

Springer Series in Statistics

Ludwig Fahrmeir
Gerhard Tutz

Multivariate Statistical Modelling Based on Generalized Linear Models

**基于广义线性模型的
多元统计建模**



Springer-Verlag
世界图书出版公司

Ludwig Fahrmeir Gerhard Tutz

Multivariate Statistical Modelling Based on Generalized Linear Models

With Contributions by Wolfgang Hennevogl

With 45 Illustrations

Springer-Verlag

世界图书出版公司

北京 · 广州 · 上海 · 西安

Ludwig Fahrmeir
Seminar für Statistik
Universität München
Ludwigstrasse 33
D-80539 München
Germany

Gerhard Tutz
Institut für Quantitative Methoden
Technische Universität Berlin
Franklinstr. 28/29
D-10587 Berlin
Germany

Mathematics Subject Classifications (1991): 62-02, 62-07, 62P10, 62P20

Library of Congress Cataloging-in-Publication Data
Fahrmeir, L.

Multivariate statistical modelling based on generalized linear models Ludwig Fahrmeir. Gerhard Tutz.

p. cm. — (Springer series in statistics)

Includes bibliographical references and index.

ISBN 0-387-94233-5

I. Multivariate analysis 2. Linear models (Statistics)

I. Tutz, Gerhard. II. Title. III. Series.

QA278.F34 1994

519.5'35 — dc20

93-50900

Printed on acid-free paper.

©1994 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

This reprint has been authorized by Springer-Verlag (Berlin/Heidelberg/New York) for sale in the People's Republic of China only and not for export therefrom.

Reprinted in China by Beijing World Publishing Corporation, 1998

9 8 7 6 5 4 3 2 (Corrected second printing, 1996)

ISBN 0-387-94233-5 Springer-Verlag New York Berlin Heidelberg

ISBN 3-540-94233-5 Springer-Verlag Berlin Heidelberg New York

Preface

Classical statistical models for regression, time series and longitudinal data provide well-established tools for approximately normally distributed variables. Enhanced by the availability of software packages these models dominated the field of applications for a long time. With the introduction of generalized linear models (GLM) a much more flexible instrument for statistical modelling has been created. The broad class of GLM's includes some of the classical linear models as special cases but is particularly suited for categorical discrete or nonnegative responses.

The last decade has seen various extensions of GLM's: multivariate and multicategorical models have been considered, longitudinal data analysis has been developed in this setting, random effects and nonparametric predictors have been included. These extended methods have grown around generalized linear models but often are no longer GLM's in the original sense. The aim of this book is to bring together and review a large part of these recent advances in statistical modelling. Although the continuous case is sketched sometimes, throughout the book the focus is on categorical data. The book deals with regression analysis in a wider sense including not only cross-sectional analysis but also time series and longitudinal data situations. We do not consider problems of symmetrical nature, like the investigation of the association structure in a given set of variables. For example, log-linear models for contingency tables, which can be treated as special cases of GLM's are totally omitted. The estimation approach that is primarily considered in this book is likelihood-based.

The book is aimed at applied statisticians, graduate students of statistics, and students and researchers with a strong interest in statistics and data

analysis from areas like econometrics, biometrics and social sciences. It is written on an intermediate mathematical level with emphasis on basic ideas. Technical and mathematical details are often deferred to starred sections, and for rigorous proofs the reader is referred to the literature.

In preliminary versions of this book Wolfgang Hennevogl was the third author. A new job and its challenges reduced his involvement. Nevertheless he made valuable contributions, in particular to parts of Section 2.3, Section 4.2, Chapter 7, Section 8.3, Appendices A3, A4 and A5, and to many of the examples. In the final stage of the manuscript Thomas Kurtz made helpful contributions, by working out examples and Appendix B.

We are grateful to various colleagues and students in our courses for discussions and suggestions. Discussions with A. Agresti were helpful when the second author visited the University of Florida, Gainesville. We would like to thank Renate Meier-Reusch and Marietta Dostert for the skilful typing of the first version. Moreover, we thank Wolfgang Schneider, Clemens Biller, Martin Krauß, Thomas Scheuchenpflug and Michael Scholz for the preparation of later versions. Further we acknowledge the computational assistance of Christian Gieger, Arthur Klinger, Harald Nase and Stefan Wagenpfeil. We gratefully acknowledge support from Deutsche Forschungsgemeinschaft. For permission to use Tables 1.5, 3.12, 3.13 and 6.1 we are grateful to the Royal Statistical Society and Biometrika Trust.

We hope you will enjoy the book.

