ and $dt = 1/2$ day, which is the same as 1/14 week.
If we measure time in days, the ratio $\frac{P(t \leq T < t+dt \mid T \geq t)}{dt} = \frac{1/3}{1/2} = 0,67$
If we measure time in weeks. the ratio $\frac{P(t \leq T < t+dt \mid T \geq t)}{dt} = \frac{1/3}{1/14} = 4,67$
It is clear that the values of the risk function derived from the upper-mentioned ratio depends on units. □

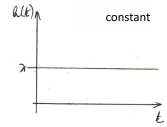
Now we introduce examples of different graphs of risk functions $h(t)$ - see picture 10.3 and survival functions $S(t)$ - see picture 10.4. Knowledge of the risk function graph enables us evaluate risk of death in whatever point in time t, and what is more, help us to identify a suitable type of mathematical model for the particular survival time. An interpretation of the survival function graph is straightforward. Its function values represent a probability of surviving t of a particular object from the population.



1.) $h(t)$ for patients with leukemia

2.) $h(t)$ for patients after a chirurgy operation

3.) $h(t)$ for patients with TBC



4.) $h(t)$ for healthy individuals

Figure 10.3: *1.) an* ***increasing Weibull model*** *where the risk of death increase with increasing time. This risk function describes duration of survival (default is a death) of leukemya patients that do not react to the care. 2.) a* ***decreasing Weibull model****, where the risk of dead decreases with increasing time. This risk function can be expected in cases after an operation. The risk of death is highest right after the operation and then decreases. 3.) a* ***lognormal model****, where the risk of death is incresing in the beggining and then start to decrease from a point in time. This can be expected in cases of TBC patients where the risk of death increases after the disease outbreak but if a patient survives a point in time after this, the risk of death starts to decrease. 4.) an* ***exponential model*** *where the risk of death is constant. This is expected in cases when patients are healthy for the all experiment duration long.*

From the pictures 10.3 and 10.4 is clear that knowledge of $h(t)$ and $S(t)$ graphs enable us interpret behavior of a random variable "survival duration". Next questions we can ask are:

• **SURVIVAL ANALYSIS OBJECTIVES:**

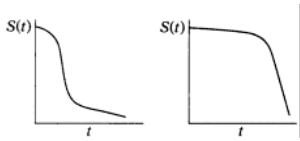


Figure 10.4: *The survival function in the left is dramatically decresing initially, then decreasing slightly. On the other hand, the survival function is almost constant for tiny t but for big t is dramatically decreasing.*

- **i.** Estimating and interpreting survival function and risk function.
- **ii.** Comparing survival functions and risk functions concerning two or more groups. (see picture 10.5.)
- **iii.** Modeling of relation between survival time and explaining variables. (We will be dealing with Cox model in the section 10.5.)

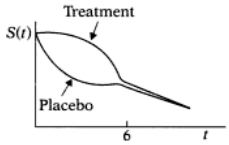


Figure 10.5: *In the picture, we can compare survival function graphs of two patients with the same desease. Concerning the one group a care was conducted and concerning control group a placebo was given. Obviously, a care increases probability of surviving in the first 6 weeks, then these groups are almost indifferent.*

10.2 Dataset assignment standards

We can assign the dataset in two ways. The first is suitable for computer processing. But for comprehension of the survival analysis priciples the latter way is better. We will show using of both ways on the data in the problem 10.2.

Example 10.2. Concerning group of $n = 42$ leukemia patients we are observing time of remision in weeks. A group of 21 patients was given a special care, the second group of 21 patients was given placebo. The sign + signs censoring.
The group I with a special care:
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+ The group II with a placebo:
1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

• **ASSIGMENT USING INDICATOR δ:**
It is suitable to assing the data in the form set in the table 10.1 in the left for computer processing. The t_i variable demonstrates survival time related to the ith object. (Index i related to the observed survival time t_i is without brackets). Variable δ_i represents that survival time of the

ith object was censored or there was death related to the ith object. It valids $\sum_{i=1}^{n} \delta_i =$ *number of defaults in the group of all objects*. Variables $X_1, \ldots, X_p$ are other following variables (regressors) which can influence the survival time.

The remision duration related to the first and the tenth patient with care in the problem 10.2 was 6 weeks. However, related to the first patient, there was a default (back to the disease) whereas the remision of the tenth patient is censored. Proto $t_1 = 6;\ \delta_1 = 1$; $t_{10} = 6;\ \delta_{10} = 0$. □

objekt	t	δ	X_1	X_2	...	X_p
1	t_1	δ_1	X_{11}	X_{12}	...	X_{1p}
2	t_2	δ_2	X_{21}	X_{22}	...	X_{2p}
⋮						
n	t_n	δ_n	X_{n1}	X_{n2}	...	X_{np}

$t_{(j)}$	m_j	q_j	n_j
$t_{(0)} = 0$	$m_0 = 0$	q_0	n_0
$t_{(1)}$	m_1	q_1	n_1
$t_{(2)}$	m_2	q_2	n_2
⋮			
$t_{(k)}$	m_k	q_k	n_k

Table 10.1: *In the left: assignment using δ. In the right: assignment using time of default.*

• ASSIGMENT USING TIME OF DEFAULT:
Pro porozumění modelům v analýze přežití je vhodné zadávat data ve formě, která je symbolicky uvedena v tabulce 10.1 vpravo. We use only those points in time for comprehension when there was a default/death. We order those points in time increasingly and assign them $t_{(1)} < t_{(2)} < \ldots < t_{(k)}, k \leq n$. (Indices are in brackets.) If there were more defaults in the same point in time, we will introduce this point in time of the increasing order only once. We add $t_{(0)}$ to the beginning of the order. Then •variable $t_{(j)}$ represents point in time when there was at least one default. •Variable m_j represents how many defaults in time $t_{(j)}$ and obviously $m_0 = 0$. Next $\sum_{j=1}^{k} m_j =$ *number of defaults in the group of all objects*. •Variable q_j represents how many censors in interval $\langle t_{(j)}; t_{(j+1)})$. This is number of objects that are known to be alive in time $t_{(j)}$ but are unknown in time $t_{(j+1)}$. So that $q_0 \geq 0$ some objects can be deleted within interval $\langle t_{(0)}; t_{(1)})$. The last •variable n_j represents *number of objects in "risk"* in time $t_{(j)}$. This is number of objects whose survival duration $\geq t_{(j)}$. There are also objects that died in $t_{(j)}$ included in this group. That is why $n_j = n_{j-1} - (m_{j-1} + q_{j-1})$. In the table 10.2, there is an example of both ways of dataset assigment for the group of leukemia patients from group I. □

• CLASSIC APPROACH VS. SURVIVAL ANALYSIS
What is behind the censoring principle? If we have objects that has become bank during the study in the dataset, we have three possibilities to handle.

i. Censored observations will be deleted from the dataset.
ii. Censored observations will consider as defaults.
iii. We take censored values into account.

