7. Methods for classification: Logistic regression

7.1 Logistic regression basics

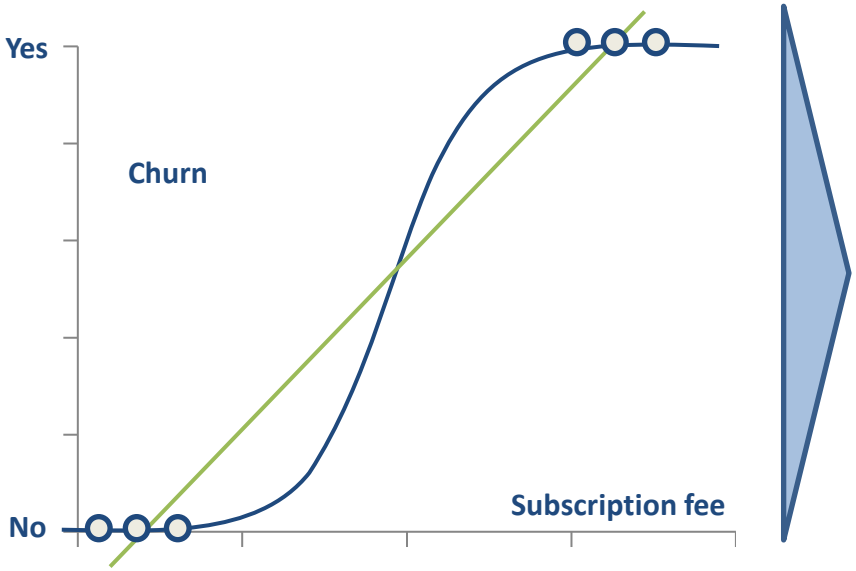
Introductory example from retention management (1)

- The task of retention management (as a part of CRM) is to prevent customers who are likely to churn from leaving with special retention offers.
- Problem: Which customers are at risk of leaving, and how can you find out?
- Procedure: Develop a prediction model from past churn information and customer data and then apply this to current customer data.
- Example: Existing customer information could be: Subscription fee, duration of customer relationship, etc.
- The target variable is: customer churns, values “yes” (1) and “no” (0).
- A linear regression may result in the following prediction model:

Customer churns = $\hat{\beta}_0$ + $\hat{\beta}_1$ * Subscription fee + $\hat{\beta}_2$ * Duration

Introductory example from retention management (2)

- Applying this linear regression model is applied to current customer data results in a problem:



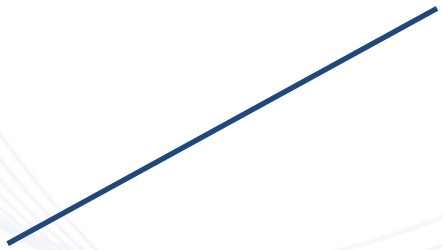
- The linear function (straight line) is obviously unsuitable for mapping the situation.
- The logistic function (S-curve) is much better suited.

- There are two steps necessary:
  - Transformation of the linear regression equation to the interval [0;1].
  - Interpretation of the target variable as a probability.

Differences between logistic regression and linear regression

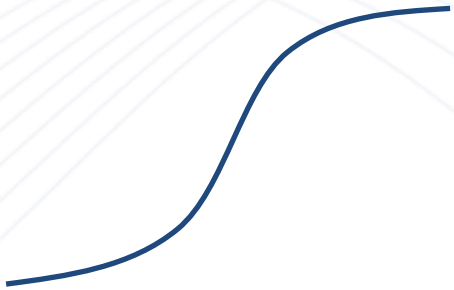
Linear regression

- Numerical target variable.
- Numerical (at least one) and kategorical independent variables.
- Linear function (straight line).



Logistic regression

- Binary (dichotomous) target variable, i. e. the target variable takes on one of two values, e. g. “yes” and “no” or 1 and 0.
- Numerical and categorical independent variables.
- Non linear (logistic) function (S-curve, sigmoid function).



Model equation of a logistic regression model

- Let Y = 0, 1 the dependent binary variable.
- Let L:R→[0;1] a link function. Then the logistic regression model equation is:

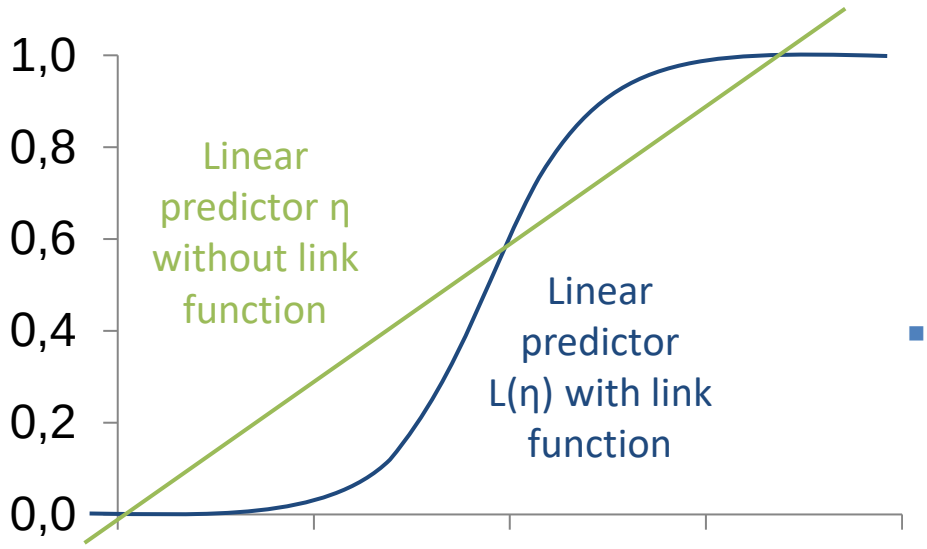
$$p_i = L\left(\overbrace{\beta_0 + \sum_{k=1}^{K} \beta_k x_{ik}}^{\eta}\right) + u_i \quad \text{für} \quad i = 1, \ldots, n; k = 1, \ldots, K$$

with:

p_i	Probability that y_i equals 1 ($p_i = P(y_i = 1)$)
x_{ik}	Observed values for the independent variables X_k
β_0	Intercept
β_k	Regression coefficients
i	Index for the observations ($i=1,\ldots,n$)
k	Index for the independent variables X ($k=1,\ldots,K$)
u_i	Residuals.

Link function

The link function L transforms the linear predictor $\eta = \beta_0 + \sum_{k=1}^{K} \beta_k x_k$ to the interval [0;1].



- The most used link function is the logit link:

$$\text{logit}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}}{1+e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}}$$

- The probit regression uses the normal distribution ϕ as a link function:

$$\text{probit}(\eta) = \Phi(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)dt$$

Logit and probit regression provide comparable results, and neither method has any particular advantages over the other.

Interpretation of the regression coefficients

Let P(customer churns = yes) = logit(1 + 0,5 * subscription fee – 2,2 * duration)

- The sign of the regression coefficients is positive:

  The independent variable has a positive effect on the target variable. In the example: the higher the subscription fee, the greater the probability of churn.

- The sign of the regression coefficient is negative:

  The independent variable has a negative effect on the target variable. In the example: The higher the duration, the smaller the probability of churn.

