

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

Editorial Board

- Z. Ghahramani (Department of Engineering, University of Cambridge)
 R. Gill (Mathematical Institute, Leiden University)
 F. P. Kelly (Department of Pure Mathematics and Mathematical Statistics, University of Cambridge)
 B. D. Ripley (Department of Statistics, University of Oxford)
 S. Ross (Department of Industrial and Systems Engineering, University of Southern California)
 M. Stein (Department of Statistics, University of Chicago)

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
978-1-107-00965-3 - Regression for Categorical Data
Gerhard Tutz
Frontmatter
[More information](#)

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

© Gerhard Tutz 2012

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

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>.

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for
external or third-party Internet Web sites referred to in this publication and does not
guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Contents

Preface	ix
1 Introduction	1
1.1 Categorical Data: Examples and Basic Concepts	1
1.2 Organization of This Book	5
1.3 Basic Components of Structured Regression	6
1.4 Classical Linear Regression	15
1.5 Exercises	27
2 Binary Regression: The Logit Model	29
2.1 Distribution Models for Binary Responses and Basic Concepts	29
2.2 Linking Response and Explanatory Variables	33
2.3 The Logit Model	37
2.4 The Origins of the Logistic Function and the Logit Model	48
2.5 Exercises	49
3 Generalized Linear Models	51
3.1 Basic Structure	51
3.2 Generalized Linear Models for Continuous Responses	53
3.3 GLMs for Discrete Responses	56
3.4 Further Concepts	60
3.5 Modeling of Grouped Data	62
3.6 Maximum Likelihood Estimation	63
3.7 Inference	67
3.8 Goodness-of-Fit for Grouped Observations	72
3.9 Computation of Maximum Likelihood Estimates	75
3.10 Hat Matrix for Generalized Linear Models	76
3.11 Quasi-Likelihood Modeling	78
3.12 Further Reading	79
3.13 Exercises	79
4 Modeling of Binary Data	81
4.1 Maximum Likelihood Estimation	82
4.2 Discrepancy between Data and Fit	87
4.3 Diagnostic Checks	93
4.4 Structuring the Linear Predictor	101
4.5 Comparing Non-Nested Models	113
4.6 Explanatory Value of Covariates	114
4.7 Further Reading	119
4.8 Exercises	120

5	Alternative Binary Regression Models	123
5.1	Alternative Links in Binary Regression	123
5.2	The Missing Link	130
5.3	Overdispersion	132
5.4	Conditional Likelihood	138
5.5	Further Reading	140
5.6	Exercises	140
6	Regularization and Variable Selection for Parametric Models	143
6.1	Classical Subset Selection	144
6.2	Regularization by Penalization	145
6.3	Boosting Methods	163
6.4	Simultaneous Selection of Link Function and Predictors	170
6.5	Categorical Predictors	173
6.6	Bayesian Approach	178
6.7	Further Reading	179
6.8	Exercises	179
7	Regression Analysis of Count Data	181
7.1	The Poisson Distribution	182
7.2	Poisson Regression Model	185
7.3	Inference for the Poisson Regression Model	186
7.4	Poisson Regression with an Offset	190
7.5	Poisson Regression with Overdispersion	192
7.6	Negative Binomial Model and Alternatives	194
7.7	Zero-Inflated Counts	198
7.8	Hurdle Models	200
7.9	Further Reading	203
7.10	Exercises	204
8	Multinomial Response Models	207
8.1	The Multinomial Distribution	209
8.2	The Multinomial Logit Model	210
8.3	Multinomial Model as Random Utility Model	215
8.4	Structuring the Predictor	215
8.5	Logit Model as Multivariate Generalized Linear Model	217
8.6	Inference for Multicategorical Response Models	218
8.7	Multinomial Models with Hierarchically Structured Response	223
8.8	Discrete Choice Models	226
8.9	Nested Logit Model	231
8.10	Regularization for the Multinomial Model	233
8.11	Further Reading	238
8.12	Exercises	239
9	Ordinal Response Models	241
9.1	Cumulative Models	243
9.2	Sequential Models	252
9.3	Further Properties and Comparison of Models	255
9.4	Alternative Models	257
9.5	Inference for Ordinal Models	261