The principle of censoring handles taking advantage of incomplete information about objects that become blank, however, a piece of information bears about T. If we would delete these objects, the partial would be lost or misrepresent the world we are researching.

	1 t_i	2 delta_i	3 skupina
1	6	1	1
2	6	1	1
3	6	1	1
4	7	1	1
5	10	1	1
6	13	1	1
7	16	1	1
8	22	1	1
9	23	1	1
10	6	0	1
11	9	0	1
12	10	0	1
13	11	0	1
14	17	0	1
15	19	0	1
16	20	0	1
17	25	0	1
18	32	0	1
19	32	0	1
20	34	0	1
21	35	0	1

$t_{(j)}$	m_j	q_j	n_j - number of people in risk
$t_{(0)} = 0$	$m_0 = 0$	$q_0 = 0$	For 21 people, the survival duration ≥ 0 weeks
$t_{(1)} = 6$	$m_1 = 3$	$q_1 = 1$	For 21 people, the survival duration ≥ 6 weeks
$t_{(2)} = 7$	$m_2 = 1$	$q_2 = 1$	For 17 people, the survival duration ≥ 7 weeks
$t_{(3)} = 10$	$m_3 = 1$	$q_3 = 2$	For 15 people, the survival duration ≥ 10 weeks
$t_{(4)} = 13$	$m_4 = 1$	$q_4 = 0$	For 12 people, the survival duration ≥ 13 weeks
$t_{(5)} = 16$	$m_5 = 1$	$q_5 = 3$	For 11 people, the survival duration ≥ 16 weeks
$t_{(6)} = 22$	$m_6 = 1$	$q_6 = 0$	For 7 people, the survival duration ≥ 22 weeks
$t_{(7)} = 23$	$m_7 = 1$	$q_7 = 5$	For 6 people, the survival duration ≥ 23 weeks

Table 10.2: *10.2 There is a table with an assignment using an indicator δ in the left and a table with a sign using points in time of a default in a group of patients.*

.

10.3 Descriptional statistics in the survival analysis

Since survival duration is a quantitative variable, the mean of survival time is a sensible variable. It is calculated as a classic arithmentic *mean* from the all times of survival $t_i, i = 1, \ldots, n$, and we do not distinguish whether the u ith objects scored default or censoring. The final number is underestimated in fact since the censored times of surviving can be actually shorter than the real ones.
Another statistics can be a
mean risk $\boxed{\bar{h} = \frac{\sum_{j=1}^{k} m_j}{\sum_{i=1}^{n} t_i}}$.
n is number of objects in the dataset; $t_i,\ i = 1, \ldots, n$ is time for surviving for the ith object; k is number of times when default (taken into consideration only once); $\sum_{j=1}^{k} m_j$ is total number of events observed suring the experiment. The higher the value of the mean risk. the smaller the chance of survining of the objects in the population.

Up to now, we have been comparing both groups using only quantitative variables. Comparing them for any time t can be provided by estimates of survival function that will be delat in section 10.4. If we have an estiamte of survival function, we can determine a *median of the survival duration*. Median is defined as the point in time t with $P(T > t) = 0,5$. For data from 10.2, there are estimates of survival function and medians for each group in the picture 10.6.

10.4 Kaplan-Meier estimates of survival function and the log-rank test

The survival function $S(t)$ is usually unknown and is being estimated fro the dataset. When we have censored data, estimating of a probability $P(T > t_{(j)})$ with this ratio would be possible:

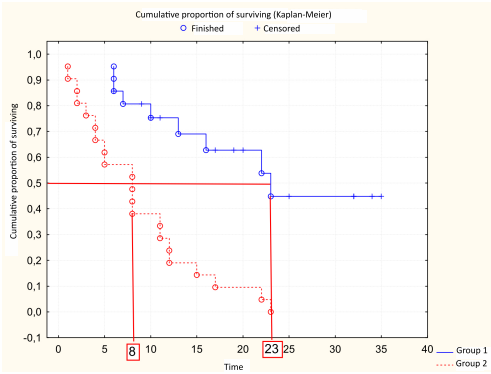


Figure 10.6: *Estimate of survival function $\hat{S}(t)$ for a group with a care and for groups with placebo. Function values for group with a care are higher than for group of placebo for any t. Median of survival duration for group with a care is 23, for group of placebo is 8. So even on a basis of medians the fact that care is more effective than placebo is apparent. Furthermore, regarding the survival function graphs we can say that the differences become higher and higher with growt in time.*

$$\frac{\text{number of objects with } T>t_{(j)}}{\text{number of all objects in the dataset}} = \frac{n_j - m_j}{n},\ j = 1, \ldots, k.$$

Should we have a look at the dta from the problem 10.1, we can notice that in the second group with placebo was no object censored. So that we can calculate the estimate of survival function, which we sign $\hat{S}(t)$, as follows - see table 10.3.

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	21/21=1,00
1	21	2	0	19/21= 0,90
2	19	2	0	17/21= 0,81
3	17	1	0	16/21=0,76
4	16	2	0	14/21=0,67
5	14	2	0	12/21=0,57
8	12	4	0	8/21=0,38
11	8	2	0	6/21=0,29
12	6	2	0	4/21=0,19
15	4	1	0	3/21=0,14
17	3	1	0	2/21=0,10
22	2	1	0	1/21=0,05
23	1	1	0	0/21=0,00

Table 10.3: *The estimate of the survival function $\hat{S}(t_{(j)})$ for group II with placebo: in this group, there was no censoring. There are estimates of survivals for times with an event in the table only. Between these times the function is constant and $\hat{S}(t) = \hat{S}(t_{(j)})$ pro $t \in \langle t_{(j)}; t_{(j+1)})$.*

• KAPLAN-MEIER ESTIMATE OF SURVIVAL FUNCTION

This procedure cannot be used when we are having censored data. For that other procedures have been implemented. We will be dealing with a *Kaplan-Meier* estimate of a survival function.