But: The effect size cannot be interpreted as slope like in linear regression.

7.2 Model fit criteria and model assessment

Model fit

- In classification methods such as logistic regression, there is no single key figure such as the coefficient of determination of linear regression that can be used to determine the model fit.
- Rather, a large number of different criteria and tests have been developed, of which only a selection will be considered below.
- The following are considered in detail: Likelihood ratio test, pseudo R-squared values, fit measures based on the confusion matrix to measure the classification properties and various concentration measures.
- The most important fit measures, which can be used not only for logistic regression but also for other classification methods (such as decision trees, neural networks, etc.), are AIC, BIC, misclassification rate, the lift curve, the ROC curve and the AUC value.

The likelihood function

- The **likelihood function L()** is a function that depends on the parameters to be estimated and produces a probability as the target value:

$$L(\beta_0, \beta_1, \ldots, \beta_k) = P(X = x).$$

- Where x is an (n x k) matrix (i.e. a typical data set) with n observations and k independent variables.
- This target value corresponds to the probability to observe the available data for given parameter estimators.
- Excursus: The maximum likelihood estimator is then just the value for the parameter estimators at which the probability for the observed data is maximum.

Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)

- The **Akaike Information Criterion (AIC)** and the **Bayes Information Criterion (BIC)** are both derived from the likelihood function.
- In addition, the number of model parameters is taken into account as a kind of penalty term.
- With the BIC, the sample size is also taken into account in the penalty term. It is more suitable for larger sample sizes than the AIC.

$$AIC = -2 \cdot \ln \mathrm{L}\left(\beta_0, \beta_1, \ldots, \beta_k\right) + 2 \cdot \left(k+1\right)$$

$$BIC = -2 \cdot \ln \mathrm{L}\left(\beta_0, \beta_1, \ldots, \beta_k\right) + \ln\left(n\right) \cdot \left(k+1\right)$$

- k denotes the number of independent variables and n denotes the sample size.
- **The smaller the value of the AIC or BIC, the better the model.**

Classification properties and confusion matrix

- The **confusion matrix** is a contingency table in which the empirically observed frequencies for the target variable are presented together with the frequencies calculated by the logistic regression model:

		From the model derived frequencies	
		1	0
Observed frequencies	1	n_{11}	n_{12}
	0	n_{21}	n_{22}

- As the logistic regression model provides probabilities for a 1 in the target variable as a result, these probabilities must still be converted to 0 or 1 to form the confusion matrix.
- This can be done as follows for each observation i= 1,...,n in the data set:
  - $y_i = 1$, if $P(Y_i = 1 | \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) > 0{,}5,$
  - $y_i = 0$, if $P(Y_i = 1 | \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) < 0{,}5.$

Misclassification rate and accuracy

- The confusion matrix can now be used to calculate various key figures to assess how well a logistic regression model is able to correctly reproduce the zeros 0 and ones 1 in the data set (classification properties).

		From the model derived frequencies	
		1	0
Observed frequencies	1	n_{11}	n_{12}
	0	n_{21}	n_{22}

- Let $n = n_{11} + n_{12} + n_{21} + n_{22}$ be the total sample size.
- The proportion of misclassified observations in all observations is called the **misclassification rate**:

$$\text{Misclassification rate} = \frac{n_{21} + n_{12}}{n}$$

- The proportion of correctly classified observations in all observations is called **accuracy**:

$$\text{Accuracy} = 1 - \text{Misclassification rate} = \frac{n_{11} + n_{22}}{n}$$

Sensitivity, specificity, true positives and false positives

- The proportion of correctly positive (i.e. classified as 1) observations out of all positive observations is called **sensitivity** or **true positive rate (TP)**:

$$\text{Sensitivity} = \frac{n_{11}}{n_{11} + n_{12}}$$

- The proportion of correctly negative (i.e. classified as 0) observations out of all negative observations is called **specificity**:

$$\text{Specificity} = \frac{n_{22}}{n_{21} + n_{22}}$$

- The proportion of observations classified as false positive out of all negative observations is called the **false positive rate (FP)**:

$$\text{FP} = 1 - \text{Specificity} = \frac{n_{21}}{n_{21} + n_{22}}$$

Interpretation of the figures

- A logistic regression model is good if, compared to a purely random assignment,...
  - the misclassification rate is low,
  - the accuracy is high,
  - the sensitivity (true positive rate) is high,
  - the specificity is high,
  - the false positive rate is low.

N.B.: The criteria are overestimated (accuracy, TP, specificity) or underestimated (misclassification rate and FP) if they are calculated on the same sample as the regression parameters. It is therefore recommended to split the sample (if it is large enough overall). The model parameters β_0, β_1, …, β_k are estimated from one subsample (training data set), and the quality criteria are then calculated from the other subsample (validation or test data set).

Example: Loan default prediction

Out of 1000 customers who were granted a loan 5 years ago, 50 have not repaid their loan. A logistic regression model was fitted to the data and the probability of default was calculated for each customer. If this was greater than 50%, the customer was marked as a loan defaulter (= 1), otherwise not (= 0). This resulted in the following confusion matrix:

		Loan defaults derived from the model	
		1	0
Observed loan defaults	1	20	30
	0	150	800

Classification properties for the loan default example

This results in the following values:

$$\text{Misclassification rate} = \frac{150 + 30}{1000} = 0{,}18$$

$$\text{Accuracy} = \frac{800 + 20}{1000} = 0{,}82$$

$$\text{Sensitivity (TP)} = \frac{20}{20 + 30} = 0{,}4$$

$$\text{Specificity} = \frac{800}{150 + 800} = 0{,}842$$

$$\text{FP} = \frac{150}{150 + 800} = 0{,}158$$

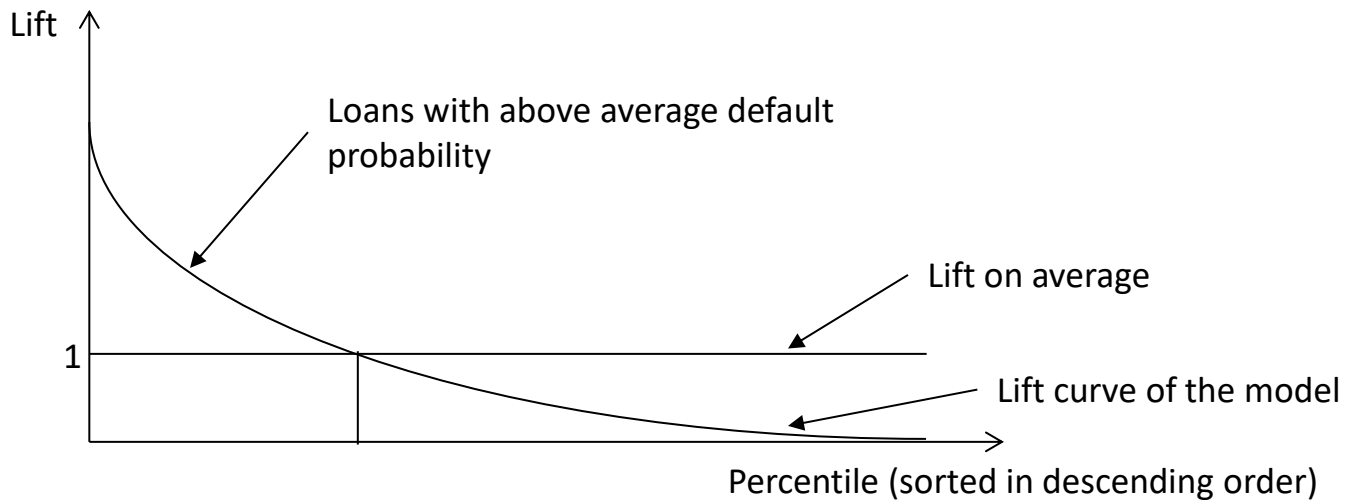
→ Although the misclassification rate is low, not even every second loan default is classified correctly.

