Institut für STATISTIK

Prof. Dr. P. Sibbertsen
Wirtschaftswissenschaftliche Fakultät
Leibniz Universität Hannover
Sommersemester 2020

Themen für Bachelorarbeiten

Hinweis: Themen mit englischer Beschreibung müssen auf Englisch bearbeitet werden.
Note: Topics with an English description have to be written in English.

Modelle der beschreibenden Statistik und Stochastik

1. **Konzentrationsmessung**

   Eine häufige Fragestellung in der Wirtschaft bezieht sich auf die Verteilung von Umsätzen oder Vermögen auf die Merkmalsträger. Anhand des Herfindahl-Indexes wird die absolute Konzentration gemessen, um zu prüfen, ob ein hoher Anteil der Merkmalssumme auf eine kleine absolute Zahl an Merkmalsträgern fällt. Damit wird z.B. Kartellrecht auf einem Markt überprüft. Die relative Konzentration, häufig gemessen mit dem Gini-Koeffizienten, gibt hingegen an, auf welchen Anteil an Merkmalsträgern ein bestimmter Anteil der Merkmalssumme verteilt ist. Darüber können z.B. Aussagen über Ungleichverteilungen von Einkommen getroffen werden.

   Einstiegsliteratur:
   - Philipp Sibbertsen und Hartmut Lehne. 2015. *Statistik: Einführung für Wirtschafts-und Sozialwissenschaftler.* Springer-Verlag (Kap. 4)
   - Jürgen Hedderich und Lothar Sachs. 2016. *Angewandte Statistik: Methodensammlung mit R.* Springer Spektrum, Berlin, Heidelberg (Kap. 3)

2. **Preisindizes**

   Monatlich gibt das Statistische Bundesamt Daten über die Preisentwicklung in Deutschland aus. Diese Preisentwicklungen basieren auf Preisindizes. Preisindizes berücksichtigen dabei sowohl die tatsächlichen Preise bestimmter Güter als auch das Verbrauchsverhalten der Konsumenten. Zur Abbildung des Kosumverhaltens gibt es verschiedene Möglichkeiten. Hierzu wird ein typischer Warenkorb gebildet, der entweder für die Basis- oder die Berichtsperiode repräsentativ sein soll. In dieser Arbeit sollen die verschiedenen Preisindizes nach Laspeyres, Paasche und Fisher vorgestellt und ihre Unterschiede und Motivation zur Bildung von Preisindizes sowie ihre Vor- und Nachteile diskutiert werden.

   Einstiegsliteratur:
   - Philipp Sibbertsen und Hartmut Lehne. 2015. *Statistik: Einführung für Wirtschafts-und Sozialwissenschaftler.* Springer-Verlag (Kap. 7)

3. **Modellierung von Extremwerten mit Extremwertverteilungen**

   In vielen Anwendungsgebieten spielt die Modellierung extremer Ereignisse eine besondere Rolle. Mithilfe der Extremwerttheorie kann z.B. das Risiko auf Finanzmärkten oder die Wahrscheinlichkeit für Überflutung eines Deichs abgebildet werden. Eine übliche Herangehensweise ist das Aufteilen des Datensatzes in Blöcke, deren Maxima bestimmten Extremwertverteilungen folgen. Dies sind die Gumbel-, Fréchet- und Weibullverteilung, die in der allgemeinen Extremwertverteilung zusammengefasst werden.

Einstiegsliteratur:

- Stuart Coles u. a. 2001. *An introduction to statistical modeling of extreme values.* Springer (Kap. 3)
- Rolf-Dieter Reiss und Michael Thomas. 2007. *Statistical analysis of extreme values.* Springer (Kap. 4)

4. **Modellierung von Extremwerten mit der Peaks-Over-Threshold-Methode**

   Bei einer anderen Herangehensweise zum Modellieren extremer Ereignisse wird ein Grenzwert festgelegt, dessen Überschreitungen betrachtet werden. Nach dem Satz von Pickands können diese Werte bei richtiger Wahl des Grenzwerts als unabhängige Realisationen einer Zufallsvariablen betrachtet werden, die der allgemeinen Pareto-Verteilung folgt. Der Grenzwert wird mit graphischen Entscheidungshilfen bestimmt, sodass anschließend mit der Maximum-Likelihood-Methode die Verteilung angepasst werden kann.

