Introduction to data analysis

Borbála Szüle

Corvinus University of Budapest

ISBN: 978-963-503-619-6
Publisher: Budapesti Corvinus Egyetem, Közgazdaságtudományi Kar
(Corvinus University of Budapest, Faculty of Economics)
2016

Contents

Preface

With the latest development in computer science, multivariate data analysis methods became increasingly popular among economists. Pattern recognition in complex economic data and empirical model construction can be more straightforward with proper application of modern softwares. However, despite the appealing simplicity of some popular software packages, the interpretation of data analysis results requires strong theoretical knowledge. This book aims at combining the development of both theoretical and application-related data analysis knowledge. The text is designed for advanced level studies and assumes acquaintance with elementary statistical terms. After a brief introduction to selected mathematical concepts, the highlighting of selected model features is followed by a practice-oriented introduction to the interpretation of SPSS[1] outputs for the described data analysis methods. Learning of data analysis is usually time-consuming and requires efforts, but with tenacity the learning process can bring about a significant improvement of individual data analysis skills.

[1]IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

1 Introduction to mathematics

1.1 Matrix calculations

Matrix calculations are often applied in data analysis. The number of rows and columns of a matrix may differ. If the number of rows and columns are equal, the matrix is called a square matrix. Square matrices are not necessarily symmetric matrices (that are symmetric about the diagonal[1]). For a matrix M the transposed matrix is denoted by M^T. The rows of the transposed matrix M^T correspond to the columns of M (and the columns of M^T correspond to the rows of M). If M is a symmetric matrix, then $M = M^T$.

An identity matrix (of order n) is a square matrix with n rows and columns having ones along the diagonal and zero values elsewhere. If the identity matrix is denoted by I, then the relationship of a square matrix M and the inverse of M (denoted by M^{-1}) is as follows (*Sydsæter-Hammond* (2008), page 591):

$$MM^{-1} = M^{-1}M = I \tag{1.1}$$

In some cases the determinant of a matrix should be interpreted in data analysis. The determinant of a matrix is a number that can be calculated based on the matrix elements. Determinant calculation is relatively simple in case of a matrix that has two rows and columns. For example assume that a matrix is defined as follows:

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \tag{1.2}$$

For the matrix M the determinant can be calculated according to *Sydsæter-Hammond* (2008), (page 574):

[1] In this case the diagonal of a square matrix with n columns is defined as containing the elements in the intersection of the i-th row and i-th column of the matrix (i=1, ..., n).

$$det(M) = m_{11}m_{22} - m_{12}m_{21} \tag{1.3}$$

In this example the determinant occurs also in the calculation of the inverse matrix. The inverse of matrix M in this example can be calculated as follows (*Sydsæter-Hammond* (2008), page 593):

$$M^{-1} = \frac{1}{det(M)} \begin{pmatrix} m_{22} & -m_{12} \\ -m_{21} & m_{11} \end{pmatrix} \tag{1.4}$$

If for example $M_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, then $det(M_1) = 1 - 0.8^2 = 0.36$. The determinant of the identity matrix is equal to one. If all elements in a row (or column) of a matrix are zero, then the determinant of the matrix is zero. (*Sydsæter-Hammond* (2008), page 583) If the determinant of a square matrix is equal to zero, then the matrix is said to be singular. A matrix has an inverse if and only if it is nonsingular. (*Sydsæter-Hammond* (2008), page 592)

The determinant can be interpreted in several ways, for example a geometric interpretation for $det(M_1)$ is illustrated by Figure 1.1, where the absolute value of the determinant is equal to the area of the parallelogram. (*Sydsæter-Hammond* (2008), page 575)

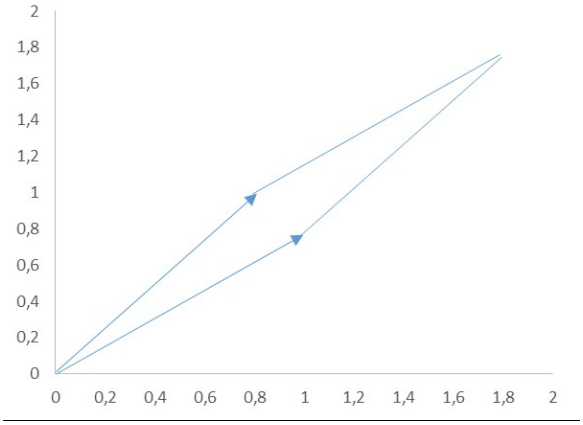


Figure 1.1: Matrix determinant

In data analysis it is sometimes necessary to calculate the determinant of a correlation matrix. If the variables in the analysis are (pairwise) uncorrelated, then the correlation matrix is the identity matrix, and in this case the determinant of the correlation matrix is equal to one. If however

the variables in an analysis are "perfectly" correlated so that the (pairwise) correlations are all equal to one, then the determinant of the correlation matrix is equal to zero. In this case (if the determinant is equal to zero) the correlation matrix is singular and does not have an inverse.

For a square matrix M a scalar λ (called „eigenvalue") and a nonzero vector x (called „eigenvector") can be found such that (*Rencher-Christensen* (2012), page 43)

$$Mx = \lambda x \tag{1.5}$$

If matrix M has n rows and n columns, then the number of eigenvalues is n, but these eigenvalues are not necessarily nonzero. Eigenvectors are unique only up to multiplication by a value (scalar). (*Rencher-Chriszensen* (2012), page 42) The eigenvalues of a positive semidefinite matrix M are positive or zero values, where the number of positive eigenvalues is equal to the rank of the matrix M. (*Rencher-Christensen* (2012), page 44) The eigenvectors of a symmetric matrix are mutually orthogonal. (*Rencher-Christensen* (2012), page 44)

According to the spectral decomposition theorem for each symmetric matrix M an orthonormal basis containing the eigenvectors of matrix M exists so that in this basis M is diagonal:

$$D = B^T M B \tag{1.6}$$

where D is a diagonal matrix and the diagonal values are the eigenvalues of M (*Medvegyev* (2002), page 454). In case of this orthonormal basis $B^T = B^{-1}$, which means that the transpose matrix and the inverse matrix are identical. (*Medvegyev* (2002), page 454) According to the spectral decomposition theorem, the symmetric matrix M can be expressed in terms of its eigenvalues and eigenvectors (*Rencher-Christensen* (2012), page 44):

$$M = BDB^T \tag{1.7}$$

Assume that the (real) matrix X has n rows and p columns and the rank of matrix X is equal to k. In this case the singular value decomposition of matrix X can be expressed as follows (*Rencher-Christensen* (2012), page 45):

$$X = UDV^T \tag{1.8}$$

where matrix U has n rows and k columns, the diagonal matrix D has k rows and k columns and matrix V has p rows and k columns. In this case the diagonal elements of the (non-singular) diagonal matrix D are the positive square roots of the nonzero eigenvalues of $X^T X$ or of XX^T. The

diagonal elements of matrix D are called the singular values of matrix X. The columns of matrix V are the normalized eigenvectors of X^TX and the columns of matrix U are the normalized eigenvectors of XX^T. (*Rencher-Christensen* (2012), page 45)

A positive semidefinite matrix M can be expressed also as follows (*Rencher-Christensen* (2012), page 38):

$$M = A^TA \tag{1.9}$$

where A is a nonsingular upper triangular matrix that can be calculated for example with Cholesky decomposition. Calculation details about Cholesky decomposition can be found for example in *Rencher-Christensen* (2012) (on pages 38-39).

1.2 Probability theory

Data analysis is usually based on (randomly selected) statistical samples. Theoretically, statistical sampling may contribute to answer quantitative research questions, since according to the Glivenko-Cantelli theorem, as the number of independent and identically distributed sample observations increases, the empirical distribution function (belonging to the sample) almost surely converges to the theoretical (population) distribution function. (*Medvegyev* (2002), page 542). This theorem is one of the theoretical reasons why data analysis is often related to probability theory. For instance, confidence intervals and empirical significance levels ("p-values") are usually calculated with assuming (theoretically explainable) probability distributions in case of certain variables.

The normal distribution is among the most frequently applied probability distributions in data analysis. In the univariate case it has two parameters (μ and σ, indicating the mean and the standard deviation, respectively). The probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma}} \tag{1.10}$$

for $-\infty < x < \infty$ where $-\infty < \mu < \infty$ and $\sigma > 0$. (*David et al.* (2009), page 465). This density function (for different standard deviations and with the mean being equal to zero) is illustrated by Figure 1.2. The histogram on Figure 1.2 belongs to a standard normal distribution (simulated data with a sample size of 1000). On Figure 1.2, the blue lines indicate the theoretical density functions belonging to the normal distributions with the standard

deviation values 1, 1.25, 1.5, 1.75 and 2, respectively (and with a theoretical mean which is equal to zero for each blue line on Figure 1.2).

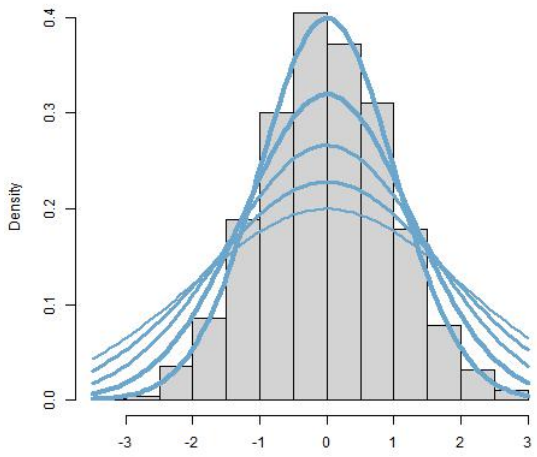


Figure 1.2: Normal distributions

Some probability distributions are related to the normal distribution. The chi-squared distribution is the sum of n independent random variables that are the squares of standard normally distributed random variables: if ξ_1, ξ_2, $\ldots \xi_n$ are (independent) random variables with standard normal distribution, then the distribution of the random variable is called χ^2-distribution with n degrees of freedom. The mean of the χ^2-distribution with n degrees of freedom is n, the variance is $2n$. (*Medvegyev* (2002), pages 263-264) Figure 1.3 illustrates χ^2-distributions with different degrees of freedom. The histogram on Figure 1.3 belongs to a χ^2_2 distribution (simulated data with a sample size of 1000), and the blue lines indicate the theoretical density functions belonging to χ^2_2, $\ldots$, χ^2_{11} distributions. It can be observed on Figure 1.3, that for higher degrees of freedom the density function of the χ^2 distribution becomes more symmetric.

Probability theory distinguishes univariate distributions from multivariate distributions (that are defined for a vector of random variables. If for example ξ_1, $\ldots$ ξ_m are random variables, then the (ξ_1, $\ldots$, ξ_m) variable (the vector of the ξ_1, $\ldots$ ξ_m random variables) has a multivariate normal distribution, if for each t_i ($i = 1, \ldots, m$) real numbers the distribution of the

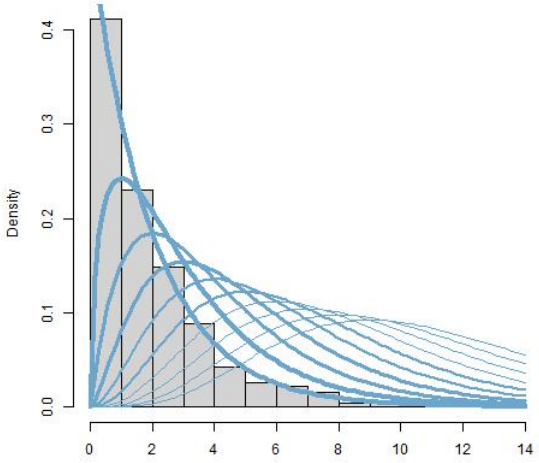


Figure 1.3: Chi-squared distributions

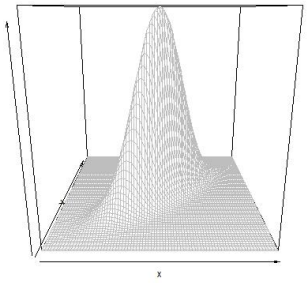
following random variable is a normal distribution (*Medvegyev* (2002), page 453):

$$\sum_{i=1}^{m} t_i \xi_i \tag{1.11}$$

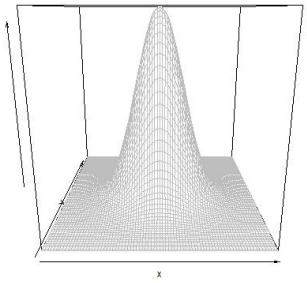
A multivariate normal distribution has a mean vector (instead of only one number as the mean of the distribution) and a covariance matrix (instead of one number as the variance of the distribution). The covariance matrix of a multivariate normal distribution is a positive semidefinite matrix (*Medvegyev* (2002), page 453).

Assume in the following that the covariance matrix of a multivariate normal distribution is denoted by C and also assume that matrix C has m rows and m columns. In that case if the rank of C is r, then matrix A (that has m rows and r columns) exists so that $C = AA^T$. This result is a consequense of the spectral decomposition theorem. (*Medvegyev* (2002), page 454)

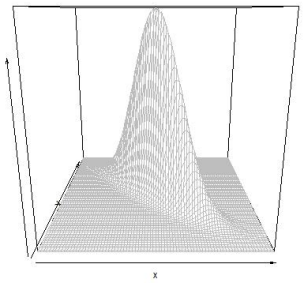
According to the spectral decomposition theorem for each symmetric matrix M an orthonormal basis containing the eigenvectors of matrix M exists so that in this basis M is diagonal:



(a) ρ=0.8



(b) ρ=0



(c) ρ=-0.9

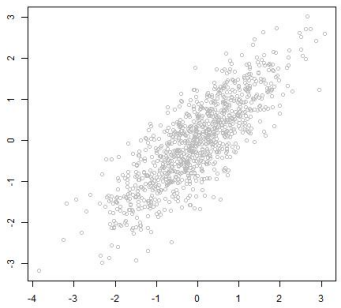
Figure 1.4: Multivariate normal distributions

$$D = B^T M B \tag{1.12}$$

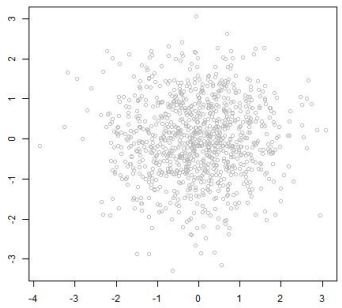
where D is a diagonal matrix and the diagonal values are the eigenvalues of M (*Medvegyev* (2002), page 454). It is worth mentioning that in case of this orthonormal basis $B^T = B^{-1}$, which means that the transposed matrix and the inverse matrix are identical. (*Medvegyev* (2002), page 454)

In case of a multivariate normal distribution the covariance matrix contains information about the independence of the univariate random variables. However it is worth emphasizing that the covariance (and thus correlation coefficient) can not automatically be applied to test whether two random variables are independent: uncorrelated random variables can only then be considered as independent if the joint distribution of the variables is a multivariate normal distribution. (*Medvegyev* (2002), page 457)

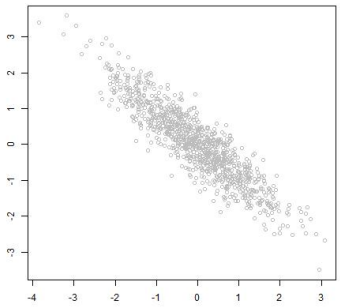
Figure 1.4 illustrates density functions of multivariate normal distribu-



(a) ρ=0.8



(b) ρ=0



(c) ρ=-0.9

Figure 1.5: Effect of correlation

tions with different theoretical correlations coefficient (indicated by ρ). It can be observed on Figure 1.4 that as the absolute value of the theoretical correlation coefficient increases, the area on the density function graph, where the density function value is not close to zero, becomes smaller. The same phenomenon can be observed on Figure 1.5, which shows simulated data on scatter plots that belong to the cases $\rho = 0.8$, $\rho = 0$ and $\rho = -0.9$, respectively (with the assumption that the univariate distributions belonging to the simulations are normal distributions). The effect of a change of sign belonging to the theoretical correlation can also be observed on Figure 1.4 and Figure 1.5.

2 | Cluster analysis

Cluster models are usually used to find groups (clusters) of similar records based on the variables in the analysis, where the similarity between members of the same group is high and the similarity between members of different groups is low. With cluster analysis it may be possible to identify relatively homogeneous groups ("clusters") of observations. There are several cluster analysis methods, in this chapter selected features of hierarchical, k-means and two-step clustering are introduced.

2.1 Theoretical background

Hierarchical, k-means and two-step cluster analysis apply different algorithms for creating clusters. The hierarchical cluster analysis procedure is usually limited to smaller data files (for example in case of thousands of objects the application of this analysis is usually related to significant computational cost). The k-means cluster analysis procedure and two-step clustering can be considered as more suitable to analyze large data files.