Lift values and lift curve

- **Lift values** relate the modeled probability of a 1 in the target variable to the proportion of 1s in the total number of observations.
- First, the modeled probabilities are ordered by size.
- Then the lift values are derived from the percentiles of the ordered modeled probabilities.
- Example: Loan default prediction. A bank has an average loan default rate of 10%. After logistic regression modeling, the bank receives a modeled loan default probability of 25% at the 90% percentile and a modeled loan default probability of 30% at the 95% percentile.
- Then the 10% lift = 25% / 10% = 2.5, and the 5% lift is 30% / 10% = 3.
- Interpretation using the example of the 10% lift: For the 10% loans with the highest default probabilities, the risk of a loan default is 2.5 times higher than the average for all loans.

Graphical representation of the lift curve

- A major advantage of lift values is the simple graphical representation:



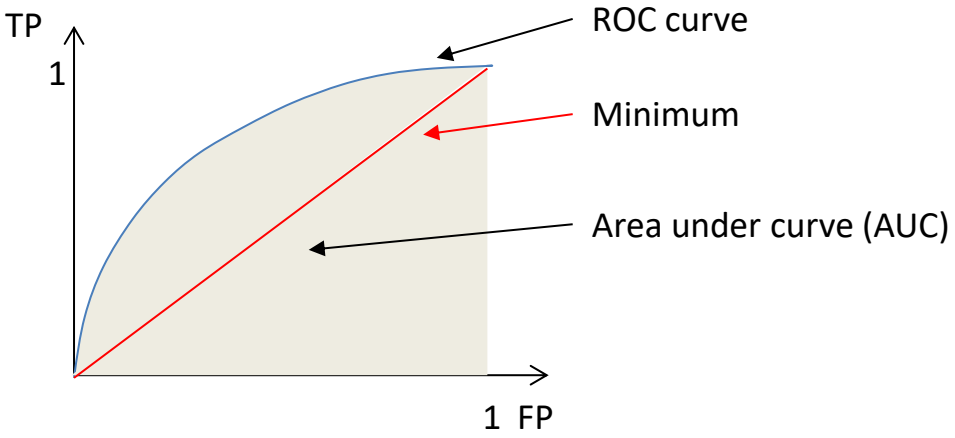
- A model is good if it has high lift values for the 10% or 5% lift, for example, and if the proportion of loans with a modeled above average probability of default is small.
- A lift value of e.g. 2 for the 10% percentile means that the probability of the modeled event occurring is at least twice as high for these 10% as for a random selection from the entire group.

The ROC curve

- The **ROC curve** (ROC = Receiver Operating Characteristic) is based on a generalization of the confusion matrix.
- The confusion matrix is based on a fixed cutpoint, i.e. with a cutpoint of 50%, modeled probabilities greater than 50% are assigned to 1, and modeled probabilities less than 50% are assigned to 0.
- To derive the ROC curve, this cutpoint is now varied. For all possible cutpoints between 0% and 100%, the modeled probabilities are assigned to the values 0 and 1. The true positive rate (sensitivity) and the false positive rate (1-specificity) are then determined for each cutpoint. These are then plotted in a coordinate grid.
- Example: For a cutpoint of 0%, TP = 1 and FP = 1. For a cutpoint of 100%, TP = 0 and FP = 0.

Graphical representation of the ROC curve and the AUC value

- The calculated TP and FP for the various cutpoints are plotted in a coordinate system:



- The **AUC value** (AUROC value) corresponds to the marked area under the ROC curve.
- The higher the AUC value, the better the model. The theoretical maximum value here is 1, the minimum value is 0.5.
- There exists a relationship between the AUC value and the Gini coefficient: Gini = 2 * AUC - 1.

Interpretation of the AUC values

Hosmer and Lemeshow propose the following classification for the AUC values for model assessment (cf. Hosmer/Lemeshow (2000), p. 162):

- AUC value = 0,5 → Poor classification
- 0,7 ≤ AUC value < 0,8 → Acceptable classification
- 0,8 ≤ AUC value < 0,9 → Excellent classification
- AUC value ≥ 0,9 → Outstanding classification

In practice, values of 0.9 or above are very rarely achieved.

Notes:

Lift values (and lift curves) depend on the underlying average and are therefore not comparable for different populations.

In contrast to the lift, AUC values of different models with different populations can be compared with each other.

7.3 Logistic regression using 

Required packages

The ROCR and pROC packages are required for the lift and ROC curves. The package car may also be required for the recode() command.

R Script

```
# Activate packages
library(car) # recode()
library(pROC) # ROC curve and AUC value
library(ROCR) # Lift curve
```

The data set

The implementation in R is explained using the example of the SwissLabor data set from the AER package.

R Script

```
# Load and inspect data set
data("SwissLabor", package = "AER")
head(SwissLabor)

# Data set description
# Quelle: http://cran.r-project.org/web/packages/AER/AER.pdf
#
# A data frame containing 872 observations on 7 variables.
# Variables:
# participation    Factor. Did individual participate in labor force?
# income           Logarithm of nonlabor income.
# age              Age in decades (years divided by 10).
# education        Years of formal education.
# youngkids        Number of young children (under 7 years of age).
# oldkids          Number of older children (over 7 years of age).
# foreign          Factor. Is individual a foreigner (i.e., not Swiss)?
```

Definition of the target variable

- It is essential to know which value of the target variable is denoted as 1 and which as 0, as the sign of the parameter estimators depend on this.
- R sorts the two values of the target variable alphanumerically and interprets the “larger” value internally as 1 and the “smaller” value as 0.
- If a different specification is required, the target variable must be recoded.
- In this data set, the target variable participation has the values yes and no.
- Yes is automatically coded internally as 1.If “no” has to be coded as 1, use the following command line:

R Script

```
# To specify participation = "no" as 1
SwissLabor$participation2 <- recode(SwissLabor$participation,
                                    '"no" = 1; "yes" = 0; ;', as.factor=TRUE)
```

Dummy variables

- R sorts the values of the categorical independent variables alphanumerically and sets the “smallest” value as the reference category.
- If a different reference category is required, it must be recoded in the same way as the binary target variable.
- In the present data set, the factor variable foreign has the values yes and no. The reference category is therefore no.

Carrying out a logistic regression

- The probability of participation in the Swiss labor market is modeled as the target variable; the variables income (numerical) and foreign (categorical) are used as independent variables.
- The categorical independent variable is automatically interpreted by R as a factor.
- For a logistic regression, the option family=binomial(logit) must be selected; for a probit regression, the option family=binomial(probit) must be selected.