München and Berlin, 26.02.1994

Ludwig Fahrmeir
Gerhard Tutz
Wolfgang Hennevogl

书 名: Multivariate Statistical Modelling Based on Generalized Linear Models
作 者: L.Fahrmeir & G.Tutz
中 译 名: 基于广义线性模型的多元统计建模
出 版 者: 世界图书出版公司北京公司
印 刷 者: 北京中西印刷厂
发 行: 世界图书出版公司北京公司 (北京朝阳门内大街 137 号 100010)
开 本: 大 32 开 850×1168 印 张: 14
版 次: 1998 年 8 月第 1 版 1998 年 8 月第 1 次印刷
书 号: 7-5062-3824-1/O·234
版权登记: 图字 01-98-1253
定 价: 68.00 元

世界图书出版公司北京公司已获得 Springer-Verlag 授权在中国
境内独家重印发行。

Contents

Preface	v
List of Examples	xv
List of Figures	xix
List of Tables	xxiii
1 Introduction	1
1.1 Outline and examples	2
1.2 Remarks on notation	13
1.3 Further reading	13
2 Modelling and analysis of cross-sectional data: a review of univariate generalized linear models	15
2.1 Univariate generalized linear models	16
2.1.1 Data	16
Coding of covariates	16
Grouped and ungrouped data	17
2.1.2 Definition of univariate generalized linear models	18
2.1.3 Models for continuous responses	22
Normal distribution	22
Gamma distribution	23
Inverse Gaussian distribution	23

AK 4

2.1.4	Models for binary and binomial responses	24
	Linear probability model	25
	Probit model	25
	Logit model	26
	Complementary log-log model	26
	Binary models as threshold models of latent linear models	27
	Parameter interpretation	29
	Overdispersion	34
2.1.5	Models for counted data	35
	Log-linear Poisson model	35
	Linear Poisson model	35
2.2	Likelihood inference	37
2.2.1	Maximum likelihood estimation	37
	Log-likelihood, score function and information matrix Computation of the MLE by iterative methods . . .	40
	Uniqueness and existence of MLE's*	41
	Asymptotic properties	42
	Discussion of regularity assumptions*	43
	Additional scale or overdispersion parameter	44
2.2.2	Hypothesis testing and goodness-of-fit statistics . .	45
	Goodness-of-fit statistics	48
2.3	Some extensions	52
2.3.1	Quasi-likelihood models	52
	Basic models	52
	Variance functions with unknown parameters	55
	Nonconstant dispersion parameter	55
2.3.2	Bayes models	57
2.3.3	Nonlinear and nonexponential family regression models*	60
2.4	Further developments	62
3	Models for multicategorical responses:	
	multivariate extensions of generalized linear models	63
3.1	Multicategorical response models	64
3.1.1	Multinomial distribution	64
3.1.2	Data	65
3.1.3	The multivariate model	66
3.1.4	Multivariate generalized linear models	68
3.2	Models for nominal responses	70
3.2.1	The principle of maximum random utility	70
3.2.2	Modelling of explanatory variables: choice of design matrix	71
3.3	Models for ordinal responses	73
3.3.1	Cumulative models: the threshold approach	75

	Cumulative logistic model or proportional odds model	76
	Grouped Cox model or proportional hazards model	78
	Extreme-maximal-value distribution model	79
3.3.2	Extended versions of cumulative models	79
3.3.3	Link functions and design matrices for cumulative models	80
3.3.4	Sequential models	84
	Generalized sequential models	87
	Link functions of sequential models	90
3.3.5	Strict stochastic ordering*	90
3.3.6	Two-step models	91
	Link function and design matrix for two-step models	94
3.3.7	Alternative approaches*	95
3.4	Statistical inference	96
3.4.1	Maximum likelihood estimation	97
	Numerical computation	98
3.4.2	Testing and goodness-of-fit	99
	Testing of linear hypotheses	99
	Goodness-of-fit statistics	99
3.4.3	Power-divergence family*	101
	Asymptotic properties under classical "fixed cells" assumptions	103
	Sparseness and "increasing-cells" asymptotics	103
3.5	Multivariate models for correlated responses	104
3.5.1	Conditional models	105
	Asymmetric models	105
	Symmetric models	108
3.5.2	Marginal models	110
	Statistical inference	113
4	Selecting and checking models	119
4.1	Variable selection	119
4.1.1	Selection criteria	120
4.1.2	Selection procedures	122
	All-subsets selection	122
	Stepwise backward and forward selection	122
4.2	Diagnostics	124
4.2.1	Diagnostic tools for the classical linear model	125
4.2.2	Generalized hat matrix	126
4.2.3	Residuals and goodness-of-fit statistics	130
4.2.4	Case deletion	138
4.3	General tests for misspecification*	140
4.3.1	Estimation under model misspecification	142
4.3.2	Hausman-type tests	144
	Hausman tests	144