Let assume point in time when there was a default during a study:
$t_{(0)} < t_{(1)} < t_{(2)} < \ldots < t_{(k)}$. Then we can derive

$$\boxed{S(t_{(j)}) = S(t_{(j-1)}) \cdot P(T > t_{(j)} | T \geq t_{(j)})} \quad \text{pro } j = 1, \ldots, k$$

where

$S(t_{(j)})$	probability of the fact that a random object would live longer than time $t_{(j)}$, thus $P(T > t_{(j)})$.
$S(t_{(j-1)})$	probability of the fact that a random object would live longer than time $t_{(j-1)}$, thus $P(T > t_{(j-1)})$.
$P(T > t_{(j)} \| T \geq t_{(j)})$	conditional probability of the fact that a random object would live longer than time $t_{(j)}$.

Derivation of this can be found in ***KM.PDF***. Since the relation is recurent, it can be turned into following formula for any $j = 1, \ldots, k$:

$S(t_{(j)}) = S(t_{(0)}) \cdot P(T > t_{(1)} | T \geq t_{(1)}) \cdot P(T > t_{(2)} | T \geq t_{(2)}) \cdot \ldots \cdot P(T > t_{(j)} | T \geq t_{(j)}) =$

$$= 1 \cdot \prod_{i=1}^{j} P(T > t_{(i)} | T \geq t_{(i)}).$$

Kaplan-Meier estimates a probability $P(T > t_{(i)} | T \geq t_{(i)})$ with a ratio $\frac{n_i - m_i}{n_i}$.
$\frac{n_i - m_i}{n_i}$ represent a ratio of those who survived $t_{(i)}$ over those who did not.
Kaplan Meier for $j = 1, \ldots, k$ follows:

$$\boxed{\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \frac{n_i - m_i}{n_i}}$$

For $t \in \langle t_{(j)}; t_{(j+1)} \rangle$ is required $\hat{S}(t) = \hat{S}(t_{(j)})$. Notice that in the last point in time when a default scored, the function $\hat{S}(t)$ does not have to be zero. (For function $S(t)$ can be written $\lim_{t \to \infty} S(t) = 0$.)

Censored observations do not cause a jump in the Kaplan-Meier estimate, but increase denominator n_i, thus "until possible" they increase the range of sample as far as a number of objects in risk is concerned.

Then we can show that if the data are not censored, the estimate using Kaplan-Meier lead to the same values as the classic approach introduced in the beginning of the chapter. For uncensored data goes $n_j = n_{j-1} - m_{j-1}$. Then

$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \frac{n_i - m_i}{n_i} = \frac{n_1 - m_1}{n_1} \cdot \frac{n_2 - m_2}{n_1 - m_1} \cdot \ldots \cdot \frac{n_j - m_j}{n_{j-1} - m_{j-1}} = \frac{n_j - m_j}{n_1} = \frac{n_j - m_j}{n}$ that corresponds with the relation in the beginning of the chapter.

There is an estimate of the survival function for the group with the treatment from the 10.1 in the table 10.4. Estimates of survival function for both groups (with and without treatment) are in the picture 10.6. We can see that the function values of the survival function are higher for the group with the treatment for any t.

Remark. Estimate of $\hat{S}(t_{(j)})$ is a statistics, so we can be questionning about its variablity $Var(\hat{S}(t_{(j)}))$. This variability can be estimated for fixed t using *Greenwood formula* which is for $t \in \langle t_j, t_{j+1} \rangle$: $\hat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i=1}^{j} \frac{m_i}{n_i(n_i - m_i)}$.

Since for fixed t the statistics $\hat{S}(t)$ is having an approximately normal distribution $\hat{S}(t) \sim N(S(t), Var(S(t)))$, we can derive for fixed t $100(1-\alpha)\%$ asymptotic confidence interval for the real value $S(t)$:

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	$1 \cdot \frac{18}{21} = 0,8571$
7	17	1	1	$0,8571 \cdot \frac{16}{17} = 0,8067$
10	15	1	2	$0,8067 \cdot \frac{14}{15} = 0,7529$
13	12	1	0	$0,7529 \cdot \frac{11}{12} = 0,6902$
16	11	1	3	$0,6902 \cdot \frac{10}{11} = 0,6275$
22	7	1	0	$0,6275 \cdot \frac{6}{7} = 0,5378$
23	6	1	5	$0,5378 \cdot \frac{5}{6} = 0,4482$

Table 10.4: *$\hat{S}(t_{(j)})$ is calculated using a recurent relation $\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \cdot \frac{n_j - m_j}{n_j}$. $\hat{S}(t_{(0)}) = 1$ allways.*

$S(t) \in (d; h)$, where

$$d = \hat{S}(t) - \hat{S}(t) \cdot \sqrt{\sum_{i=1}^{j} \frac{m_i}{n_i(n_i - m_i)}} \cdot u_{1-\alpha/2}$$

$$h = \hat{S}(t) + \hat{S}(t) \cdot \sqrt{\sum_{i=1}^{j} \frac{m_i}{n_i(n_i - m_i)}} \cdot u_{1-\alpha/2}$$

This interval is symmetric around $\hat{S}(t)$, so it can happen that it holds $d < 0$ for tiny value of $\hat{S}(t)$ or it holds $h > 1$ for great values of $\hat{S}(t)$. In this case, we replace d with null and h with one[2]. □

• LOG-RANK TEST

If we have a look at estimates of survival functions from the both groups (with treatment and with placebo) in the picture 10.6, we can see that the function values are higher with the group with the treatment for any t. Is this statistically significant or is this thanks to random influences only? The most usual test concerning the significance between two groups is the *Log-Rank* test. It is suitable when the survival functions does not cross each other.[3] This test is based on the χ^2 statistic and, as lots of χ^2 tests, compares the difference between observed and expected frequencies. In this case we are dealing with frequencies of deaths in different points in time t_j that are (or are not) influenced by the membership of the group. Let sign:
$e_{1j} = \left(\frac{n_{1j}}{n_{1j}+n_{2j}}\right) \cdot (m_{1j} + m_{2j}) = \frac{n_{1j}}{n_j} \cdot m_j$, where

n_{ij} represent number of objects in risk in the ith group in time $t_{(j)},\ i = 1, 2$;
m_{ij} represent number of objects that have died in the ith group in time $t_{(j)},\ i = 1, 2$
Statistics e_{1j} demonstrates which ratio of all died in time $t_{(j)}$ should be in favor objects from the first group if the survival duration is not dependent on the group membership. ($\frac{n_{1j}}{n_j}$ represents relative object frequency of the first group related to the complete dataset that are in risk in time $t_{(j)}$.) We will compare the statistics e_{1j} with the observed frequencies of deaths m_{1j} in time $t_{(j)}$. If the values e_{1j} and m_{1j} are similar, a membership to a group apparently does not matter in terms of survival time.
Analogically $e_{2j} = \frac{n_{2j}}{n_j} \cdot m_j$.

