Institut für STATISTIK

Prof. Dr. P. Sibbertsen
Wirtschaftswissenschaftliche Fakultät
Leibniz Universität Hannover
Sommersemester 2022

# 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.*

## Grenzen der klassischen linearen Regression

1. **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:*
   - J.M. Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2013 (Chap. 8 + 12)
   - W.H. Greene. *Econometric analysis*. Pearson Education, 2012 (Chap. 9)

2. **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:*
   - J.M. Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2013 (Chap. 12)

3. **Specification Tests: RESET**

   Consider the model specification of a linear regression model, where the independent regressors are linearly related to the dependent variable. 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:*
   - W. Krämer und H. Sonnberger. *The linear regression model under test*. Physica-Verlag Heidelberg, 1986 (Chap. 4)

4. **Endogenität: Instrumentalvariablen**

Im klassischen linearen Regressionsmodell ist eine Voraussetzung für die Konsistenz des OLS-Schätzers, dass die Kovarianz zwischen der Regressormatrix und dem Fehlerterm Null ist. Falls diese Annahme verletzt ist, liegt sogenannte Endogenität vor. Eine Folge davon ist, dass der OLS-Schätzer einen Bias besitzt. Das Vorhandensein von Endogenität kann durch sogenannte Instrumentalvariablen gelöst werden, die im sogenannten Two Stage Least Squares (2SLS) verwendet werden, um eine konsistente Schätzung der Koeffizienten zu erlangen.

*Einstiegsliteratur:*

- J.H. Stock und M.W. Watson. *Introduction to Econometrics*. Pearson Education, 2011 (Kap. 12)
- J.M. Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2013 (Kap. 15)

5. **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:*

- W.H. Greene. *Econometric analysis*. Pearson Education, 2012 (Chap. 13)
- J.M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010 (Chap. 8)
- F. Hayashi. "Econometrics". In: *Princeton University Press* (2000) (Chap. 8)
- J.M. Wooldridge. "Applications of generalized method of moments estimation". In: *Journal of Economic perspectives* 15.4 (2001), S. 87–100

6. **Ridge Regression**

Beim Vorliegen von Multikollinearität ist der OLS-Schätzer unzuverlässig. Die Ridge Regression bietet bei Multikollinearität einen effizienteren Schätzer. Bei dieser Regularisierungsmethode werden die Koeffizienten mit Hilfe eines Strafterms für die Wertegröße geschrumpft. Dies erhöht zwar den Bias der Schätzung, verringert dafür aber dafür ihre Varianz (Verzerrung-Varianz-Dilemma). Die Gewichtung des Strafterms wird durch eine Kreuzvalidierung bestimmt.

*Einstiegsliteratur:*

- J. Friedman, T. Hastie, R. Tibshirani u. a. *The elements of statistical learning*. Springer series in statistics New York, 2001 (Kap. 3)
- G. James u. a. *An introduction to statistical learning*. Springer, 2013 (Kap. 6)
- A.E. Hoerl und R.W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (1970), S. 55–67

7. **Simultane Gleichungssysteme**

Ein einfaches simultanes Gleichungssystem lässt sich dadurch charakterisieren, dass die abhängige Variable in der einen Gleichung als erklärende Variable in der anderen Gleichung vorkommt und ungekehrt. Daher entsteht ein Endogenitätsproblem. Zwei Probleme sollen in dieser Arbeit näher erläutert werden: Zum einen das Problem der Identifikation, d.h. unter welchen Umständen können die Koeffizienten beider Gleichungen geschätzt werden. Zum zweiten sollen Schätzer vorgestellt werden, die unter Endogenität funktionieren und die die Koeffizienten des Systems Gleichung für Gleichung schätzen.