	Information matrix test	145
4.3.3	Tests for non-nested hypotheses	146
	Tests based on artificial nesting	146
	Generalized Wald and score tests	147
5	Semi- and nonparametric approaches to regression analysis	151
5.1	Smoothing techniques for continuous responses	152
5.1.1	Simple neighbourhood smoothers	152
5.1.2	Spline smoothing	153
	Cubic smoothing splines	153
	Regression splines	155
5.1.3	Kernel smoothing	156
	Relation to other smoothers	158
	Bias-variance trade-off	158
5.1.4	Selection of smoothing parameters*	160
5.2	Kernel smoothing with multicategorical response	162
5.2.1	Kernel methods for the estimation of discrete distributions	162
5.2.2	Smoothed categorical regression	167
5.2.3	Choice of smoothing parameters*	172
5.3	Spline smoothing in generalized linear models	175
5.3.1	Cubic spline smoothing with a single covariate	175
	Fisher scoring for generalized spline smoothing*	176
	Choice of smoothing parameter	177
5.3.2	Generalized additive models	180
	Fisher scoring with backfitting*	181
6	Fixed parameter models for time series and longitudinal data	187
6.1	Time series	188
6.1.1	Conditional models	188
	Generalized autoregressive models	188
	Quasi-likelihood models and extensions	191
6.1.2	Statistical inference for conditional models	194
6.1.3	Marginal models	200
	Estimation of marginal models	202
6.2	Longitudinal data	204
6.2.1	Conditional models	205
	Generalized autoregressive models, quasi-likelihood models	205
	Statistical inference	206
	Transition models	207
	Subject-specific approaches and conditional likelihood	208

6.2.2	Marginal models	211
	Statistical inference	213
7	Random effects models	219
7.1	Linear random effects models for normal data	221
7.1.1	Two-stage random effects models	221
	Random intercepts	222
	Random slopes	223
	Multilevel models	224
7.1.2	Statistical inference	224
	Known variance-covariance components	225
	Unknown variance-covariance components	225
	Derivation of the EM-algorithm*	227
7.2	Random effects in generalized linear models	228
7.3	Estimation based on posterior modes	233
7.3.1	Known variance-covariance components	234
7.3.2	Unknown variance-covariance components	235
7.3.3	Algorithmic details*	235
	Fisher scoring for given variance-covariance components	235
	EM-type algorithm	237
7.4	Estimation by integration techniques	238
7.4.1	Maximum likelihood estimation of fixed parameters	238
7.4.2	Posterior mean estimation of random effects	240
7.4.3	Algorithmic details*	241
	Direct maximization	241
	Indirect maximization	243
	Posterior mean estimation	247
7.5	Examples	249
7.6	Marginal estimation approach to random effects models	252
7.7	Further approaches	254
8	State space models	257
8.1	Linear state space models and the Kalman filter	258
8.1.1	Linear state space models	258
8.1.2	Statistical inference	263
	Linear Kalman filtering and smoothing	263
	Kalman filtering and smoothing as posterior mode estimation*	265
	Unknown hyperparameters	267
	EM-algorithm for estimating hyperparameters*	268
8.2	Non-normal and nonlinear state space models	269
8.2.1	Dynamic generalized linear models	270
	Categorical time series	271
8.2.2	Nonlinear and nonexponential family models*	274

8.3	Non-normal filtering and smoothing	275
8.3.1	Posterior mode estimation	276
	Generalized extended Kalman filter and smoother*	277
	Gauss-Newton and Fisher-scoring filtering and smoothing*	279
	Estimation of hyperparameters*	281
	Some applications	281
8.3.2	Posterior mean estimation	286
	A Gibbs sampling approach*	287
	Integration-based approaches*	290
8.4	Longitudinal data	293
8.4.1	State space modelling of longitudinal data	293
8.4.2	Filtering and smoothing	295
	Generalized Kalman filter and smoother for longitudinal data*	296
9	Survival models	305
9.1	Models for continuous time	305
9.1.1	Basic models	305
	Exponential distribution	306
	Weibull distribution	307
9.1.2	Parametric regression models	307
	Location-scale models for log T	308
	Proportional hazards models	308
	Linear transformation models and binary regression models	309
9.1.3	Censoring	310
	Random censoring	310
	Type I censoring	311
9.1.4	Estimation	312
	Exponential model	312
	Weibull model	313
9.2	Models for discrete time	314
9.2.1	Life table estimates	315
9.2.2	Parametric regression models	318
	The grouped proportional hazards model	318
	A generalized version: the model of Aranda-Ordaz	320
	The logistic model	321
	Sequential model and parameterization of the baseline hazard	321
9.2.3	Maximum likelihood estimation	322
9.2.4	Time-varying covariates	325
	Internal covariates*	328
	Maximum likelihood estimation*	329
9.3	Discrete models for multiple modes of failure	331