   Einstiegsliteratur:

   - Stuart Coles u. a. 2001. *An introduction to statistical modeling of extreme values.* Springer (Kap. 4)
   - Rolf-Dieter Reiss und Michael Thomas. 2007. *Statistical analysis of extreme values.* Springer (Kap. 5)

5. **Markov-Ketten**

   Eine Markov-Kette ist ein stochastischer Prozess, bei dem die möglichen Merkmalsausprägungen als Zustände aufgefasst werden, die der Prozess annehmen kann. Es werden Übergangswahrscheinlichkeiten zwischen diesen Zuständen modelliert, sodass ermittelt werden kann, mit welcher Wahrscheinlichkeit der Prozess zu einem Zeitpunkt in einem bestimmten Zustand ist. Eine besondere Eigenschaft von Markov-Ketten ist die sogenannte Gedächtnislosigkeit, bei der vergangene Zustände keinen Einfluss auf die zukünftige Entwicklung haben. Übliche Anwendungen sind z.B. Modellierung von Geburts- und Sterbeprozessen einer Population, von Aktienkursen, von Ausfallrisiken oder Warteschlangenmodelle.

   Einstiegsliteratur:

   - Karsten Webel und Dominik Wied. 2016. *Stochastische Prozesse.* Springer (Kap. 4)
   - David Meintrup und Stefan Schäffler. 2006. *Stochastik: Theorie und Anwendungen.* Springer-Verlag (Kap. 9)

6. **Anpassungstests an die Normalverteilung**

   Die Normalverteilung von Daten ist eine zentrale Annahme vieler statistischer Verfahren, wie beispielsweise des t-Tests oder der linearen Regression. Zum Prüfen der Normalverteilungsannahme dienen Tests, die auf unterschiedlichen Prinzipien basieren: beispielsweise der Chi-Quadrat-Anpassungstest und der Kolmogorov-Smirnov-Test zum Vergleich mit der theoretischen Verteilungsfunktion, der Jarque-Bera-Test, basierend auf Schiefe und Wölbung, und der Shapiro-Wilk-Test zur Analyse der Varianz.

   Einstiegsliteratur:

   - Henry C Thode. 2011. "Normality tests". *International Encyclopedia of Statistical Science:* 999–1000
   - Nornadiah Mohd Razali und Yap Bee Wah. 2011. "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests". *Journal of statistical modeling and analytics* 2 (1): 21–33
   - Jürgen Hedderich und Lothar Sachs. 2016. *Angewandte Statistik: Methodensammlung mit R.* Springer Spektrum, Berlin, Heidelberg (Kap. 7)

Grenzen der klassischen linearen Regression

7. **Heteroscedasticity**

   In the least squares method it is assumed that the variance of the disturbance terms is constant. However, if the variance of the disturbance varies, the LS estimator is no longer efficient. This can be proven with tests such as White-Test or Godfrey LM Test. Solutions are offered by heteroscedasticity-resistant standard errors or the weighted LS method. If autocorrelation is also present, HAC (heteroscedasticity and autocorrelation consistent) estimators must be used.

   Introductory Literature:
   - Jeffrey M Wooldridge. 2013. *Introductory econometrics: A modern approach.* Nelson Education (Chap. 8 + 12)
   - William H Greene. 2012. *Econometric analysis.* Pearson Education (Chap. 9)

8. **Autocorrelation**

   Regression with time series data could cause classical assumptions about the OLS estimator to be violated, rendering it ineffective. Autocorrelation is an example of this. When autocorrelation exists, the errors of a linear regression are time dependent. In this thesis the AR(1) error model should be presented. In addition, a test on autocorrelation should be presented and shown how to estimate linear regression models efficiently despite autocorrelation.