*Einstiegsliteratur:*

- W.H. Greene. *Econometric analysis*. Pearson Education, 2012 (Kap. 10)
- F. Hayashi. "Econometrics". In: *Princeton University Press* (2000) (Kap. 8)
- J.M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010 (Kap. 8+9)

## Spezielle Regressionsmodelle

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

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

*Einstiegsliteratur:*

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

9. **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:*

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

10. **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. Beispiele 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:*

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

11. **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 berücksichtigt. Das Tobit-Modell adressiert den Informationsverlust durch die Zensierung, indem es allen beobachtbaren Werten per Definition den latenten Wert zuordnet, solange dieser 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:*

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

12. **Modelle für zensierte Daten: Das Hurdle-Modell**

Zensierte Daten sind Daten, die „abgeschnitten" sind. Wenn wir Daten über Arbeitsentgelte erheben, so sind die Daten nicht-negativ, d.h. sie sind ab dem Wert Null abgeschnitten. Das Hurdle-Modell modelliert Auswirkungen von unabhängigen Variablen auf die beobachtete Variable. Es interpretiert den Prozess der Erzeugung von den beobachteten Variablen als zweiteiliges Modell, welches die Hurdle und die einer bestimmten Verteilung folgenden Größe beinhaltet. Die Hurdle zeigt an, ob die beobachtete Variable zensiert wird, während die andere Größe anzeigt, wie groß die nicht-zensierte Variable ist.

*Einstiegsliteratur:*

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

13. **Treatment Effects: Matching Estimator**

With average treatment effects, social scientists try to assess the effects of measures such as participation in job training for the unemployed. In particular, the question arises how these effects can be measured if no randomized experiment is available, i.e. the participating group may be different from the non-participating group. Nevertheless, matching estimators are used to establish comparability between these two groups, by focusing on the external characteristics. In this paper, covariate matching should be presented. Different matching functions and distance measures are to be compared.

*Introductory Literature:*

- G. Cerulli. *Econometric evaluation of socio-economic programs.* Springer, 2015 (Chap. 2)
- J.M. Wooldridge. *Econometric analysis of cross section and panel data.* MIT Press, 2010 (Chap. 21)

14. **Treatment Effects: Propensity Score**

With average treatment effects, social scientists try to assess the effects of measures such as participation in job training for the unemployed. In particular, the question arises how these effects can be measured if no randomized experiment is available, i.e. the participating group may be different from the non-participating group. Nevertheless, the propensity score can be used to establish comparability between these two groups, which indicates the probability of belonging to a group. In this paper, the propensity score should be presented. Different properties of the propensity score are to be described.

*Introductory Literature:*

- G. Cerulli. *Econometric evaluation of socio-economic programs.* Springer, 2015 (Chap. 2)
- J.M. Wooldridge. *Econometric analysis of cross section and panel data.* MIT Press, 2010 (Chap. 21)

15. **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:*

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

16. **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:*

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

17. **Random Forests**

Entscheidungsbäume stellen ein leicht zu interpretierendes nichtparametrisches Verfahren dar. Allerdings sind sie in der Praxis oft zu variabel, weswegen meist auf eine Erweiterung, die sogenannten Random Forests zurückgegriffen wird. Diese basieren auf der Idee des Bootstraps. Aus der ursprünglichen Stichprobe wird mit Zurücklegen eine neue Stichprobe gezogen, für die dann ein neuer Entscheidungsbaum bestimmt wird. Dabei wird in jedem Schritt zufällig ausgewählt auf Grundlage welcher Regressoren Entscheidungen getroffen werden können. Dieser Vorgang wird viele Male wiederholt und die Vorhersagen der so entstandenen Bäume werden durch Durchschnittsbildung zu einem Modell zusammengefügt.

*Einstiegsliteratur:*

- G. James u. a. *An introduction to statistical learning*. Springer, 2013 (Kap. 8)
- L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), S. 5–32
- E. Scornet. "On the asymptotics of random forests". In: *Journal of Multivariate Analysis* 146 (2016), S. 72–83

18. **Perzeptron**

Das Perzeptron stellt den Grundbaustein moderner neuronaler Netze dar und wird zur Klassifikation verwendet. In seiner grundlegenden Funktionalität kommt das Perzeptron dem multiplen linearen Regressionsmodell gleich. Im Bereich der neuronalen Netze werden die unabhängigen Variablen des Modells als Eingabe in das Perzeptron interpretiert, welche abhängig von den gelernten Gewichten des Perzeptrons zu einer bestimmten Ausgabe führen. Das Lernen der Gewichte erfolgt über einen iterativen Trainingsprozess, dessen Funktionsweise und Limitationen im Rahmen dieser Arbeit vorgestellt werden sollen. In der Arbeit soll weiter auf das Problem der linearen Separierbarkeit der zu klassifizierenden Daten eingegangen und Lösungsmöglichkeiten wie das mehrlagige Perzeptron oder der Maxover-Algorithmus vorgestellt werden.