[2] or we can derive a confidence interval for $-\log\{\hat{S}(t)\}$ or $\log\left(-\log\{\hat{S}(t)\}\right)$. In both cases the functions have an asymptotic normal distribution (for fixed t)

[3] For differently intensive differences between the two groups the strenght of the log-rank test is different. So that, it have been implemented weighted log-rank tests.

We test a hypothesis H_0 claiming there is no difference in survival time for the objects from the first and second group against an alternative H_1 stating the time is different concerning both groups.

Let sign $O_i - E_i = \sum_{j=1}^{k} (m_{ij} - e_{ij})$, where $i = 1, 2$ represent the membership to the group.

For the statistics $O_i - E_i$ variabililty can be said $Var(O_i - E_i) = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} m_j (n_j - m_j)}{n_j^2 (n_j - 1)}$, $i = 1, 2$, so it is the same for both groups[4].

Test statistics LR is of $\boxed{LR = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}}$, $LR \approx \chi^2(1)$ when valid null hypothesis. Too large values of the LR statistics turn down the null hypothesis.

Example 10.3. Let's get back to the problem 10.2. Test whether a treatment does have an impact on survival duration.

$(O_2 - E_2) = 10,26$, $(O_1 - E_1) = -10,26$, $Var(O_2 - E_2) = Var(O_1 - E_1) = 6,2685$.
$LR = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)} = 16,793; \quad \chi^2_{0,95}(1) = 3,841$
Null hypothesis about insignificance of the difference between survival curves of groups is being declined. We have shown that a treatment does have an impact on the duration of live. □

10.5 Cox model of a proportional risk

Coxův model belongs to regression models in the survival analysis where the dependent variable "survival duration" is dependent on some variables (regressors). If we cannot use a parametric model (briefly introduced in the chapter 10.6), one possibility how to model dependency of the survival duration is the Cox model.

• **MODEL SPECIFICATION**

Mostly, the Cox model is in formula stemming from the risk function but can be rewritten as a function of surival function)

$$\boxed{h(t, \boldsymbol{X}) = h_0(t) \cdot e^{\sum_{i=1}^{p} \beta i X_i} = h_0(t) \cdot \exp\{\boldsymbol{X}'\boldsymbol{\beta}\} \quad \text{pro } t \geq 0}$$

$\boldsymbol{X} = (X_1, \ldots, X_p)'$	is a vector of predictors.
	(The particular levels of predictors are setup by a user.)
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$	is a vector of unknown parameters that needs estimating.
$h_0(t)$	is called *baseline hazard function*.
	It is an unknown function which plays a role of an "intercept".
	Pro $(X_1, \ldots, X_p) = (0, \ldots, 0)$ is $h(t, \boldsymbol{X}) = h_0(t)$.
	This function is not being estimated in the Cox modelu.
$e^{\sum_{i=1}^{p} \beta i X_i}$	This is not dependent on time t.

Coxův model is semiparametric as the baseline hazard function $h_0(t)$ is not specified. (There

[4] Derivation of the variability for $O_i - E_i$ stems from the fact the m_{1j} statististics does have a hypergeometric distribution with parameters n_j, n_{1j}, m_j

is no assumption about its distribution.) Thus, the model contains a paramteric regression part and imparamteric baseline hazard function.

Cox model is very popular thanks to:

1. As $\exp\{\boldsymbol{X}'\boldsymbol{\beta}\}$ is always positive, the model ensures $0 \leq h(t, \boldsymbol{X}) < \infty$.

2. Cox model approximates parametric models well.

3. Parameters $(\beta_1, \ldots, \beta_p)$ and risk ratios can be estimated even when unspecified baseline hazard function $h_0(t)$.

4. By Cox model it is possible to be given estimates of baseline hazard function $h_0(t)$, risk function $h(t, \boldsymbol{X})$ and survival function $S(t, \boldsymbol{X})$ even with unspecified baseline hazard function $h_0(t)$.

• REGRESSION PARAMETERS INTERPRETATION

Before introducing how to estimates parameters $(\beta_1, \ldots, \beta_p)$, we can show their interpretation. Imagine we research survival time that is influenced with following regressors: X_1 demonstrates a membership to the group with a treatment (0), or with placebo (1); X_2 represents sex (1 female, 0 male) and X_3 represents age. Now suppose two patients differentiating only in having a treatment or placebo. We want to research whether their risk of death differs. Let sign a risk function of the patient with a treatment $h_1(t, \boldsymbol{x})$ and a risk function of a patient with placebo $h_2(t, \boldsymbol{x})$.
$h_1(t, \boldsymbol{x}) = h_0(t) \cdot \exp\{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}$ is a risk function for an object with predictors: *treatment* t.j. $x_1 = 0, x_2 = 1, x_3 = 50$

$h_2(t, \boldsymbol{x}) = h_0(t) \cdot \exp\{\beta_1(x_1 + 1) + \beta_2 x_2 + \beta_3 x_3\}$ is a risk function for ana object with predictors: *placebo*, $x_2 = 1, x_3 = 50$

Then the ratio of their risk is:
$\frac{h_2(t,\boldsymbol{x})}{h_1(t,\boldsymbol{x})} = \frac{h_0(t)\cdot\exp\{\beta_1(x_1+1)+\beta_2 x_2+\beta_3 x_3\}}{h_0(t)\cdot\exp\{\beta_1 x_1+\beta_2 x_2+\beta_3 x_3\}} = e^{\beta_1}$.
or $h_2(t, \boldsymbol{x}) = e^{\beta_1} \cdot h_1(t, \boldsymbol{x})$