   Introductory Literature:
   - Jeffrey M Wooldridge. 2013. *Introductory econometrics: A modern approach.* Nelson Education (Chap. 12)

9. **Specification tests: RESET**

   Consider the model specification of a linear regression model, where the independent regressors x_i are linearly related to the dependent variable y. This assumption about the functional form of a regression can be tested and these tests should be content of this work. The best-known test is the so-called RESET test. In addition, for example, the Rainbow and the Harvey-Collier Test can be presented.

   Introductory Literature:
   - Walter Krämer und Harald Sonnberger. 1986. *The linear regression model under test.* Physica-Verlag Heidelberg (Chap. 4)

10. **Endogeneity: Instrumental Variables**

    Consider a linear regression model, where a prerequisite for the consistency of the OLS estimator is that the independent variable x and the error term e are uncorrelated. If this assumption is violated, there is so-called endogeneity. One consequence is that the OLS estimator has a bias. The presence of endogeneity can be resolved through so-called instrumental variables, which are used in the Two Stage Least Squares (2SLS) in order to obtain a consistent estimate of the effect β.

    Introductory Literature:
    - Jeffrey M Wooldridge. 2013. *Introductory econometrics: A modern approach.* Nelson Education (Chap. 15)
    - James H Stock und Mark W Watson. 2011. *Introduction to Econometrics.* Pearson Education (Chap. 12)

11. **Simultaneous systems of equations**

Consider a simple simultaneous system of equations, where it is characteristic is that $y_{1,t}$ and $y_{2,t}$ appear both on the left in an equation and on the right in an equation. Therefore, an endogeneity problem arises. Two problems should be explained in detail in this work: First, the problem of identification, i.e. under which circumstances γ_1 and γ_2 can be estimated. Second, one should introduce an estimator that works under endogeneity and estimates the coefficients of the system equation by equation.

Introductory Literature:

- William H Greene. 2012. *Econometric analysis.* Pearson Education (Chap. 10)
- Fumio Hayashi. 2000. "Econometrics". *Princeton University Press* (Chap. 8)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Chap. 8 + 9)

12. **Endogeneity: Generalized Method of Moments (GMM)**

Endogeneity bias can lead to inconsistent estimates and incorrect inferences, which may provide misleading conclusions and inappropriate theoretical interpretations. GMM is a statistical method that combines economic data with the information in population moment conditions and is able to estimate all coefficients simultaneously. The idea behind GMM must be explained and then applied to solve the system.

Introductory Literature:

- William H Greene. 2012. *Econometric analysis.* Pearson Education (Chap. 13)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Chap. 8)
- Fumio Hayashi. 2000. "Econometrics". *Princeton University Press* (Chap. 8)
- Jeffrey M Wooldridge. 2001. "Applications of generalized method of moments estimation". *Journal of Economic perspectives* 15 (4): 87–100

13. **Paneldaten: Hausmann Test**

Paneldaten liegen vor, wenn für jede Beobachtung $i = 1, \ldots, N$ Beobachtungen zu verschiedenen Zeitpunkten $t = 1, \ldots, T$ vorliegen, d.h. die abhängige Variable y besitzt zwei Indizes. Nun nehmen wir an, dass jede Beobachtung y einen eigenen Achsenabschnitt a besitzt, d.h. $y = x'b + a + u$. Es gibt nun zwei Fälle: Falls a eine Konstante ist, so liegt ein fixed effect vor. Falls a eine Zufallsvariable ohne Korrelation mit x ist, so liegt ein random effect vor. In dieser Arbeit soll vorgestellt werden, wie mittels eines statistischen Tests entschieden werden kann, welche Art von Achsenabschnitt vorliegt.

