

LMU München, Institut für Statistik, Seminar für angewandte Stochastik

Seminar im Sommersemester 2012

Modellierung kategorialer Daten

Prof. Dr. G. Tutz;

Dipl.-Stat. M. Oelker; Dipl.-Stat. F. Heinzl;

Dipl.-Stat. W. Pössnecker; Dipl.-Stat. S. Rubenbauer;

G. Schauburger, M.Sc.;

Kategoriale Datenanalyse ist in den Pflichtveranstaltungen des Statistikstudiums vergleichsweise wenig präsent. Im Seminar werden daher klassische Verfahren zur Analyse kategorialer Daten sowie neuere Ansätze behandelt. Einige Themen sind geeignet für Studierende des Bachelor-Studiengangs, die die Vorlesung *Kategoriale Daten* nicht besuchen. Die meisten Themen sind für das Masterstudium geeignet. Die Themen sind entsprechend durch (B) bzw. (M) gekennzeichnet. Vorausgesetzt wird die Veranstaltung Generalisierte Regressionsmodelle. Als Einführungsliteratur empfehlen sich Tutz (2012): *Regression for Categorical Data*, sowie Agresti (2002): *Categorical Data Analysis*.

Das Seminar wird dienstags von 16 - 18 Uhr im Seminarraum des Instituts für Statistik, Ludwigstr. 33, stattfinden. Die Anmeldung zum Seminar erfolgt über das LSF. Das Seminar ist auf zwölf Teilnehmer begrenzt.

Erster Seminartermin ist Dienstag, 17.4.2012. (Einführung, Themenvergabe, etc.)

Themen und Literatur:

Analyse von Kontingenztafeln:

- Zusammenhangsmaße in Kontingenztafeln (B).
 - Benninghaus (1990): Einführung in die sozialwissenschaftliche Datenanalyse, *Oldenbourg Verlag*
- Loglineare Modelle (B).
 - Goodman (1970): The Multivariate Analysis of Qualitative Data: Interactions among Multiple Classifications, *Journal of the American Statistical Association* 65.
 - Tutz (2012): *Regression for Categorical Data*, Cambridge: Cambridge University Press.
 - Agresti (2002): *Categorical Data Analysis*, 2nd ed., New York: Wiley.
- Graphische Modelle (B)
 - Goodman (1970): The Multivariate Analysis of Qualitative Data: Interactions among Multiple Classifications, *Journal of the American Statistical Association* 65.
 - Whittaker (1990): *Graphical Models in Applied Multivariate Statistics*, Chichester: Wiley.
 - Agresti (2002): *Categorical Data Analysis*, 2nd ed., New York: Wiley.
- Visualisierung in Kontingenztafeln (B).
 - Meyer et al. (2006): The strucplot framework: Visualizing multiway contingency tables with vcd, *Journal of Statistical Software*, 17(3), 1–48.
 - Friendly (1994): Mosaic displays for n-way contingency tables, *Journal of the American Statistical Association* 89, 190-200.

Modellierung für kategorialen Response:

- Visualisierung kategorialer Responsemodelle I (B)
 - Fox & Andersen (2006): Effect displays for multinomial and proportional-odds logit models, *Sociological Methodology* 36(1), 225–255.
 - Fox & Hong (2009): Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package, *Journal of Statistical Software* 32(1), 1–24.
- Visualisierung kategorialer Responsemodelle II (B)
 - Tutz & Schaubberger (2012): Visualization of Categorical Response Models – from Data Glyphs to Parameter Glyphs, *Department of Statistics, LMU Munich, Technical Report 117*.
- CARTs - Classification and Regression Trees (B)
 - Tutz (2012): *Regression for Categorical Data*, Cambridge: Cambridge University Press.
 - Breiman et al. (1984): *Classification and Regression Trees*, Chapman & Hall/CRC.
 - Breiman (2001): Random forests, *Machine Learning* 45(1), 5–32.
- Genestete Logit-Modelle (M).
 - McFadden (1978): Modeling the choice of residential location, in A. Karlqvist et al. (Hrsg.): *Spatial interaction theory and planning models*, Amsterdam: North-Holland.
 - Koppelman & Wen (1998): Alternative nested logit models: structure, properties and estimation, *Transportation Research B* 32.
 - Koppelman & Wen (2000): The paired combinatorial logit model: properties, estimation and application, *Transportation Research B* 34.

- Additive Modelle (M).
 - Yee & Wild (1996): Vector Generalized Additive Models, *Journal of the Royal Statistical Society B* 58.
 - Tutz (2012): *Regression for Categorical Data*, Cambridge: Cambridge University Press.
 - Agresti (2002): *Categorical Data Analysis*, 2nd ed., New York: Wiley.
 - Yee (2009): VGAM: Vector Generalized Linear and Additive Models, *R package version 0.7-9*.

Kategoriale Prädiktoren:

- Selektion gruppierter Variablen: Group Lasso (M).
 - Yuan & Lin (2006): Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society B* 68.
 - Meier et al. (2008): The group lasso for logistic regression, *Journal of the Royal Statistical Society B* 70.
- Glättung bei ordinalen Prädiktoren (M).
 - Gertheiss & Tutz (2009): Penalized Regression with Ordinal Predictors, *International Statistical Review* 77, 345–365.
- Selektion bei ordinalen Prädiktoren (M).
 - Gertheiss et al. (2009): Selection of Ordinally Scaled Independent Variables, *Department of Statistics, LMU Munich, Technical Report* 62.
- Cluster-Bildung und Selektion bei kategorialen Prädiktoren (M).
 - Bondell & Reich (2009): Simultaneous Factor Selection and Collapsing Levels in ANOVA, *Biometrics*.
 - Gertheiss & Tutz (2009): Sparse Modeling of Categorical Explanatory Variables, *Department of Statistics, LMU Munich, Technical Report* 60.

Penalisierung und Modellselektion:

- Penalisierung in graphischen Modellen (B/M).
 - Dahinden, Parmigiani, Emerick, Bühlmann (2006): Penalized likelihood for sparse contingency tables. Research Report 132, ETH Zürich (auch BMC Bioinformatics, 2007).
- Penalisierung im multinomialen Logit-Modell (B/M).
 - Zhu & Hastie (2004): Classification of gene microarrays by penalized logistic regression, *Biostatistics*, 5, 3, 427–443.
 - Friedman, Hastie & Tibshirani (2010): Regularization paths for generalized linear models via coordinate Descent, *Journal of Statistical Software*, 33, 1, 1–22.

Zählraten:

- Zero-inflated Regression (B).
 - Tutz (2012): *Regression for Categorical Data*, Cambridge: Cambridge University Press.
 - Cameron & Trivedi (1998): *Regression Analysis of Count Data*, Econometric Society Monograph No.30, Cambridge University Press.
 - Lambert (1992): Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, 1–14.
- Hürden Modelle (B).
 - Tutz (2012): *Regression for Categorical Data*, Cambridge: Cambridge University Press.
 - Cameron & Trivedi (1998): *Regression Analysis of Count Data*, Econometric Society Monograph No.30, Cambridge University Press.
 - Mullahy (1986): Specification and Testing of Some Modified Count Data Models, *Journal of Econometrics*, 33, 341–365.