Theoretically, in case of hierarchical cluster analysis the aggregation from individual points to the most high-level cluster (agglomerative approach, bottom-up process) or the division from a top cluster to atomic data objects (divisive hierarchical clustering, top-down approach) can also be solved from a computational point of view. (*Bouguettaya et al.* (2015)) In the following selected features of the agglomerative hierarchical clustering are introduced. As opposed to hierarchical clustering, k-means cluster analysis is related to a partitional clustering algorithm, which repeatedly assigns each object to its closest cluster center and calculates the coordinates of new cluster centers accordingly, until a predefined criterion is met. (*Bouguettaya et al.* (2015)) In case of two-step clustering procedure it may be possible to pre-cluster the observations into many small subclusters in the first step and group the subclusters into final clusters in the second step. (*Steiner-Hudec* (2007))

One of the differences between hierarchical and k-means clustering is that

as long as all the variables are of the same type, the hierarchical cluster analysis procedure can analyze both "scale" or "categorical" (for example also binary) variables, but the k-means cluster analysis procedure is basically limited to "scale" variables. In case of two-step clustering it may be possible that "scale" and "categorical" variables can be combined. In cluster analysis, usually the standardization of "scale" variables should be considered. Cluster analysis is usually applied for grouping cases, but clustering of variables (rather than cases) is also possible (in case of hierarchical cluster analysis). (*George–Mallery* (2007), page 262)

The algorithm used in (agglomerative) hierarchical cluster analysis starts by assuming that each case (or variable) can be considered as a separate cluster, then the algorithm combines clusters until only one cluster is left. If the number of cases is n, then the number of steps in the analysis is $n-1$ (it means that after $n-1$ steps all cases are in one cluster). The distance or similarity measures used in the hierarchical cluster analysis should be appropriate for the variables in the analysis. In case of "scale" variables (and assuming that the value of the jth variable in case of the ith observation is indicated by x_{ij}), for example the following distance or similarity measures can be used in the analysis (*Kovács* (2011), page 45):

- Euclidean distance: $\sqrt{\sum_j (x_{ij} - x_{kj})^2}$
- squared Euclidean distance: $\sum_j (x_{ij} - x_{kj})^2$
- Chebychev method for the calculation of distance: $\max_j |x_{ij} - x_{kj}|^2$
- City-block (Manhattan): $\sum_j |x_{ij} - x_{kj}|^2$
- „customized": $(\sum_j |x_{ij} - x_{kj}|^p)^{\frac{1}{r}}$
- etc.

The cluster methods in a hierarchical cluster analysis (methods for agglomeration in $n-1$ steps) can also be chosen, available alternatives are for example (*Kovács* (2011), pages 47-48):

- nearest neighbor method (single linkage): the distance between two clusters is equal to the smallest distance between any two members in the two clusters

- furthest neighbor method (complete linkage): the distance between two clusters is equal to the largest distance between any two members in the two clusters

- Ward's method: clusters are created in such a way to keep the within-cluster „variability" as small as possible

- within-groups linkage

- between-groups linkage

- centroid clustering

- median clustering.

One of the most important graphical outputs of a hierarchical cluster analysis is the dendrogram that displays distance levels at which objects and clusters have been combined.

The k-means cluster analysis is a non-hierarchical cluster analysis method that attempts to identify relatively homogeneous groups of cases in case of a specified number of clusters (this number of clusters is indicated by k). The distances in this analysis are computed based on Euclidean distance. The k-means cluster analysis applies iteration when cases are assigned to clusters. Iteration in this analysis starts with the selection of initial cluster centers (number of cluster centers is equal to k). Iteration stops when a complete iteration does not move any of the cluster centers by a distance of more than a given value. (*Kovács* (2014), page 61)

Some important outputs of the k-means cluster analysis:

- final cluster centers

- distances between cluster centers

- ANOVA-table (the F tests are only descriptive in this table)

- number of cases in clusters.

The maximum number of clusters is sometimes calculated as $\sqrt{\frac{n}{2}}$ (where n indicates the number of observations in the cluster analysis). (*Kovács* (2014), page 62) There are several methods how to choose the number of clusters in a cluster analysis. For example in addition to the studying of the dendrogram (if it is possible) the "cluster elbow method" can also provide information about the "optimal" number of clusters (*Kovács* (2014), page 62) The measurement of silhouette can also contribute to the selection of

an „appropriate" number of clusters (*Rousseeuw* (1987)). For each object it is possible to define values (for example for the ith observation $s(i)$) with an absolute value smaller or equal to one (in case of $s(i)$ the minimum can be minus one and the maximum can be one), so that a higher $s(i)$ value indicates a "better" clustering result. The "silhouette" of a cluster can be defined as a plot of $s(i)$ values (ranked in decreasing order). The average of the $s(i)$ values can be calculated and that number of cluster could be considered as "appropriate", for which the average of the $s(i)$ values is the largest. (*Rousseeuw* (1987))

2.2 Cluster analysis examples

In the following selected information society indicators (belonging to European Union member countries, for the year 2015) are analyzed: data is downloadable from the homepage of Eurostat[1] and it is also presented in the Appendix. In the following cluster analysis examples are presented with the application of the following three variables:

- "ord": individuals using the internet for ordering goods or services
- "ord_EU": individuals using the internet for ordering goods or services from other EU countries
- "enterprise_ord": enterprises having received orders online.

Question 2.1. *Conduct hierarchical cluster analyis (with squared Euclidean distance and Ward's method). Do Spain and Luxembourg belong to the same cluster, if the number of clusters is equal to 2?*

Solution of the question.

In case of cluster analysis "scale" variables are usually standardized. Standardization of a variable can be considered as a (relatively) straightforward procedure: the average value is subtracted from the value (for each case separately) and then this difference is divided by the standard deviation. As a result of this calculation, the mean of a standardized variable is equal to zero and the standard deviation is equal to one. Standardized variables can

[1]Data source: homepage of Eurostat (http://ec.europa.eu/eurostat/web/information-society/data/main-tables)

be created in SPSS by performing the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Descriptive Statistics → Descriptives...

In this example, in the appearing dialog box the variables "ord", "ord_EU" and "enterprise_ord" should be selected as "Variable(s):", and the option "Save standardized values as variables" should also be selected. After clicking "OK" the standardized variables are created.

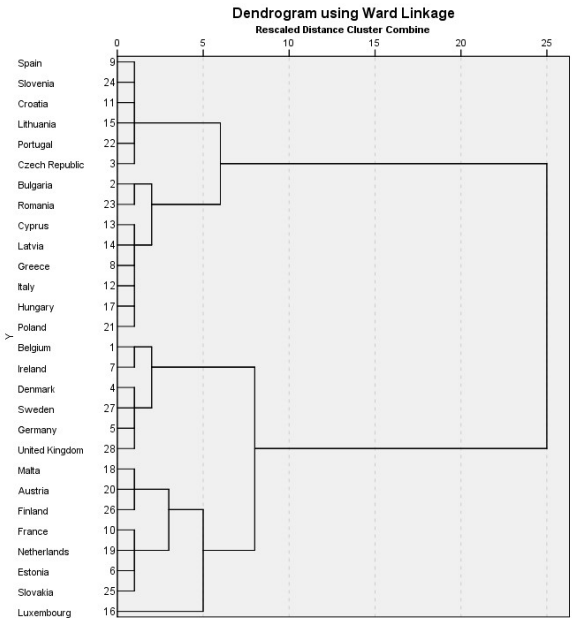


Figure 2.1: Dendrogram with Ward linkage

To conduct a hierarchical cluster analysis in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Classify → Hierarchical Cluster...

As a next step, in the appearing dialog box select the three standardized variables as "Variable(s):", and in case of "Method..." button select "Ward's method" as "Cluster Method:" and "Squared Euclidean distance" as "Measure". The dendrogram is displayed as an output, if the "Dendrogram" option is selected in case of "Plots..." button.

The dendrogram is shown by Figure 2.1. It can be observed, that if the number of clusters is equal to 2, then the number of countries in both clusters is equal to 14, and it can also be observed that Spain and Luxembourg do not belong to the same cluster.

Question 2.2. *Conduct hierarchical cluster analyis (with Euclidean distance and nearest neighbor method). How many countries are in the clusters, if the number of clusters is equal to 2?*

Solution of the question.

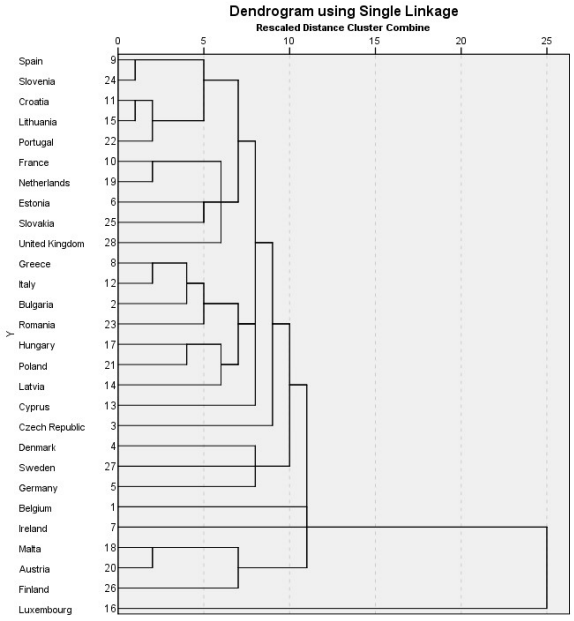


Figure 2.2: Dendrogram with single linkage

In case of the dialog box (belonging to hierarchical cluster analysis in SPSS) similar options can be selected as for Question 2.1, the only difference is in case of the "Method..." button: to solve Question 2.2 "Nearest neighbor" should be selected as "Cluster method", and "Euclidean distance" should be selected as "Measure".

Figure 2.2 shows the dendrogram (belonging to nearest neighbor method and Euclidean distance). It can be observed on Figure 2.2 that if the number of clusters is equal to 2, then in one of the clusters there is only one country

(Luxembourg), thus (since the number of countries in the analysis is 28), the number of countries in the clusters are 27 and 1, respectively.

Question 2.3. *Conduct hierarchical cluster analyis (with Euclidean distance and nearest neighbor method). Which two countries belong to the first cluster (in the process of agglomeration in hierarchical cluster analysis) that has at least two elements?*

Solution of the question.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	24	,219	0	0	10
2	11	15	,300	0	0	5
3	10	19	,334	0	0	12
4	8	12	,341	0	0	8
5	11	22	,376	2	0	10
6	18	20	,380	0	0	17
7	17	21	,505	0	0	13
8	2	8	,508	0	4	9
9	2	23	,569	8	0	16
10	9	11	,593	1	5	15
11	6	25	,612	0	0	12
12	6	10	,628	11	3	14
13	14	17	,629	0	7	16
14	6	28	,665	12	0	15
15	6	9	,714	14	10	20
16	2	14	,749	9	13	19
17	18	26	,786	6	0	26
18	4	27	,796	0	0	21
19	2	13	,809	16	0	20
20	2	6	,836	19	15	22
21	4	5	,843	18	0	23
22	2	3	,911	20	0	23
23	2	4	,953	22	21	24
24	1	2	1,067	0	23	25
25	1	7	1,081	24	0	26
26	1	18	1,109	25	17	27
27	1	16	2,248	26	0	0

Table 2.1: Agglomeration schedule in hierarchical cluster analysis

In this case the same options should be selected in the dialog box (that belongs to hierarchical cluster analysis in SPSS) as in case of Question 2.2. One of the outputs is the "Agglomeration schedule" that summarizes information about the process of agglomeration in hierarchical cluster analysis. Table 2.1 shows this "Agglomeration schedule". It is worth noting that this "Agglomeration schedule" contains information about 27 steps in the process of agglomeration (since the number of cases in the analysis is equal to 28).

In the first row of this table it can be observed that the countries indicated by "9" and "24" are in the first cluster that has two elements (before the first step in the agglomeration process each case can be considered to be a cluster that contains one element). Thus, in this example Spain (the country indicated by "9") and Slovenia (the country that is indicated by "24" in this case) belong to the first cluster that has at least two elements.

Question 2.4. *Conduct k-means cluster analyis with k=2. Which variables should be omitted from the analysis?*

Solution of the question.

To conduct a k-means cluster analysis in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Classify → K-Means Cluster...

As a next step, in the appearing dialog box select the three standardized variables as "Variables:". As "Number of Clusters:" 2 should be written (it is also the default value) and in case of the "Options..." button select "ANOVA table".

Table 2.2 shows the ANOVA table that is one of the outputs of the k-means cluster analysis. In the last column of this ANOVA table all values can be considered as relatively small (smaller than 0.05, but in this case these values can not be interpreted exactly in the same way as in case of a "classical" hypothesis testing, since the results of the presented F tests should only be applied for descriptive purposes). The conclusion is that no variables should be omitted from the analysis in this example.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(ord)	19,903	1	,273	26	72,918	,000
Zscore(ord_EU)	13,132	1	,533	26	24,622	,000
Zscore(enterprise_ord)	6,141	1	,802	26	7,655	,010

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 2.2: ANOVA in k-means cluster analysis

Question 2.5. *Which can be considered as the optimal number of clusters in k-means cluster analysis according to the "cluster elbow" method?*

Solution of the question.

In the following, "cluster elbow" calculations are introduced based on *Kovács* (2014) (page 62). Since the solution of Question 2.4 indicates that none of the variables should be omitted from the analysis, these three (standardized) variables are applied in k-means cluster analyses (so that the cluster membership variables are saved). The (standardized) variables and the cluster membership variables are applied in one-way ANOVA (the results for $k = 2$ are shown by Table 2.3).

To conduct one-way ANOVA in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Compare Means → One-Way ANOVA...

As a next step, in the appearing dialog box select the three standardized variables in case of the "Dependent List:" and the saved cluster membership variable as "Factor".

The "cluster elbow" graph (demonstrated by Figure 2.3) plots certain ratios against k values, for example the ratio (plotted on the "cluster elbow" graph) for $k = 2$ can be calculated as follows (based on the values in Table 2.3):

$$\frac{19.903 + 13.132 + 6.141}{27 + 27 + 27} = \frac{39.176}{81} = 0.48 \tag{2.1}$$

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Zscore(ord)	Between Groups	19,903	1	19,903	72,918	,000
	Within Groups	7,097	26	,273		
	Total	27,000	27			
Zscore(ord_EU)	Between Groups	13,132	1	13,132	24,622	,000
	Within Groups	13,868	26	,533		
	Total	27,000	27			
Zscore(enterprise_ord)	Between Groups	6,141	1	6,141	7,655	,010
	Within Groups	20,859	26	,802		
	Total	27,000	27			

Table 2.3: ANOVA with cluster membership variable

Figure 2.3 shows the "cluster elbow" graph. According to *Kovács* (2014) (page 62) the "optimal" number of cluster corresponds to that k value, where (on the graph) the slope of the graph becomes smaller. In this example it can not be considered as obvious which k corresponds to this requirement. Both $k = 2$ and $k = 3$ could be chosen ($k = 4$ should not be chosen, since $\sqrt{\frac{28}{2}} < 4$), thus it could depend on the other features of the analysis, which k is considered as "optimal" (for example the number of cases in the clusters could be compared for the solutions where $k = 2$ or $k = 3$).

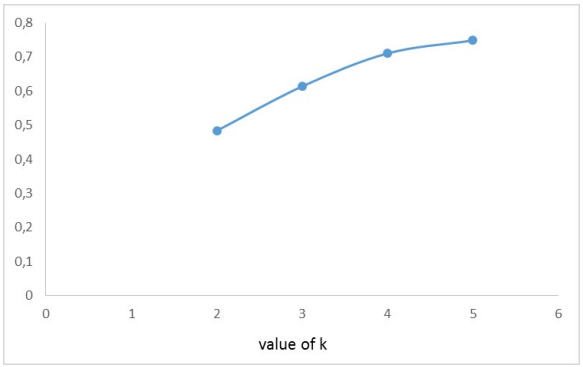


Figure 2.3: Cluster elbow method calculation results

Question 2.6. *Conduct k-means cluster analysis with k=2. How can the result about the final cluster centers be interpreted?*

Solution of the question.

In this case the same options should be selected in the dialog box (that belongs to hierarchical cluster analysis in SPSS) as in case of Question 2.4.

Table 2.4 contains information about the final cluster centers (in case of $k = 2$). Since the analysis has been carried out with standardized variables (when the average value of each variable is equal to zero), thus in Table 2.4 positive numbers can be interpreted as "above average" values (and negative numbers refer to "below average" values). For example in case of Cluster 1 all values are above average, and thus the "name" of this cluster (if a "name" should be given to the cluster) should refer to the names of the variables: in this example the "name" Cluster 1 should somehow express that the use of internet for "online ordering" is more widespread in the countries that belong to this cluster (compared to the countries belonging to the other cluster).

Final Cluster Centers

	Cluster	
	1	2
Zscore(ord)	,90564	-,78489
Zscore(ord_EU)	,73564	-,63756
Zscore(enterprise_ord)	,50307	-,43599

Table 2.4: Final cluster centers in k-means cluster analysis

Question 2.7. *Which could be considered as an "appropriate" number of clusters in two-step cluster analysis?*

Solution of the question.

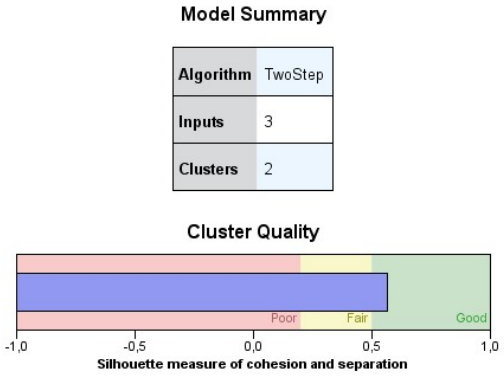


Figure 2.4: Results of two-step clustering

One of the advantages of two-step clustering is that it makes the combination of "scale" and "categorical" variables possible ("scale" and "categorical" variables can be applied simultaneously in a two-step cluster analysis). An other advantage of two-step cluster analysis is that it can "recommend" an "ideal" number of clusters. (*Pusztai* (2007), pages 325-326) To conduct a two-step cluster analysis in SPSS perform the following sequence (beginning

with selecting "Analyze" from the main menu):

Analyze → Classify → TwoStep Cluster...

As a next step, in the appearing dialog box select the three (standardized) variables as "Continuous Variables:" and then click "OK".

Figure 2.4 shows some results of the two-step cluster analysis, and according to these results 2 can be considered as an "appropriate" number of clusters. A silhouette measure is also calculated and Figure 2.4 shows that this silhouette measure is higher than 0.5.