Introductory Literature:

- James H Stock und Mark W Watson. 2011. *Introduction to Econometrics.* Pearson Education (Kap. 10)
- Jeffrey M Wooldridge. 2013. *Introductory econometrics: A modern approach.* Nelson Education (Kap. 13)

Spezielle Regressionsmodelle

14. **Modelle für Binärvariablen: Logit und Probit**

Binärvariablen sind Variablen, die nur die Werte 0 und 1 realisieren. Sie werden typischerweise dazu verwendet, um zu zeigen, ob ein bestimmtes Ereignis eingetreten ist oder nicht. Da die beobachteten y Werte in diesem Fall als Wahrscheinlichkeiten interpretiert werden und ein OLS-Modell für y Werte voraussagen könnte, die Eins über- oder Null unterschreiten, ist eine Interpretation nicht gewährleistet. Um mit diesem und anderen Problemen umgehen zu können, können die nichtlinearen Logit- und Probit-Modelle verwendet werden. Hierbei werden die Reaktionswahrscheinlichkeiten bezüglich x restringiert. Das Ziel der Arbeit ist die Vorstellung und Interpretation dieser Modelle. Daneben können auch Tests oder Endogenität vorgestellt werden.

Einstiegsliteratur:

- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 15)
- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 17)

15. **Modelle für kategoriale Variablen: Multinomial Logit**

Eine kategoriale, oder auch nominale, Variable ist eine Variable, die in eine bestimmte Kategorie fällt und keine sinnvolle Ordnung aufweist. Das multinomiale Logit Modell wird verwendet, um eine Wahrscheinlichkeit einer bestimmten Entscheidung unter zwei oder mehr Alternativen zuzuweisen. Zum Beispiel ist die Wahl des Verkehrsmittels, um zur Arbeit zu gelangen, gegeben durch: das Auto, den Bus, den Zug oder das Fahrrad. In dieser Arbeit sollen das standard Logit sowie die Erweiterung zum multinomialen Modell definiert werden. Des Weiteren sollen die Interpretation dieser Modelle und auch Tests vorgestellt werden. Eine Erweiterung bezüglich des Nested Logits ist auch möglich.

Einstiegsliteratur:

- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 18)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 16)

16. **Modelle für ordinalskalierte Variablen: Ordered Probit**

Eine ordinalskalierte Variable ist eine Variable, die eine bestimmte Reihenfolge (oder auch Ordnung) der Variablenwerte aufweist. Das bedeutet, sie können zwar in eine auf- oder absteigende Reihenfolge gebracht werden, allerdings geben diese Variablen keinerlei Auskunft über die Abstände zwischen den Rangplätzen. Als Beispiel für solche geordneten multinomialen Entscheidungsvariablen sind Bond Ratings oder die Ergebnisse aus einem Geschmackstest. Um diesen Variablen dann Wahrscheinlichkeiten zuordnen zu können, wird das Ordered Probit-Modell verwendet. Ziel dieser Arbeit ist die Motivation dieses Modells, sowie die Vorstellung der standard Probit- und Ordered Probit-Modelle. Des Weiteren sollen Tests und die Interpretation der Koeffizienten erklärt werden.

Einstiegsliteratur:

- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 18)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 16)

17. **Modelle für zensierte Daten: Tobit**

Zensierte Daten sind Daten, die "abgeschnitten" sind. Seien wir interessiert an der Nachfrage von Eintrittskarten für ein bestimmtes Event, haben aber nur die Anzahl der verkauften Karten als Maßzahl gegeben, so ist die Variable der Nachfrage limitiert, wenn z.B. das Konzert ausverkauft ist. Das Tobit-Modell ist so aufgebaut, dass es die Zensierung der latenten Variable y^* berücksichtigt. Das Tobit-Modell adressiert den Informationsverlust durch die Zensierung, indem es allen beobachtbaren Werten per Definition den latenten Wert zuordnet, solange y^* größer als Null ist. Ziel dieser Arbeit ist die Motivation der Modelle, hauptsächlich basierend auf Zensierung und Corner Solutions, sowie die Vorstellung des Tobit-Modells. Darüber hinaus soll auf Spezifikationsprobleme und die Interpretation der Koeffizienten eingegangen werden.