R Script

```
# Carrying out a logistic regression
logreg <- glm(participation ~ income + factor(foreign),
              family=binomial(logit), data=SwissLabor)
# Summary of the results
summary(logreg)
```

Summary of the results

Output

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           6.7717     1.9905   3.402 0.000669 ***
income               -0.6743     0.1857  -3.632 0.000281 ***
factor(foreign)[T.yes]  1.0961     0.1699   6.450 1.12e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1132.2  on 869  degrees of freedom
AIC: 1138.2

Number of Fisher Scoring iterations: 4
```

p values (Wald test)

Parameter estimators for the regression coefficients.

Akaike Information Criterion

Given the significant regression coefficients from above, the following regression equation is obtained (significance level $\alpha = 0.05$):

$$P(participation = 1) = \frac{\exp(6.7717 - 0.6743 \cdot income + 1.0961 \cdot foreign)}{1 + \exp(6.7717 - 0.6743 \cdot income + 1.0961 \cdot foreign)}$$

Model fit: AIC and BIC

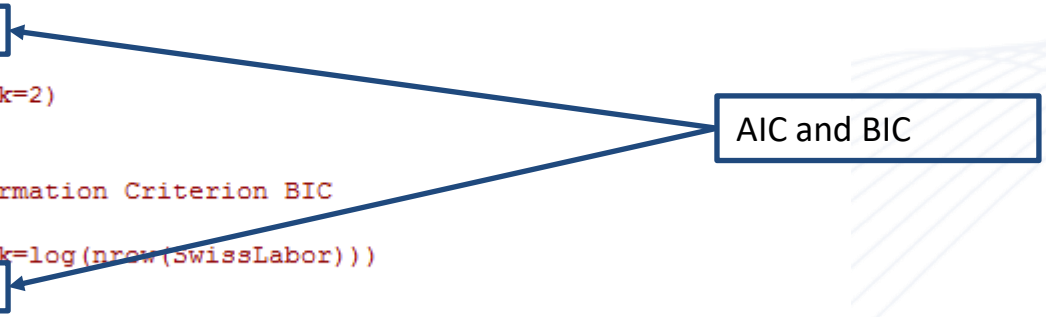
There are at least two ways to determine AIC and BIC:

R Script

```
# Aikaike Information Criterion AIC
AIC(logreg)
AIC(logreg, k=2)
# Bayes Information Criterion BIC
AIC(logreg, k=log(nrow(SwissLabor)))
BIC(logreg)
```

Output

```
> # Aikaike Information Criterion AIC

> AIC(logreg)
[1] 1138.242

> AIC(logreg, k=2)
[1] 1138.242

> # Bayes Information Criterion BIC

> AIC(logreg, k=log(nrow(SwissLabor)))
[1] 1152.555

> BIC(logreg)
[1] 1152.555
```



Classification properties: Preparation of the confusion matrix

- To calculate the classification figures for the confusion matrix, the confusion matrix must first be calculated. This requires the values predicted by the model for the probability as well as a cutpoint.
- If you want to eliminate those parameters from the model that are not significant, a new logistic regression must be carried out without these parameters. Otherwise, non-significant parameters are also used to calculate the modeled probabilities.
- The cutpoint used is 0.5 - if the probability calculated according to the model is greater than this cutpoint, the modeled target value is set to 1, otherwise to 0.

R Script

```
# Determination of the modelled target variable for a given cutpoint
cutpoint <- 0.5
SwissLabor$predicted <- ((logreg$fitted.values) > cutpoint)*1
```

Classification properties: Calculation of the confusion matrix

R Script

```
# Calculation and display of the confusion matrix
km <- xtabs(~participation+predicted, data=SwissLabor)
km
```

Output

```
> # Berechnung und Ausgabe der Konfusionsmatrix

> km <- xtabs(~participation+predicted, data=SwissLabor)

> km
             predicted
participation   0   1
          no  384  87
          yes 244 157
```

The rows contain the observed values, and the columns contain the values calculated by the model.

Classification properties: Calculation of the classification figures

R Script

```
# Misclassification rate
(km[2]+km[3])/sum(km[1:4])

# Accuracy
(km[1]+km[4])/sum(km[1:4])

# Sensitivity / True Positive Rate
km[4]/(km[2]+km[4])

# Specifity
km[1]/(km[1]+km[3])

# False Positive Rate
km[3]/(km[1]+km[3])
```

R 4.4.2 · ~/

```
> # Misclassification rate
> (km[2]+km[3])/sum(km[1:4])
[1] 0.3795872
> # Accuracy
> (km[1]+km[4])/sum(km[1:4])
[1] 0.6204128
> # Sensitivity / True Positive Rate
> km[4]/(km[2]+km[4])
[1] 0.3915212
> # Specificity
> km[1]/(km[1]+km[3])
[1] 0.8152866
> # False Positive Rate
> km[3]/(km[1]+km[3])
[1] 0.1847134
>
```

The misclassification rate is relatively high, with low sensitivity. On the other hand, the high specificity and low false positive rate are positive.

Calculation of the concentration figures

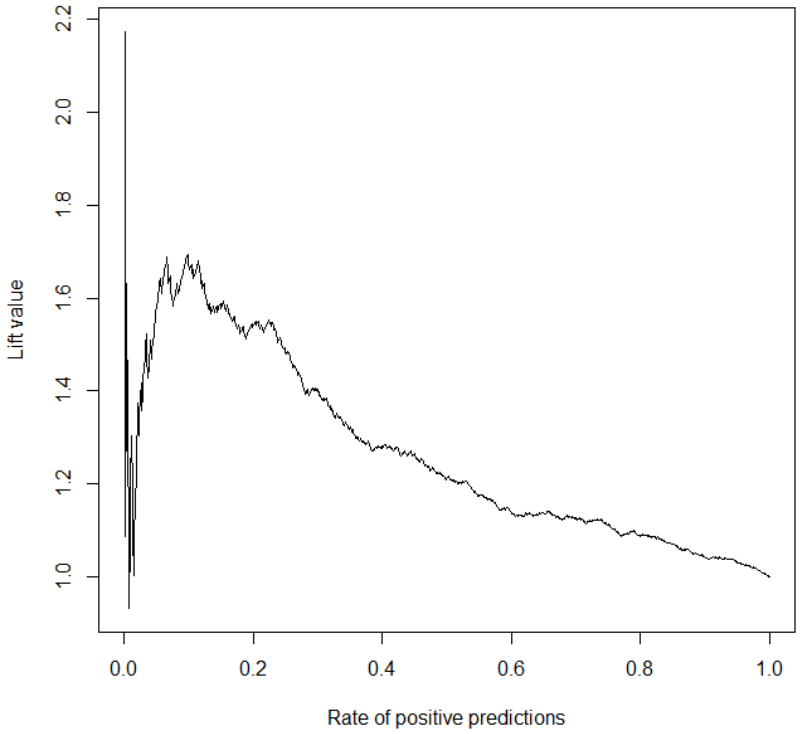
- First, an object of the prediction class must be created. The first argument of the prediction() function is the modeled probabilities, the second argument is the true values of the target variable.
- With the ROC curve, it is exactly the opposite. The first argument of the roc() function is the true values of the target variable, the second argument is the modeled probabilities.

R Script

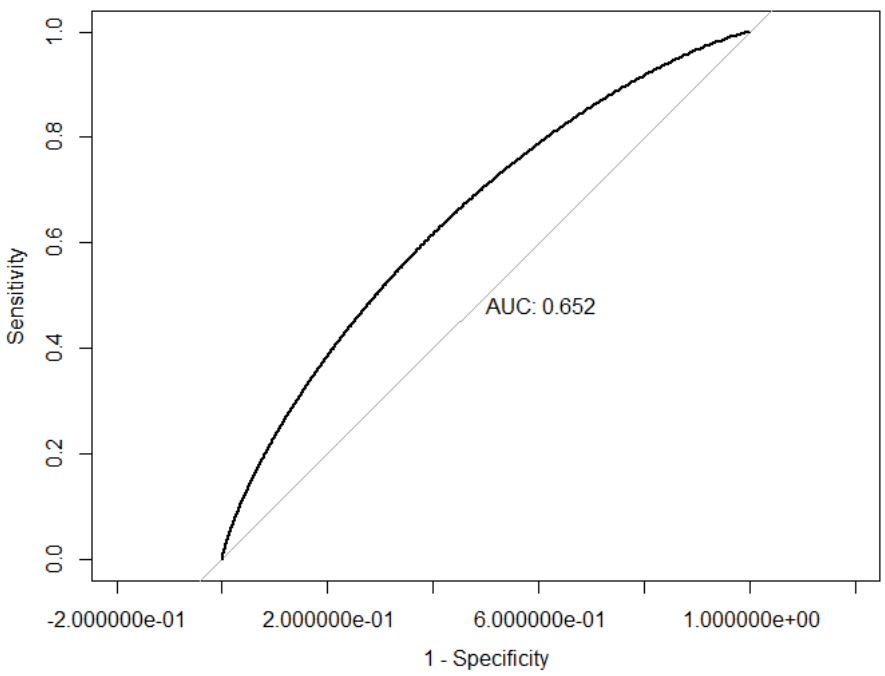
```
# Create an object of the prediction class
pred <- prediction(logreg$fitted.values, SwissLabor$participation)