3 | Factor analysis

As a dimension reduction method, factor analysis is widely applied in econometric model building. (*McNeil et al.* (2005), page 103) Factor analysis refers to a set of multivariate statistical methods aimed at exploring relationships between variables. The methods applied in factor analysis can be grouped into two categories: exploratory factor analysis (aimed at creating new factors) and confirmatory factor analysis (applicable for testing an existing model). (*Sajtos-Mitev* (2007), pages 245-247) In this chapter only selected features of exploratory factor analysis are discussed.

3.1 Theoretical background

Factor analysis attempts to identify underlying variables (factors) that explain most of the variance of variables (X_j, $j = 1, \ldots, p$). The factor analysis model assumes that variables are determined by common factors and unique factors (so that all unique factors are uncorrelated with each other and with the common factors). The factor analysis model can be described as follows (*Kovács* (2011), page 95):

$$X = FL^T + E \tag{3.1}$$

where matrix X has n rows and p columns, matrix F has n rows and k columns (where the number of common factors is indicated by $k < p$), matrix L contains the factor loadings and matrix E denotes the "errors". (*Kovács* (2011), page 95) It belongs to the assumptions of the factor analysis model that (*Kovács* (2011), page 95):

- $\frac{F^T F}{n} = I$, where I denotes the indentity matrix
- $F^T E = E^T F = 0$
- $\frac{E^T E}{n}$ is the covariance matrix of the „errors" and it is assumed that this matrix is diagonal.

An important equation in factor analysis is related to the reproduction of the correlation matrix (*Kovács* (2011), page 95):

$$R = \frac{X^T X}{n} = \frac{(FL^T + E)^T (FL^T + E)}{n} = LL^T + \frac{E^T E}{n} \tag{3.2}$$

In case of factor analysis (if $\frac{E^T E}{n}$ is known) usually the eigenvalue-eigenvector decomposition of the reduced correlation matrix (LL^T) is calculated. In principal component analysis (that is one of the methods for factor extraction in factor analysis) the variance values of the "errors" however usually have to be estimated. In a factor analysis it is possible that the eigenvalues of matrix LL^T are negative values.

Correlation coefficients are important in the interpretation of factor analysis results:

- a (simple) linear correlation coefficient describes the linear relationship between two variables (if the relationship is not linear, this correlation coefficient is not an appropriate statistic for the measurement of the strength of the relationship of variables).
- a partial correlation coefficient describes the linear relationship between two variables while controlling for the effects of one or more additional variables.

Correlation coefficients are used for example in the calculation of the KMO (Kaiser-Meyer-Olkin) measure of sampling adequacy as follows (*Kovács* (2011), page 95):

$$\frac{\sum_{i=1}^{p} \sum_{j \neq i} r_{ij}^2}{\sum_{i=1}^{p} \sum_{j \neq i} r_{ij}^2 + \sum_{i=1}^{p} \sum_{j \neq i} q_{ij}^2} \tag{3.3}$$

where r_{ij} indicates the (Pearson) correlation coefficients (in case of the variables in an analysis) and q_{ij} denotes the partial correlation values. The KMO value shows whether the partial correlations among variables (X_j $j = 1, \ldots, p$) are small "enough", because relatively large partial correlation coefficients are not advantageous in case of factor analysis. For example if the KMO value is smaller than 0.5, then the data should not be analyzed with factor analysis (*George–Mallery* (2007), page 256) If the KMO value is

above 0.9, then sample data can be considered as excellent (from the point of view of applicability in case of factor analysis). (*Kovács* (2014), page 156)

The Bartlett's test of sphericity also can be used to assess adequacy of data for factor analysis. Bartlett's test of sphericity tests whether the correlation matrix is an identity matrix (in that case the factor model is inappropriate). (*Kovács* (2014), page 157)

Data about partial correlation coefficients can also be found in the anti-image correlation matrix. The off-diagonal elements of the anti-image correlation matrix are the negatives of the partial correlation coefficients (in a good factor model, the off-diagonal elements should be small), and on the diagonal of the anti-image correlation matrix the measure of sampling adequacy for a variable is displayed. (*Kovács* (2014), page 156)

There are numerous methods for factor extraction in a factor analysis, for example (*Kovács* (2011), pages 106-107):

- Principal Component Analysis: uncorrelated linear combinations of the variables in the analysis are calculated

- Unweighted Least-Squares Method: minimizes the sum of the squared differences between the observed and reproduced correlation matrices (when the diagonals are ignored)

- Principal Axis Factoring: extracts factors from the correlation matrix (iterations continue until the changes in the communalities satisfy a given convergence criterion)

- Maximum Likelihood method: it can be applied if the variables in the analysis follow a multivariate normal distribution

- etc.

Exploratory factor analysis methods can be grouped into two categories: common factor analysis and principal component analysis. (*Sajtos-Mitev* (2007), page 249) In the following principal component analysis is discussed.

Assume that the (standardized) variables in the analysis are denoted by $X_1, \ldots, X_p$, where p is the number of variables in the analysis. The matrix where the columns correspond to the $X_1, \ldots, X_p$ variables is denoted by X in the following. In the principal component analysis the variables Y_i $(i = 1, \ldots, p)$ should be calculated as linear combinations of the variables $X_1, \ldots, X_p$:

$$Y = XA \tag{3.4}$$

It means that for example Y_1 is calculated as follows:

$$Y_1 = Xa_1 \tag{3.5}$$

where (according to the assumptions) $a_1^T a_1 = 1$ (the sum of squares of coefficients is equal to 1). (*Kovács*(2014), page 150)

The correlation matrix of X_j $(j = 1, \ldots, p)$ variables is denoted by R. In case of standardized X_j $(j = 1, \ldots, p)$ variables the variance of the first component is (as described for example in *Kovács* (2011), pages 90-93):

$$Var(Y_1) = a_1^T R a_1 = \lambda_1 \tag{3.6}$$

This result means that the variance of the first component depends also on the values in vector a. The variance of the first component has its maximum value if (by assuming that $a_1^T a_1 = 1$):

$$Ra_1 = \lambda_1 a_1 \tag{3.7}$$

It means that the maximum value of $Var(Y_1) = a_1^T R a_1 = \lambda_1$ can be calculated based on the eigenvalue-eigenvector decomposition of the matrix R. In this eigenvalue-eigenvector decomposition the λ_i $(i = 1, \ldots, p)$ values are the eigenvalues and the a_i $(i = 1, \ldots, p)$ vectors are the eigenvectors. In case of the eigenvalue-eigenvector decomposition of the correlation matrix R the sum of eigenvalues is equal to p (the number of X_j variables). It is worth emphasizing that the variance of the component is the eigenvalue: for example $a_1^T R a_1 = \lambda_1$. (*Kovács* (2014), pages 150-151)

The condition $a_1^T a_1 = 1$ means that the length of a_i $(i = 1, \ldots, p)$ eigenvectors is equal to 1. Eigenvectors with length not equal to 1 also can be calculated:

$$c_i = a_i \sqrt{\lambda_i} \tag{3.8}$$

The elements of the vectors c_i can be interpreted as correlation coefficients between the j-th variable and the i-th component. (*Kovács* (2011), page 93) In the following assume that a matrix is created so that the columns correspond to the c_i vectors (assume that this matrix is denoted by C). Matrix C is not necessarily a symmetric matrix. The correlation matrix (R) can be "reproduced" with the application of matrix C.

Matrix C can be called "component matrix" and it is possible that in a calculation output the component matrix shows only those components that have been extracted in the analysis. Based on the component matrix, the eigenvalues and communality values can also be calculated. The communality is that part of the variance of a variable X_j $(j = 1, \ldots, p)$ that is explained

by the (extracted) components. (*Kovács* (2014), page 157) If in a principal component analysis all components are extracted, then the communality values are equal to one. However, in other factor analysis methods the maximum value of communality can be smaller than one: for example in case of a factor analysis with Principal Axis Factoring eigenvalue-eigenvector decomposition is related to a "reduced" correlation matrix (and not the correlation matrix) that is calculated so that the diagonal values of the correlation matrix (that are equal to one) are replaced by estimated communality values. Thus, in case of a factor analysis with Principal Axis Factoring the calculated eigenvalues (that belong to the "reduced" correlation matrix) theoretically may be negative values. (*Kovács* (2014), pages 165-167)

As a result of principal component analysis, in some cases "names" can be given to the components (based on the component matrix). Sometimes rotation of the component matrix is needed in order to achieve a "simple structure" (in absolute values high component loadings on one component and low loadings on all other components, in an optimal case for all variables). (*George – Mallery* (2007), page 248)

3.2 Factor analysis examples

In the following (similar to Chapter 2) selected information society indicators (belonging to European Union member countries, for the year 2015) are analyzed: data is downloadable from the homepage of Eurostat[1] and it is also presented in the Appendix. Factor analysis examples are presented with the application of the following five variables:

- "ord": individuals using the internet for ordering goods or services
- "ord_EU": individuals using the internet for ordering goods or services from other EU countries
- "reg_int": individuals regularly using the internet
- "never_int": individuals never having used the internet
- "enterprise_ord": enterprises having received orders online

Question 3.1. *Conduct principal component analysis with the five variables and calculate (and interpret) the KMO value.*

[1]Data source: homepage of Eurostat (http://ec.europa.eu/eurostat/web/information-society/data/main-tables)

Solution of the question.

To conduct principal component analysis in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Dimension Reduction → Factor...

As a next step, in the appearing dialog box select the variables "ord", "ord_EU", "reg_int", "never_int" and "enterprise_ord" as "Variables:", select the "Descriptives..." button, and then select the "KMO and Bartlett's test of sphericity" option. Table 3.1 shows the calculation results: the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is equal to 0.779.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,779
Bartlett's Test of Sphericity	Approx. Chi-Square	170,568
	df	10
	Sig.	,000

Table 3.1: KMO measure of sampling adequacy

This result can be interpreted so that data is suitable for principal component analysis, since the KMO value is higher than 0.5. More precisely, the suitability of data for principal component analysis can be assessed as "average", since KMO measure is between 0.7 and 0.8. According to *Kovács* (2014) (pages 155-156) the suitability of data for principal component analysis can be assessed as follows:

Table 3.2: Assessment of data suitability in principal component analyis

KMO value	data suitability
smaller than 0.5	data not suitable
between 0.5 and 0.7	weak
between 0.7 and 0.8	average
between 0.8 and 0.9	good
higher than 0.9	excellent

Question 3.2. *Conduct principal component analysis with the five variables and calculate (and interpret) the anti-image correlation matrix.*

Solution of the question.

In this case the solution of Question 3.1 can be applied with the difference that in case of the "Descriptives..." button the "Anti-image" option should also be selected. Table 3.3 shows the anti-image correlation matrix.

Anti-image Matrices

		ord	ord_EU	reg_int	never_int	enterprise_ord
Anti-image Covariance	ord	,082	,001	-,013	,012	-,028
	ord_EU	,001	,393	-,053	-,042	,131
	reg_int	-,013	-,053	,024	,022	-,010
	never_int	,012	-,042	,022	,029	-,002
	enterprise_ord	-,028	,131	-,010	-,002	,714
Anti-image Correlation	ord	,940[a]	,007	-,286	,249	-,115
	ord_EU	,007	,720[a]	-,543	-,391	,248
	reg_int	-,286	-,543	,708[a]	,825	-,080
	never_int	,249	-,391	,825	,736[a]	-,013
	enterprise_ord	-,115	,248	-,080	-,013	,896[a]

a. Measures of Sampling Adequacy(MSA)

Table 3.3: Anti-image correlation matrix

The elements in the main diagonal of the anti-image correlation matrix correspond to the "individual" KMO values (calculated for each variable separately). The "individual" KMO for the ith variable can be calculated (based on the r_{ij} Pearson correlation coeffients and the q_{ij} partial correlation coefficients) as follows (*Kovács* (2014), page 156):

$$\frac{\sum\limits_{j\neq i} r_{ij}^2}{\sum\limits_{j\neq i} r_{ij}^2 + \sum\limits_{j\neq i} q_{ij}^2} \tag{3.9}$$

If a KMO value in the main diagonal of the anti-image correlation matrix is smaller than 0.5, then the given variable should be omitted from the analysis (*Kovács* (2014), page 156). In this example none of the variables should be omitted from the analysis as a consequence of low KMO values. The off-diagonal elements of the anti-image correlation matrix correspond to

the negatives of the partial correlations. In a good factor model the partial correlations should be close to zero. (*Kovács* (2014), page 156)

Question 3.3. *Assume that principal component analysis is conducted with the five variables. How many components are extracted?*

Solution of the question.

In SPSS, the same options should be selected as in case of the solution of Question 3.1. Tabel 3.4 contains information about the extracted components: in this case only one component is extracted.

The default option in SPSS is to extract those components for which the calculated eigenvalue (that belongs to the component) is at least one. (*Kovács* (2014), page 157) It may be easier to understand this default option, if it is emphasized that in this principal component analysis the eigenvalue-eigenvector decomposition of the correlation matrix is analyzed. The correlation matrix belonging to the unstandardized and standardized variables is the same. The eigenvalues of the correlation matrix can be interpreted as variance values (belonging to the components), and the variance of a standardized variable is one. Thus, the default option for extracting components can be interpreted so that only those components are extracted, for which the calculated variance (eigenvalue) is higher (or maybe equal) to the variance of a standardized variable. In this case (with the extraction of one component) $\frac{3.692}{5} = 73.832\%$ of total variance is explained.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,692	73,832	73,832	3,692	73,832	73,832
2	,866	17,319	91,151			
3	,372	7,448	98,599			
4	,056	1,116	99,715			
5	,014	,285	100,000			

Extraction Method: Principal Component Analysis.

Table 3.4: Total variance explained (5 variables)

Question 3.4. *Assume that principal component analysis is conducted in two cases: with the five variables and without the "enterprise_ ord" variable (with four variables). Compare the communality values in these two cases!*

Communalities

	Initial	Extraction
ord	1,000	,931
ord_EU	1,000	,546
reg_int	1,000	,966
never_int	1,000	,941
enterprise_ord	1,000	,307

Extraction Method: Principal Component Analysis.

(a) 5 variables

Communalities

	Initial	Extraction
ord	1,000	,928
ord_EU	1,000	,608
reg_int	1,000	,973
never_int	1,000	,940

Extraction Method: Principal Component Analysis.

(b) 4 variables

Table 3.5: Comparison of communalities

Solution of the question.

To solve this question, the same options should be selected (in SPSS) as in case of the solution of Question 3.1. Table 3.5 shows the communality values in the two cases (for the principal component analysis with 5 and 4 variables). In the first case (the principal component analysis with 5 variables) the communality value belonging to the variable "enterprise_ord" is relatively low (compared to the other communality values): the communality value belonging to "enterprise_ord" is equal to 0.307. According to *Kovács* (2011) (page 99) it may be considered to omit variables with a communality value of less than 0.25 from the principal component analysis. Although the variable "enterprise_ord" could remain in the analysis, Table 3.5 shows that if the variable "enterprise_ord" is omitted from the principal component analysis, then the lowest communality value is 0.608 (belonging to the variable "ord_EU"). It is also worth mentioning that the communality values belonging to the four variables in the second principal component analysis changed (compared to the first principal component analysis with five variables): for example the communality value belonging to the variable "ord" is 0.931 in the first principal component analysis (with 5 variables) and 0.928 in the second principal component analysis (with 4 variables).

Question 3.5. *Assume that principal component analysis is conducted in two cases: with the five variables and without the "enterprise_ord" variable (with four variables). Compare the component matrices in these two cases!*

Solution of the question.

In case of this question the same options should be selected (in SPSS) as in case of the solution of Question 3.1. Table 3.6 shows the two component matrices that contain the correlation values between the variables in

Component Matrix[a]

	Component
	1
ord	,965
ord_EU	,739
reg_int	,983
never_int	-,970
enterprise_ord	,554

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

(a) 5 variables

Component Matrix[a]

	Component
	1
ord	,963
ord_EU	,780
reg_int	,986
never_int	-,969

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

(b) 4 variables

Table 3.6: Component matrices

the analysis and the components. The component matrix can contribute to interprete the components (maybe to give a "name" to a component). It can be observed that in the principal component analysis with 5 variables the correlation between the variable "enterprise_ord" and the first (extracted) component is relatively low (in absolute value, compared to the other values in the component matrix). This result is associated with the results of Question 3.4: the communality value belonging to the variable "enterprise_ord" is relatively low (compared to the other communality values in the principal component analysis with 5 variables). After omitting the variable "enterprise_ord" from the principal component analysis it could be easier to interpret the extracted component. Since the variables "ord", "ord_EU" and "reg_int" are positively and the "never_int" variable is negatively correlated with the first component (and the absolute values of correlations in the component matrix are relatively high), the extracted component (in case of the principal component analysis with 4 variables) may be interpreted for example as an indicator of the state of development of information society (of course, other interpretations may also be possible).

Question 3.6. *Conduct principal component analysis with the variables "ord", "ord_EU", "reg_int" and "never_int", and calculate the reproduced correlation matrix. How can the diagonal values in the reproduced correlation matrix be interpreted?*

Solution of the question.

In this case the solution of Question 3.1 can be applied with the difference that in case of the "Descriptives..." button the "Reproduced" option should also be selected. Table 3.7 shows the reproduced correlation matrix. The diagonal values of the reproduced correlation matrix are the communality values (for example 0.928 is the communality value belonging to the variable "ord").

Reproduced Correlations

		ord	ord_EU	reg_int	never_int
Reproduced Correlation	ord	,928[a]	,751	,950	-,934
	ord_EU	,751	,608[a]	,769	-,756
	reg_int	,950	,769	,973[a]	-,956
	never_int	-,934	-,756	-,956	,940[a]
Residual[b]	ord		-,117	,004	-,018
	ord_EU	-,117		-,069	,129
	reg_int	,004	-,069		-,025
	never_int	-,018	,129	-,025	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 3 (50,0%) nonredundant residuals with absolute values greater than 0.05.