Einstiegsliteratur:

- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 17)
- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 19)

18. **Modelle für zensierte Daten: Das Lognormal Hurdle Modell**

Zensierte Daten sind Daten, die „abgeschnitten" sind. Wenn wir Daten y über Arbeitsentgelte erheben, so sind die Daten nicht-negativ, d.h. sie sind am Wert $y = 0$ abgeschnitten. Das Lognormal Hurdle Modell modelliert Auswirkungen von unabhängigen Variablen x auf die beobachtete Variable y. Es interpretiert den Prozess der Erzeugung von y als zweiteiliges Modell: $y = sw$, wobei s entweder 0 oder 1 und $w > 0$ ist. Die Variable s (die Hurdle) zeigt an, ob y zensiert wird, während die Variable w (die lognormalverteilte Größe) anzeigt, wie groß das nicht-zeniserte y ist.

Einstiegsliteratur:

- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 17)
- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 19)

19. **Treatment Effects: Matching-Schätzer**

Mit Average Treatment Effects versuchen Sozialwissenschaftler die Auswirkungen von Maßnahmen wie z.B. der Teilnahme an einem Job-Training für Arbeitslose zu messen. Insbesondere stellt sich die Frage, wie diese Auswirkungen gemessen werden können, wenn kein randomisiertes Experiment vorliegt, d.h. die teilnehmende von der nicht-teilnehmenden Gruppe unterschiedlich sein kann. Mit Matching-Schätzern wird versucht dennoch eine Vergleichbarkeit dieser beiden Gruppen herzustellen, indem auf die äußeren Merkmale abgestellt wird. In dieser Arbeit soll das Covariate Matching vorgestellt werden. Verschiedene Matching-Funktionenen und Abstandsmaße sollen verglichen werden.

Einstiegsliteratur:

- Giovanni Cerulli. 2015. *Econometric evaluation of socio-economic programs.* Springer (Kap. 2)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 21)

20. **Treatment Effects: Propensity Score**

Mit Average Treatment Effects versuchen Sozialwissenschaftler die Auswirkungen von Maßnahmen wie z.B. der Teilnahme an einem Job-Training für Arbeitslose zu messen. Insbesondere stellt sich die Frage, wie diese Auswirkungen gemessen werden können, wenn kein randomisiertes Experiment vorliegt, d.h. die teilnehmende von der nicht-teilnehmenden Gruppe unterschiedlich sein kann. Mit dem Propensity Score wird versucht dennoch eine Vergleichbarkeit dieser beiden Gruppen herzustellen, indem auf den sogenannten Propensity Score abgestellt wird, der die Wahrscheinlichkeit zu einer Gruppe zu gehören angibt. In dieser Arbeit soll das Propensity-Score Matching vorgestellt werden. Verschiedene Eigenschaften des Propensity-Scores sollen beschrieben werden.

Einstiegsliteratur:

- Giovanni Cerulli. 2015. *Econometric evaluation of socio-economic programs.* Springer (Kap. 2)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 21)

21. **Treatment Effects: Sample Selection Schätzer**

Mit Average Treatment Effects versuchen Sozialwissenschaftler die Auswirkungen von Maßnahmen wie z.B. der Teilnahme an einem Job-Training für Arbeitslose zu messen. Insbesondere stellt sich die Frage, wie diese Auswirkungen gemessen werden können, wenn kein randomisiertes Experiment vorliegt, d.h. die teilnehmende von der nicht-teilnehmenden Gruppe unterschiedlich sein kann. Der Ansatz des Sample Selection Schätzers ist es, die sogenannte Selbstauswahl in eine dieser Gruppen explizit zu modellieren und damit die Schätzung des Effekts von dieser Selbstauswahl zu trennen. In dieser Arbeit soll das 2-Stufen Modell bestehend aus Probit und OLS-Schätzung vorgestellt werden.

Einstiegsliteratur:

- Giovanni Cerulli. 2015. *Econometric evaluation of socio-economic programs.* Springer (Kap. 3)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 19)

22. **Überlebenszeitanalyse: Nichtparametrik und Cox-Regression**

In der Medizin werden zum Vergleich von zwei Therapien häufig Verteilungen der Überlebenszeiten herangezogen. In dieser Arbeit sollen grundlegende Konzepte und Methoden der Überlebenszeitanalyse vorgestellt und auf Daten angewendet werden. Das nichtparametrische Kaplan-Meier-Verfahren, der Log-Rank-Test und die Cox-Regression sollen untersucht werden.