# Lift curve
perf <- performance(pred,"lift","rpp")
plot(perf)

# ROC curve and AUC value (area under curve)
roc(SwissLabor$participation,logreg$fitted.values,
    plot=TRUE, legacy.axes=TRUE, print.auc=TRUE,
    smooth=TRUE)
```

Lift curve, ROC curve and AUC value

Lift curve



ROC curve



Both curves and the AUC value indicate a poor model fit.

Exercise

Please exercise the logistic regression.

8. Methods for dimension reduction: Factor analysis

Based on Backhaus, K.; Erichson, B.; Gensler, S.; Weiber, R.; Weiber, T. (2023): Multivariate Analysis, Berlin. Some images may be under fair use guidelines (educational purposes).

8.1 Problem definition

Objectives and basic assumption

- Factor analysis is a multivariate method that can be used for analyzing large data sets with two main goals:
  1. To reduce a large number of correlating variables to a fewer number of factors (main objective in data science → principal component analysis).
  2. To structure the data with the aim of identifying dependencies between correlating variables and examining them for common causes (factors) in order to generate a new construct (factor) on this basis (main objective in social sciences → exploratory factor analysis).
- For both objectives it is assumed that the data set to be analyzed consists mainly of highly correlated (dependent) variables.
- Factor analysis therefore primarily analyzes the correlation matrix (or the variance-covariance matrix) of the data, which represents the interrelations between the variables.
- Therefore, the correlation matrix must have certain properties, i.e. there should be sufficient interrelation (correlation) between the variables. Otherwise, a factor analysis is not possible.

Examples

Discipline	Exemplary research questions
Environmental	A broad factor analysis assesses and summarizes the 4 micro-environmental factors (political, economic, social and technological) which have a significant impact on the business's operating environment
Marketing	A chocolate company is interested in its image. 26 variables were reduced to 3 factors: reputation, competence, brand
Medicine	Factor analysis can improve medical diagnostics. Compared to several other attempts to solve the diagnostic problem, this concept has the advantage that it does not depend on the condition of mutual independence of symptoms
Psychology	The Big Five of the OCEAN project (openness, conscientiousness, extraversion, agreeableness, neuroticism) are factors that describe the construct 'personality'. These 5 factors were derived from 18,000 variables
Operations	In many modern manufacturing processes, large quantities of multivariate data are available through automated in-process sensing. Factor analysis is suitable for extracting and interpreting information from these data for the purpose of diagnosing root causes of process reliability

Backhaus et al. (2023), p. 382.

Fundamental theorem of factor analysis

- Factor analysis assumes that each observed value (z_{pi}) of a standardized variable i in observation p can be represented as a linear combination of several (unobserved) factors:

$$z_{pi} = \lambda_{i1} * f_{p1} + \ldots + \lambda_{iK} * f_{pK} + \varepsilon_{pi}$$

z_{pi}	Standardized value of observation p for variable i
λ_{ik}	Weight of factor k for variable i (factor loading)
f_{pk}	(Not observable) factor value of factor k for observation p
ε_{pi}	Measurement error ob observation p for variable i

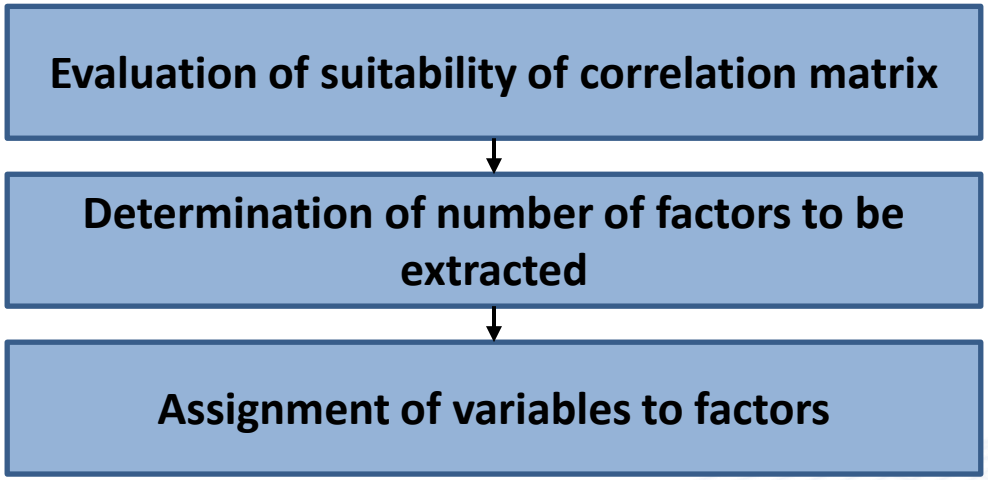