Table 3.7: Reproduced correlation matrix

The communality values can also be calculated based on the component matrix, for example the communality value belonging to the variable "ord" can be calculated in this example as $0.963^2 = 0.928$. The reproduced correlation matrix can be calculated based on the component matrix as follows:

$$\begin{pmatrix} 0.963 \\ 0.780 \\ 0.986 \\ -0.969 \end{pmatrix} \begin{pmatrix} 0.963 & 0.780 & 0.986 & -0.969 \end{pmatrix} = \\ = \begin{pmatrix} 0.928 & 0.751 & 0.950 & -0.934 \\ 0.751 & 0.608 & 0.769 & -0.756 \\ 0.950 & 0.769 & 0.973 & -0.956 \\ -0.934 & -0.756 & -0.956 & 0.940 \end{pmatrix} \tag{3.10}$$

The eigenvalues may also be calculated based on the component matrix. In this example (with one extracted component) the first (highest) eigenvalue can be calculated as follows:

$$\begin{pmatrix} 0.963 & 0.780 & 0.986 & -0.969 \end{pmatrix} \begin{pmatrix} 0.963 \\ 0.780 \\ 0.986 \\ -0.969 \end{pmatrix} = 3.448 \tag{3.11}$$

It is also possible to display all columns of the component matrix (not only the column that belongs to the extracted component). In this case the solution of Question 3.1 can be applied with the difference that in case of the "Extraction..." button the "Fixed number of factors" option should be selected (instead of the "Based on Eigenvalue" option), with selecting 4 as the number of factors to extract. The resulting component matrix has 4 columns, and the eigenvalues (belonging to the components) can be calculated based on the component matrix as follows:

$$\begin{pmatrix} 0.928 & 0.780 & 0.986 & -0.969 \\ -0.189 & 0.626 & -0.108 & -0.206 \\ 0.191 & 0.007 & -0.086 & 0.108 \\ -0.004 & -0.011 & 0.090 & 0.078 \end{pmatrix}$$
$$\begin{pmatrix} 0.963 & -0.189 & 0.191 & -0.004 \\ 0.780 & 0.626 & 0.007 & -0.0011 \\ 0.986 & -0.108 & -0.086 & 0.090 \\ -0.969 & 0.206 & 0.108 & 0.078 \end{pmatrix} = \tag{3.12}$$
$$= \begin{pmatrix} 3.448 & 0 & 0 & 0 \\ 0 & 0.481 & 0 & 0 \\ 0 & 0 & 0.056 & 0 \\ 0 & 0 & 0 & 0.014 \end{pmatrix}$$

The result of multiplying the transpose of the component matrix with the component matrix is a diagonal matrix, in which the diagonal values correspond to the eigenvalues (of the correlation matrix in this example).

4 | Multidimensional scaling

Multidimensional scaling is a methodology that can be applied to reduce dimensionality using only the information about similarities or dissimilarities of objects (for example similarities of cases in an analysis). With multidimensional scaling (MDS) it may be possible to represent objects (for example cases in an analysis) in a low dimensional space. (*Bécavin et al.* (2011) Multidimensional scaling methods can be grouped into two categories: classical (metric) scaling and non-metric scaling. (*Kovács* (2011), page 142) Classical scaling may be applied to embed a set of objects in the simplest space possible, with the constraint that the Euclidean distance between data points is preserved. (*Bécavin et al.* (2011)) Non-metric multidimensional scaling assumes that the proximities (used to assess similarities) represent ordinal information about distances (*Balloun-Oumlil* (1988)), and it aims at producing a configuration of points in a (usually Euclidean) space of low dimension, where each point represents an object (for example a case in the analysis). (*Cox-Ferry* (1993))

4.1 Theoretical background

Distance measurement has a central role in multidimensional scaling. At the beginning of the analysis the distances between pairs of items should be measured (these distances are indicated by δ_{ij} in the following). These distances can be compared to other distance values between pairs of items (indicated by for example d_{ij}) that can be calculated in a low-dimensional coordinate system. The original distances δ_{ij} may be "proximity" or "similarity" values, but the distances d_{ij} (that can be calculated in a low-dimensional coordinate system) are usually Euclidean distances. (*Rencher-Christensen* (2012), page 421)

One of the most important outputs in multidimensional scaling is a plot that shows how the items in the analysis relate to each other. Either variables or cases can be considered as "items" in multidimensional scaling. The

"level of measurement" can be "interval" or "ratio" (in metric multidimensional scaling) or "ordinal" (in nonmetric multidimensional scaling). (*Kovács* (2011), pages 141- 142)

In metric multidimensional scaling (also known as the "classical solution") an important element in the calculation of the results is the spectral decomposition of a symmetric matrix (indicated by M), that can be calculated based on the originally calculated distance matrix (where the elements of this distance matrix are indicated by δ_{ij}). If this symmetric matrix M is positive semidefinite of rank q, then the number of positive eigenvalues is q and the number of zero eigenvalues is $n - q$. In multidimensional scaling the preferred dimension in the analysis (indicated by k) is often smaller than q, and in this case the first k eigenvalues and the corresponding eigenvectors can be applied to calculate "coordinates" for the n items in the analysis so that the "interpoint" distances (indicated by d_{ij}, in case of k dimensions) are approximately equal to the corresponding δ_{ij} values. If the symmetric matrix M is not positive semidefinite, but the first k eigenvalues are positive and relatively large, then these eigenvalues and the corresponding eigenvectors may sometimes be applied to calculate "coordinates" for the n items in the analysis. (*Rencher-Christensen* (2012), pages 421-422) It is worth mentioning that it is possible that principal component analysis and classical scaling give the same results (*Bécavin et al.* (2011))

Instead of metric multidimensional scaling it is worth applying nonmetric multidimensional scaling if the original distances δ_{ij} are only "proximity" or "similarity" values. In this case in nonmetric multidimensional scaling only the rank order among the "similarity" or "proximity" values are preserved by the final spatial representation. (*Rencher-Christensen* (2012), page 421) In nonmetric multidimensional scaling it is assumed that the original δ_{ij} "dissimilarity" values can be ranked in order and the goal of the analysis is to find a low-dimensional representation of the „points" (related to the items in the analysis) so that the rankings of the distances d_{ij} match exactly the ordering of the original δ_{ij} "dissimilarity" values. (*Rencher-Christensen* (2012), page 425)

Results for nonmetric multidimensional scaling can be calculated with an iteration process. With a given k value and an initial configuration the d_{ij} "interitem" distances and the corresponding $\hat{d_{ij}}$ values (as a result of a monotonic regression) can be calculated. The $\hat{\delta_{ij}}$ values can be estimated by monotonic regression with the minimization of the following scaled sum of squared differences (*Rencher-Christensen* (2012), page 426):

$$S^2 = \frac{\sum_{i<j}(d_{ij} - \hat{d_{ij}})^2}{\sum_{i<j}(d_{ij})^2} \tag{4.1}$$

For a given dimension (k value) the minimum value of S^2 is called STRESS. In the iteration process a new configuration of points (related to the "items" in the analysis) should be calculated so that this S^2 value is minimized with respect to the given $\hat{d_{ij}}$ values and then for this new configuration (and the corresponding new d_{ij} "interitem" distance values) the corresponding new $\hat{\delta_{ij}}$ values should be calculated with monotonic regression. This iterative process should continue until STRESS value converges to a minimum. The $\hat{d_{ij}}$ values are sometimes referred to as disparities. (*Rencher-Christensen* (2012), page 426) The Stress value may be applied to measure the "goodness" of the fit of the model, depending on the value of S in the following equation (*Kovács (2011)*, page 146):

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d_{ij}})^2}{\sum_{i<j}(d_{ij})^2}} \tag{4.2}$$

If for example $S < 0.05$, then the solution can be evaluated as good, while for $S > 0.2$ the solution can be evaluated as weak. (*Kovács* (2011), page 146)

With an individual difference model (INDSCAL) it is possible to use more than one "dissimilarity" matrix in one multidimensional scaling analysis (*George – Mallery* (2007), page 236) In an individual difference model weights can be calculated that show the importance of each dimension to the given subjects. (*George–Mallery* (2007), page 243) In an INDSCAL analysis MDS coordinates can be calculated in a "common" space and in "individual" spaces so that the relationship between the "common" space and the "individual" spaces is described by the individual weights. (*Kovács (2011)*, pages 155-156)

4.2 Multidimensional scaling examples

In the following (similar to Chapter 3) five variables (selected information society indicators belonging to European Union member countries) are analyzed: data is downloadable from the homepage of Eurostat[1] and it is also

[1]Data source: homepage of Eurostat (http://ec.europa.eu/eurostat/web/information-society/data/main-tables)

presented in the Appendix. For ALSCAL analysis data for 2015 is analyzed, and INDSCAL analysis is carried out with data for both 2010 and 2015. Multidimensional scaling examples are presented with the application of the following five variables:

- "ord": individuals using the internet for ordering goods or services
- "ord_EU": individuals using the internet for ordering goods or services from other EU countries
- "reg_int": individuals regularly using the internet
- "never_int": individuals never having used the internet
- "enterprise_ord": enterprises having received orders online

Question 4.1. *Conduct multidimensional scaling (with ALSCAL method) with the five (standardized) variables (in case of variables, level of measurement: ordinal). How can the model fit be evaluated if the number of dimensions is equal to 1 or 2?*

Solution of the question.

To conduct multidimensional scaling in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Scale → Multidimensional Scaling (ALSCAL)...

As a next step, in the appearing dialog box select the variables "ord", "ord_EU", "reg_int", "never_int" and "enterprise_ord" as "Variables:". In the dialog box the option "Create distances from data", and then the "Measure..." button should be selected. In the appearing dialog box in case of "Standardize:" the "Z scores" option should be selected. After clicking on "Continue" the previous dialog box appears, and then the "Model" button should be selected. After clicking on the "Model" button "Ordinal" should be selected in case of the "Level of Measurement", and in case of "Dimensions" the minimum value should be 1 and the maximum value should be equal to 2.

Figure 4.1 shows the Stress value if the number of dimensions is equal to 2 (and it also shows the coordinates in the two-dimensional space). Since the Stress value is lower than 0.05, the model fit can be evaluated as "good". In case of the one-dimensional solution the Stress value is equal to 0.05727, thus the model fit in case of the one-dimensional solution can not be evaluated

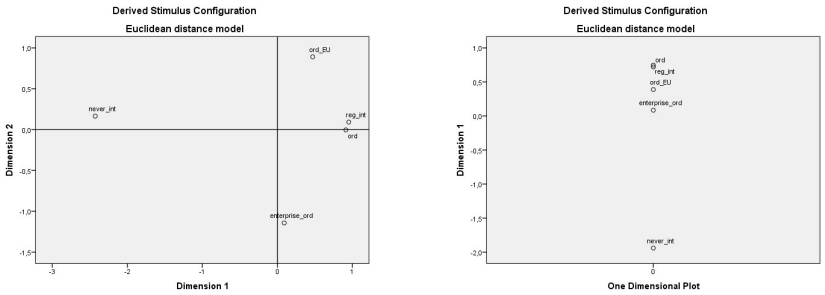
For matrix
Stress = ,00000 RSQ = 1,00000

Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	Dimension 2
1	ord	,9132	-,0039
2	ord_EU	,4713	,8908
3	reg_int	,9517	,0917
4	never_in	-2,4279	,1642
5	enterpri	,0917	-1,1429

Figure 4.1: Numerical MDS results (for variables)



(a) 2 dimensional solution (b) 1 dimensional solution

Figure 4.2: Graphical MDS results (for variables)

as "good" (although it can also not be evaluated as "weak", since the Stress value is not higher than 0.2). (*Kovács* (2011), page 146)

Figure 4.2 shows the multidimensional scaling results in the two-dimensional and one-dimensional case. Since in this example the "objects" in the analysis are the variables, thus the points on Figure 4.2 represent the variables (theoretically, the "objects" could also be the cases in an analysis). It can be observed on Figure 4.2 that in case of the first axis the sign belonging to the variable "never_int" differs from the sign belonging to the other variables (the sign of the variable "never_int" is negative, while the sign of the other variables is positive). This result is similar to the results of the principal component analysis (about the component matrix, described in Chapter 3).

Question 4.2. *Conduct multidimensional scaling (with ALSCAL method) with the five (standardized) variables (for the cases in the analysis, level of measurement: ordinal, number of dimensions: 2). How can the model fit be*

evaluated?

Solution of the question.

For matrix
Stress = ,06113 RSQ = ,98426

Configuration derived in 2 dimensions

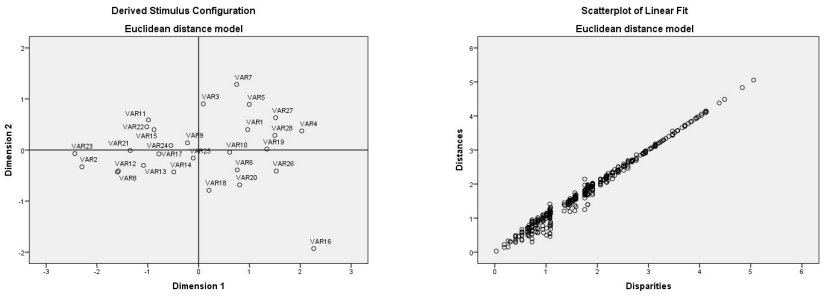
Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	Dimension 2
1	VAR1	,9652	,3996
2	VAR2	-2,2957	-,3299
3	VAR3	,0915	,9033
4	VAR4	2,0305	,3740
5	VAR5	,9959	,8949
6	VAR6	,7592	-,3894
7	VAR7	,7483	1,2856
8	VAR8	-1,5985	-,4311
9	VAR9	-,2185	,1421
10	VAR10	,6111	-,0433
11	VAR11	-,9867	,5892
12	VAR12	-1,5782	-,4109
13	VAR13	-1,0849	-,3009
14	VAR14	-,4873	-,4295
15	VAR15	-,8774	,3987
16	VAR16	2,2658	-1,9301
17	VAR17	-,7819	-,0767
18	VAR18	,2033	-,7900
19	VAR19	1,3429	,0189
20	VAR20	,8063	-,6829
21	VAR21	-1,3457	-,0104
22	VAR22	-1,0196	,4535
23	VAR23	-2,4357	-,0696
24	VAR24	-,5456	,0876
25	VAR25	-,1067	-,1578
26	VAR26	1,5271	-,4126
27	VAR27	1,5149	,6322
28	VAR28	1,5003	,2854

Figure 4.3: Numerical MDS results (for cases)

In this case the solution of Question 4.1 can be applied with the difference that after selecting the option "Create distances from data" (in the dialog box belonging to the multidimensional scaling) the "Between cases" option should be selected. Figure 4.3 shows the Stress value (and the two-dimensional coordinates that belong to the cases in the analysis). Since the Stress value is not smaller than 0.05 (the Stress value is equal to 0.06113), the model fit should not be assessed as "good". Figure 4.4 illustrates the results of multidimensional scaling in this case.

Question 4.3. *Assume that the values belonging to the five variables in the analysis are available for both 2010 and 2015, and the data is organised in such a way that the variable "year" can have two values (2010 and 2015), thus indicating the year (2010 or 2015) that belongs to a given case. Conduct multidimensional scaling (with INDSCAL method) with the five (standardized) variables (for the cases in the analysis, level of measurement: ordinal,*



(a) 2 dimensional plot (b) disparities and distances

Figure 4.4: Graphical MDS results (for cases)

number of dimensions: 2), and assume that the groups in the analysis correspond to the two categories of the variable "year". Which dimension (the first or the second dimension) can be considered as more important?

Solution of the question.

The solution of this question (related to INDSCAL) is similar to the solution of Question 4.1: the solution of Question 4.1 may be applied with the difference that in the dialog box (belonging to multidimensional scaling) the variable "year" should be selected in case of "Individual Matrices for:", and after selecting the "Model..." button "Individual differences Euclidean distance" should be selected as "Scaling Model". Figure 4.5 shows some of the results related to INDSCAL. According to *Kovács* (2011) (page 158) the importance of the first dimension can be calculated as follows (in this example):

$$\frac{0.9667^2 + 0.9276^2}{2} = 0.8974 \tag{4.3}$$

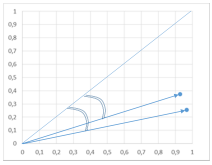
The overall importance of the dimensions in the analysis can be calculated based on the subject weights. Figure 4.5 indicates that in this example the first dimension can be considered as more important than the second dimension (0.8974>0.1026).

Based on the subject weights it may also be assessed, whether the weights belonging to a given group can be considered as "proportional" with the average weights. The weights (belonging to the groups in the INDSCAL analysis) may be plotted in a space (that is two-dimensional in this example). If the weights (belonging to a given group) are proportional with the average weights, then (when the weights can be plotted in a two-dimensional graph,

Subject Weights

Subject Number	Weird-ness	Dimension 1	Dimension 2
1	,1434	,9667	,2560
2	,1215	,9276	,3736
Overall importance of each dimension:		,8974	,1026

(a) subject weights: values



(b) subject weights: graph

Figure 4.5: INDSCAL results

similar to the graph on Figure 4.5) the point (belonging to a given group) is close to the 45° line. (*Kovács* (2011), page 158)

5 | Correspondence analysis

Correspondence analysis is a method that can be applied to analyze contingency tables. In this chapter "simple" („classical") correspondence analysis is discussed. As opposed to "multiple" correspondence analysis (which is related to the studying of more than two categorical variables) "simple" correspondence analysis can be applied to explore the relationship of variable categories in a two-way contingency table. (*Beh* (2004)) Similar to principal component analysis (that decomposes total variance into components), mathematically "simple" correspondence analysis decomposes the Pearson χ^2 measure of association into components. (*Hajdu* (2003), page 136)

5.1 Theoretical background

Correspondence analysis can be applied to graphically analyze data in a contingency table (for example data in a cross table analysis). Rows and columns of a contingency table are usually interpreted in a low-dimensional (usually two-dimensional) space. Relationship of different categories can be explored with outputs of the correspondence analysis (for example based on the graphical results).