Apparently, the regressors' values X_2 and X_3 in the Cox modelu do not needed to be specified; being the same is sufficient. Thus the number e^{β_1} represents how much higher/lower the risk of death of the patient with placebo than the case of the patient with a treatment if everything is the same. Let:
$\bullet h_1(t, \boldsymbol{x}) = h_0(t) \cdot \exp\{\beta_1 x_1 + \ldots + \beta_p x_p\}$ is a risk function for an object with predictors $x_1, \ldots, x_p$

$\bullet h_2(t, \boldsymbol{x}) = h_0(t) \cdot \exp\{\beta_1(x_1 + 1) + \ldots + \beta_p x_p\}$ is a risk function for an object with predictors $x_1 + 1, \ldots, x_p$

Then the hazard ratio is: $HR = \frac{h_2(t,\boldsymbol{x})}{h_1(t,\boldsymbol{x})} = e^{\beta_1}$. Analogically, we can increase a unit of regressor x_j and fix the values of other regressors. Then, e^{β_j} represents how much the risk of death increases/decreases when we increase a value of theu jth regressor of one unit, $j = 1, \ldots, p$.

The number e^{β_j} is called a *relative risk related to a predictor* x_j. This relative risk is *adjusted*.

Using partial likehood function we can get $\hat{\beta}_j$ and $e^{\hat{\beta}_j}$. Confidence interval for $\hat{\beta}_j$ and $e^{\hat{\beta}_j}$ can be derived using a Wald statistics.

Sometimes, we can be interested in change of the relative risk if age increase of 10 years and we will be dealing with a man instead of women.
In general, we will be dealing with an object with predictor values $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*)$ and an object with predictor values $\boldsymbol{x} = (x_1, \ldots, x_p)$. Then the ratio of risk functions:

$$HR = \frac{h(t,\boldsymbol{x}^*)}{h(t,\boldsymbol{x})} = \frac{h_0(t)\cdot\exp\{\sum_{j=1}^{p}\beta_j x_j^*\}}{h_0(t)\cdot\exp\{\sum_{j=1}^{p}\beta_j x_j\}} = \exp\{\sum_{j=1}^{p}\beta_j(x_j^* - x_j)\}$$

If we replace the "betas" with their estimates, we will be given an estimate of the risk functions ratio and subsequently can state confidence intervals using the Wald statistics.

• PARAMETER ESTIMATES USING A PARTIAL LIKEHOOD FUNCTION
Usually, we know the simultaneous density $f(x_1, \ldots, x_p)$ for parameter estimates using maximum likehood and we estimate only the unknown parameters of this density. Nonetheless, the density in the Cox model is not specified and we cannot use the maximum likehood method without a trick.

The estimate of the **partial** likehood designed by Cox is based on observed order of deaths for specified regressors. We are asking for what parameters $(\beta_1, \ldots, \beta_p)$ in the model is that Jan (smoker, man, above 60 years) died in the first since the outbreak most probably. (in the time of outbreak), the second was Pavel and the least was Eva. In other words, for what model parameters is the observed death order list the most probable? Actually, we are not interested if the first death man was Jan but whether it was an above 60 year man, a smoker.

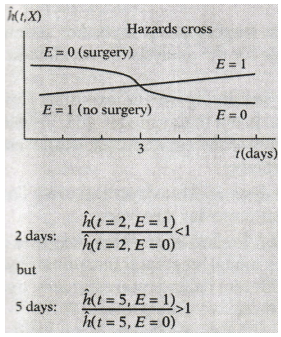
Put $L_{(j)}(\boldsymbol{\beta})$ as a probability that the first dead-man would be Jan from the set of people that were in risk in time $t_{(j)}$ whose death has been observed in time $t_{(j)}$. Then we cam derive

$$L_{(j)}(\boldsymbol{\beta}) = \frac{h_j(t_{(j)})\,dt_{(j)}}{\sum_{r\in R(t_{(j)})} h_r(t_{(j)})\,dt_{(j)}} = \frac{\exp\{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}\}}{\sum_{r\in R(t_{(j)})}\exp\{\boldsymbol{x}'_{(r)}\boldsymbol{\beta}\}},$$

where $R(t_{(j)})$ is a group of people that were in risk in time $t_{(j)}$.
Let sign $L(\boldsymbol{\beta})$ a probability that the patients would be dying in the order we have observed for set parameters $\boldsymbol{\beta}$. For what paramters $\boldsymbol{\beta}$ the observed order of deaths would be the most probable? The function $L(\boldsymbol{\beta})$ is called a *partial likehood*.
It valids: $L(\boldsymbol{\beta}) = \prod_{j=1}^{k} L_{(j)}(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{\exp\{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}\}}{\sum_{r\in R(t_{(j)})}\exp\{\boldsymbol{x}'_{(r)}\boldsymbol{\beta}\}}$
When searching for maximum of the argument $L(\boldsymbol{\beta})$, it is more convenient to find the maximum from the logarithms of the partial likehood. An estimate of the regression coeficients: $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p) = \operatorname{argmax} L(\boldsymbol{\beta}) = \operatorname{argmax}\log(L(\boldsymbol{\beta}))$.
• NON-PARAMETRIC ESTIMATES OF THE BASELINE HAZARD FUNCTION AND THE ADJUSTED SURVIVAL FUNCTION
Using partial likehood is possible to estimate paramters $\beta_1, \ldots, \beta_p$ in the Cox model. If we were able to estimate the baseline hazard function $h_0(t)$ (that Cox originally did not get), we could draw estimates of survival function for specifically set values of regressors.

• PROPORTIONAL RISK MODEL ASSUMPTIONS
Coxův model doe not claim any requirements for the baeline hzard function $h_0(t)$ specification, but assumes that the particluar predictors meet an assumption of risk proportionality. In other words it assumes that the relative risk HR is stable related to particular regressors in time t. So that:

$HR = \frac{h(t,\boldsymbol{x}^*)}{h(t,\boldsymbol{x})} = \text{constant}$
Let introduce situations when the assumption is not met and the Cox model is not suitable. See picture 10.7.



There are risk functions for two patients with a cancer that differs in a regressor value E. Patients with $E = 0$ underwent an operation, patients with $E = 1$ underwent radiotherapy without operation. It can be seen that for small values of t, the risk function do have bigger values for patients with operation and for big values of t the risk function have higher values for patients without operation. If we put these two into a ratio, the ratio would change in time and the risk functions are not proportional.