9.3.1	Basic models	331
9.3.2	Maximum likelihood estimation	333
9.4	Smoothing in discrete survival analysis	337
9.4.1	Dynamic discrete time survival models	337
	Posterior mode smoothing	338
9.4.2	Kernel smoothing	340
Appendix A		345
A.1	Exponential families and generalized linear models	345
A.2	Basic ideas for asymptotics	350
A.3	EM-algorithm	355
A.4	Numerical integration	357
A.5	Monte Carlo methods	363
Appendix B Software for fitting generalized linear models		367
References		379
Author Index		413
Subject Index		419

List of Examples

1.1	Caesarean birth study	2
1.2	Credit-scoring	3
1.3	Cellular differentiation	4
1.4	Job expectation	5
1.5	Breathing test results	5
1.6	Visual impairment	7
1.7	Rainfall data	8
1.8	Polio incidence	9
1.9	IFO business test	9
1.10	Ohio children	10
1.11	Duration of unemployment	12
2.1	Caesarean birth study	29
2.2	Credit-scoring	31
2.3	Cellular differentiation	35
2.4	Caesarean birth study	48
2.5	Credit-scoring	50
2.6	Cellular differentiation	51
2.7	Cellular differentiation	56
3.1	Caesarean birth study	67
3.2	Breathing test results	74

3.3	Job expectation	74
3.4	Breathing test results	82
3.5	Job expectation	83
3.6	Tonsil size	85
3.7	Tonsil size	88
3.8	Breathing test results	89
3.9	Rheumatoid arthritis	91
3.10	Rheumatoid arthritis	93
3.11	Caesarean birth study	100
3.12	Reported happiness	107
3.13	Visual impairment	116
4.1	Credit-scoring	124
4.2	Vaso-constriction	127
4.3	Job expectation	130
4.4	Vaso-constriction	133
4.5	Job expectation	134
4.6	Vaso-constriction	140
4.7	Job expectation	140
4.8	Credit-scoring	148
5.1	Motorcycle data	159
5.2	Memory	165
5.3	Vaso-constriction data	168
5.4	Unemployment data	170
5.5	Rainfall data	177
5.6	Vaso-constriction data	183
6.1	Polio incidence in USA	197
6.2	Polio incidence in USA	203
6.3	IFO business test	209
6.4	Ohio children	215
7.1	Ohio children data	228
7.2	Bitterness of white wines	228
7.3	Ohio children data	249
7.4	Bitterness of white wines	250
8.1	Rainfall data	281
8.2	Advertising data	283
8.3	Phone calls	284

8.4	Rainfall data	292
8.5	Business test	298
9.1	Duration of unemployment	317
9.2	Duration of unemployment	325
9.3	Duration of unemployment	335
9.4	Head and neck cancer	341

List of Figures

1.1	Number of occurrences of rainfall in the Tokyo area for each calendar day during 1983–1984	9
1.2	Monthly number of polio cases in USA from 1970 to 1983	10
2.1	The gamma distribution: $G(\mu = 1, \nu)$	24
2.2	Response functions for binary responses	26
2.3	Response functions for binary responses adjusted to the logistic function (that means linear transformation yielding mean zero and variance $\pi^2/3$).	28
2.4	Log-likelihood (—) and quadratic approximation (- - -) for Wald test and slope for score test	47
3.1	Densities of the latent response for two subpopulations with different values of x (logistic, extreme-minimal-value, extreme-maximal-value distributions)	77
4.1	Index plot of h_{ii} for vaso-constriction data	129
4.2	Index plot of $\text{tr}(H_{ii})$ and $\det(H_{ii})$ for grouped job expectation data	131
4.3	Index plot of $r_{i,s}^P, r_{i,s}^D$ and r_i^D for vaso-constriction data	134
4.4	$N(0, 1)$ -probability plot of $r_{i,s}^P$ for vaso-constriction data	135