*Einstiegsliteratur:*

- W. Ertel und N.T. Black. *Grundkurs Künstliche Intelligenz*. Springer, 2016 (Kap. 8.2)
- C.M. Bishop u. a. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995 (Kap. 3.5)
- F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological review* (1958), S. 386

## Multivariate Methoden

### 19. 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:*

- Olive Jean Dunn und Virginia A Clark. *Applied statistics: analysis of variance and regression*. John Wiley & Sons, Inc., 1986
- J.F. Hair u. a. *Multivariate Data Analysis*. Pearson Education Limited, 2014 (Chap. 14)
- A.C. Rencher und W.F. Christensen. *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., 2012 (Chap. 6)
- Henry Scheffe. *The analysis of variance*. Bd. 72. John Wiley & Sons, 1999

### 20. 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:*

- R.A. Johnson, D.W. Wichern u. a. *Applied multivariate statistical analysis*. Prentice Hall, NJ, 2002 (Chap. 8)
- A.J. Izenman. "Multivariate regression". In: *Modern Multivariate Statistical Techniques*. Springer, 2013, S. 159–194 (Chap. 7)
- W.J. Krzanowski. *Recent advances in descriptive multivariate analysis*. Clarendon Press, 1995 (Chap. 5)
- M. Ringnér. "What is principal component analysis?" In: *Nature biotechnology* 26.3 (2008), S. 303

### 21. Factor Analysis

In factor analysis, we take multiple observed variables that have similar response patterns. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain. 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:*

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

22. **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:*

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

23. **k-Nearest-Neighbors**

k-Nearest-Neighbors (k-NN) ist eine nicht-parametrische Klassifikationsmethode. Der Grundgedanke ist, einzelne Datenpunkte basierend auf der Klassenzugehörigkeit ihnen ähnlicher Datenpunkte - ihrer Nachbarn - zu klassifizieren. Neben der Definition von Entfernung spielt die Wahl des Parameters k, welcher die Größe der zu berücksichtigenden Nachbarschaft steuert, eine wichtige Rolle. In dieser Arbeit soll zunächst das Prinzip der Nächste-Nachbarn-Klassifikation und ihrer verschiedenen Ausprägungen vorgestellt werden, um darauf aufbauend die Wahl des Parameters k und der daraus folgenden Konsequenzen sowie die Evaluationsmöglichkeiten der resultierenden Klassifikation diskutieren zu können.

*Introductory Literature:*

- W. Ertel und N.T. Black. *Grundkurs Künstliche Intelligenz*. Springer, 2016 (Kap. 8.3)
- C.M. Bishop u. a. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995 (Kap. 2.5)

24. **Gaußsche Mischmodelle**

Gaußsche Mischmodelle (engl.: Gaussian Mixture Models, GMMs) werden genutzt, um die Verteilung eines Datensatzes durch einen Mix aus verschiedenen Normalverteilungen zu modellieren. Dazu müssen sowohl die Parameter der einzelnen Normalverteilungen als auch deren korrespondierenden Gewichte für den Mix geschätzt werden. Da für das Schätzproblem keine analytische Lösung existiert, wird auf den Erwartungs-Maximierungs-Algorithmus (EM) zurückgegriffen. In dieser Arbeit sollen die Grundannahmen und die Funktionsweise der Gaußschen Mischmodelle vorgestellt und der EM-Algorithmus diskutiert werden.