Figure 10.7: *Proportional risk assumption break.*

For brief check of the assumption we can use cummulative survival risk function graphs against time and residual analysis.

1. **Graph analysis:**
   We make a group of graphs with time (or log time) on the x axis and a logarith of cummulative risk function for different levels of X_j regressors estimates on the y axis. If the regressor is qualitative, we can categorize. Estimates of the cummulative risk function for the particular levels of X_j can be given by Kaplan-Meier method and must not be based on the Cox model.

   If the Cox model assumption is met, the difference between the logarithms of the cummulative risk functions for the different levels of regressor X_j is not changing in time.

2. **Residual analysis:**
   There have been designed different types of residuals for the Cox model. For the assumption assessment, the Schoenfeld residuals are suitable. Their expected value is asymptotically zero if the assumption is met. Schoenfeld residuals are computed for each regressor. For meeting an assumption, we determine residuals of all n objects and a line corresponding with the x axis should be approximated through the data.

What if a qualatative regressor breaks an assumption of the proportionality? We have to devide the dataset into strats according to levels of this regressor and we create a different Cox model in each strat.

10.6 Final reccomendations

In majority of statistical analyses cases it is truth that the bigger the dataset, the more precise findings. Concerning survival analysis, it is not essential how big is n but number of deaths. If there are almost all censored data in the dataset, we will not be given any findings related to the survival function.

A brief reccomendation: at least ten deaths for each regressor.

Chapter 11

Linear models classification

The unique clasification does not exist. We will be dealing only with models with one dependent variable.

11.1 Multiple linear regression model

Y is a dependent variable; $x_1, \ldots, x_p$ are random regressors or known functions of explaining variables (predictors) $z_1, \ldots, z_r$.
We transform these predictors for dependent variable modeling by a function of $f_1, \ldots, f_p$:
Let $\boldsymbol{z} = (z_1, \ldots, z_r)'$; $\quad x_1 = f_1(\boldsymbol{z}),\ x_2 = f_2(\boldsymbol{z}), \ldots,\ x_p = f_p(\boldsymbol{z})$.
The multiple linear regression model then takes the form:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

Example 11.1. Let's assume $r = 2$ predictors z_1, z_2 and $p = 5$ regressors $x_1, \ldots x_5$ that $x_1 = z_1,\ x_2 = z_2,\ x_3 = z_1^2,\ x_4 = z_2^2,\ x_5 = z_1 z_2$.
We derived the regressors from the predictors by following functions:
$f_1(z_1, z_2) = z_1,\ f_2(z_1, z_2) = z_2, \ldots, f_5(z_1, z_2) = z_1 z_2$. □

If $x_1 = z_1,\ x_2 = z_2,\ \ldots,\ x_p = z_p$, we are tallking about a *complete linear model*.

11.2 General Linear Model

If we are not taking into only x continuous variables in the $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$ model but also categorical variables (or their interactions), we are talking about a general linear model.

If there are all the regressors continuous, we are talking about simple linear or multinominal linear regression. If all the regressors corresponds with the classes of the categorical variables, we are talking about ANOVA. If we substitute into xs both types, then we are talking about ANCOVA. The continuous regressors are called *covariates* and classes of categorical

factors are called *factors*. In **R**, we can get all these models by the `lm(formula,data,...)`. For example:

`lm(y~x)`	simple regression
`lm(y~x1 + x2 + x3)`	multidimensional regression
`lm(y~f)`	if f is a factor, this is one-way ANOVA
`lm(y~f1 + f2)`	if f1 and f2 are factors, this is two-way ANOVA
`lm(y~f + x)`	if f is a factor, this is ANCOVA
`lm(y~f1+f2+f3 + f1:f2 + f1:f3 + f2:f3 + f1:f2:f3)`	3 main effects, 3 double interactions, 1 tripple interaction
`lm(y~f1*f2*f3)`	$-\|\|-$ 3 main effects, 3 double interactions, 1 tripple interaction

11.3 Generalized Linear Model

We can formulate the **general** linear model
•as $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$
•or $\mu = E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \quad Y \sim N(\mu, \sigma^2)$

The **generalized** linear model uses a trick that there is a function instead the estimated value $\mu = E(Y)$. The full spefication of the generalized linear model contains":
1) deterministic part (linear combination of predictors),
2) known distribution of Y random variable
3) linking function.

The model goes:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

It is possible to model the estimated value by the invesion of the linking function altenatively, so $\mu = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$. The right side is not linear in general.

The aim of the linking function is to transform the estimated value μ in a suitable way so that the estimated value could be modeled by a linear function. We can use more linking functions for particular distributions of Y but there are typical linking functions that are called *canonical* linking functions.

If the Y is alternatitive for example, the canonical linking function is a function called *logit*, $g(\mu) = log(\frac{\mu}{1-\mu})$. If the Y is a random variable with a Gama distribution, the canonical linking function is the invesion $g(\mu) = \frac{1}{\mu}$. For Y with a normal distribution, the linking function is an identity. For Y with a Poisson distribution, the linking function is a logarithm, $g(\mu) = log(\mu)$. The generalized linear models estimate parameters by a maximum likehood (the distribution of Y must be known and the particular realizations $y_1, \ldots, y_n$ of the random variable Y must be independent).

Chapter 12

Logistic, multinomical and ordinal regression

Logistic regression models the binar variable Y with only either 0 or 1 values. *Multinomical* regression models categorical variable with more than two values. If there is a need to order these values, then we are talking about an *ordinal* regression.

12.1 Logistic regression

`glm(formula, binomial)`
Logistic regression is a specific case of the generalized linear models where the Y is of alternative (Bernoulli) distribution and the linking function is logit. We will sign:
Random variable $Y \sim A(\mu)$ is having an estimated value $E(Y) = \mu$ and variance $D(Y) = \mu(1-\mu)$. The interpretation of the estimated value is a probability of the success; and the variance is a function of the estimated value. Since we interpret the μ as a prabibality of the success $P(Y=1)$, it is common to write p intead of μ in the logistic regression.
Motivation to linking function choosing:
Whether we model the alternative random variable Y by a classical regression $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$, we will be facing a problem that there is a binary variable on the left side and the right side can be of values from the minus infinity to infinity.