- Let R be the correlation matrix for the (standardized) variables and let L be the matrix of the factor loadings of the I variables on the K factors. Then the correlation matrix can be decomposed (under assumptions not specified here) as follows:

$$R = L * L' + V$$

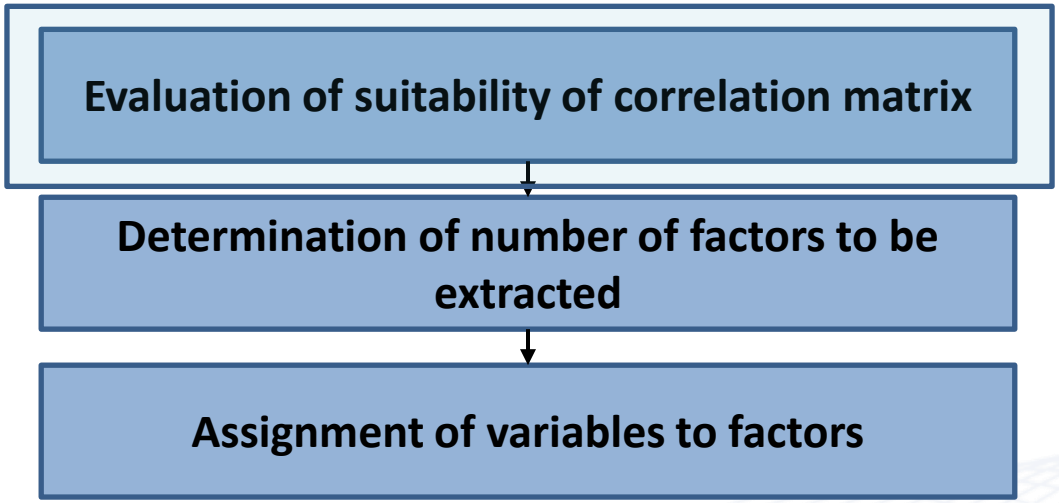
- Under the assumption of no variable-specific measurement error ε_{pi}, V = 0, and the correlations of the indicators can be fully explained by the extracted factors: R = L * L‘

8.2 Carrying out a factor analysis

Procedure


Evaluation of suitability of correlation matrix
Determination of number of factors to be extracted
Assignment of variables to factors

Procedure


Evaluation of suitability of correlation matrix
Determination of number of factors to be extracted
Assignment of variables to factors

Criteria for the suitability of the correlation matrix

- Critical assumption for a factor analysis is a suitable correlation matrix, i.e. a correlation matrix, that contains sufficiently high correlations between the variables.
- The **correlation matrix** itself gives a first impression about the correlations in the data set, but it is not possible the derive any conclusions about its suitability for factor analysis.
- The **Kaiser-Meyer-Olkin (KMO) criterion** and the **Bartlett test of sphericity** enable the assessment of the entire correlation matrix.
- The **Measure of Sampling Adequacy (MSA)** allows the assessment of each single variable and its suitability for factor analysis.

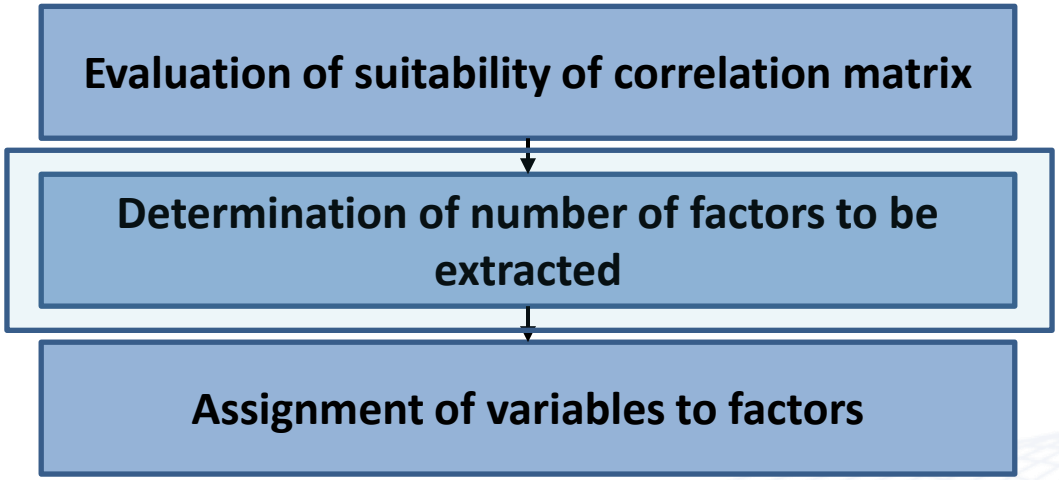
Kaiser-Meyer-Olkin criterion and Bartlett test of sphericity

- The **Kaiser-Meyer-Olkin (KMO) criterion** and indicates to what extend a set of variables share a common variance, i.e. to what extend these variables “measure the same thing”.
- The KMO ranges between 0.5 and 1, the closer to 1, the larger the share of a common variance.
- KMO < 0.5 is not acceptable, KMO > 0.8 is desirable.
- The **Bartlett test of sphericity** enable tests the null hypothesis of no correlation between the variables against the alternative, that the variables are correlated.
- The test assumes normal distributed variables, and the test statistic follows a Chi squared distribution. Under H_0, the test statistic becomes 0.
- If H_0 is rejected, then the variables can be assumed as correlated. However, no statement is made about the extend of the correlation.

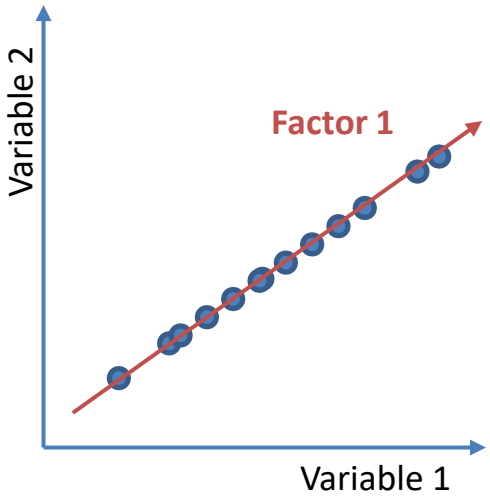
Measure of Sampling Adequacy

- While the KMO criterion evaluates the suitability of all variables together, the **measure of sampling adequacy (MSA)** assesses the suitability of single variables.
- The MSA indicates to what extend one single variable shares a common variance with the other variables.
- The MSA ranges between 0 and 1, the closer to 1, the larger the share of a common variance.
- MSA < 0.5 is not acceptable, MSA > 0.8 is desirable.

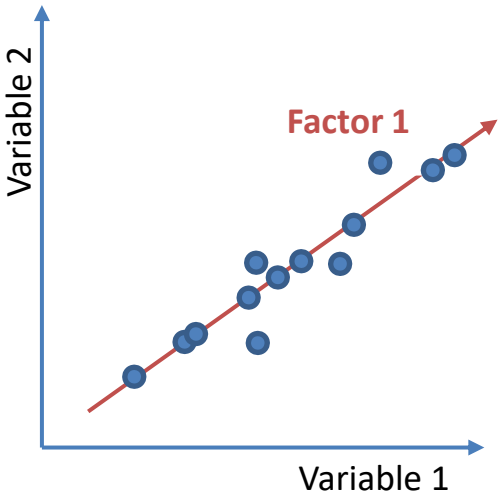
Procedure


Evaluation of suitability of correlation matrix
Determination of number of factors to be extracted
Assignment of variables to factors

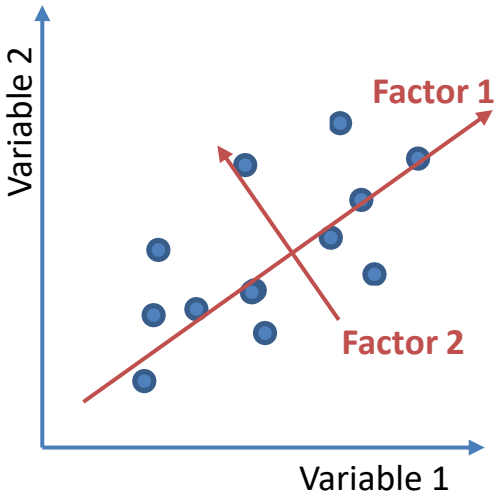
Example: When to extract one or two factors from two variables



One factor captures the entire variance.
→ Extract 1 factor with no loss of information.



One factor captures most of the variance.
→ Extract 1 factor with only little loss of information.



One factor is not sufficient, the loss in information is too big.

→ Extract 2 factors.

The number of factors to be extracted depends on the share of the variance captured by the factors.

Eigenvalues