The frequency values in a contingency table can be converted to relative frequency values by dividing by the total number of cases (n) in the analysis, and in this matrix (containing relative frequency values) the row sum values and the column sum values are sometimes referred to as row mass values and column mass values, respectively. (*Rencher-Christensen* (2012), page 431) The ith row profile is defined by dividing the ith row in the contingency table by the sum of the row values. The jth column profile is defined similarly (by dividing the elements in the jth column in the contingency table by the sum of the column values). (*Rencher-Christensen* (2012), pages 431-432)

In the correspondence analysis a "point" is plotted for each row and each column (in a contingency table) so that the relationship of the rows (or columns) are preserved as good as possible. (*Rencher-Christensen* (2012),

page 430) The coordinates belonging to the rows and the columns (of a contingency table) can be calculated based on singular value decomposition. It is important to emphasize that the singular values are calculated based on a matrix that is not necessarily symmetric. (*Rencher-Christensen* (2012), page 435)

Assume that matrix X has n rows and p columns. In case of singular value decomposition, matrix X can be reproduced with matrices A and B so that if $A^T \cdot A = B^T \cdot B = I$ (the identity matrix) and matrix D is diagonal (*Rencher-Christensen* (2012), page 435), then the following equation holds:

$$X = A \cdot D \cdot B^T \tag{5.1}$$

The diagonal elements of matrix D are the singular values. (*Rencher-Christensen* (2012), page 435) In this case the following results indicate that the (positive) square root values of the eigenvalues of matrix $X^T \cdot X$ are equal to the singular values:

$$X^T \cdot X = B \cdot D \cdot (A^T \cdot A) \cdot D \cdot B^T = B \cdot D \cdot D \cdot B^T \tag{5.2}$$

$$X \cdot X^T = A \cdot D \cdot D \cdot A^T \tag{5.3}$$

One of the results of the correspondence analysis is a plot in which the coordinates belonging to the rows and columns of a contingency table are plotted. The amount of "information" belonging to the dimensions shown by this plot is referred to as inertia. (*Rencher-Christensen* (2012), page 436)

Total inertia can be calculated based on the singular values that are calculated in a correspondence analysis. If the singular values in a correspondence analysis are indicated by $\lambda_1, \ldots, \lambda_k$, then total inertia can be calculated as follows (*Rencher-Christensen* (2012), page 436):

$$\sum_{i=1}^{k} \lambda_i^2 \tag{5.4}$$

If r denotes the number of rows and c refers to the number of columns of the contingency table in the correspondence analysis, then the maximum number required to graphically depict the association between the row and column responses can be calculated as follows:

$$k = max(r, c) - 1 \tag{5.5}$$

However, usually only the first two dimensions are applied to construct a graph that summarizes the results of the correspondence analysis (*Beh*

(2004)). Based on the singular values, the contribution of the dimensions of the plot (that can be created in a correspondence analysis) to the total inertia can be measured. For example the contribution of the first dimension to the total inertia can be calculated as follows (*Rencher-Christensen* (2012), page 436):

$$\frac{\lambda_1^2}{\sum_{i=1}^{k} \lambda_i^2} \tag{5.6}$$

In "simple" correspondence analysis, the decomposition of total inertia (for example with singular value decomposition) can be applied to identify important sources of information that contribute to describe association between two categorical variables. (*Beh* (2004))

5.2 Correspondence analysis examples

The file data1.xlsx contains (simulated) data that can be imported into SPSS. The following questions are related to this dataset, in which there are two categorical variables (X_1 and X_2) that are assumed to be measured on a nominal level of measurement.

Question 5.1. *Conduct correspondence analysis with the variables X_1 and X_2 and calculate column mass values (assume that columns are related to the categories of variable X_2).*

Solution of the question.

Before conducting correspondence analysis, first the relationship of the two categorical variables is analyzed in the following. Frequency tables for the variables can be calculated in SPSS by performing the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Descriptive Statistics → Frequencies...

In the appearing dialog box select both variables and click "OK". The frequency tables for X_1 and X_2 are shown in Table 5.1 and Table 5.2, respectively. In this example, the number of observations is 5000, in case of X_1 the number of categories is 18 and in case of X_2 the number of categories is 3 (the categories are indicated with integer numbers). The relationship of these two variables can be analyzed with cross table analysis. In SPSS, cross

table analysis results can be calculated if the following sequence is performed (beginning with selecting "Analyze" from the main menu):

Table 5.1: Frequency table for X_1

X1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	114	2,3	2,3	2,3
	2	91	1,8	1,8	4,1
	3	126	2,5	2,5	6,6
	4	183	3,7	3,7	10,3
	5	287	5,7	5,7	16,0
	6	312	6,2	6,2	22,3
	7	421	8,4	8,4	30,7
	8	462	9,2	9,2	39,9
	9	485	9,7	9,7	49,6
	10	499	10,0	10,0	59,6
	11	452	9,0	9,0	68,6
	12	391	7,8	7,8	76,5
	13	390	7,8	7,8	84,3
	14	252	5,0	5,0	89,3
	15	191	3,8	3,8	93,1
	16	135	2,7	2,7	95,8
	17	97	1,9	1,9	97,8
	18	112	2,2	2,2	100,0
	Total	5000	100,0	100,0	

Table 5.2: Frequency table for X_2

X2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	787	15,7	15,7	15,7
	2	1697	33,9	33,9	49,7
	3	2516	50,3	50,3	100,0
	Total	5000	100,0	100,0	

Analyze → Descriptive Statistics → Crosstabs...

In the appearing dialog box for example X_1 can be selected as "Row(s)" and X_2 can be selected as "Column(s)". To calculate a chi-squared test statistic value (associated with the null hypothesis that the two categorical variables are independent) the "Chi-square" option can be selected in the dialog box that appears after clicking on the "Statistics..." button.

Table 5.3 shows that the chi-squared test statistic value (related to the null hypothesis that the two categorical variables are independent) is equal

Table 5.3: Cross table analysis results

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8601,974[a]	34	,000
Likelihood Ratio	8636,255	34	,000
Linear-by-Linear Association	3817,179	1	,000
N of Valid Cases	5000		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 14,32.

to 8601.974, and the related p-value (in the last column of Table 5.3) is smaller than 0.05, thus the null hypothesis about the independence of the two variables in the analysis can not be accepted on a 5% significance level.

To conduct a correspondence analysis in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Dimension Reduction → Correspondence Analysis...

As a next step, in the appearing dialog box select X_1 as "Row" variable and X_2 as "Column" variable. After clicking on "Define Range..." button in case of X_1 set the category range for row variable as follows: the minimum value should be equal to 1 and the maximum value should be equal to 18 (and then click on the "Update" button). In case of X_2 follow similar steps: click on the "Define Range..." button and as caategory range for column variable set 1 as the minimum value and 3 as the maximum value (and then click on the "Update" button). Table 5.4 shows the column mass values that can be calculated based on the frequency values in Tabe 5.2: in this example for example the first column mass value can be calculated as $0.157 = \frac{787}{5000}$.

Table 5.4: Column mass values

Overview Column Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
X2	Mass	1	2	Inertia	1	2	1	2	Total
1	,157	1,830	-1,290	,741	,549	,293	,684	,316	1,000
2	,339	,425	1,254	,536	,064	,597	,110	,890	1,000
3	,503	-,859	-,442	,444	,387	,110	,802	,198	1,000
Active Total	1,000			1,720	1,000	1,000			

a. Symmetrical normalization

Question 5.2. *Calculate the singular values belonging to the correspondence analysis with the variables X_1 and X_2.*

Solution of the question.

In this example the maximum number of singular values that can be calculated is $max(18, 3) - 1 = 2$. These two singular values can be found in Table 5.5: the singular values are 0.96 and 0.894.

Table 5.5: Singular values and inertia

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,960	,922			,536	,536	,002	,897
2	,894	,799			,464	1,000	,006	
Total		1,720	8601,974	,000[a]	1,000	1,000		

a. 34 degrees of freedom

Question 5.3. *Calculate total inertia belonging to the correspondence analysis with the variables X_1 and X_2.*

Solution of the question.

The total inertia in this correspondence analysis is equal to 1.72 (this value that can be found in Table 5.5). This value can be calculated based on the singular values in the correspondence analysis as follows:

$$0.96^2 + 0.894^2 = 1.72 \tag{5.7}$$

Total inertia in this example can also be calculated based on the test statistic value in the cross table analysis that is discussed in Question 5.1 (*Beh* (2004)):

$$\frac{8601.974}{5000} = 1.72 \tag{5.8}$$

Question 5.4. *Calculate the contribution of the first dimension to the total inertia in the correspondence analysis that is carried out with the variables X_1 and X_2.*

Solution of the question.

In this example the contribution of the first dimension to the total inertia can be calculated as follows (this result is also shown in Table 5.5.):

$$\frac{0.96^2}{1.72} = 0.536 \tag{5.9}$$

Question 5.5. *Create a two-dimensional plot that graphically illustrates the results of the correspondence analysis with the variables X_1 and X_2. How can this plot be interpreted?*

Solution of the question.

In "simple" correspondence analysis it may be possible to create a two-dimensional plot on which each "point" represents rows and columns of the contingency table in the analysis. In this example Figure 5.1 illustrates the relationship of the categories belonging to the two (categorical) variables in the correspondence analysis.

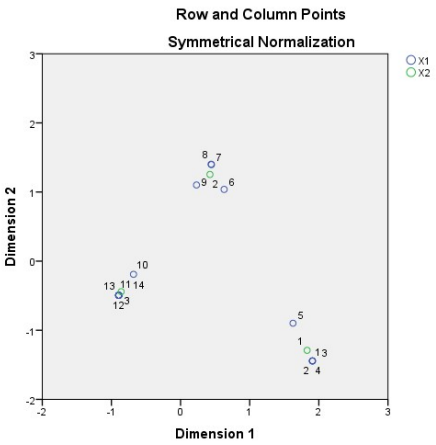


Figure 5.1: Two-dimensional plot of the results

On Figure 5.1, it is possible to observe a certain type of relationship between the variables X_1 and X_2: for example it can be observed that the category indicated by "2" in case of variable X_2 is (to some extent) related to the categories indicated by "6", "7", "8" and "9" in case of variable X_1.

6 | Logistic regression

Logistic regression (sometimes also referred to as logit analysis) is similar to ordinary linear regression with the difference that the regressand is not a continuous variable but a state which may or may not hold. (*Cramer* (2003), page 1) Logistic regression is thus a binary response model, in which the dependent (response) variable can take on the values zero and one. (*Wooldridge* (2010), page 561) The logit model may be applied for classification of cases in an analysis.

6.1 Theoretical background

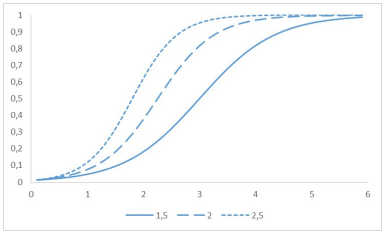
In logistic regression the dependent variable may have only two values, and one of the results is the prediction of values that represent "probability". Predictor variables in a logistic regression may be "categorical" or "scale" variables as well. (*George-Mallery* (2007), page 322) Assume that the dependent variable in a logistic regression is indicated by Y and the two categories of Y are indicated by 0 and 1, respectively. If p denotes the probability that $Y = 1$, then "probability" values can be estimated in a logistic regression with the predictor variables $X_1, \ldots, X_p$ as follows (*Kovács* (2011), pages 162-165):

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \ldots + b_p X_p)}} \tag{6.1}$$

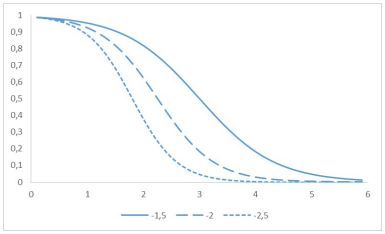
Based on the estimated probability values the odds can also be calculated (*Kovács* (2011), page 165):

$$\frac{p}{1-p} = e^{b_0} e^{b_1 X_1} \ldots e^{b_p X_p} \tag{6.2}$$

An important concept in logistic regression is the natural logarithm of the odds, this value is called "logit" (*George-Mallery* (2007), page 323):

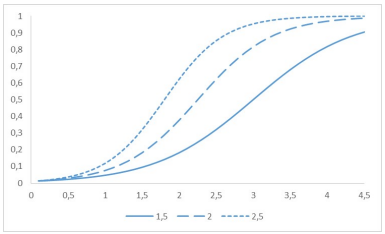


(a) examples for positive b_1

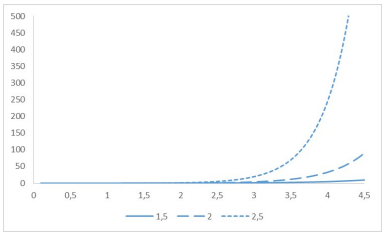


(b) examples for negative b_1

Figure 6.1: Estimated probability functions



(a) probability function



(b) odds function

Figure 6.2: Probability and odds functions

$$ln\frac{p}{1-p} = b_0 + b_1X_1 + \ldots + b_pX_p \tag{6.3}$$

In logistic regression, "logit" is a linear function of the predictor variables. In the following assume that only one predictor variable is applied in a logistic regression model. In this case the relationship of the predictor variables and the estimated probability values is nonlinear, as shown on Figure 6.1 (for different b_1 parameter values).

Figure 6.1 illustrates estimated probability values as a function of a predictor variable. The examples for a positive b_1 are calculated with the equation $p = \frac{1}{1+e^{-(-4.5+b_1X_1)}}$, while the examples for a negative b_1 are calculated with the equation $p = \frac{1}{1+e^{-(4.5+b_1X_1)}}$, for 1, 1.5 and 2 as b_1 values. On Figure 6.1 it can be observed that if the sign of b_1 is positive, then an increase in X_1 increases the estimated probability value, while for a b_1 value with negative sign an increase in X_1 decreases the estimated probability value. Figure 6.2 illustrates the difference between estimated probability and odds functions (with the equation $p = \frac{1}{1+e^{-(4.5+b_1X_1)}}$, for different b_1 values).

Coefficients in a logistic regression may be estimated with Maximum Likelihood method. (*Kovács* (2011), page 163) *Hajdu* (2004) emphasizes that

some advantageous features of the Maximum Likelihood method (for example minimum variance) occur asymptotically, in case of large samples. For small samples the application of logistic regression may be associated with some estimation problems. (*Hajdu* (2004)) The "separation" of cases (if there is a value that separates the values in the two groups) in case of at least one of the „predictor" variables may also be problematic, since then it is possible that a maximum likelihood estimate does not exist. (*Hajdu* (2004))

The interpretation of estimated coefficient values can be based on the odds values. For example assume that in a logistic regression only one predictor variable is entered into the analysis and the estimated coefficient belonging to this predictor variable is equal to b_1. In this case e^{b_1} shows the value by which odds is multiplied if X_1 increases by one unit. (*Kovács* (2011), page 165)

One of the results in a logistic regression is the classification table that compares the predicted values for the dependent variable with the actual observed data. In the prediction of group membership the estimated probability values are compared to a "cut value" (this "cutoff" may be for example 0.5. (*George-Mallery* (2007), page 329)

The "goodness" of model fit may be evaluated based on several measures, for example Hosmer-Lemeshow test or R-square measures (for example Cox and Snell R-square value or Nagelkerke R-square value) may be applied to evaluate the "goodness" of model fit in a logistic regression analysis. (*Kovács* (2011), pages 166-171) Goodness-of-fit tests aim at measuring whether the predicted values are an accurate representation of the observed values (for example omitted predictor variables or a misspecified form of a predictor variable can result in poor predictions). (*Xie et al.* (2008)) In case of the Hosmer-Lemeshow test the individual cases (observations) are ordered into groups by their estimated probability, and in each group the expected and actual frequency of successes is compared (the expected frequency of successes is the sum of the estimated probabilities). (*Cramer* (2003), page 63) *Fliszár* (2011) points out that the result of the Hosmer-Lemeshow test may be sensitive to the number of groups in the test: if for example the number of categories is "too low", then it may be easier to conclude that the estimation results in logistic regression are good. Hosmer-Lemeshow test results indicate a good model fit if the p-value (belonging to the null hypothesis of Hosmer-Lemeshow test) is relatively high (for example higher than 0.05), while in case of the pseudo-R^2 values a value close to 0.5 may be considered as an indicator of good model fit (*Paefgen et al.* (2014)) [1]

[1] *Paefgen et al.* (2014) assessed a Cox and Snell pseudo-R^2 value of approximately 0.38 as an indicator of good model fit.