So that we aim to make the left side more continous by modeling not the binar variable Y but the chance of success (success/unsuccess), $\frac{p}{1-p} = \frac{P(Y=1)}{P(Y=0)}$. This ratio is now continuous on the interval $(0, \infty)$. Now we can only have to spread that interval on the complete function domain by a logarithm: $log(\frac{p}{1-p}) \in (-\infty, \infty)$. So that, the linear regression function is not modeled by a binar variable Y, but by the logarithm of a chance that is called *logit* and we sign $\boxed{\text{logit}(p) := \log(\frac{p}{1-p})}$.

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{x}'.\boldsymbol{\beta}$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and $p(\boldsymbol{x}) = P(Y = 1 | x_1, \ldots, x_p)$.

We can also model an estiamted value by an invesion of the logit function, $p(\boldsymbol{x})$:

$$\begin{aligned}
\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} &= e^{\boldsymbol{x}'\boldsymbol{\beta}} \\
&= e^{\boldsymbol{x}'\boldsymbol{\beta}} - p(\boldsymbol{x}).e^{\boldsymbol{x}'\boldsymbol{\beta}} \\
p(\boldsymbol{x}).(1+e^{\boldsymbol{x}'\boldsymbol{\beta}}) &= e^{\boldsymbol{x}'\boldsymbol{\beta}} \\
p(\boldsymbol{x}) &= \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}} = \frac{1}{1+e^{-\boldsymbol{x}'\boldsymbol{\beta}}}
\end{aligned}$$

So that $\text{logit}^{-1}(p(\boldsymbol{x})) = p(\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{x}'\boldsymbol{\beta}}}$ and is called a *logistic function*.

12.2 Simple logistic regression with one continuous predictor

Now we are about to model a probability of the success by one continuous predictor x, so the logistic function of the model is going to be $p(x) = P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$.
The parameters' effects β_0 a β_1 on the logistic function with one predictor x can be seen on the picture 12.1.

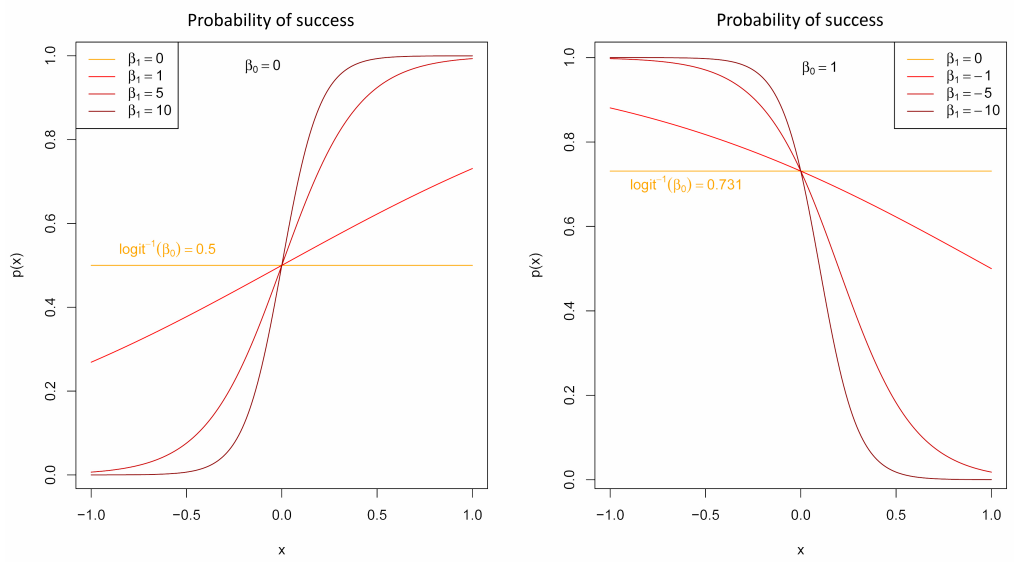


Figure 12.1: *Logistic curves: success probability modeling by one predictor x for different paratemer values β_0 a β_1.*
Picture on the left: an orange constant function with value of 0.5 models the success probability with the dependence on x for $\beta_0 = 0$ a $\beta_1 = 0$. However, for β_1 positive the more increasing value of x predictor, the higher probablity of the success.
Picture on the right: for β_1 negative, the more increasing value of x predictor, the lower probablity of the success.

12.2.1 Interpretation of the constant

Further, we are going to sign the odds of the success over the odds of the unsuccess as "odds", so

$$\text{odds}(x) = \frac{p(x)}{1 - p(x)}$$

And we will sign the odds ratio as "or", so

$$\text{or}(x + \delta, x) = \frac{\text{odds}(x + \delta)}{\text{odds}(x)}$$

How can be interpreted the constant β_0 in the logistic regression model? The better for interpretation is e^{β_0} instead of β_0.

For the null value of the predictor, $x = 0$, it goes:

$$\text{logit}(p(0)) = \log(\text{odds}(0)) = \log\left(\frac{p(0)}{1 - p(0)}\right) = \beta_0 + \beta_1.0 = \beta_0$$

•β_0 can be interpreted as a "logaritmic" odds on the success if there is a zero value of the predictor x.

$$\frac{p(0)}{1 - p(0)} = e^{\beta_0 + \beta_1.0} = e^{\beta_0}$$

•e^{β_0} can be interpreted as an odds on the success if there is zero value of the predictor x which is easier to interpret then β_0.

$$p(0) = \frac{e^{\beta_0 + \beta_1.0}}{1 + e^{\beta_0 + \beta_1.0}} = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

•$\frac{e^{\beta_0}}{1+e^{\beta_0}}$ can be interpreted as a probability of the success in case of zero value of the predictor x.

12.2.2 Slope interpretation

It is more suitable to interpret e^{β_1} than the parameter β_1 again.

$$\log(\text{or}(x+1,x)) = \log\left(\frac{\text{odds}(x+1)}{\text{odds}(x)}\right) = \log(\text{odds}(x+1)) - \log(\text{odds}(x))$$

•β_1 can be interpreted as a "logaritmical" odds ratio if we are going to increase the predictor value x by a unit.

$$\text{or}(x+1,x) = \frac{\text{odds}(x+1)}{\text{odds}(x)} = e^{\beta_1}$$

•e^{β_1} can be interpreted as the odds ratio if we are going to increase the value of predictor x by a unit. In other words - what is the odds on success higher/lower when we increase the value of predictor x by a unit?