- The **eigenvalues** describe the explanatory contribution of a factor to the variance of all variables.
- They are calculated as the sum of the squared factor loadings of a factor across all variables.
- If a high share of the variance of the variables is explained (i.e. the eigenvalue is large), this means that the corresponding factor reflects the differences in the observed values well.
- A maximum of as many factors can be extracted as there are variables. This means that the maximum number of possible eigenvalues corresponds to the number of factors.
- For standardized variables, i.e. variables with mean = 0 and variance = 1, the sum of the variances equals the number of variables.
- In the case of standardized variables, if the number of factors extracted equals the number of variables, then the sum of their eigenvalues equals also the number of variables.

Kaiser criterion and elbow criterion

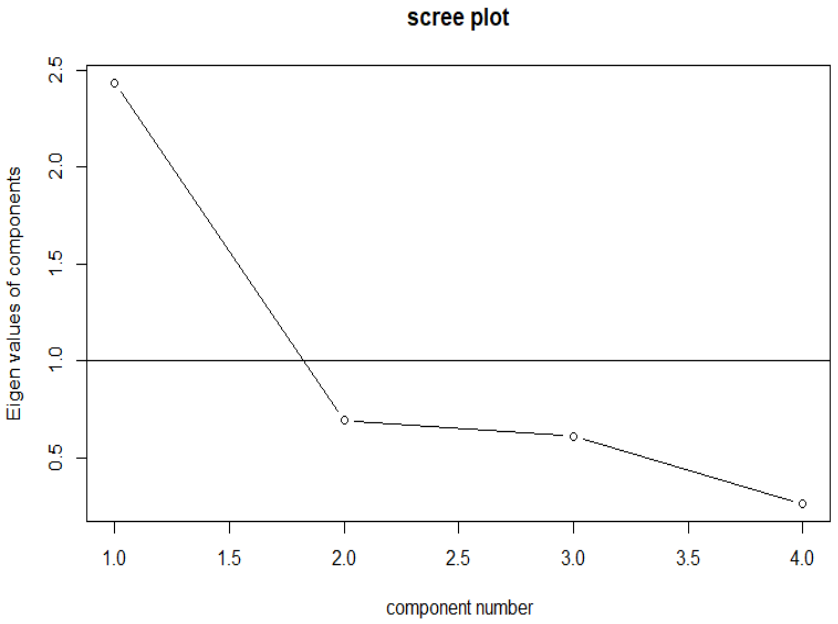
Eigenvalue criterion (Kaiser criterion)

- Extract as many factors as there are eigenvalues >1.
- Reason: Factors with an eigenvalue <1 have less variance and therefore less explanatory power than the original variables.
- This means that they explain the differences in the observed values less well than the variables themselves.

Scree plot (Elbow criterion)

- Graphically displays the eigenvalues in order of magnitude.
- Consider as many factors as there are eigenvalues left from the elbow.
- If there are several elbows, consider the elbow most to the right.

Example: Scree plot for four variables

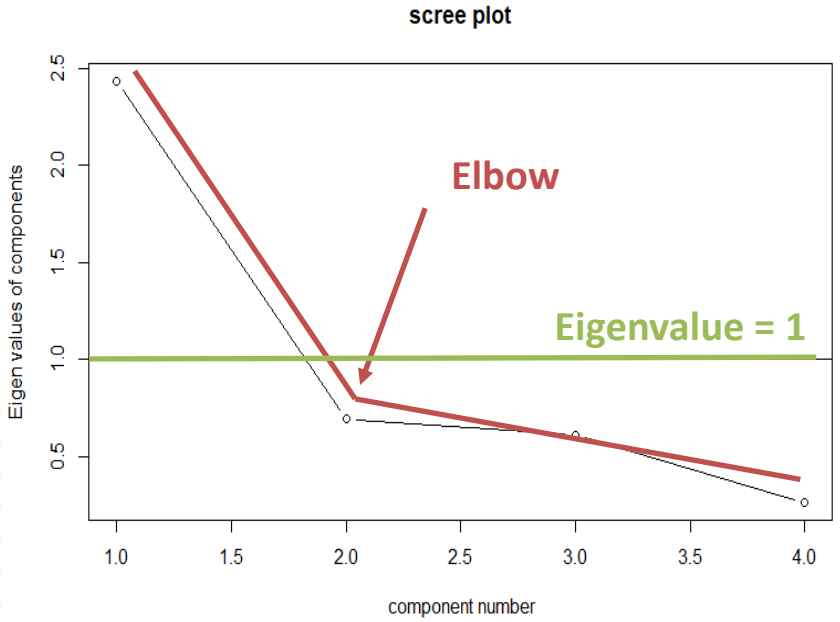


- The factors (component number) are plotted on the x-axis according to the size of the eigenvalues.
- The eigenvalues are plotted on the y-axis.

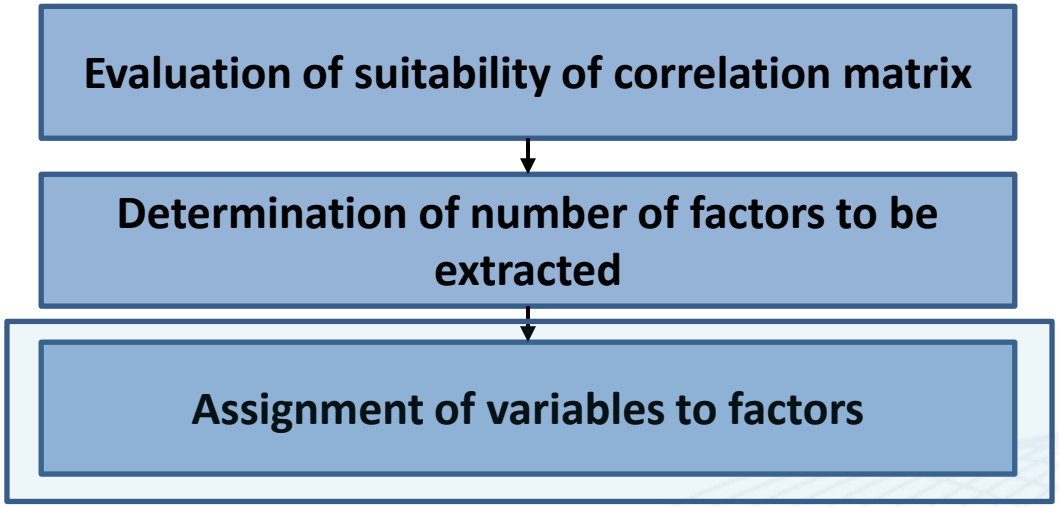
Interpretation

- One eigenvalue exceeds 1.
- Left from the elbow is one eigenvalue.

→ Extract one factor.



Procedure


Evaluation of suitability of correlation matrix
Determination of number of factors to be extracted
Assignment of variables to factors

Factor loading and average variance extracted

- It follows from the fundamental theorem, that the **factor loadings** can be understood as regression coefficients of the individual factors on the variables. The values for the factor loadings range between 0 and 1.
- A variable contributes to a factor if its factor loading is sufficiently high. Factor loadings from 0.4 are considered sufficient.
- Further, the reduction of many variables to a few factors should not result in a loss of too much information, i.e. variance.
- The sum of the eigenvalues of the extracted factors divided by the number of variables yields the share of the variance captured by all the extracted factors. This figure is called the **average variance extracted (AVE)**.
- The AVE ranges between 0 and 1, whereby values of at least 0.5 are considered as sufficient. That means, at least 50% of the original information (variance) shall be captured by the extracted factors.

Example: Obstacles to the growth of new SMEs in South Africa

- In 2010, Fatoki Olawale and David Garwe from the University of Fort Hare, South Africa, examine several obstacles to the growth of new SMEs in South Africa.
- By literature analysis and by questionnaires, they identify 30 obstacles, who shall be summarized to a few main factors.
- The results are published here: https://academicjournals.org/journal/AJBM/article-full-text-pdf/A1AFDEC23302

8.3 Factor analysis using 

Required packages and example data set

R Script