*Introductory Literature:*

- S. Richter. *Statistisches und maschinelles Lernen*. Springer, 2019 (Kap. 9.2)
- C.M. Bishop u. a. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995 (Kap. 2.6)
- G.J. McLachlan, S.X. Lee und S.I. Rathnayake. "Finite Mixture Models". In: *Annual review of statistics and its application* 6 (2019), S. 355–378

## Modelle der Stochastik

### 25. 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: unter anderem 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:*

- H.C. Thode. "Normality tests". In: *International Encyclopedia of Statistical Science* (2011), S. 999–1000
- N.M. Razali und Y.B. Wah. "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests". In: *Journal of statistical modeling and analytics* 2.1 (2011), S. 21–33
- J. Hedderich und L. Sachs. *Angewandte Statistik: Methodensammlung mit R.* Springer Spektrum, Berlin, Heidelberg, 2016 (Kap. 7)

### 26. Bootstrap

Der Bootstrap ist ein Resampling-Verfahren, bei dem aus einer gegebenen Stichprobe eine Reihe von Unterstichproben mit Zurücklegen gezogen wird. Mit jeder dieser Unterstichproben wird die interessierende Statistik berechnet, um anschließend die Verteilung dieser Statistik beschreiben zu können. Generell approximieren Boostrap-Methoden also bei vorliegenden Daten die exakte Verteilung eines Schätzers oder einer Teststatistik ohne zusätzliche strukturelle Annahmen an den zugrundeliegenden Prozess. Die ursprüngliche Boostrap-Methode beruht dabei auf unabhängigen Zufallsvariablen.

*Einstiegsliteratur:*

- B. Efron und R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994
- B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), S. 1–26

### 27. Autoregressive Prozesse

Eines der wichtigsten Modelle in der Zeitreihenanalyse ist der autoregressive Prozess (AR), bei dem Beobachtungen anhand von vergangenen Beobachtungen und einem Zufallsschock modelliert werden. Wenn die passende Modellordnung bekannt ist oder geschätzt wurde, also die Anzahl an zu berücksichtigenden vergangenen Beobachtungen, kann mit unterschiedlichen Methoden das Modell angepasst und zur Prognose genutzt werden. Interessant ist besonders die Eigenschaft der Stationarität des Prozesses.

*Einstiegsliteratur:*

- M. Deistler und W. Scherrer. *Modelle der Zeitreihenanalyse.* Springer, 2018 (Kap. 5)
- K. Neusser. *Zeitreihenanalyse in den Wirtschaftswissenschaften.* Springer, 2009 (Kap. 2 + 5)

28. **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:*

- K. Webel und D. Wied. *Stochastische Prozesse.* Springer, 2016 (Kap. 4)
- D. Meintrup und S. Schäffler. *Stochastik: Theorie und Anwendungen.* Springer-Verlag, 2006 (Kap. 9)

29. **Poisson-Prozesse**

Der Poisson-Prozess ist ein stochastischer Prozess, der oft zur Modellierung von Zähldaten genutzt wird. Diese bilden das Auftreten von vorher definierten Ereignissen ab, wie z.B. den eingehenden Anrufen in einem Callcenter, Erdbeben bestimmter Stärke oder Schadensfälle einer Versicherung. Die Prozesswerte folgen dabei einer Poisson-Verteilung, deren Parameter angibt, mit welcher Rate oder Wahrscheinlichkeit innerhalb der nächsten Zeiteinheit ein Ereignis auftritt. Da es sich beim Poisson-Prozess um eine spezielle zeitstetige Markov-Kette handelt, spielt auch hier die Eigenschaft der Gedächtnislosigkeit eine Rolle.

*Einstiegsliteratur:*

- K. Webel und D. Wied. *Stochastische Prozesse.* Springer, 2016 (Kap. 3)
- D. Meintrup und S. Schäffler. *Stochastik: Theorie und Anwendungen.* Springer-Verlag, 2006 (Kap. 10)

30. **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:*

- S. Coles u. a. *An introduction to statistical modeling of extreme values.* Springer, 2001 (Kap. 3)
- R.-D. Reiss und M. Thomas. *Statistical analysis of extreme values.* Springer, 2007 (Kap. 4)