Regression for Categorical Data

This book introduces basic and advanced concepts of modern categorical regression with a focus on the structuring constituents of regression. Meant for statisticians, applied researchers, and students, it includes many topics not normally included in books on categorical data analysis, including recent developments in flexible and high-dimensional regression.

In addition to standard methods such as logit and probit models and their extensions to multivariate settings, the book presents more recent developments in regularized regression with a focus on the selection of predictors; tools for flexible nonparametric regression that yield fits that are closer to the data; advanced models for count data; nonstandard tree-based ensemble methods; and tools for the handling of both nominal and ordered categorical predictors. Issues of prediction are explicitly considered in a chapter that introduces standard and newer classification techniques.

Software including an R package that contains datasets and code for most of the examples is available from http://www.stat.uni-muenchen.de/∼tutz/catdata.

Dr. Gerhard Tutz is a Professor of Statistics in the Department of Statistics at Ludwig-Maximilians University, Munich. He was formerly a Professor at the Technical University Berlin. He is the author or co-author of nine books and more than 100 papers.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at http://www.cambridge.org/uk/series/sSeries.asp?code=CSPM.

Recent titles include the following:

8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics Using R*, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by David Ruppert, M. P. Wand, and R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison, and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort et al.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer, and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability*, by Rick Durrett

Regression for Categorical Data

GERHARD TUTZ
Ludwig-Maximilians Universität

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9781107009653

First published 2012

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Tutz, Gerhard.
Regression for categorical data / Gerhard Tutz.
p. cm. – (Cambridge series in statistical and probabilistic mathematics)
ISBN 978-1-107-00965-3 (hardback)
1. Regression analysis. 2. Categories (Mathematics) I. Title. II. Series.
QA278.2.T88 2011
519.5′36–dc22 2011000390

ISBN 978-1-107-00965-3 Hardback

Additional resources for this publication at http://www.stat.uni-muenchen.de/~tutz/catdata.

Contents

Preface

The focus of this book is on applied structured regression modeling for categorical data. Therefore, it is concerned with the traditional problems of regression analysis: finding a parsimonious but adequate model for the relationship between response and explanatory variables, quantifying the relationship, selecting the influential variables, and predicting the response given explanatory variables.

The objective of the book is to introduce basic and advanced concepts of categorical regressions with the focus on the structuring constituents of regressions. The term "categorical" is understood in a wider sense, including also count data. Unlike other texts on categorical data analysis, a classical analysis of contingency tables in terms of association analysis is considered only briefly. For most contingency tables that will be considered as examples, one or more of the involved variables will be treated as the response. With the focus on regression modeling, the generalized linear model is used as a unifying framework whenever possible. In particular, parametric models are treated within this framework.

In addition to standard methods like the logit and probit models and their extensions to multivariate settings, more recent developments in flexible and high-dimensional regressions are included. Flexible or non-parametric regressions allow the weakening of the assumptions on the structuring of the predictor and yield fits that are closer to the data. High-dimensional regression has been driven by the advance of quantitative genetics with its thousands of measurements. The challenge, for example in gene expression data, is in the dimensions of the datasets. The data to be analyzed have the unusual feature that the number of variables is much higher than the number of cases. Flexible regression as well as high-dimensional regression problems call for regularization methods. Therefore, a major topic in this book is the use of regularization techniques to structure predictors.

Special topics that distinguish it from other texts on categorical data analysis include the following:

- Non-parametric regressions that let the data determine the shape of the functional relationships with weak assumptions on the underlying structure.
- Selection of predictors by regularized estimation procedures that allow one to apply categorical regression to higher dimensional modeling problems.
- The focus on regression includes alternative models like the hurdle model and zero-inflated regression models for count data, which are beyond generalized linear models.
- Non-standard tree-based ensemble methods that provide excellent tools for prediction.
- Issues of prediction are explicitly considered in a chapter that introduces standard and newer classification techniques, including the prediction of ordered categorical responses.
- The handling of categorical predictors, nominal as well as ordered ones. Regularization provides tools to select predictors and to determine which categories should be collapsed.

The present book is based on courses on the modeling of categorical data that I gave at Technical University Berlin and my home university, Ludwig-Maximilians Universität München. The students came from different fields – statistics, computer science, economics, business, sociology – but most of them were statistics students. The book can be used as a text for such courses that include students from interface disciplines. Another audience that might find the text helpful is applied researchers and working data analysts from fields where quantitative analysis is indispensable, for example, biostatisticians, econometricians, and social scientists. The book is written from the perspective of an applied statistician, and the focus is on basic concepts and applications rather than formal mathematical theory. Since categorical data analysis is such a wide field, not all approaches can be covered. For topics that are neglected, for example, exact tests and correlation models, an excellent source is always Alan Agresti's book *Categorical Data Analysis* (Wiley, 2002).

Most of the basic methods for categorical data analysis are available in statistical packages like SAS and SPSS or the free package R (R Development Core Team, 2010). Software including an R package that contains most of the datasets and code for the examples is available from http://www.stat.uni-muenchen.de/~tutz/catdata. Some references to R packages are given in the text, but code is available in the package only. When using the package one should be familiar with R. One of the many tutorials available at the R site might help. Also, introductory books that explicitly treat the use of R with some applications to categorical data like Everitt and Hothorn (2006) and Faraway (2006) could be helpful.

I had much help with computational issues in the examples; thanks to Jan Gertheiss, Sebastian Petry, Felix Heinzl, Andreas Groll, Gunther Schauberger, Sarah Maierhofer, Wolfgang Pößnecker, and Lorenz Uhlmann. I also want to thank Barbara Nishnik and Johanna Brandt for their skillful typing and Elise Oranges for all the corrections in grammar – thanks for all the commas. It was a pleasure to work with Lauren Cowles from Cambridge University Press.

Gerhard Tutz