CONTENTS	vii
9.6 Further Reading	265
9.7 Exercises	265
10 Semi- and Non-Parametric Generalized Regression	269
10.1 Univariate Generalized Non-Parametric Regression	269
10.2 Non-Parametric Regression with Multiple Covariates	285
10.3 Structured Additive Regression	289
10.4 Functional Data and Signal Regression	307
10.5 Further Reading	313
10.6 Exercises	314
11 Tree-Based Methods	317
11.1 Regression and Classification Trees	317
11.2 Multivariate Adaptive Regression Splines	328
11.3 Further Reading	329
11.4 Exercises	329
12 The Analysis of Contingency Tables: Log-Linear and Graphical Models	331
12.1 Types of Contingency Tables	332
12.2 Log-Linear Models for Two-Way Tables	335
12.3 Log-Linear Models for Three-Way Tables	338
12.4 Specific Log-Linear Models	341
12.5 Log-Linear and Graphical Models for Higher Dimensions	345
12.6 Collapsibility	348
12.7 Log-Linear Models and the Logit Model	349
12.8 Inference for Log-Linear Models	350
12.9 Model Selection and Regularization	354
12.10 Mosaic Plots	357
12.11 Further Reading	358
12.12 Exercises	359
13 Multivariate Response Models	363
13.1 Conditional Modeling	365
13.2 Marginal Parametrization and Generalized Log-Linear Models	370
13.3 General Marginal Models: Association as Nuisance and GEEs	371
13.4 Marginal Homogeneity	385
13.5 Further Reading	392
13.6 Exercises	393
14 Random Effects Models and Finite Mixtures	395
14.1 Linear Random Effects Models for Gaussian Data	396
14.2 Generalized Linear Mixed Models	402
14.3 Estimation Methods for Generalized Mixed Models	407
14.4 Multicategorical Response Models	416
14.5 The Marginalized Random Effects Model	419
14.6 Latent Trait Models and Conditional ML	420
14.7 Semiparametric Mixed Models	420
14.8 Finite Mixture Models	422
14.9 Further Reading	426
14.10 Exercises	427

15 Prediction and Classification	429
15.1 Basic Concepts of Prediction	430
15.2 Methods for Optimal Classification	438
15.3 Basics of Estimated Classification Rules	445
15.4 Parametric Classification Methods	451
15.5 Non-Parametric Methods	457
15.6 Neural Networks	468
15.7 Examples	471
15.8 Variable Selection in Classification	473
15.9 Prediction of Ordinal Outcomes	474
15.10 Model-Based Prediction	480
15.11 Further Reading	481
15.12 Exercises	482
A Distributions	485
A.1 Discrete Distributions	485
A.2 Continuous Distributions	487
B Some Basic Tools	490
B.1 Linear Algebra	490
B.2 Taylor Approximation	491
B.3 Conditional Expectation, Distribution	493
B.4 EM Algorithm	494
C Constrained Estimation	496
C.1 Simplification of Penalties	496
C.2 Linear Constraints	498
C.3 Fisher Scoring with Penalty Term	499
D Kullback-Leibler Distance and Information-Based Criteria of Model Fit	500
D.1 Kullback-Leibler Distance	500
E Numerical Integration and Tools for Random Effects Modeling	504
E.1 Laplace Approximation	504
E.2 Gauss-Hermite Integration	505
E.3 Inversion of Pseudo-Fisher Matrix	507
List of Examples	509
Bibliography	513
Author Index	545
Subject Index	554

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 Schauburger, Sarah Maierhofer, Wolfgang Pöbnecker, 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