Exam Questions

Data Analytics

Achim Zeileis, Janette Walde

From this list of questions the theory questions for the exam are taken. The questions are to be answered on the basis of the course material. In addition to the theory questions, the test contains application examples in R. For these application tasks R-output and graphs are provided and must be interpreted along questions. The R tasks are always based on the problems discussed in the exercises and tutorials during the course.

1 Multivariate Analysis & Cluster Analysis

Univariate and bivariate exploratory data analysis

- Which possibilities for a numerical description of a qualitative or quantitative variable do you know?
- Which possibilities for a graphical description of a qualitative or quantitative variable do you know? Sketch for each type of variable a typical graph.
- Give possibilities for a numerical and graphical description of the relationship between a dependent qualitative or quantitative variable and an explanatory qualitative or quantitative variable. Sketch a typical praph in each case.

Multivariate analysis

- Which possibilities of standardizing a data matrix with n observations of p variables were discussed? What is the idea of the respective standardization?
- What is a principal component analysis (PCA)? What does the first principal component indicate?
- How is a biplot constructed? What does it reveal?
- How can you measure distances between observations? How are Manhattan, Euclidean and the maximum distance defined?
- What is the basic objective of a cluster analysis, which properties should the clusters have?
- What is a partition, what is a hierarchy of partitions?
- How is a hierarchy of partitions constructed in agglomerative clustering? Illustrate the approach using the single/complete/average linkage method.
- What property defines the k-means partition? Describe an algorithm for construction.
- How do you choose the number of main components?
- How do you choose the number of clusters in a hierarchical cluster analysis?
- How do you choose the number of clusters in k-means?

2 Associations & Multi-Sample Comparisons

- What is the idea behind the construction of the t test statistic in the 2-sample problem?
- Which test distributions for the t-test statistic were discussed? What are the ideas/assumptions for the respective distribution?
- What is the idea behind the construction of the F-test statistic in the 2-sample problem?
- Describe the necessary steps to perform a permutation test for investigating the association between two variables X and Y.
- Describe how a robust test statistic for a test on the difference in the central tendency can be constructed in the 2-sample problem.

3 Linear Models

Linear regression and Analysis of Variance

- What is a simple regression model?
- Given the classical linear regression model

  $$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1, ..., n$$

  (a) What is the interpretation of the regression coefficients $\beta_j, j \geq 2$?

  (b) Which change is expected for the response variable Y if X_j increases/decreases by 1 unit and $\beta_j < 0$ or $\beta_j > 0$ or $\beta_j = 0$?

- Given the classical linear regression model in semi-logarithmic form

  $$\log(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1, ..., n$$

  (a) What is the interpretation of the regression coefficients $\beta_j, j \geq 2$?

  (b) Which change is expected for the response variable Y approximately if X_j increases/decreases by 1 unit (or fällt) and $\beta_j < 0$ or $\beta_j > 0$ or $\beta_j = 0$?

- Given the classical linear regression model in log-log specification

  $$\log(y_i) = \beta_1 + \beta_2 \log(x_{i2}) + \beta_3 \log(x_{i3}) + ... + \beta_k \log(x_{ik}) + \varepsilon_i, i = 1, ..., n$$

  (a) What is the interpretation of the regression coefficients $\beta_j, j \geq 2$?

  (b) Which change is expected for the response variable Y approximately if X_j increases/decreases by 1 percent and $\beta_j < 0$ or $\beta_j > 0$ or $\beta_j = 0$?

- What are the standard assumptions in the classical linear regression model?
- Why is in the classical linear regression model the assumption $E(\varepsilon_i) = 0$ important? What are the consequences for the estimation of the response variable if that assumption is not fulfilled?

- What is a homoskedastic and what is a heteroskedastic error in the classical linear regression model?
- What is meant by autocorrelated errors in the classic linear regression model?
- What is the Gauss-Markov Theorem?
- What can be assessed by the coefficient of determination in the simple linear regression model?
- In the classical linear regression model the response variable Y is explained by the regressors $X_2, \dots, X_k$. For what testing problem can the F statistic be used?
- In the classical linear regression model the response variable Y is explained by the regressors $X_2, \dots, X_k$. What is the meaning of the t statistic of the regression parameters?
- Given a classical linear regression model with response variable Y and regressors $X_2, \dots, X_k$. Based on the data $y_i, x_{i2}, \dots, x_{ik}$ $(i = 1, \dots, n)$ the parameters are estimated. What is the point forecast for a new scenario $x_{n+1,2}, \dots, x_{n+1,k}$?
- What is meant by nested models? Based on which measures can the model comparison of nested classical linear regression models be done?
- Given are n observations $y_1, \dots, y_n$, for which the mean value μ is computed using the least squares method. What is the residual sum of squares $RSS(\mu)$ and how can this sum be used to determine the model parameter? (This corresponds to the trivial model with only an intercept.)
- How can qualitative variables be included in the linear model and what role do contrasts play? Illustrate the problem based on a variance analysis.

Model selection

- Which two fundamentally different methods for comparing two models do you know? Can you briefly explain the idea for both methods.
- How is AIC or BIC defined? What is the motivation for these information criteria and how can you compare models with them?
- What kind of strategies are available for step-by-step model selection? Explain each of these strategies. How can they be used to find a suitable model?

4 Generalized Linear Models

- What is the model equation of GLM? Explain briefly the role of the individual symbols.
- How can the link function be interpreted in the logistic regression? What change occurs in the dependent variable if a regressor x_j with coefficient β_j changes one unit and all other regressors stay the same?
- How are the Pearson residuals defined in the GLM? Can you briefly explain the meaning of the symbols. In addition, give the specific formula for the normal/binomial/Poisson distributed model.