```
# Activation of the required packages
library(corrplot) # corrplot()
library(REdaS) # KMOS()
library(psych) # cortest.bartlett(), VSS.scree(), principal()
```

The sample data set contains 170 observations for 55 variables. Either modify this path: “E:/mydirectory/“ or use file.choose():

R Script

```
# Import example data set
divorce <- read.csv("E:/mydirectory/divorce.csv")
# alternatively:
divorce <- read.csv(file.choose())
```

A data set description can be found here (QR code):
https://www.kaggle.com/datasets/csafrit2/predicting-divorce/data

Choose the variables for factor analysis and correlation plot

Of the 55 variables in the data set, not all shall be used for factor analysis, the last variable shall be omitted:

R Script

```
# Choose columns for factor analysis
pca <- divorce[,-55]
```

A first overview can be gained with a plot of the correlations:

R Script

```
# Graphical representation of the correlations
corrplot(cor(pca))
```

The result is too tiny to be included into the slides.

Suitability of the correlation matrix – KMO, Bartlett test and MSA

KMO and MSA can be calculated using the KMOS() function from the REdaS package. For the Bartlett test, the cortest.bartlett() function from the psych package can be used. The cortest.bartlett() function requires the correlation matrix instead of the data set, and it requires also the data set's number of observations n.

R Script

```
# Kaiser-Mayer-Olkin criterion
KMOS(pca)$KMO
# --> KMO = 0.96 > 0.8 --> FA can be carried out.

# Bartlett test of sphericity
cortest.bartlett(cor(pca), n=nrow(pca))
# --> p=0, i.e. there are correlations between the variables

# Measure af sampling adequacy MSA
KMOS(pca)$MSA
# --> All MSAs exceed 0.7, almost all MSAs exceed 0.9
# --> All variables can be used for FA
```

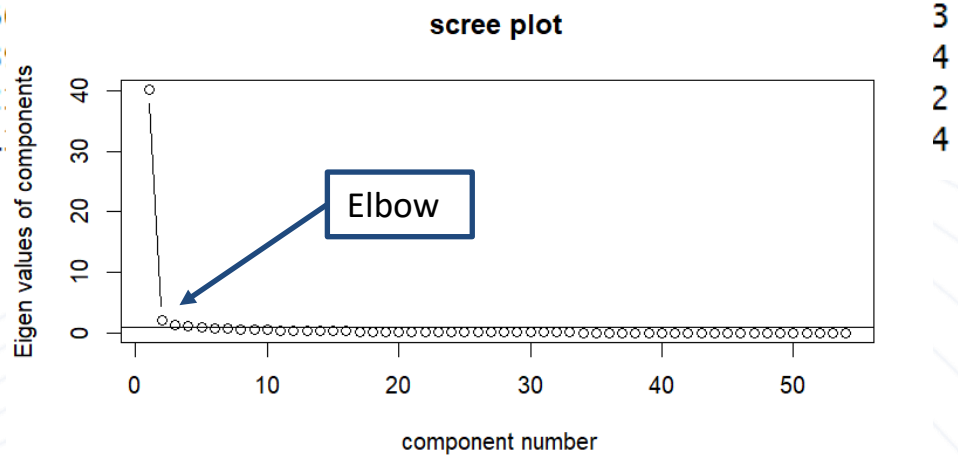
All criteria are met, the correlation matrix is suitable for factor analysis

Number of factors – Scree plot and eigenvalues

R Script

```
# Scree plot
VSS.scree(pca)

# Eigenvalues
eigen(cor(pca))$values
```

```
> eigen(cor(pca))$values
 [1] 40.175686675  2.165315888  1.416514361  1.194053683  0.896249766  0.788897467
 [7]  0.698636632  0.595362778  0.568366461  0.529115848  0.428671399  0.390369411
[13]  0.362469358  0.319198119  0.283031788  0.270341457  0.252520084  0.220402276
[19]  0.202203198  0.191133304  0.175596029  0.165126304  0.138913064  0.132199503
[25]  0.119176457  0.107826952  0.10                                             7
[31]  0.078281900  0.071081992  0.06                                             3
[37]  0.050131110  0.046403026  0.03                                             4
[43]  0.028048802  0.024985265  0.02                                             2
[49]  0.015062529  0.013119694  0.01                                             4
```



Difficult to decide. The Kaiser criterion suggests four factors, the scree plot suggests one factor.

Assignment of variables to factors – Factor loadings and AVE

The principal() function from the psych package carries out the factor analysis. The print function can be used to format the output. E.g., all factor loadings below 0.4 will be suppressed (option cut=0.4).

R Script

```
# One factor
result <- principal(pca, 1) # Number of factors is 1
print(result, cut=.4, sort=F, digits=2)

# Four factors
result <- principal(pca, 4) # Number of factors is 4
print(result, cut=.4, sort=F, digits=2)
```

All but one loading exceeds 0.4, and AVE = 0.74 (proportion variance).

There is no clear assignment of the variables to the four factors.
AVE = 0.83 (cumulative variance)

Exercise

Exercise the factor analysis.