If we will increase the value of x predictor by δ, than

$$\begin{aligned}\log(\text{or}(x+\delta,x)) &= \log\left(\frac{\text{odds}(x+\delta)}{\text{odds}(x)}\right) = \log(\text{odds}(x+\delta)) - \log(\text{odds}(x)) \\ &= \beta_0 + \beta_1(x+\delta) - (\beta_0 + \beta_1 x) = \beta_1\delta\end{aligned}$$

•$\beta_1\delta$ can be interpreted as a "logaritmic" odds ratio if we have increased the value of the x predictor by δ.

$$\text{or}(x+\delta,x) = \frac{\text{odds}(x+\delta)}{\text{odds}(x)} = e^{\beta_1\delta} = (e^{\beta_1})^{\delta}$$

•$e^{\beta_1\delta}$ states what is the odds on success higher/lower when we increase the value of predictor x by δ.

We can summarize (see picture 12.1):

$e^{\beta_1} = 1 \Leftrightarrow \beta_1 = 0$	•x does not influence odds ratio
	•x does not influence also probability of the success
$e^{\beta_1} > 1 \Leftrightarrow \beta_1 > 0$	•with increasing value of x the odds ratio is also increasing
	•with increasing value of x the probability of success is also increasing
$e^{\beta_1} < 1 \Leftrightarrow \beta_1 < 0$	•with increasing value of x the odds ratio is decreasing
	•with increasing value of x the probability of success is decreasing

12.2.3 When the logistic regression is suitable?

Example 12.1. TODO dataset `GermanCredit`
Binary variable `Class` contains only values of 0 or 1, where 1 is a bad (undued) credit and 0 is good (dued) one; we are going to model the bads. The continuous variable `Amount100` demonstrates the amount of credit in 100 DEM. When can we use the logistic regression model with regard to parameter interpretaion? □

Since $e^{\beta_1 \delta} = \mathrm{or}(x+\delta, x)$, we can see that the odds ratio is not influenced by the absolute values of $x+\delta$ and x, but only by the difference between δ.

Since that, the odds ratio in case of credits of 100 and 200 DEM must be the same as in case of 10 000 and 10 100 DEM.

Model assumes that the logaritmic chance of success is dependent on the predictor (or predictors) linearly and the derived odds ratio is not dependent on the absolute value of a predictor but only is dependent on the differences in the values of that predictor. We can test this assumption by the Pearson's chi-squared test (or by the Hosmer-Lemeshow test for multinomical logistic regression).

12.2.4 Maximum likehood

Usual method in linear models is the least squares method. If there is an assumtion of the normal distribution of the dependent variable, then the estimates derived by the LS have good properties, known distribution and we can then proceed the statistical inference.

However, in cases of generalized linear models, we often use maximum likehood for parameter estimation.

Subsequently, it is possible to proceed only asymptotic inference, thus we have to take findings with reserve for small datasets.

We will sign:
$*$ vector of predictors with one: $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$;
$*$ vector of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$;
$*$ binary dependent variable $Y \in \{0, 1\}$
$*$ n independent observations Y: $(y_1, y_2, \ldots, y_n)$
$*$ probability of success conditioned by the values of predictors $p(\boldsymbol{x}) = P(Y = 1 | x_1, \ldots, x_p) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}$

Furthermore, for $\begin{cases} y = 1 \text{ we go} & P(Y = y | x_1, \ldots, x_p) = p(\boldsymbol{x}) & = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}} = \left(\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}\right)^y \\ y = 0 \text{ we go} & P(Y = y | x_1, \ldots, x_p) = 1 - p(\boldsymbol{x}) & = \frac{1}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}} = \left(\frac{1}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}\right)^{1-y} \end{cases}$

Likehood function L demonstrates probability of the observed data. In this probability function we can find also the unknown values of paramters $\boldsymbol{\beta}$. Then we ask: for what vector of paramters $\boldsymbol{\beta}$ the observed data $(y_1, y_2, \ldots, y_n)$ are most probable? For this only numerical methods are possible, thus the findings can be different using different software.

$$\bullet L(\boldsymbol{\beta}) = P(Y_1 = y_1 \wedge \ldots \wedge Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{y_i} \cdot \left(\frac{1}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i} =$$
$$= \ldots = \prod_{i=1}^{n} \frac{1}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \cdot \prod_{i=1}^{n} (e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^{y_i}$$

•ML odhad $\hat{\boldsymbol{\beta}} = \operatorname{argmax} L(\boldsymbol{\beta})$.

From the practical point of viewm the estimates are not being computed from $L(\boldsymbol{\beta})$, but from $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$, thus

$$\begin{aligned} \bullet l(\boldsymbol{\beta}) &= \log\left(\prod_{i=1}^{n} \tfrac{1}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \cdot \prod_{i=1}^{n} (e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^{y_i}\right) = \sum_{i=1}^{n} \log \tfrac{1}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} + \sum_{i=1}^{n} \log(e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^{y_i} = \\ &= -\sum_{i=1}^{n} \log(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}) + \sum_{i=1}^{n} y_i \boldsymbol{x}_i'\boldsymbol{\beta} \end{aligned}$$

Bibliography

[A1] ANDĚL, J. (1993): Statistické metody. 1. vydání., MATFYZPRESS Praha.
[A2] ANDĚL, J. (2002): Základy matematické statistiky Preprint. MFF UK, Praha
[BKM] BUDÍKOVÁ, M., KRÁLOVÁ, M., MAROŠ, B. (2010): Průvodce základními statistickými metodami, Grada Publishing, Praha.
[H] HEBÁK, P. za kolektiv (2007): Vícerozměrné statistické metody I-III, Informatorium, Praha.
[Hen] Přehled statistických metod zpracování dat :analýza a metaanalýza dat. Edited by Jan HENDL. 1. vyd. Praha: Portál, 2004.
[KK] KOLÁČEK, J., KONEČNÁ, K. (2011): Jak pracovat s jazykem R .
[MM] MELOUN, Milan a Jiří MILITKÝ. (2005) Počítačová analýza vícerozměrných dat v příkladech, Academia, Praha.
[KA] KOMÁREK Arnošt. Nepublikované studijní materiály