Einstiegsliteratur:

- Mario Cleves u. a. 2010. *An introduction to survival analysis using Stata.* Stata press (Kap. 8 + 9)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 22)

23. **Überlebenszeitanalyse: Parametrische Ansätze**

In der Medizin werden zum Vergleich von zwei Therapien häufig Verteilungen der Überlebenszeiten herangezogen. In dieser Arbeit sollen parametrische Methoden der Überlebenszeitanalyse vorgestellt und auf Daten angewendet werden. Dabei wird unterstellt, dass die Verteilung der Überlebenszeit einer parametrischen Verteilungsfunktion folgen, die mit Maximum-Likelihood geschätzt werden können. Es sollen verschiedene Modelle (z.B. Exponential-, Weibull-, Lognormal-, Generalized-Gamma-Modell) vorgestellt und miteinander verglichen werden.

Einstiegsliteratur:

- Mario Cleves u. a. 2010. *An introduction to survival analysis using Stata.* Stata press (Kap. 12 + 13)
- Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data.* MIT Press (Kap. 22)
- William H Greene. 2012. *Econometric analysis.* Pearson Education (Kap. 19)

Multivariate Methoden

24. Analysis of Variance (ANOVA)

ANOVA is the extension of the t- and z-tests where the means of two samples (or a sample and population) are compared relative to the standard error of the mean or pooled standard deviation. ANOVA is best applied where more than two populations are meant to be compared. The different test procedures as well as the motivation of the various test statistics should be presented.

Introductory Literature:

- Alvin C Rencher und William F Christensen. 2012. *Methods of Multivariate Analysis.* John Wiley & Sons, Inc. (Chap. 6)
- Joseph F Hair u. a. 2014. *Multivariate Data Analysis.* Pearson Education Limited (Chap. 14)

25. Principal Component Analysis

A principal component analysis (PCA) is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. Its general objectives are, first, data reduction and second, interpretation. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. Therefore, it is extensively used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics and environmental studies.

Introductory Literature:

- Richard A Johnson, Dean W Wichern u. a. 2002. *Applied multivariate statistical analysis.* Prentice Hall, NJ (Chap. 8)
- Alan J Izenman. 2013. "Multivariate regression". In *Modern Multivariate Statistical Techniques,* 159–194. Springer (Chap. 7)
- Wojtek J Krzanowski. 1995. *Recent advances in descriptive multivariate analysis.* Clarendon Press (Chap. 5)
- Markus Ringnér. 2008. "What is principal component analysis?" *Nature biotechnology* 26 (3): 303

26. Factor Analysis

In factor analysis, we represent the p elements of the vector y as linear combinations of a smaller number of m random variables, where $m < p$, called factors. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question. If the p elements of vector y are at least moderately correlated, the basic dimensionality of the system is less than p. The goal of factor analysis is to reduce the redundancy (needlessness) among the variables by using a smaller number of factors. Motivation for factor models, model definition and assumptions as well as the estimation procedure should be covered.

Introductory Literature:

- Alvin C Rencher und William F Christensen. 2012. *Methods of Multivariate Analysis.* John Wiley & Sons, Inc. (Chap. 13)
- Joseph F Hair u. a. 2014. *Multivariate Data Analysis.* Pearson Education Limited (Chap. 3)

27. Cluster Analysis

In cluster analysis we search for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other. To group the observations into clusters, many techniques begin with similarities between all pairs of observations. In many cases the similarities are based on some measure of distance. Other cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability. The techniques of cluster analysis have been extensively applied to data in many fields, such as medicine, psychiatry, sociology, criminology, anthropology, archaeology, geology, geography, remote sensing, market research, economics, and engineering.

Introductory Literature:

- Alvin C Rencher und William F Christensen. 2012. *Methods of Multivariate Analysis.* John Wiley & Sons, Inc. (Chap. 15)
- Joseph F Hair u. a. 2014. *Multivariate Data Analysis.* Pearson Education Limited (Chap. 8)