The value of the area under the ROC (Receiver Operating Characteristics) curve may also be applied in the evaluation of the "goodness" of logistic regression results. (*Oravecz* (2007)) In the following assume a classification problem with two classes ("positive" and "negative"). If a case is positive and it is classified as positive, then it is counted as a "true positive", and if the case is negative and it is classified as positive, then the case is counted as a "false positive". On a (two-dimensional) ROC graph the "true positive" rate is plotted on the vertical axis and the "false positive" rate is plotted on the horizontal axis. (*Fawcett* (2005)) The probability of a true positive is referred to as sensitivity and the probability of a true negative is referred to as specificity. If the classification of a case in an analysis depends on a given threshold value (for example if a case is classified as belonging to one of the classes if a diagnostic marker value is higher than a given threshold value), then the theoretical ROC curve can be considered as a plot of sensitivity versus (1-specificity) for all possible threshold values. The definition of an optimal threshold value (sometimes referred to as "cutoff" value) may depend on the individual classification problem, for example profit-based performance measurement can contribute to calculate an optimal cutoff value. (*Verbraken et al.* (2014)

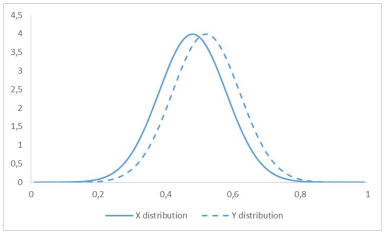
On a ROC graph the diagonal line represents a strategy of randomly guessing a class for the cases in the analysis. (*Fawcett* (2005)) If the area under the ROC curve is closer to one, then it indicates a higher diagnostic accuracy. (*Faraggi-Reiser* (2002)) Assume in the following that the diagnostic test results are available for two classes in an analysis, then if the two distributions (belonging to the test results in the two classes) are assumed to be independent normal distributions, then the area under the ROC curve (AUC) can be calculated as follows (*Faraggi-Reiser* (2002)):

$$AUC = \Phi(\frac{\mu_X - \mu_Y}{\sqrt{(\sigma_X^2 + \sigma_Y^2)}}) \tag{6.4}$$

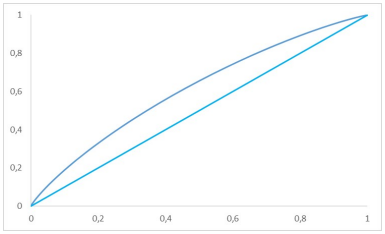
with μ_X and μ_Y indicating the theoretical mean values and assuming that σ_X and σ_Y indicate the standard deviations of the two distributions.

Figure 6.3 illustrates density functions and ROC curve for two classes that are relatively similar (the diagnostic test results are assumed to be similar). In this case the ROC curve is close to the diagonal line in the two-dimensional graph, and the area under the ROC curve (AUC) is relatively close to 0.5 (the standard deviations of the two distributions are assumed to be equal).

Figure 6.4 illustrates a classification problem with higher diagnostic accuracy. On Figure 6.4 the density functions (belonging to the diagnostic test results in the two classes) are relatively different (the standard deviations

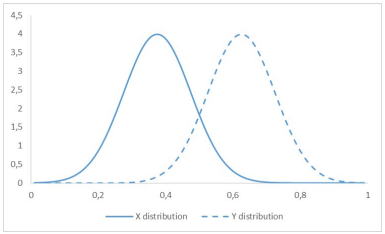


(a) theoretical density functions

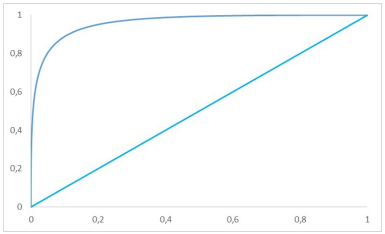


(b) theoretical ROC curve

Figure 6.3: Density functions and ROC curve



(a) theoretical density functions



(b) theoretical ROC curve

Figure 6.4: Density functions and ROC curve

are assumed to be equal). The higher diagnostic accuracy is also shown by the AUC value that is higher than in case of Figure 6.3 (and the AUC is relatively close to one).

6.2 Logistic regression examples

In the following 6 variables are analyzed: one of the variables has two categories (indicated by 0 and 1), and the other 5 variables are measured on a "scale" level of measurement. The five "scale" variables can be considered as possible "explanatory" variables, while the binary variable can be considered as dependent variable (in a logistic regression). Data belonging to the five "scale" variables (selected information society indicators of European Union member countries, for the year 2015) is downloadable from the homepage of Eurostat[2] and it is also presented in the Appendix. The values of the binary variable (name of the variable: "after2000") are associated with the European Union entry date:

[2]Data source: homepage of Eurostat (http://ec.europa.eu/eurostat/web/information-society/data/main-tables)

$$\text{“}after2000\text{''} = \begin{cases} 1 & \text{if the EU entry date is after 2000} \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

The five “scale” variables in the analysis are:

- “ord”: individuals using the internet for ordering goods or services
- “_EU”: individuals using the internet for ordering goods or services from other EU countries
- “reg_int”: individuals regularly using the internet
- “never_int”: individuals never having used the internet
- “enterprise_ord”: enterprises having received orders online

Question 6.1. *Conduct logistic regression with the five variables (by applying forward Wald method). How many variables are entered into the analysis?*

Solution of the question.

To conduct logistic regression in SPSS perform the following sequence (beginning with selecting “Analyze” from the main menu):

Analyze → Regression → Binary Logistic...

As a next step, in the appearing dialog box select the variables “ord”, “ord_EU”, “reg_int”, “never_int” and “enterprise_ord” as “Covariates:” and “after2000” as “Dependent”.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ord	-,081	,031	6,946	1	,008	,923
	Constant	3,648	1,471	6,148	1	,013	38,388

a. Variable(s) entered on step 1: ord.

Table 6.1: Estimated coefficients

As "Method" "Forward: Wald" should be selected. After clicking on the "Save..." button "Probabilities" should be selected. The Hosmer-Lemeshow test results can be calculated by selecting "Hosmer-Lemeshow goodness-of-fit" in case of the "Options..." button.

Table 6.1 shows that (by applying the forward Wald method for variable selection) only one step has been performed, which means that only one variable is entered into the analysis.

Question 6.2. *Conduct logistic regression with the five variables (by applying forward Wald method). How can the estimated coefficient(s) belonging to the "explanatory" variables be interpreted?*

Solution of the question.

According to the solution of Question 6.1 only one variable is entered with the application of the forward Wald method, and the probability values can be estimated with the following equation:

$$p = \frac{1}{1 + e^{-(3.648 - 0.081 \cdot \text{“}ord\text{”})}} \tag{6.6}$$

The estimated coefficient of the variable "ord" can be interpreted as follows: if the value of the variable "ord" increases by 1 unit, then the odds (that the EU entry happened after 2000) is multiplied by $e^{-0.081}$. [3]

Question 6.3. *Conduct logistic regression with the five variables (by applying forward Wald method). How can the model fit be evaluated?*

Solution of the question.

There are several methods how the goodness of model fit can be evaluated in a logistic regression model, for example:

- pseudo R-square values (for example Nagelkerke R-square value)
- results of Hosmer-Lemeshow test
- the area under the ROC curve

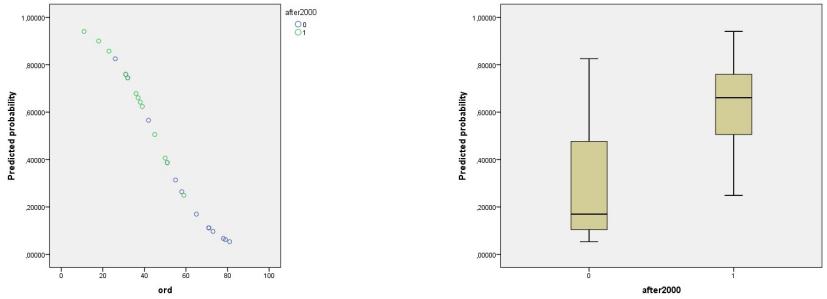
[3] It is worth mentioning that this is only a calculation example that aims at contributing to the learning of logistic regression. In practical applications, the binary variable that is analyzed in logistic regression is often related to an economic event (for example to default of a loan, etc.).

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27,807[a]	,322	,430

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Table 6.2: Pseudo R-square values



(a) scatter plot (b) boxplot

Figure 6.5: Comparison of predicted probability values in the two classes

Figure 6.5 shows that the predicted probability values do (relatively) differ in the two classes, which can indicate a (relatively) good model fit. Table 6.3 shows the classification table (belonging to the cut value that is equal to 0.5). According to the results in Table 6.3 75% of all cases has been correctly classified with a cutoff value equal to 0.5 (when all cases with a predicted probability value of higher than 0.5 has been classified as belonging to the class indicated by "1"). The results in Table 6.3 also show that 7 $(= 3 + 4)$ cases (countries) could not be correctly classified in this example.

Classification Table[a]

			Predicted		
			after2000		Percentage Correct
	Observed		0	1	
Step 1	after2000	0	11	4	73,3
		1	3	10	76,9
	Overall Percentage				75,0

a. The cut value is ,500

Table 6.3: Classification results

Table 6.2 shows the calculated pseudo R-square values: the Cox and Snell R-square value is equal to 0.322, while the Nagelkerke R-square value is equal to 0.43. The null hypothesis belonging to the Hosmer-Lemeshow test can be accepted (the p-value is equal to 0.447).

To create a ROC curve and calculate the area under the ROC curve in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → ROC Curve...

As a next step, in the appearing dialog box select the (previously saved) "Predicted probability" as "Test Variable:", and "after2000" as "State Variable:" (the "Value of State Variable" should be equal to 1 in this example). In this dialog box (in case of "Display:") the following options should also be selected: "ROC Curve", "With diagonal reference line", "Standard error and confidence interval".

Figure 6.6 shows the ROC curve, and in Table 6.4 the value of the area under the ROC curve (0.818) is presented.

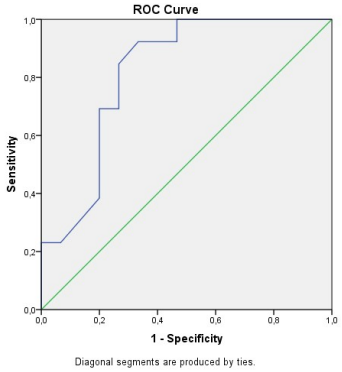


Figure 6.6: ROC curve

According to *Kovács* (2014) (page 146) the model fit in logistic regression can be assessed as follows (based on the ROC AUC values): if the AUC value is between 0.5 and 0.6, then the logistic regression model is not applicable for classification, if AUC is between 0.6 and 0.7, then the model fit is weak, if AUC is between 0.7 and 0.8, then the model fit can be considered as average, if the AUC value is between 0.8 and 0.9, then the model fit is good, and if the ROC AUC value is higher than 0.9, then the model fit of the logistic regression model can be considered as excellent.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,818	,082	,004	,656	,980

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Table 6.4: Area under the ROC curve

Based on the results, the model fit in case of the logistic regression model in this example can be evaluated as good.

7 | Discriminant analysis

Discriminant analysis is a method that can be applied for classification. One of the differences between logistic regression and discriminant analysis is that discriminant analysis can not only be applied in case of binary response (dependent) variables. In the following linear discriminant analysis is discussed, in which the classification can be solved by finding linear functions of the "predictor" variables that best separate the groups. (*Cramer* (2003), page 89)

7.1 Theoretical background

Discriminant analysis describes group separation, in which linear functions of the original ("independent") variables are applied to describe the differences between two or more groups. (*Rencher-Christensen* (2012), page 226) Results of discriminant analysis may be applied to predict membership in groups (indicated by categories of the "grouping" variable). (*George-Mallery* (2007), page 280) The main assumptions in the discriminant analysis (*Kovács* (2011), pages 115-123) are as follows:

- the original variables ("independent", "predictor" variables) should have a multivariate normal distribution
- within-group covariance matrices should be equal across groups (a test for the equality of the group covariance matrices is based on Box's M value).

Lee-Wang (2015) points out that Fisher's linear discriminant analysis can be considered as optimal in minimizing the misclassification rate under the normality and equal covariance assumptions. It may be possible to compare logistic regression and discriminant analysis (for example if the number of groups in an analysis is equal to two). *Press-Wilson* (1978) point out that (for the discriminant analysis problem) discriminant analysis estimators may be

preferred to logistic regression estimators in case of normal distribution with identical covariance matrices. However, it has to be mentioned that under nonnormality logistic regression model with maximum likelihood estimators may be preferred for solving a classification problem. (*Press-Wilson* (1978))

Classification in discriminant analysis can be based on for example (*Kovács* (2011), pages 127-128):

- distance in the "canonical space": a case is assigned to the group where the distance between the group centroid and the case is the smallest in the canonical space
- Fisher's classification functions: for each group a classification function is constructed and a case is assigned to that group for which the largest classification function value can be calculated.

The mathematical background of discriminant analysis is based on the eigenvalue-eigenvector decomposition of a matrix. In the following the number of cases in the analysis is indicated by n, the number of original ("independent", "predictor") variables is p and the number of groups is indicated by g. Let X denote the matrix of the original ("independent", "predictor") centered variables (in that case when the average of each variable is zero). Then X^TX can be considered as the sum of two matrices (*Kovács* (2011), page 115):

$$X^TX = K + B \tag{7.1}$$

where $B = \sum_{i=1}^{g}(n_i - 1)S_i$, $n = \sum_{i=1}^{g} n_i$ and S_i is the group covariance matrix for group i. (*Kovács* (2011), pages 115-116) Discriminant functions are linear combinations of the original ("independent", "predictor") variables (*Kovács* (2011), page 116):

$$y = Xc \tag{7.2}$$

where $c^Tc = 1$. (*Kovács* (2011), page 116) Based on the previous assumptions (*Kovács* (2011), pages 116):

$$y^Ty = (Xc)^T(Xc) = c^TX^TXc = c^T(K + B)c = c^TKc + c^TBc \tag{7.3}$$

The coefficients (c) should be calculated so that (*Kovács* (2011), page 116):

$$\underset{c}{max}\frac{c^T K c}{c^T B c} \tag{7.4}$$

The solution to this problem is (*Kovács* (2011), pages 116-117), where I refers to the identity matrix:

$$(B^{-1}K - \lambda I)c = 0 \tag{7.5}$$

It means that the eigenvectors and eigenvalues of the matrix $B^{-1}K$ should be calculated in a discriminant analysis. (*Kovács* (2011), pages 116-117) The matrix $B^{-1}K$ is not symmetric, and it can be shown that the eigenvalues of the matrix $B^{-1}K$ are equal to the eigenvalues of the symmetric matrix $(U^{-1})^T K U^{-1}$ if $B = U^T U$ is the Cholesky factorization of matrix B. (*Rencher-Christensen* (2012), page 232) It can also be shown that if v is an eigenvector of $(U^{-1})^T K U^{-1}$, then $y = U^{-1}v$ is an eigenvector of $B^{-1}K$. (*Rencher-Christensen* (2012), page 232)

In discriminant analysis the maximum number of discriminant functions is (*Kovács* (2011), page 117):

$$min(g-1, p) \tag{7.6}$$

Let λ_j $(j = 1, \ldots, k)$ denote the eigenvalues of the matrix $B^{-1}K$, where $k = min(g-1, p)$. The λ_j $(j = 1, \ldots, k)$ eigenvalues refer to the "goodness" of classification based on the discriminant functions. A Wilks' Lambda value can be calculated also for discriminant functions and this measure shows how good the given discriminant functions together separate the groups in the analysis (*Kovács* (2011), page 117):

$$\prod_{j=1}^{k}\frac{1}{1+\lambda_j} \tag{7.7}$$

In case of this Wilks' Lambda a smaller value refers to a better separation of the groups. (*Kovács* (2011), page 117) Beside the Wilks' Lambda values the canonical correlation values can also be calculated based on the eigenvalues of the matrix $B^{-1}K$. The canonical correlation measures the association between the discriminant scores and the groups (*Kovács* (2011), page 124):

$$\sqrt{\frac{\lambda_j}{1+\lambda_j}} \tag{7.8}$$

Theoretically the value of the canonical correlation can be between 0 and 1, and a higher value refers to a better separation result.

7.2 Discriminant analysis examples

Similar to Chapter 6 (about logistic regression), six variables are analyzed in the following: a binary variable (that has two categories, indicated by 0 and 1), and five "scale" variables. Data belonging to the five "scale" variables (selected information society indicators of European Union member countries, for the year 2015) is downloadable from the homepage of Eurostat[1] and it is also presented in the Appendix. The values of the binary variable "after2000" are associated with the European Union entry date:

$$\text{“}after2000'' = \begin{cases} 1 & \text{if the EU entry date is after 2000} \\ 0 & \text{otherwise} \end{cases} \tag{7.9}$$

The five "scale" variables in the analysis are:

- "ord": individuals using the internet for ordering goods or services
- "ord_EU": individuals using the internet for ordering goods or services from other EU countries
- "reg_int": individuals regularly using the internet
- "never_int": individuals never having used the internet
- "enterprise_ord": enterprises having received orders online

Question 7.1. *Conduct discriminant analysis (with stepwise method and selecting "Use probability of F" option) with the 5 scale variables (as "independent" variables) and "after2000" (as grouping variable). Can the covariance matrices in the groups be considered as equal?*

[1]Data source: homepage of Eurostat (http://ec.europa.eu/eurostat/web/information-society/data/main-tables)

Solution of the question.

To conduct discriminant analysis in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Classify → Discriminant...

As a next step, in the appearing dialog box select the variables "ord", "ord_EU", "reg_int", "never_int" and "enterprise_ord" as "Independents:" and "after2000" as "Grouping Variable". After selecting the "Define Range..." button the "Minimum" should be equal to 0 and the "Maximum" should be equal to 1 (because in this example the variable "after2000" has two categories, indicated by 0 and 1). In case of the "Statistics..." button the "Box's M" option should be selected.
In order to carry out a discriminant analysis with stepwise method (instead of enter method) the "Use stepwise method" option should be selected. Details belonging to the applied stepwise method can be selected after clicking on the "Method..." button: as "Criteria" the "Use probability of F" option should be selected. In case of discriminant analysis the multivariate normality of the variables and the equality of covariance matrices in the groups belong to the application assumptions. The equality of covariance matrices can be examined based on the Box's M value (and a related test statistic). Table 7.1 shows the p-value that belongs to the null hypothesis that the covariance matrices are equal in the groups. Since this p-value is higher than 0.05 (0.266>0.05), the null hypothesis about the equality of covariance matrices (in the groups) can be accepted.

Test Results

Box's M		1,285
F	Approx.	1,237
	df1	1
	df2	1996,189
	Sig.	,266

Tests null hypothesis of equal population covariance matrices.

Table 7.1: Box's M value

Question 7.2. *Conduct discriminant analysis (with stepwise method and selecting "Use probability of F" option) with the 5 scale variables (as "independent" variables) and "after2000" (as grouping variable). How can the model fit be evaluated?*

Solution of the question.

In SPSS, the same options should be selected as in case of the solution of Question 7.1. In this example only one variable is entered ("ord"), as also shown by the structure matrix (Table 7.2). The elements of the structure matrix are (pooled within-groups) correlations (between the variables and the standardized canonical discriminant functions). Table 7.2 shows that the correlation of the variable "ord" and the first (standardized) canonical discriminant function is equal to 1 (which is related to that solution in the discriminant analysis that only one variable is entered).

Structure Matrix

	Function
	1
ord	1,000
never_int[a]	-,938
reg_int[a]	,937
ord_EU[a]	,536
enterprise_ord[a]	,366

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Table 7.2: Structure matrix

In a discriminant analysis, for example Wilks' Lambda or canonical correlation values may be applied to evaluate the model fit (and if the number

of groups in the analysis is equal to two, then the area under the ROC curve may also be appropriate to assess the "goodness" of the model fit). The Wilks' Lambda and canonical correlation values can be calculated based on the eigenvalues of the matrix $B^{-1}K$. In this example the number of canonical discriminant functions is $min(p, g-1) = min(1, 2-1) = 1$, thus the matrix $B^{-1}K$ has only one eigenvalue (0.51).

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,510[a]	100,0	100,0	,581

a. First 1 canonical discriminant functions were used in the analysis.

Table 7.3: Canonical correlation

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,662	10,510	1	,001

Table 7.4: Wilks' lambda value

Table 7.3 shows the canonical correlation, that can be calculated in this example based on the eigenvalue of the matrix $B^{-1}K$ as follows:

$$\sqrt{\frac{0.51}{1+0.51}} = 0.581 \tag{7.10}$$

The canonical correlation can be interpreted in this case so that 58.1% of the variability of the discriminating "scores" is explained by the grouping of the cases in the analysis. (*Kovács* (2011), page 124) The Wilks' Lambda (0.662) can also be calculated based on the eigenvalue of the matrix $B^{-1}K$:

$$\frac{1}{1+0.51} = 0.662 \tag{7.11}$$

The Wilks' Lambda value can be interpreted so that the heterogeneity that is not explained by the discriminating function is 0.662. (*Kovács* (2011), page 124) In case of a good model fit in discriminant analysis the Wilks' Lambda value should be close to zero, thus in this example the model fit can not be considered as good (this conclusion is also confirmed by the canonical

correlation value). If the number of groups in a discriminant analysis is equal to 2, then only one Wilks' Lambda and canonical correlation value can be calculated based on the eigenvalue of the matrix $B^{-1}K$. Since in this example the variable "after2000" has only two categories, the Wilks' Lambda value can be calculated based on the canonical correlation:

$$0.662 = 1 - 0.581^2 \tag{7.12}$$

Question 7.3. *Conduct discriminant analysis (with enter method) based on the variables "ord" and "enterprise_ ord" (as "independent" variables) and "after2000" (as grouping variable). How can the estimated canonical discriminant function coefficients be interpreted?*

Solution of the question.

In SPSS, the same options should be selected as in case of the solution of Question 7.1, with the following differences:

- in the dialog box (belonging to discriminant analysis) the "Enter independents together" option should be selected (instead of the "Use stepwise method" option)

- after clicking on the "Statistics" button the option "Unstandardized" should be selected (in case of "Function Coefficients").

As the solution of Question 7.1 indicates, with stepwise method only one variable is entered into the analysis, thus in this calculation example two ("independent") variables are entered together, so that the discriminant function can also be examined in a two-dimensional graph (in case of the scatter plot belonging to the two "independent" variables).

Table 7.5 shows the canonical discriminant function coefficients. Based on these results the coefficients of the linear line that best separates the groups in the two-dimensional space (in this example on the scatter plot that belongs to the two "independent" variables) can be calculated, since the following equation holds in case of the linear line that best separates the groups:

$$0.056 \cdot \text{“}ord'' + 0.027 \cdot \text{“}enterprise\_ord'' - 3.13 = 0 \tag{7.13}$$

Figure 7.1 shows the linear line that best separates the two groups (that belong to the two categories of the variable "after2000"). The equation belonging to this linear line can be written as follows:

Canonical Discriminant Function Coefficients

	Function
	1
ord	,056
enterprise_ord	,027
(Constant)	-3,130

Unstandardized coefficients

Table 7.5: Canonical discriminant function coefficients

$$\text{“}enterprise\_ord\text{”} = 115.9 - 2.07 \cdot \text{“}ord\text{”} \tag{7.14}$$

In this case the points on Figure 7.1 are not “perfectly” separated by the linear line (this result is also indicated by the canonical correlation and Wilks' Lambda values), but it can be observed on Figure 7.1 that most points, that are located on the same side of the linear line, belong to the same class.

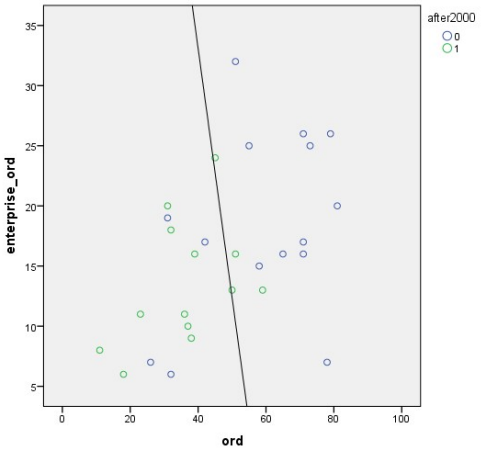


Figure 7.1: Separation of the two classes (in case of two entered variables)

In this example, the “canonical space” in the discriminant analysis is only one-dimensional ($min(p, g-1) = min(2, 2-1) = 1$), and the zero point in this dimension is associated with the linear line (that best separates the

groups) on Figure 7.1. The centroids of the two classes in this example are located on different sides of the linear (separating) line, thus in the canonical space the signs of the centroids differ, as indicated by Table 7.6.

Functions at Group Centroids

	Function
after2000	1
0	,650
1	-,750

Unstandardized canonical discriminant functions evaluated at group means

Table 7.6: Function value at group centroids

8 | Survival analysis

In survival modeling the focus of the analysis is on the estimation of certain time distributions from observed data. This time distribution can be interpreted as "failure time distribution" and failure time random variables may represent time to a certain event, for example time to insurance policy termination. (*Robinson* (2014), page 481) There are several survival models, this chapter discusses the Kaplan-Meier model and the Cox regression model.

8.1 Theoretical background

The failure time random variable is defined on the non-negative real numbers in survival analysis. (*Robinson* (2014), page 481) If t indicates (failure) time random variable values in the analysis, then $S(t)$ survival function indicates that probability that a given case has still not yet quit the analysis until time t. (*Vékás* (2011), page 176)

Assume in the following that t indicates the time until a certain event occurs and the distribution function belonging to t is indicated by $F(t)$. The survival function in this case can be calculated as $S(t) = 1 - F(t)$ and the density function is $f(t) = \frac{dF(t)}{dt}$. The hazard rate can be defined as follows (*Vékás* (2011), pages 180-181):

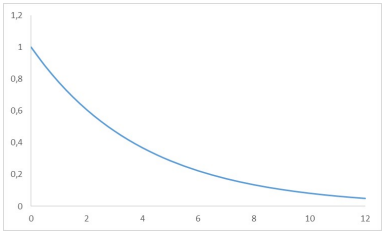
$$h(t) = \frac{f(t)}{S(t)} \tag{8.1}$$

where $S(t) \neq 0$.

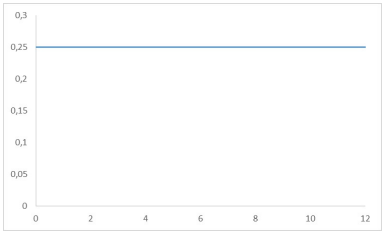
The relationship between the survival function and the cumulative hazard rate (denoted by $H(t)$) can be described as follows (*Vékás* (2011), page 182):

$$S(t) = e^{-H(t)} \tag{8.2}$$

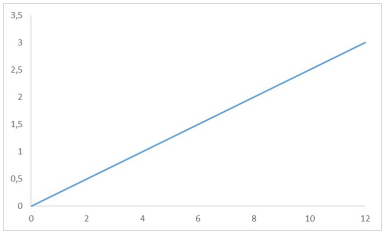
The distribution of the (failure) time variable can be described with the survival function, hazard function and the density function, which may theoretically be discrete, continuous or mixture. These functions (for example



(a) survival function



(b) hazard rate function



(c) cumulative hazard rate function

Figure 8.1: Functions in survival analysis (an example)

the survival function) may also be parametric, nonparametric or mixture. (*Robinson* (2014), page 481)

As an example, assume that the survival function is described by the equation $S(t) = e^{-\lambda t}$ (with $\lambda > 0$). Figure 8.1 shows that in this case the hazard rate is constant and the cumulative hazard rate function is linear.

Data preparation is an important step in survival analysis. Assume that in a database it is recorded whether and when a given case (for example a given person) experiences a specific event. In a survival analysis the following variables have a central role (*Vékás* (2011), page 174):

- status variable: shows whether or not the event has occurred (usually this variable has two categories: 0 indicates that the event has not yet occurred and 1 indicates that the event has occurred)

- time variable: shows how long the given case has been "observed".

In some respect, the observed values of the (failure) time random variable are usually incomplete in a survival analysis, because for example at the end of the observation period the event (studied in the survival analysis) has not occurred in case of some of the observations. If the final value of the (failure) time random variable is not observed by the end of the observation period, then the observation is referred to as "right-censored". (*Robinson* (2014), page 482) In the following in this chapter only right-censored data

is analyzed, thus "censoring" will be mentioned if data is "right-censored". In survival analysis, the value of state variable is zero for a censored case. (*Vékás* (2011), page 173)

Depending on the features of variables in a survival analysis, several models can be applied to model the time to the given event. For example in case of a Kaplan-Meier model "scale" [1] variables can not be applied as "explanatory" variables (without transformation of the "scale" variable). In a Cox regression "explanatory" variables may be measured on a scale, ordinal or nominal level of measurement.

In a Kaplan-Meier model the survival function is estimated from the sample, and it is possible to estimate the expected survival time. An other concept in the Kaplan-Meier model is the median survival time which is the time period after which half of the original cases (at the beginning of the analysis) is expected to quit the analysis. (*Vékás* (2011), pages 178-179)

With the following tests it is possible to test whether the survival functions belonging to different subsamples can be considered as equal in a Kaplan-Meier model (*Vékás* (2011), page 179):

- log-rank test
- Breslow test
- Tarone-Ware test.

In some survival analysis models it is possible to apply "scale" variables as explanatory variables. In a Cox regression analysis the hazard rate can be estimated as a function of the explanatory variables in the analysis as follows (*Vékás* (2011), page 182):

$$h(t) = h_0(t)e^{b_1X_1+\ldots+b_pX_p} \tag{8.3}$$

where p indicates the number of "explanatory" variables in the model and $h_0(t)$ is the baseline hazard rate. The Cox regression model is a proportional hazards model, and it belongs to the testing of the proportional hazards assumption to assess the constancy of the estimated coefficients over time. (*Robinson* (2014), pages 497-499) The baseline hazard rate is a function of the time in the Cox regression model. In case of "scale" explanatory variables the estimated coefficients (belonging to the explanatory variables) can be interpreted so that if the value of the j-th explanatory variable increases by one unit (ceteris paribus) then the hazard rate is multiplied by e^{b_j} for each t. (*Vékás* (2011), page 183) Explanatory variables may also be "categorical"

[1] „Scale" variables are considered to be measured on a scale level of measurement

variables in a Cox regression model (for example *Szepesváry* (2015) applies "categorical" explanatory variables in a Cox regression model).

Maximum likelihood method may be applied in the estimation of coeffients in a Cox regression model (*Cox* (1972), *Vékás* (2011), page 183) Similar to the Kaplan-Meier model, it is possible to calculate mean and median for survival time also in a Cox regression model.

There are several methods that may be applied to assess the adequacy of model results. In case of the omnibus test the null hypothesis is that theoretically all coefficients (belonging to the explanatory variables) are equal to zero. If this null hypothesis is accepted, then the given model should be restructured (for example new variables should be entered into the model). The Wald test (with the null hypothesis that theoretically the coefficient is equal to zero) can be applied individually to each entered explanatory variable, and if the null hypothesis is accepted then the given variable should not be an explanatory variable in the Cox regression model. (*Vékás* (2011), pages 183-184) To test the proportional hazards assumption the partial residuals (for each uncensored case and for each explanatory variable) can be calculated and plotted against time, and if on these plots no trend can be observed then the proportional hazards assumption can be (approximately) accepted. (*Vékás* (2011), pages 185-186)

8.2 Survival analysis examples

In the first data analysis example assume that a firm registers data about customer churn: it is registered when the customer relationship began and whether (and when) it ended. Figure 8.2 summarizes these assumed data.

On Figure 8.2 it is assumed that data for 10 months are available (on the horizontal axis 10 is the highest number). According to the assumed data the firm currently (10 months after the beginning of data collection) has 3 customers. Data for other (previous) customers (for example the first customer with a 1 month long customer relationship) can be referred to as censored data. Data on Figure 8.2 is presented in Table 8.1 in a tabular form (data for each customer corresponds to an observation in the analysis).

Based on the simple database in Table 8.1 a new dataset can be created in SPSS. Assume that in SPSS the name of the status variable is "status" and the name of the time variable is "time". The following two questions are related to this database.

Question 8.1. *Estimate and plot data for the survival function with Kaplan-Meier model.*

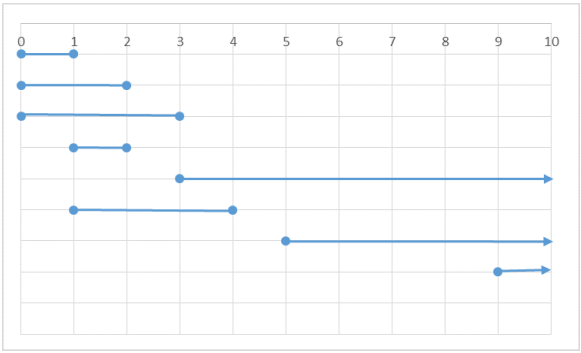


Figure 8.2: Survival data example

Solution of the question.

To conduct a survival analysis with Kaplan-Meier model in SPSS perform the following sequence (beginning with selecting "Analyze" from the main menu):

Analyze → Survival → Kaplan-Meier...

As a next step, in the appearing dialog box select "time" as "Time" variable and "status" as "Status" variable. In case of the "status" variable select "Define Event ..." button and set "Single value:" equal to one. After clicking on the "Options" button the "Survival" plot should be selected in the dialog box. The resulting estimated survival function is illustrated by Figure 8.3.

This estimated survival function is not a continuous function in this example. When the time variable is equal to zero, then the survival function value is equal to one. Other estimated values belonging to the survival function are shown in Table 8.2, in the fourth column.

Table 8.3 illustrates how survival function values can be estimated in this example (the presented calculations follow the calculation method described in *Vékás* (2011), pages 176-179) In Table 8.3 the columns can be interpreted as follows:

- quitting time: the quitting customers have quitted after 1,2 and 3 month, respectively
- "nr.of q.cust.": the number of quitting customers, belonging to the quitting times, for example there are 2 customers in the example who quitted after 1 month

Table 8.1: Sample data for Kaplan-Meier model

	survival analysis variables	
	status variable	time variable
1. observation	1	1
2. observation	1	2
3. observation	1	3
4. observation	1	1
5. observation	0	7
6. observation	1	3
7. observation	0	5
8. observation	0	1

Table 8.2: Survival table

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	1,000	1,00	.	.	1	7
2	1,000	1,00	,750	,153	2	6
3	1,000	,00	.	.	2	5
4	2,000	1,00	,600	,182	3	4
5	3,000	1,00	.	.	4	3
6	3,000	1,00	,300	,175	5	2
7	5,000	,00	.	.	5	1
8	7,000	,00	.	.	5	0

- "nr.of cust.": the number of customers who have stayed at least so long as the given quitting time: for example all (8) customers in the example stayed at least 1 month

- "probability": the quitting probability: it can be calculated as a ratio of the values in the previous two columns, for example in the first row $0.25 = \frac{2}{8}$

Based on the values in Table 8.3, the survival function values can be calculated as $(1-\frac{1}{4}) = \frac{3}{4}$, $(\frac{3}{4}-(1-\frac{1}{5}) = \frac{3}{5}$ and $(\frac{3}{5}-(1-\frac{1}{2}) = \frac{3}{10}$, respectively. These estimated values are presented on Figure 8.3.

Question 8.2. *Calculate mean and median for survival time.*

Solution of the question.

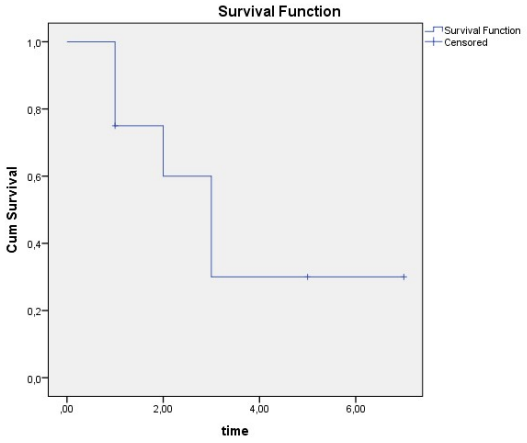


Figure 8.3: Estimated survival function

Table 8.3: Kaplan-Meier model calculations

"quitting time"	nr.of q.cust.	nr.of cust.	probability
1	2	8	0.25
2	1	5	0.2
3	2	4	0.5

In the SPSS output that belongs to solution of Question 8.1 the following table shows the mean and median for survival time.

The mean for survival time can be calculated with the application of quitting probabilities (that can be calculated based on the quitting probabilities shown in Table 8.3). Table 8.5 shows the quitting probabilities belonging to the calculation of the mean survival time. Based on data in Table 8.5, the mean for survival time can be calculated as follows:

$$\frac{1}{4} \cdot 1 + \frac{3}{20} \cdot 2 + \frac{3}{10} \cdot 3 + (1 - 0.25 - 0.15 - 0.3) \cdot 7 = 3.55 \qquad (8.4)$$

The median for survival time is 3 (months), since the survival function value is 0.5 when the time value on the horizontal axis of Figure 8.3 is equal to 3.

The previous example data contained only a few observations. In practice, it

Table 8.4: Survival time data

Means and Medians for Survival Time

Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
3,550	,890	1,805	5,295	3,000	,585	1,854	4,146

a. Estimation is limited to the largest survival time if it is censored.

Table 8.5: Kaplan-Meier model calculations

"quitting time"	probability
1	$1 \cdot \frac{1}{4} = \frac{1}{4}$
2	$\frac{3}{4} \cdot \frac{1}{5} = \frac{3}{20}$
3	$\frac{3}{5} \cdot \frac{1}{2} = \frac{3}{10}$

is sometimes possible to analyze larger datasets. The file data2.xlsx contains (simulated) data that can be imported into SPSS. The following questions are related to this dataset, which contains a time variable, a status variable and three other variables (X_1 and X_2 are "scale" variables, while X_3 is a variable measured on a nominal level of measurement). Since this dataset contains simulated data (and aims only at highlighting selected survival analysis concepts), no additional "names" and interpretation are given to the variables.

Question 8.3. *In a Kaplan-Meier model can the survival functions be considered as identical in the categories of the variable X_3?*

Solution of the question.

After selecting the time and status variable in the dialog box that belongs to Kaplan-Meier model, the variable X_3 should be "Factor" in the analysis (this option can be found in the dialog box belonging to the Kaplan-Meier model). In this example, the variable X_3 has 3 categories (indicated by 0, 1 and 2). Figure 8.4 illustrates that the estimated survival functions differ in these groups. In SPSS the equality of survival distribution can be tested by applying three test statistics. The results of these analyses can be found in Table 8.6. In case of all tests (log rank, Breslow, Tarone-Ware) the null hypotesis (about the equality of survival distributions) can be rejected, since the p-value (indicated by "Sig." in Table 8.6) is small (it is smaller than 0.05, thus the null hypoteses can be rejected at a 5 % significance level).

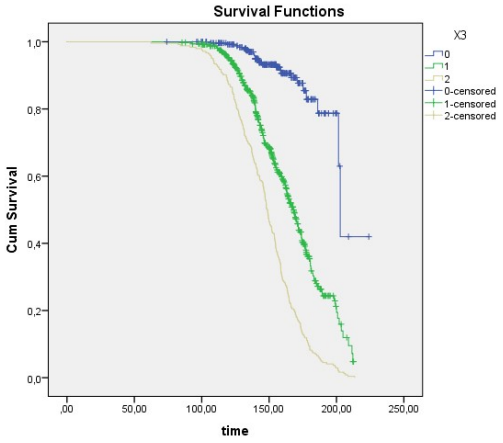


Figure 8.4: Survival functions

Table 8.6: Differences between survival functions

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	229,874	2	,000
Breslow (Generalized Wilcoxon)	176,018	2	,000
Tarone-Ware	206,820	2	,000

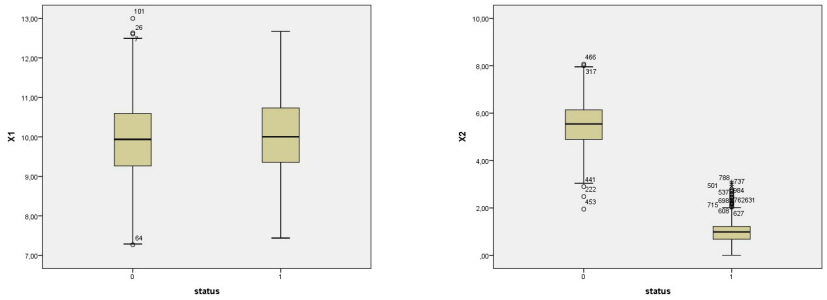
Test of equality of survival distributions for the different levels of X3.

Question 8.4. *Conduct Cox regression applying forward Wald method with the variables X_1 and X_2. How can the estimated coefficient(s) of the entered variable(s) be interpreted?*

Solution of the question.

Before conducting a survival analysis with Cox regression model the relationship of the explanatory variables and the status variable is worth analyzing. With a comparison of boxplots (illustrated by Figure 8.5) it can be observed that the values of X_2 differ more in the two groups than the values of X_1.

To conduct a survival analysis with Cox regression model in SPSS, perform the following sequence (beginning with selecting "Analyze" from the



(a) variable X_1 (b) variable X_2

Figure 8.5: Boxplots of variables

main menu):

Analyze → Survival → Cox Regression...

As a next step, in the appearing dialog box select "time" as "Time" variable and "status" as "Status" variable. In case of the "status" variable select "Define Event ..." button and set "Single value:" equal to one. In the dialog box belonging to Cox regression the variables X_1 and X_2 should be "Covariates", and "Forward: Wald" should be selected as method. Similar to the Kaplan-Meier model, the survival function can also be estimated with Cox regression: after clicking on the "Plots" button the "Survival" plot can be selected in the dialog box.

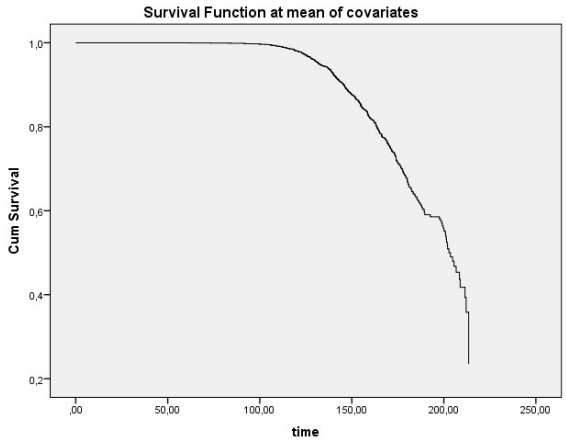


Figure 8.6: Survival function

Table 8.7: Estimated coefficients

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
Step 1 X2	-,737	,047	246,375	1	,000	,479	,436	,525

Table 8.7 shows that only one of the variables (X_2) is entered into the analysis. The estimated coefficient of X_2 is -0.737, which means that the hazard rate function in this example is as follows:

$$h(t) = h_0(t) \cdot e^{-0.737 \cdot X_2} \tag{8.5}$$

This result can be interpreted so that if the value of X_2 increases by one unit, then the hazard rate is multiplied by $e^{-0.737}$ for each t.

Question 8.5. *How can the partial residuals in the model be interpreted?*

Solution of the question.

Partial residuals belonging to the uncensored observations are calculated for each (entered) explanatory variable (*Vékás* (2011), page 185) Figure 8.7 shows the relationship of partial residuals (belonging to variable X_2) and the time variable. This plot can be considered as showing no significant trend, thus it may be concluded that (approximately) the proportional hazards assumption can not be rejected (based on Figure 8.7).

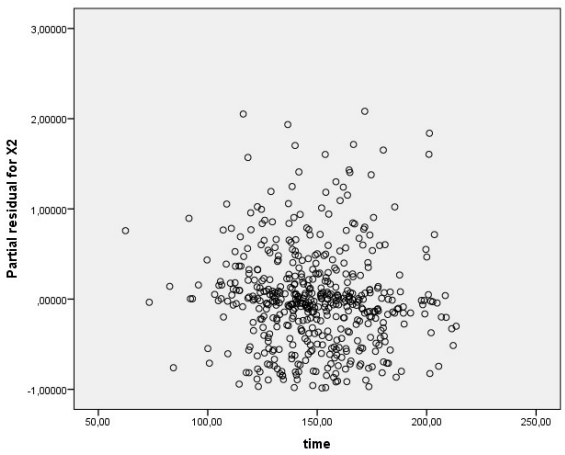


Figure 8.7: Partial residuals

References

[1] Balloun, J. L. – Oumlil, A. B. (1988): Comparing the results of nonmetric multidimensional scaling and principal component analysis. Journal of Business Research, 17, pp. 5-14.

[2] Beh, E. J. (2004): Simple correspondence analysis: a bibliographic review. International Statistical Review, 72, 2, pp. 257-284.

[3] Bécavin, C. – Tchitchek, N. – Mintsa-Eya, C. – Lesne, A. – Benecke, A. (2011): Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. Bioinformatics, Vol. 27., No. 10., pp. 1413-1421.

[4] Bouguettaya, A. – Yu, Q. – Liu X. – Zhou, X. – Song, A. (2015): Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 42, pp. 2785-2797.

[5] Cox, D. R. (1972): Regression models and life-tables. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 34, No.2., pp. 187-202.

[6] Cox, T. F. – Ferry, G. (1993): Discriminant analysis using non-metric multidimensional scaling. Pattern Recognition, Vol. 26., No. 1., pp. 145-153.

[7] Cramer, J. S. (2003): Logit models from economics and other fields. Cambridge University Press

[8] Faraggi, D. – Reiser, B. (2002): Estimation of the area under the ROC curve. Statistics in Medicine, Vol. 21., pp. 3093-3106.

[9] Fawcett, T. (2006): An introduction to ROC analysis. Pattern Recognition Letters, Vol. 27., pp. 861-874.

[10] Fliszár, V. (2011): Az ügyfélminősítés statisztikai módszerei (in Hungarian), In: Kovács, E. (2011): Pénzügyi adatok statisztikai elemzése (chapter 11), Tanszék Kft., Budapest

[11] George, D. – Mallery, P. (2007): SPSS for Windows Step by step. Pearson Education, Inc.

[12] Hajdu, O. (2003): Többváltozós statisztikai számítások (in Hungarian). Központi Statisztikai Hivatal, Budapest

[13] Hajdu, O. (2004): A csődesemény logit-regressziójának kismintás problémái (in Hungarian). Statisztikai Szemle, Vol. 82., pp. 392-422.

[14] Kovács, E. (2011): Pénzügyi adatok statisztikai elemzése (in Hungarian). Tanszék Kft., Budapest

[15] Kovács, E. (2014): Többváltozós adatelemzés (in Hungarian). Typotex

[16] Lee, Y. – Wang, R. (2015): Does modeling lead to more accurate classification?: A study of relative efficiency in linear classification. Journal of Multivariate Analysis, Vol 133., pp. 232-250.

[17] McNeil, A. J. - Frey, R. - Embrechts, P. (2005): Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press

[18] Medvegyev, P. (2002): Valószínűségszámítás, Fejezetek a matematikai analízisből és a valószínűségszámításból (in Hungarian). Aula, 2002

[19] Oravecz, B. (2007): Credit scoring modellek és teljesítményük értékelése (in Hungarian). Hitelintézeti Szemle, 2007/6. pp. 607-627.

[20] Paefgen, J. - Staake, T. - Fleisch, E. (2014): Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. Transportation Research Part A, Vol. 61., pp. 27-40.

[21] Press, S. J. – Wilson, S. (1978): Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, Vol. 73., pp. 699-705.

[22] Pusztai, T. (2007): Gyakorlati tapasztalatok (in Hungarian), In: Sajtos, L. – Mitev, A. (2007): SPSS kutatási és adatelemzési kézikönyv (pp. 321-327.), Alinea Kiadó, Budapest

[23] Rencher, A. C. - Christensen, W. F. (2012): Methods of Multivariate Analysis. Third Edition, Wiley, John Wiley & Sons, Inc.

[24] Robinson, J. (2014): Survival models In: Predictive modeling applications in actuarial science, Volume I: Predictive modeling techniques, edited by: Frees, E. W. – Derrig, R. A. – Meyers, G. (2014), Cambridge University Press, Chapter 19 (pp. 481-514)

[25] Rousseeuw, P. J. (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, pp. 53-65.

[26] Sajtos, L. – Mitev, A. (2007): SPSS kutatási és adatelemzési kézikönyv. Alinea Kiadó, Budapest

[27] Steiner, P. M. – Hudec, M. (2007): Classification of large data sets with mixture models via sufficient EM. Computational Statistics & Data Analysis, 51, pp. 5416-5428.

[28] Sydsæter, K. - Hammond, P. (2008): Essential mathematics for economic analysis, Third edition, Prentice Hall, Pearson Education Limited

[29] Szepesváry, L. (2015): Dinamikus modellek alkalmazása életbiztosítások cash flow előrejelzésére. (in Hungarian) In: Keresztes, G. (editor) (2015): Tavaszi szél 2015 Konferenciakötet II. kötet, Líceum Kiadó, Eger, Doktoranduszok Országos Szövetsége, Budapest (pp. 581-599.) Retrieved from: http://dosz.hu/dokumentumfile/tsz2015_2.pdf, Download date: 2016.04.03.

[30] Verbraken, T. - Bravo, C. - Weber, R. - Baesens, B. (2014): Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, Vol. 238., pp. 505-513.

[31] Vékás, P. (2011): Túlélési modellek (in Hungarian), In: Kovács, E. [2011]: Pénzügyi adatok statisztikai elemzése (chapter 9), Tanszék Kft., Budapest

[32] Wooldridge, J. M. (2010): Econometric analysis of cross section and panel data, Second Edition. The MIT Press

[33] Xie, X.- Pendergast, J. - Clarke, W. (2008): Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. Computational Statistics & Data Analysis, Vol. 52., pp. 2703-2713.

Appendix

Data description

Individuals using the internet for ordering goods or services
(percentage of individuals aged 16 to 74)

	year	
	2010	2015
Belgium	38	55
Bulgaria	5	18
Czech Republic	27	45
Denmark	68	79
Germany	60	73
Estonia	17	59
Ireland	36	51
Greece	12	32
Spain	24	42
France	54	65
Croatia	14	31
Italy	15	26
Cyprus	18	23
Latvia	17	38
Lithuania	11	32
Luxembourg	60	78
Hungary	18	36
Malta	38	51
Netherlands	67	71
Austria	42	58
Poland	29	37
Portugal	15	31
Romania	4	11
Slovenia	27	39
Slovakia	33	50
Finland	59	71
Sweden	66	71
United Kingdom	67	81

Data source: Eurostat (Retrieved from:
http://ec.europa.eu/eurostat/web/information-society/data/main-tables)
Download date: 2016.03.23.

Individuals using the internet for ordering goods or services from other EU countries
(percentage of individuals aged 16 to 74)

	year	
	2010	2015
Belgium	20	35
Bulgaria	2	7
Czech Republic	2	9
Denmark	28	35
Germany	8	13
Estonia	8	26
Ireland	18	30
Greece	4	10
Spain	7	18
France	15	21
Croatia	3	10
Italy	4	11
Cyprus	15	20
Latvia	7	19
Lithuania	3	11
Luxembourg	53	68
Hungary	3	11
Malta	35	44
Netherlands	12	21
Austria	29	44
Poland	2	4
Portugal	6	16
Romania	1	2
Slovenia	10	17
Slovakia	9	20
Finland	21	38
Sweden	13	25
United Kingdom	10	20

Data source: Eurostat (Retrieved from:
http://ec.europa.eu/eurostat/web/information-society/data/main-tables)
Download date: 2016.03.23.

Individuals regularly using the internet
(percentage of individuals aged 16 to 74)

	year	
	2010	2015
Belgium	75	83
Bulgaria	42	55
Czech Republic	58	77
Denmark	86	93
Germany	75	84
Estonia	71	86
Ireland	63	78
Greece	41	63
Spain	58	75
France	72	81
Croatia	51	66
Italy	48	63
Cyprus	50	70
Latvia	62	75
Lithuania	58	69
Luxembourg	86	97
Hungary	60	72
Malta	60	74
Netherlands	88	91
Austria	70	81
Poland	55	65
Portugal	47	65
Romania	34	52
Slovenia	65	71
Slovakia	73	74
Finland	83	91
Sweden	88	89
United Kingdom	80	90

Data source: Eurostat (Retrieved from:
http://ec.europa.eu/eurostat/web/information-society/data/main-tables)
Download date: 2016.03.23.

Individuals never having used the internet
(percentage of individuals aged 16 to 74)

	year	
	2010	2015
Belgium	18	13
Bulgaria	51	35
Czech Republic	28	13
Denmark	9	3
Germany	17	10
Estonia	22	9
Ireland	27	16
Greece	52	30
Spain	32	19
France	20	11
Croatia	42	26
Italy	41	28
Cyprus	45	26
Latvia	29	18
Lithuania	35	25
Luxembourg	8	2
Hungary	33	21
Malta	36	22
Netherlands	8	4
Austria	23	13
Poland	35	27
Portugal	46	28
Romania	57	32
Slovenia	28	22
Slovakia	17	16
Finland	11	5
Sweden	7	5
United Kingdom	13	6

Data source: Eurostat (Retrieved from:
http://ec.europa.eu/eurostat/web/information-society/data/main-tables)
Download date: 2016.03.23.

Enterprises having received orders online (at least 1%)
(percentage of enterprises with at least 10 persons employed in the given NACE sectors, by size class, all enterprises, without financial sector)

	year	
	2010	2015
Belgium	26	25
Bulgaria	4	6
Czech Republic	20	24
Denmark	25	26
Germany	22	25
Estonia	10	13
Ireland	21	32
Greece	9	6
Spain	12	17
France	12	16
Croatia	22	20
Italy	4	7
Cyprus	7	11
Latvia	6	9
Lithuania	22	18
Luxembourg	14	7
Hungary	8	11
Malta	16	16
Netherlands	22	17
Austria	14	15
Poland	8	10
Portugal	19	19
Romania	6	8
Slovenia	10	16
Slovakia	7	13
Finland	16	16
Sweden	24	26
United Kingdom	14	20

Data source: Eurostat (Retrieved from:
http://ec.europa.eu/eurostat/web/information-society/data/main-tables)
Download date: 2016.03.23.